Eyke Hüllermeier
Rudolf Kruse
Frank Hoffmann (Eds.)

Communications in Computer and Information Science 80

Information Processing and Management of Uncertainty in Knowledge-Based Systems

Theory and Methods

13th International Conference, IPMU 2010
Dortmund, Germany, June/July 2010
Proceedings, Part I

Part 1

Springer

Communications
in Computer and Information Science 80

Eyke Hüllermeier Rudolf Kruse
Frank Hoffmann (Eds.)

Information Processing and Management of Uncertainty in Knowledge-Based Systems

Theory and Methods

13th International Conference, IPMU 2010
Dortmund, Germany, June 28 – July 2, 2010
Proceedings, Part I

Springer

Volume Editors

Eyke Hüllermeier
Philipps-Universität Marburg
Marburg, Germany
E-mail: eyke@mathematik.uni-marburg.de

Rudolf Kruse
Otto-von-Guericke-Universität Magdeburg
Magdeburg, Germany
E-mail: kruse@iws.cs.uni-magdeburg.de

Frank Hoffmann
Technische Universität Dortmund
Dortmund, Germany
E-mail: frank.hoffmann@tu-dortmund.de

Library of Congress Control Number: Applied for

CR Subject Classification (1998): I.2, H.3, F.1, H.4, I.5, I.4

ISSN 1865-0929
ISBN-10 3-642-14054-8 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-14054-9 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

The International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU, is organized every two years with the aim of bringing together scientists working on methods for the management of uncertainty and aggregation of information in intelligent systems. Since 1986, this conference has been providing a forum for the exchange of ideas between theoreticians and practitioners working in these areas and related fields. The 13^{th} IPMU conference took place in Dortmund, Germany, June 28–July 2, 2010.

This volume contains 79 papers selected through a rigorous reviewing process. The contributions reflect the richness of research on topics within the scope of the conference and represent several important developments, specifically focused on theoretical foundations and methods for information processing and management of uncertainty in knowledge-based systems.

We were delighted that Melanie Mitchell (Portland State University, USA), Nihkil R. Pal (Indian Statistical Institute), Bernhard Schölkopf (Max Planck Institute for Biological Cybernetics, Tübingen, Germany) and Wolfgang Wahlster (German Research Center for Artificial Intelligence, Saarbrücken) accepted our invitations to present keynote lectures. Jim Bezdek received the Kampé de Fériet Award, granted every two years on the occasion of the IPMU conference, in view of his eminent research contributions to the handling of uncertainty in clustering, data analysis and pattern recognition.

Organizing a conference like this one is not possible without the assistance and continuous support of many people and institutions. We are particularly grateful to the organizers of sessions on dedicated topics that took place during the conference—these 'special sessions' have always been a characteristic element of the IPMU conference. Frank Klawonn and Thomas Runkler helped a lot to evaluate and select special session proposals. The special session organizers themselves rendered important assistance in the reviewing process, that was furthermore supported by the Area Chairs and regular members of the Programme Committee. Thomas Fober was the backbone on several organizational and electronic issues, and also helped with the preparation of the proceedings. In this regard, we would also like to thank Alfred Hofmann and Springer for providing continuous assistance and ready advice whenever needed.

Finally, we gratefully acknowledge the support of several organizations and institutions, notably the German Informatics Society (Gesellschaft für Informatik, GI), the German Research Foundation (DFG), the European Society for Fuzzy Logic and Technology (EUSFLAT), the International Fuzzy Systems Association (IFSA), the North American Fuzzy Information Processing Society (NAFIPS) and the IEEE Computational Intelligence Society.

April 2010

Eyke Hüllermeier
Rudolf Kruse
Frank Hoffmann

Organization

Conference Committee

General Chair	Eyke Hüllermeier (Philipps-Universität Marburg)
Co-chairs	Frank Hoffmann (Technische Universität Dortmund)
	Rudolf Kruse (Otto-von-Guericke Universität Magdeburg)
	Frank Klawonn (Hochschule Braunschweig-Wolfenbüttel)
	Thomas Runkler (Siemens AG, Munich)
Web chair	Thomas Fober (Philipps-Universität Marburg)
Executive Directors	Bernadette Bouchon-Meunier (LIP6, Paris, France)
	Ronald R. Yager (Iona College, USA)

International Advisory Board

G. Coletti, Italy	C. Marsala, France	L. Valverde, Spain
M. Delgado, Spain	M. Ojeda-Aciego, Spain	J.L. Verdegay, Spain
L. Foulloy, France	M. Rifqi, France	M.A. Vila, Spain
J. Gutierrez-Rios, Spain	L. Saitta, Italy	L.A. Zadeh, USA
L. Magdalena, Spain	E. Trillas, Spain	

Special Session Organizers

P. Angelov	F. Hoffmann	B. Prados Suárez
A. Antonucci	S. Kaci	M. Preuß
C. Beierle	J. Kacprzyk	A. Ralescu
G. Beliakov	G. Kern-Isberner	D. Ralescu
G. Bordogna	C. Labreuche	E. Reucher
A. Bouchachia	H. Legind Larsen	W. Rödder
H. Bustince	E. William De Luca	S. Romaní
T. Calvo	E. Lughofer	G. Rudolph
P. Carrara	E. Marchioni	G. Ruß
J. Chamorro Martínez	N. Marin	D. Sanchez
D. Coquin	M. Minoh	R. Seising
T. Denoeux	G. Navarro-Arribas	A. Skowron
P. Eklund	H. Son Nguyen	D. Slezak
Z. Elouedi	V. Novak	O. Strauss
M. Fedrizzi	P. Melo Pinto	E. Szmidt
J. Fernandez	E. Miranda	S. Termini
T. Flaminio	V.A. Niskanen	V. Torra
L. Godo	D. Ortiz-Arroyo	L. Valet
M. Grabisch	I. Perfilieva	A. Valls
A.J. Grichnik	O. Pons	R.R. Yager

International Programme Committee

Area Chairs

P. Bosc, France
O. Cordon, Spain
G. De Cooman, Belgium
T. Denoeux, France
R. Felix, Germany
L. Godo, Spain
F. Gomide, Spain
M. Grabisch, France
F. Herrera, Spain
L. Magdalena, Spain
R. Mesiar, Slovenia
D. Sanchez, Spain
R. Seising, Spain
R. Slowinski, Poland

Regular Members

P. Angelov, UK
J.A. Appriou, France
M. Baczynski, Poland
G. Beliakov, Australia
S. Ben Yahia, Tunisia
S. Benferat, France
H. Berenji, USA
J. Bezdek, USA
I. Bloch, France
U. Bodenhofer, Austria
P.P. Bonissone, USA
C. Borgelt, Spain
H. Bustince, Spain
R. Casadio, Italy
Y. Chalco-Cano, Chile
C.A. Coello Coello, Mexico
I. Couso, Spain
B. De Baets, Belgium
G. De Tré, Belgium
M. Detyniecki, France
D. Dubois, France
F. Esteva, Spain
M. Fedrizzi, Italy
J. Fodor, Hungary
D. Fogel, USA
K. Fujimoto, Japan
P. Gallinari, France
B. Gerla, Italy
M.A. Gil, Spain
S. Gottwald, Germany
S. Grossberg, USA
P. Hajek, Czech Republic
L. Hall, USA
E. Herrera-Viedma, Spain
C. Noguera, Spain
K. Hirota, Japan
A. Hunter, UK
H. Ishibuchi, Japan
Y. Jin, Germany
J. Kacprzyk, Poland
A. Kandel, USA
G. Kern-Isberner, Germany
E.P. Klement, Austria
L. Koczy, Hungary
V. Kreinovich, USA
T. Kroupa, Czech Republic
C. Labreuche, France
J. Lang, France
P. Larranaga, Spain
H. Larsen, Denmark
A. Laurent, France
M.J. Lesot, France
C.J. Liau, Taiwan
W. Lodwick, USA
J.A. Lozano, Spain
T. Lukasiewicz, UK
F. Marcelloni, Italy
J.L. Marichal, Luxembourg
N. Marin, Spain
T. Martin, UK
L. Martinez, Spain
J. Medina, Spain
J. Mendel, USA
E. Miranda, Spain
P. Miranda, Spain
J. Montero, Spain
S. Moral, Spain
M. Nachtegael, Belgium
Y. Nojima, Japan
V. Novak, Czech Republic
H. Nurmi, Finland
E. Pap, Serbia
W. Pedrycz, Canada
F. Petry, USA
V. Piuri, Italy
O. Pivert, France
P. Poncelet, France
H. Prade, France
A. Ralescu, USA
D. Ralescu, USA
M. Ramdani, Morocco
M. Reformat, Canada
D. Ruan, Belgium
E. Ruspini, USA
R. Scozzafava, Italy
P. Shenoy, USA
G. Simari, Argentina
P. Sobrevilla, Spain
U. Straccia, Italy

T. Stutzle, Belgium
K.C. Tan, Singapore
R. Tanscheit, Brazil
S. Termini, Italy
V. Torra, Spain
I.B. Turksen, Canada
B. Vantaggi, Italy
P. Vicig, Italy
Z. Wang, USA
M. Zaffalon, Switzerland
H.J. Zimmermann, Germany
J. Zurada, USA

Table of Contents – Part I

Reasoning with Uncertainty

Decomposable Models

Imprecise Probabilities

Logics for Reasoning

Preference Modeling

Data Analysis and Knowledge Processing

Belief Functions

Rough Sets

Machine Learning

Probabilistic Methods

Fuzzy Methods

Fuzzy Sets and Fuzzy Logic

Fuzzy Measures and Integrals

Fuzzy Inference

Intuitionistic Fuzzy Sets

Aggregation Functions

Evolutionary Agorithms

Table of Contents – Part II

Data Analysis Applications

Intelligent Databases

Information Fusion

Decision Support

Modelling Collective Choices

Fuzzy Decision Theory

Applications in Finance

Fuzzy Systems

Philosophical Aspects

Fuzzy Numbers

Fuzzy Arithmetic

Fuzzy Equations

Soft Computing Applications

Image Processing

Privacy and Security

The Sensing Web

Manufacturing and Scheduling

An Algorithm to Find a Perfect Map for Graphoid Structures

Marco Baioletti[1], Giuseppe Busanello[2], and Barbara Vantaggi[2]

[1] Dip. Matematica e Informatica, Università di Perugia, Italy
baioletti@dipmat.unipg.it
[2] Dip. Metodi e Modelli Matematici, Università "La Sapienza" Roma, Italy
{busanello,vantaggi}@dmmm.uniroma1.it

Abstract. We provide a necessary and sufficient condition for the existence of a perfect map representing an independence model and we give an algorithm for checking this condition and drawing a perfect map, when it exists.

Keywords: Conditional independence models, Inferential rules, Acyclic directed graphs, Perfect map.

1 Introduction

Graphical models [11,12,14,15,16,20] play a fundamental role in probability and multivariate statistics and they have been deeply developed as a tool for representing conditional independence models. The usefulness of graphical models is not limited to the probabilistic setting, in fact they have been extended to other frameworks (see, e.g. [6,7,8,13,17]). Among graphical structures, we consider graphoids that are induced, for example, by a strictly positive probability under the classical notion of independence [9].

A relevant problem is to represent a set J of conditional independence relations, provided by an expert, by a directed acyclic graph (DAG), where independencies are encoded by d–separation. Such a graph is called perfect map for J (see [14]). A DAG gives a very compact and human–readable representation, unfortunately it is known that there exist sets of independencies which admit no perfect maps. The problem of the existence of a perfect map has been studied by many authors (see for instance [14]) by providing only partial answers in terms of necessary or sufficient conditions.

In [2] we have introduced a sufficient condition for the existence of a perfect map in terms of existence of a certain ordering among the random variables, and we describe BN–DRAW procedure in order to build the corresponding independence map given an ordering. The sufficient condition, as well as BN–DRAW, uses the "fast" closure J_* of J [1]. From J_* it is possible to solve the implication problem for J and to extract independence maps with fast algorithms. The set J_* can be computed in a reasonable amount of time, as shown in [1,3] and it is extremely smaller than the complete closure $\bar{J}$ of J with respect to graphoid properties, even if it gathers the same information of $\bar{J}$.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 1–10, 2010.

A similar construction has been given in [15], essentially for the semi–graphoids, and used in [10] to describe a necessary condition for the existence of a perfect map for semi–graphoid structures.

In this paper we provide a necessary and sufficient condition for the existence of a perfect map for graphoid structures. This condition relies on some constraints among the triples of the set J_* and their components. Moreover, we give an algorithm to check the existence of a perfect map based on the provided condition[1]. In the positive case, the algorithm returns a relevant perfect map.

2 Graphoid

Let $\tilde{S} = \{Y_1, \ldots, Y_n\}$ be a finite not empty set of variables and $S = \{1, \ldots, n\}$ the set of indices associated to $\tilde{S}$. Furthermore, $S^{(3)}$ is the set of all (ordered) triples (A, B, C) of disjoint subsets of S, such that A and B are not empty.

A conditional independence model $\mathcal{I}$ is a suitable subset of $S^{(3)}$. We refer to graphoid structure $(S, \mathcal{I})$, with $\mathcal{I}$ ternary relation on the set S, satisfying the following properties (where A, B, C, D are pairwise disjoint subsets of S):

G1 if $(A, B, C) \in \mathcal{I}$, then $(B, A, C) \in \mathcal{I}$ (Symmetry);
G2 if $(A, B \cup C, D) \in \mathcal{I}$, then $(A, B, D) \in \mathcal{I}$ (Decomposition);
G3 if $(A, B \cup C, D) \in \mathcal{I}$, then $(A, B, C \cup D) \in \mathcal{I}$ (Weak Union);
G4 if $(A, B, C \cup D) \in \mathcal{I}$ and $(A, C, D) \in \mathcal{I}$, then $(A, B \cup C, D) \in \mathcal{I}$ (Contraction);
G5 if $(A, B, C \cup D) \in \mathcal{I}$ and $(A, C, B \cup D) \in \mathcal{I}$, then $(A, B \cup C, D) \in \mathcal{I}$ (Intersection).

Given a triple $\theta = (A, B, C)$ we denote with $\theta^T(B, A, C)$ the transpose triple obtained by applying G1 to θ.

Given a set J of conditional independence statements, a relevant problem about graphoids is to find efficiently the closure of J with respect to G1–G5

$$\bar{J} = \{\theta \in S^{(3)} : \theta \text{ is obtained from } J \text{ by } G1 - G5\}.$$

A related problem, called implication, concerns to establish whether a triple $\theta \in S^{(3)}$ can be derived from J.

This implication problem can be easily solved once the closure has been computed. But, the computation of the closure is infeasible because its size is exponentially larger than the size of J. In [1] we have described how it is possible to compute a smaller set of triples having the same information as the closure.

Now we recall some definitions and properties introduced and studied in [1], which are used in the rest of the paper.

Given a pair of triples $\theta_1, \theta_2 \in S^{(3)}$, we say that θ_1 is *generalized–included* in θ_2 (briefly g–included), in symbol $\theta_1 \sqsubseteq \theta_2$, if θ_1 can be obtained from θ_2 by a finite number of applications of G1, G2 and G3.

[1] An implementation of the proposed algorithms is available at `http://www.dmi.unipg.it/baioletti/graphoids`

Proposition 1. *Given $\theta_1 = (A_1, B_1, C_1)$ and $\theta_2 = (A_2, B_2, C_2)$, then $\theta_1 \sqsubseteq \theta_2$ if and only if the following conditions hold*

(i) $C_2 \subseteq C_1 \subseteq A_2 \cup B_2 \cup C_2$;
(ii) *either* $A_1 \subseteq A_2$ *and* $B_1 \subseteq B_2$ *or* $A_1 \subseteq B_2$ *and* $B_1 \subseteq A_2$.

Generalized inclusion is strictly related to the concept of dominance [15].

In [1] we introduce a particular subset J_* of $\bar{J}$ (called "fast closure"), which can be computed from $\bar{J}$ by discarding the not "maximal" triples $\tau \in \bar{J}$, i.e. those g–included in some other triple of J. Moreover, in [1], we describe and compare two different algorithms to compute J_*, called FC2 and FC1. In particular, FC2 iteratively uses two inferential rules G4* and G5*, related to G4 and G5, introduced always in [1], and discards not maximal triples, until the set of independence relations is closed. FC1 has a similar structure, but uses a single inference rule U, which corresponds to compute at once the fast closure of a couple of triples.

For some considerations and experimental results (see also [3]) FC1 appears to be faster than FC2.

3 Graphs

In the following, we refer to the usual graph definitions (see [14]): we denote by $G = (\mathcal{U}, \mathcal{E})$ a graph with set $\mathcal{U}$ of nodes and oriented arcs $\mathcal{E}$ (ordered pairs of nodes). In particular, we consider directed graphs having no cycles, i.e. acyclic directed graphs (DAG). As usual, we denote by $pa(u)$, for any $u \in \mathcal{U}$, the parent set of u.

Definition 1. *If A, B and C are three disjoint subsets of nodes in a DAG G, then C is said to d–separate A from B, denoted $(A, B, C)_G$, if for each non–directed path between a node in A and a node in B, there exists a node x in the path which satisfies one of these two conditions*

1. *x is a collider (i.e. both edges point to x), $x \notin C$ and no descendant of x is in C;*
2. *x is not a collider and belongs to C.*

In order to study the representation of a conditional independence model, we need to distinguish between dependence map and independence map, since there are conditional independence models that cannot be completely represented by a DAG (see e.g. [12,14]).

In the following we denote with J (analogously for $\bar{J}$, J_*) both a set of triples and a set of conditional independence relations, obviously, the triples are defined on the set S and the independence relations on $\tilde{S}$.

Definition 2. *Let J be a set of conditional independence relations on a set S. A DAG $G = (S, \mathcal{E})$ is a* dependence map *(briefly a D–map) if for all triples $(A, B, C) \in S^{(3)}$*

$$(A, B, C) \in \bar{J} \Rightarrow (A, B, C)_G.$$

Moreover, $G = (S, \mathcal{E})$ is an independence map *(briefly an I–map) if for all triples $(A, B, C) \in S^{(3)}$*

$$(A, B, C)_G \Rightarrow (A, B, C) \in \bar{J}.$$

G is a minimal *I–map of J if deleting any arc, G is no more an I-map.*

G is said to be a perfect map *(briefly a p–map) if it is both a I–map and a D–map.*

The next definition and theorem [14] provide a tool to build a DAG given an independence model $\bar{J}$.

Definition 3. *Let $\bar{J}$ be an independence model defined on S and let $\pi =< \pi_1, \ldots, \pi_n >$ be an ordering of the elements of S. The* boundary strata *of $\bar{J}$, relative to π, is an ordered set of subsets $< B_{(1)}, B_{(2)}, \ldots, B_{(m)} >$ of S (with $m \leq n$), such that each $B_{(i)}$ is a minimal set satisfying $B_{(i)} \subseteq S_{(i)} = \{\pi_1, \ldots, \pi_{i-1}\}$ and $\gamma_i = (\{\pi_i\}, S_{(i)} \setminus B_{(i)}, B_{(i)}) \in \bar{J}$. The DAG obtained by setting each $B_{(i)}$ as parent set of the node π_i is called* boundary DAG *of J, relative to π.*

The introduced triple γ_i is known as *basic triple*.

The next theorem is an extension of Verma's Theorem [18] stated for conditional independence relations (see [14]).

Theorem 1. *Let J be a independence model closed with respect to the semi–graphoid properties. If G is a boundary DAG of J, relative to any ordering π, then G is a minimal I–map of J.*

Theorem 1 helps to build a DAG for an independence model $\bar{J}$ (induced by a probability P) given an ordering π on indices of S. It is well known (see [14]) that the boundary DAG of J relative to π is a minimal I–map. In the following, given an ordering π on S, G_π is the corresponding I–map of $\bar{J}$.

4 Perfect Map

In [2] we have introduced some sufficient conditions for the existence of a perfect map, given the fast closure J_*, and described the algorithm Backtrack which checks these conditions and, in the affirmative case, builds a perfect map. Since these conditions are only sufficient, this algorithm can fail also in the cases where a perfect map exists.

In [4] we have improved the previous result by introducing conditions which, under a suitable hypothesis, are necessary and sufficient for the existence of a perfect map. This *partial* characterization relies on some constraints among the triples of the set J_* and their components.

In this paper we provide a necessary and sufficient condition valid also in the case where the previously cited hypothesis fails (for the proof see [5]). These condition fully characterizes the ordering from which a perfect map can be built. An algorithm able to check this condition and, in the positive case, to find a perfect map will be described in the next section.

In the following, we review the procedure BN–DRAW introduced in [2], which builds the minimal I–map G_π of $\bar{J}$ (see Definition 2) given the fast closure J_* of J and an ordering π on S. This procedure is used by the algorithms described in [2] and in this paper.

Note that, given the fast closure set J_*, it is not possible to apply the standard procedure (see [11,14]), described in Definition 3, to draw an I–map. In fact, the basic triples, related to an arbitrary ordering π, might not be elements of J_*, but they could be just g–included in some triples of J_* (see Example in [2]).

However, in [2] we have shown that it is easy to find the basic triples in the fast closure by using the following result, where, as in the rest of the paper, $S_{(x)}$ denotes the set of elements of S preceding $x \in S$, with respect to a given ordering π.

Proposition 2. *Let J be a set of independence relations on S, J_* its fast closure and π an ordering on S. For each $x \in S$, the set*

$$\mathcal{B}_x = \{(\{x\}, B, C) \in S^{(3)} : B \cup C = S_{(x)},\ \exists\, \theta \in J_* \text{ with } (\{x\}, B, C) \sqsubseteq \theta\}$$

is not empty if and only if the basic triple $\gamma_x = (\{x\}, S_{(x)} \setminus B_{(x)}, B_{(x)})$ exists, and coincides with the unique maximal triple $\bar{\gamma}_x$ of $\mathcal{B}_x$.

In this paper, we describe a new version of BN–DRAW which uses the following operation. For each $\theta = (A, B, C) \in S^{(3)}$, let $X = (A \cup B \cup C)$ and for any $x \in S$ and $P \subseteq S$, define

$$\Pi(\theta, P, x) = \begin{cases} P \cap (A \cup C) & \text{if } C \subseteq P \subseteq X \text{ and } x \in A \\ P \cap (B \cup C) & \text{if } C \subseteq P \subseteq X \text{ and } x \in B \\ P & \text{otherwise.} \end{cases}$$

Algorithm 1. The set of parents of x

```
function PARENTS(x, P, K)
    pa ← P
    for all θ ∈ K do
        p ← Π(θ, P, x)
        if |p| < |pa| then pa ← p
    end for
    return pa
end function
```

The procedure BN–DRAW calls for each π_i the function PARENTS and uses its results as parent set of π_i.

Given π, BN–DRAW builds the minimal I–map G_π in linear time with respect to the cardinality m of J_* and the number of variables n. In fact, it is based on the function PARENTS which computes the set of parents of a given variable

Algorithm 2. DAG from J_* given an ordering π of S

```
function BN–DRAW(n, π, J*)
    P ← ∅
    G ← a graph with S as vertex set and no edges
    for i ← 2 to n do
        P ← P ∪ {π_{i−1}}
        pa ← PARENTS(π_i, P, J*)
        draw an arc in G from each index in pa to π_i
    end for
    return G
end function
```

in $O(m)$ steps. In each step, some set operations must be executed and this can be efficiently performed by using a compact representations for sets (e.g., as bit vectors). The space needed in memory by BN–DRAW is almost exclusively used to store the fast closure (see [4]).

The introduction of the function Π is important also for the definition of the necessary and sufficient condition for the existence of a p–map.

Theorem 2. *A set J_* is representable with a p–map if and only if there exists an ordering π such that for each $\theta = (A, B, C) \in J_*$, let $X = A \cup B \cup C$,*

C1 for each $c \in C$ such that $S_{(c)} \cap A \neq \emptyset$ and $S_{(c)} \cap B \neq \emptyset$, there exists a triple $\theta_c \in J_*$ such that $\Pi(\theta_c, S_{(c)}, c) \cap A = \emptyset$ or $\Pi(\theta_c, S_{(c)}, c) \cap B = \emptyset$;

C2 for each $a \in A$ such that $S_{(a)} \cap B \neq \emptyset$ or $S_{(a)} \cap (S \setminus X) \neq \emptyset$ there exists a triple $\theta_a \in J_*$ such that $\Pi(\theta_a, S_{(a)}, a) \cap [B \cup (S \setminus X)] = \emptyset$;

C3 for each $b \in B$ such that $S_{(b)} \cap A \neq \emptyset$ or $S_{(b)} \cap (S \setminus X) \neq \emptyset$ there exists a triple $\theta_b \in J_*$ such that $\Pi(\theta_b, S_{(b)}, b) \cap [A \cup (S \setminus X)] = \emptyset$;

C4 for each $c \in C$ such that $S_{(c)} \cap (S \setminus X) \neq \emptyset$, there exists a triple $\theta'_c \in J_*$ such that $\Pi(\theta'_c, S_{(c)}, c) \cap (S \setminus X) = \emptyset$.

Proof. We give a sketch of the proof, for a complete proof see [5].

($\Rightarrow$) Suppose that G_π is a p–map for J_*, we need to prove that π satisfies the condition *C1* (the other conditions follow similarly). Let $\theta = (A, B, C)$ be in J_*, if *C1* were not satisfied, then there would exists an element $c \in C$, such that $S_{(c)} \cap A \neq \emptyset$ and $S_{(c)} \cap B \neq \emptyset$. However, for any $\theta' \in J_*$ ones has $\Pi(\theta', S_{(c)}, c) \cap A \neq \emptyset$ and $\Pi(\theta', S_{(c)}, c) \cap B \neq \emptyset$. Hence, there exists $\alpha \in pa(c) \cap A$ and $\beta \in pa(c) \cap B$, so the path $\alpha \to c \leftarrow \beta$ would not be blocked by C. This is absurd since A is d–separated from B by C.

($\Leftarrow$) Conditions *C1*–*C4* imply that for each $x \in X$, $pa(x) \subseteq X$. Let $\rho = (u_1, \ldots, u_l)$ be a path and consider $j = \max\{i : u_i \in A\}$ and $l = \min\{i : u_i \in B\}$. Then, $j + 1 \leq l - 1$, otherwise there would be an element of A having parents in B or vice versa. If $u_{j+1} \in pa(u_j)$, then $u_{j+1} \in C$ and, since it is not a collider, it blocks ρ. Similarly, if $u_{l-1} \in pa(u_l)$. Now, suppose that $u_{j+1} \in ch(u_j)$ and $u_{l-1} \in ch(u_l)$, let r be such that any u_i $(i = j, \ldots, l)$ precedes (according to π) u_r. Thus, u_r is a collider. If $u_r \in C$, then $j + 1 = r = l - 1$ cannot be, otherwise

u_r would have parents both in A and in B. So, u_{r-1} or u_{r+1} is a parent of u_r, belongs to C and blocks ρ. Otherwise, no descendent of u_r belongs to C, so u_r blocks ρ. □

The conditions *C1–C4* are not so easy to check from the computational point of view, because they require for each triple in J_* and for $x \in X$ to verify some constraints and, when some of them do not hold, a suitable triple in J_* needs to be found. In the worst case, this process requires $O(m^2)$ steps, for each possible ordering. In the next section we will describe a more efficient way of achieving the same result.

5 The Algorithm

In this section we show how to use Theorem 2 to check whether J_* is representable by a graph, and in the affirmative case to find a perfect map.

Algorithm 3. Main function for representability

```
function REPRESENT(J_*)
    PREPROCESS(J_*)
    return SEARCH([], 1, S, J_*)
end function
```

The main procedure is REPRESENT where [ ] denotes an empty sequence of integers. The function PREPROCESS will be described in the following.

The recursive function SEARCH incrementally tries to build an ordering π satisfying conditions *C1–C4* of Theorem 2. It returns the element $\perp$ if it fails into finding such an ordering. At the i–th recursive call it attempts to fix the i–th element in π, by selecting each of the remaining variables. For each possible variable x, the procedure CHECK–CONDS checks whether the conditions *C1–C4* are not violated by setting π_i as x. In the positive case, it calls itself until a complete ordering is obtained. If no variable can be set at the i–th place of π, then the recursive call fails and a revision of the previously chosen variables is performed (backtracking).

To check whether the choice of π_i as x is correct we must verify whether the conditions *C1–C4* are satisfied for all the triples in which x appears. Note that we know all the variables preceding x, in fact $S_{(x)}$ is exactly the set $\{\pi_1, \pi_2, \ldots, \pi_{i-1}\}$. Hence, it is possible to compute the set of parents Q of x in the graph candidate to be a perfect map.

Let $\theta = (A, B, C)$ be a triple containing x. If x appears in C, then only conditions *C1* and *C4* must be checked. Let us see how to handle condition *C1*. It basically requires that if P intersects both A and B, there must exist a triple $\tau \in K$ such that $\Pi(\tau, S_{(x)}, x)$ does not intersect both A and B.

But, since we know that the set Q, the parents of x, is the smallest set among the sets $\Pi(\tau, S_{(x)}, x)$, for $\tau \in K$, then it is sufficient to check if Q does not intersect A and B at the same time. For conditions *C2*, *C3*, and *C4* the situation

Algorithm 4. Backtracking procedure

```
function SEARCH(π, i, V, K)
    if V = ∅ then
        return BN–DRAW(π, K)
    else
        for all x ∈ V do
            π_i ← x
            if CHECK–CONDS(π, i, K) then
                G ← SEARCH(π, i + 1, V \ {x}, K)
                if G ≠ ⊥ then return G
            end if
        end for
        return ⊥
    end if
end function
```

is much easier. In fact, before starting the search process, we can compute, first of all, for each $x \in S$ the set $NP(x)$ of *non–parents*, i.e. those elements of S which cannot be parents of x, otherwise one of the conditions *C2, C3, or C4* would be violated, by means of the function PREPROCESS. Hence, to check the above mentioned conditions it is sufficient to verify whether Q does not intersect $NP(x)$. Unfortunately, this preprocessing cannot work for condition *C1*.

Algorithm 5. Preprocessing for conditions C2–C4

```
function PREPROCESS(K)
    for all x ∈ S : NP(x) ← ∅
    for all θ = (A, B, C) ∈ K do
        X ← A ∪ B ∪ C
        R ← S \ X
        for all x ∈ A : NP(x) ← NP(x) ∪ B ∪ R
        for all x ∈ B : NP(x) ← NP(x) ∪ A ∪ R
        for all x ∈ C : NP(x) ← NP(x) ∪ R
    end for
end function
```

The cost of the entire procedure can be estimated as follows. Let us recall that n is the number of the variables and m is the cardinality of J_*.

The function CHECK–CONDS requires at most $O(m)$ steps. The number of the steps of SEARCH is in the worst case exponential in n, but backtracking can hopefully perform an early pruning on not promising orderings, so to avoid many useless computation steps. A great impact, as in other backtracking procedures, is given by the order in which the variable are chosen in the instruction for all $x \in V$. We will discuss this point in the conclusion. Finally, note that SEARCH can avoid at all to call BN–DRAW by storing, for each $x \in S$, the sets Q computed in the function CHECK–CONDS.

Algorithm 6. Checking conditions C1–C4

```
function CHECK–CONDS(π, i, K)
    P ← π[1, ..., i − 1]
    x ← π_i
    Q ← PARENTS(x, P, K)
    if NP(x) ∩ Q ≠ ∅ then
        return FALSE
    end if
    for all θ = (A, B, C) ∈ K do
        if (x ∈ C) ∧ (Q ∩ A ≠ ∅) ∧ (Q ∩ B ≠ ∅) then return FALSE
    end for
    return TRUE
end function
```

6 Conclusions

We provide a necessary and sufficient condition for the existence of a perfect map representing a set of conditional independence relations and we provide an algorithm which finds a perfect map when it exists.

This algorithm can be improved in many ways. First of all, by using suitable data structures we can reduce the time for searching variables occurring in set of triples, for instance representing J_* as a bipartite graph, where each variable is linked to the triples in which appears and each triple is linked to the variables which contains.

Second, we will investigate the use of some heuristic rules that help the procedure SEARCH. For instance, the well known CSP techniques, like *fail–first* or *min–conflicts*, could be used to order the variables and to reduce the number of attempts. A simple way to have a sort of the fail–first heuristic is to choose the variables in decreasing order with respect to their corresponding $|NP(x)|$. Other useful CSP technique could be a non–chronological backtracking, in which the cause of the failure is detected and all the choices, which led to the failure, are undone. Moreover, another technique is learning, in which the forbidden ordering constraints are learned from the failures.

Third, we could introduce a further preprocessing phase, in which it would be possible to deduce, from the triple of the fast closure, a list of impossible ordering constraints among variables.

Another aspect that will be worth to be investigated is when a set J_* is not representable by a p–map, how to determine a subset J' of J, hopefully as large as possible, such that J'_* is representable.

Finally, another possible way of enhancing this result is to find a new characterization for the existence of a p–map, which can generate a faster algorithm.

References

1. Baioletti, M., Busanello, G., Vantaggi, B.: Conditional independence structure and its closure: inferential rules and algorithms. Int. J. of Approx. Reason. 50, 1097–1114 (2009)
2. Baioletti, M., Busanello, G., Vantaggi, B.: Acyclic directed graphs to represent conditional independence models. In: Sossai, C., Chemello, G. (eds.) ECSQARU 2009. LNCS (LNAI), vol. 5590, pp. 530–541. Springer, Heidelberg (2009)
3. Baioletti, M., Busanello, G., Vantaggi, B.: Closure of independencies under graphoid properties: some experimental results. In: Proc. 6th Int. Symp. on Imprecise Probability: Theories and Applications, pp. 11–19 (2009)
4. Baioletti, M., Busanello, G., Vantaggi, B.: Acyclic directed graphs representing independence models. Int. J. of Approx. Reason (2009) (submitted)
5. Baioletti, M., Busanello, G., Vantaggi, B.: Necessary and sufficient conditions for the existence of a perfect map. Tech. Rep. 01/2010. Univ. of Perugia (2010)
6. Coletti, G., Scozzafava, R.: Probabilistic logic in a coherent setting. Kluwer, Dordrecht (2002) (Trends in logic n.15)
7. Cozman, F.G., Seidenfeld, T.: Independence for full conditional measures, graphoids and Bayesian networks, Boletim BT/PMR/0711 Escola Politecnica da Universidade de Sao Paulo, Sao Paulo, Brazil (2007)
8. Cozman, F.G., Walley, P.: Graphoid properties of epistemic irrelevance and independence. Ann. of Math. and Art. Intell. 45, 173–195 (2005)
9. Dawid, A.P.: Conditional independence in statistical theory. J. Roy. Stat. Soc. B 41, 15–31 (1979)
10. de Waal, P.R., van der Gaag, L.C.: Stable Independence in Perfect Maps. In: Proc. 21st Conf. Uncertainty in Artificial Intelligence, UAI 2005, Edinburgh, pp. 161–168 (2005)
11. Jensen, F.V.: An introduction to Bayesian networks. UCL Press, Springer (1996)
12. Lauritzen, S.L.: Graphical models. Clarendon Press, Oxford (1996)
13. Moral, S., Cano, A.: Strong conditional independence for credal sets. Ann. of Math. and Art. Intell. 35, 295–321 (2002)
14. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, Los Altos (1988)
15. Studený, M.: Semigraphoids and structures of probabilistic conditional independence. Ann. of Math. Artif. Intell. 21, 71–98 (1997)
16. Studený, M.: Probabilistic conditional independence structures. Springer, London (2005)
17. Vantaggi, B.: Conditional independence structures and graphical models. Int. J. Uncertain. Fuzziness Knowledge-Based Systems 11(5), 545–571 (2003)
18. Verma, T.S.: Causal networks: semantics and expressiveness. Tech. Rep. R–65, Cognitive Systems Laboratory, University of California, Los Angeles (1986)
19. Verma, T.S., Pearl, J.: Equivalence and synthesis of causal models. Uncertainty in Artificial Intelligence 6, 220–227 (1991)
20. Witthaker, J.: Graphical models in applied multivariate statistic. Wiley & Sons, New York (1990)

An Empirical Study of the Use of the Noisy-Or Model in a Real-Life Bayesian Network

Janneke H. Bolt and Linda C. van der Gaag

Department of Information and Computing Sciences, Utrecht University,
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
{janneke,linda}@cs.uu.nl

Abstract. The use of the noisy-OR model is advocated throughout the literature as an approach to lightening the task of obtaining all probabilities required for a Bayesian network. Little evidence is available, however, as to the effects of using the model on a network's performance. In this paper, we construct a noisy-OR version of a real-life hand-built Bayesian network of moderate size, and compare the performance of the original network with that of the constructed noisy-OR version. Empirical results from using the two networks on real-life data show that the performance of the original network does not degrade by using the noisy-OR model.

1 Introduction

When building a Bayesian network, the task of obtaining all probabilities required is generally acknowledged to be the most daunting among the engineering tasks involved [1]. A Bayesian network of realistic size easily requires hundreds and sometimes even thousands of probabilities for its conditional probability tables. While for some application domains these probabilities are readily available or can be estimated from data, for other domains they need to be provided by domain experts. Over the last decades, researchers have taken two different approaches to lightening this quantification task. On the one hand, researchers have focused on methods and tools to support the elicitation of many probabilities from domain experts in little time [2]. On the other hand, researchers have designed causal independency models to support the quantification [3,4].

A causal independency model is a parameterised specification of a conditional probability table for a Bayesian network. It requires the assessment of just a small number of conditional probabilities for a table under construction; these probabilities are the model's parameters. The other probabilities required to arrive at a fully specified table then are computed from the model's parameters through simple mathematical functions. Well-known examples of causal independency models are the noisy-OR model which was originally devised by J. Pearl [5], and its generalisation, the noisy-MAX model [6].

To provide for parameterisation, causal independency models assume a specific type of interaction of the causes of a common effect. Before using such a model for a network under construction, therefore, a network engineer has to

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 11–20, 2010.

verify whether the assumed interaction actually holds in reality. Yet, the assumption's validity is not always easily checked; engineers, moreover, may not always be acquainted with the full details of the interaction assumption underlying the model. As a consequence, causal independency models are used in practice where they may not be entirely appropriate. Little is known about the effects of simply using these models on the performance of a Bayesian network in a real-life setting, however. Some researchers studied how using noisy-MAX approximations influence posterior probabilities computed from three different Bayesian networks [7]. Using randomly generated evidence, they found that the posterior probabilities computed from the noisy-MAX versions quite closely approximated the ones computed from the original networks. In another study, the noisy-OR model was used upon learning conditional probability tables from data [8]. The researchers found that the performance of the original network was improved upon by using noisy-OR approximations instead of uniform distributions for the tables for which sufficient data were lacking.

In this paper, we present the results from an empirical study of the effects of applying the noisy-OR model without verification of its underlying assumption. For the study, we used our CSF network for the early detection of Classical Swine Fever (CSF) in pigs, which was constructed and quantified in close collaboration with two domain experts. This network includes 32 stochastic variables, for which over 1100 conditional probabilities are specified; 470 of these probabilities were estimated directly by one of the experts. The basic idea of the study was to substitute new probability tables for the expert-provided ones, where these new tables were obtained from applying the leaky noisy-OR model, and to compare the performance of the original network with that of its leaky noisy-OR version. More specifically, the conditional probability tables of 11 of the 32 variables in total were replaced by noisy-OR approximations, which resulted in 122 of the 470 directly estimated probabilities being replaced by computed ones. Since the performance of a Bayesian network for diagnostic reasoning in biomedical applications is commonly expressed in terms of its sensitivity and specificity, we studied the performances of the two versions of the CSF network in terms of these characteristics, using real data collected from veterinary practice and from experimental infections. We found that use of the noisy-OR model had little effect on the network's performance, even though some of the substituted probability tables differed substantially from the original expert-provided ones.

Even though little is known about the effects of their inadvertent application, the use of causal independency models, and of the noisy-OR and noisy-MAX models more specifically, is advocated throughout the literature as an approach to substantially lighten the task of obtaining all probabilities required for a Bayesian network. Although more fundamental research is required to corroborate our findings, the results from our empirical study warrant the cautious conclusion that the noisy-OR model can indeed be applied without extensive knowledge elicitation efforts for Bayesian networks for diagnostic applications.

The paper is organised as follows. In Section 2, we briefly describe our CSF network. Section 3 illustrates the noisy-OR model and its use in Bayesian networks

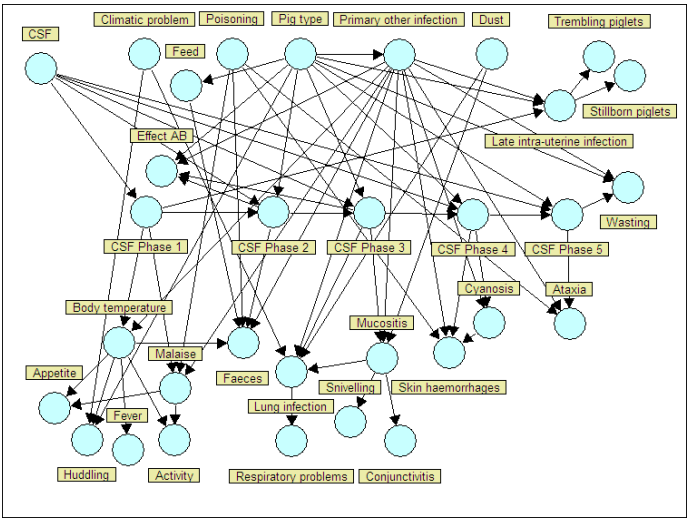


Fig. 1. The graphical structure of the CSF network for the early detection of Classical Swine Fever in individual pigs

in general. In Section 4, we describe the construction of a leaky noisy-OR version of the CSF network. The performances of the two networks are compared in Section 5. The paper ends in Section 6 with our concluding observations.

2 A Bayesian Network for Classical Swine Fever

A Bayesian network describes a joint probability distribution over a collection of stochastic variables. The variables and the qualitative (in)dependency relationships between them are modelled by nodes and arcs respectively, in a graphical structure. The strengths of the dependencies between the variables are expressed through probabilities. More specifically, for each variable X in the graphical structure, a conditional probability table is specified which describes the probability distributions $\Pr(X \mid \mathbf{pa}(X))$ over the values of X for each possible combination of values for the parents $\mathbf{pa}(X)$ of X in the structure; note that a conditional probability table thus specifies an exponential number of probability distributions, that is, exponential in the number of parents involved.

As an example, Figure 1 depicts the graphical structure of our CSF network for the early detection of Classical Swine Fever in pigs. Classical Swine Fever is a highly infectious viral disease of pigs which has a tendency to spread rapidly, both within a herd and between herds, and which can develop quite aggressively with large proportions of affected animals dying. Because of the major socio-economical consequences that an outbreak may have, reducing the time between first infection of a herd and first detection is of major importance. In close collaboration with experts from the Central Veterinary Institute in the

Table 1. The conditional probability table of the variable *Appetite* in the CSF network; the boldfaced probabilities were provided directly by our domain expert

Body temperature	*Malaise*	*Appetite*	
		reduced	normal
elevated	yes	**0.9000**	0.1000
	no	**0.2500**	0.7500
normal	yes	**0.8500**	0.1500
	no	**0.0050**	0.9950

Netherlands, we developed the depicted network to support veterinarians in the early detection of the disease. The network takes for its input the clinical signs observed in an individual pig and returns the posterior probability of these signs being caused by a CSF infection.

The CSF network currently includes 32 stochastic variables. To describe the strengths of the dependencies among these variables, the network's graphical structure is supplemented with over 1100 probabilities, organised in conditional probability tables. As an example, Table 1 shows the conditional probability table for the variable *Appetite*, which models whether or not a pig has a reduced appetite. Since the probabilities from a single distribution sum to 1, only some of the probabilities from the table were estimated explicitly by our domain expert; these probabilities are shown in boldface. All in all, 470 of the network's more than 1100 probabilities were estimated directly by one of the experts. We would like to note that some of the network's dependency relationships are deterministic, which accounts for the observation that fewer than half of the probabilities were expert provided. We further note that we will specify the probabilities in all tables throughout the paper with a precision of four decimals, since the domain expert provided his probability assessments up to this high precision.

3 The Noisy-OR Model

For a Bayesian network of realistic size, hundreds or even thousands of conditional probabilities are required. One of the approaches to lightening the task of obtaining all these probabilities is the use of causal independency models. A causal independency model is a parameterised specification of a conditional probability table. The model requires the assessment of just a small number of conditional probabilities to arrive at a fully specified probability table; these probabilities are the model's parameters. The other probabilities for the table are computed from these parameters through simple functions. Several researchers have studied different types of causal independency model along with their properties, and by now an entire family of models have been described [3,4].

The best-known causal independency model is the noisy-OR model for binary variables [5]. This model pertains to a causal mechanism composed of an effect variable E and parent variables C_i, $i = 1, \ldots, n$, $n \geq 2$, which model possible

Table 2. The conditional probability tables that result from application of the noisy-OR model (*left*) and the leaky noisy-OR model (*right*), respectively, for the variable *Appetite* in the CSF network; the parameters for the models are shown in boldface

Body temp.	*Malaise*	*Appetite* reduced	*Appetite* normal
elevated	yes	0.8875	0.1125
	no	**0.2500**	0.7500
normal	yes	**0.8500**	0.1500
	no	0.000	1.000

Body temp.	*Malaise*	*Appetite* reduced	*Appetite* normal
elevated	yes	0.8881	0.1119
	no	**0.2500**	0.7500
normal	yes	**0.8500**	0.1500
	no	**0.0050**	0.9950

causes of the effect. The effect variable has values e for the effect being present and $\bar{e}$ for the effect being absent; each cause variable C_i has the values c_i and $\bar{c}_i$ for the presence and absence of the cause, respectively. The variable *Appetite* in the CSF network constitutes an example of such a causal mechanism, along with its parent variables *Body temperature*, which models an elevated body temperature as a possible cause for a reduced appetite, and *Malaise*, modelling a sense of malaise as another cause for such a finding.

The noisy-OR model now provides a parameterised probability table for the effect variable E given its parent variables in the causal mechanism. The parameters of the model are the probabilities $\Pr(e \mid \bar{c}_1, \ldots, \bar{c}_{j-1}, c_j, \bar{c}_{j+1}, \ldots, \bar{c}_n)$ of the effect occurring in the presence of a single cause c_j. Note that the model thus has n parameters, that is, the number of parameters to be assessed explicitly is linear in the number of possible causes of the modelled effect. The probability $\Pr(e \mid \bar{c}_1, \ldots, \bar{c}_n)$ of the effect occurring spontaneously in the absence of all causes, is taken to be zero by the noisy-OR model. To arrive at a fully specified probability table for the variable E, the conditional probabilities of the effect e occurring given all possible combinations of values $\mathbf{c}$ involving multiple causes are taken to be

$$\Pr(e \mid \mathbf{c}) = 1 - \prod_{j \in \mathcal{J}} (1 - \Pr(e \mid \bar{c}_1, \ldots, \bar{c}_{j-1}, c_j, \bar{c}_{j+1}, \ldots, \bar{c}_n))$$

where $\mathcal{J}$ is the set of indices of the cause variables C_j which have the value c_j in the combination of values $\mathbf{c}$ under consideration. As an example, Table 2 shows, on the left, the conditional probability table that would result from application of the noisy-OR model for the variable *Appetite* in the CSF network.

The noisy-OR model assumes a disjunctive interaction of the causes of a common effect, which implies that the properties of exception independence and accountability are assumed to hold. The property of exception independence states that the presence of any single cause essentially suffices to produce the effect and that the hidden processes that may inhibit the occurrence of the effect are mutually independent. The property of accountability states that the effect of the causal mechanism is absent if none of the possible causes is present.

For many causal mechanisms in practice, the assumption of accountability underlying the noisy-OR model is not met: the effect of these mechanisms can occur even if all modelled causes are absent, that is, $\Pr(e \mid \bar{c}_1, \ldots, \bar{c}_n) \neq 0$. If the

property of exception independence does hold for such a mechanism, the leaky noisy-OR model can be employed for constructing a conditional probability table for the effect variable. This model is closely related to the noisy-OR model yet includes the probability $\Pr(e \mid \bar{c}_1, \ldots, \bar{c}_n)$ as a leak probability. In the literature, two different views of the leaky noisy-OR model have been described, resulting in two variants of the model. The first view assumes that all estimated parameter probabilities, given each cause separately, implicitly include the leak probability [9,10]. The second view assumes that the leak probability needs to be explicitly assessed and incorporated into all other parameter probabilities [6]. It has been argued that the first view applies to probabilities derived from data and that the second view is more appropriate for probabilities provided by experts [3]. Since the CSF network used in our current study was constructed by hand and all probabilities were provided by an expert, we focus here on the second view. With this view, the leaky noisy-OR model has $n+1$ parameters; these are the n conditional probabilities $\Pr(e \mid \bar{c}_1, \ldots, \bar{c}_{j-1}, c_j, \bar{c}_{j+1}, \ldots, \bar{c}_n)$ from the noisy-OR model and the leak probability $\Pr(e \mid \bar{c}_1, \ldots, \bar{c}_n)$. To arrive at a fully specified probability table for the effect variable E, the conditional probabilities of e for all combinations of values $\mathbf{c}$ involving multiple causes are taken to be

$$\Pr(e \,|\, \mathbf{c}) = 1 - (1 - \Pr(e \,|\, \bar{c}_1, \ldots, \bar{c}_n)) \cdot \prod_{j \in \mathcal{J}} (1 - \Pr(e \,|\, \bar{c}_1, \ldots, \bar{c}_{j-1}, c_j, \bar{c}_{j+1}, \ldots, \bar{c}_n))$$

where $\mathcal{J}$ again denotes the set of indices of the cause variables C_j which have the value c_j in the combination of values $\mathbf{c}$ under consideration.

The noisy-OR model being specified for binary variables only, researchers have introduced the noisy-MAX model as a generalisation to multiple-valued variables [6,9]. A general introduction to this model is beyond the scope of the present paper. In the sequel, however, we exploit the property that for a causal mechanism involving a binary effect variable and multiple-valued cause variables, the computation rule of the (leaky) noisy-MAX model is a simple generalisation of the function of the (leaky) noisy-OR model. As an example from the CSF network, we consider the causal mechanism involving the effect variable *Body temperature* (B), modelling whether or not a pig has an increased body temperature, and the cause variables *Primary other infection* (P), modelling the presence of an infection other than a CSF infection, and *CSF Phase 1* (C), modelling whether or not a pig has entered the first phase of an infection with the CSF virus. The variables *Body temperature* and *CSF Phase 1* are binary; the variable *Primary other infection* has the four values none ($\bar{p}$), respiratory infection (p_1), intestinal infection (p_2), and respiratory+intestinal infection (p_3). The leaky noisy-MAX model for this mechanism has for its parameters the conditional probabilities $\Pr(b \mid \bar{p}c)$, $\Pr(b \mid p_i\bar{c})$, $i = 1, 2, 3$, and the leak probability $\Pr(b \mid \bar{p}\bar{c})$. The probability $\Pr(b \mid p_1 c)$ of an increased body temperature given both a respiratory and a first-phase CSF infection now is computed to be

$$\Pr(b \mid p_1 c) = 1 - (1 - \Pr(b \mid \bar{p}\bar{c})) \cdot (1 - \Pr(a \mid p_1\bar{c})) \cdot (1 - \Pr(a \mid \bar{p}c))$$

In our empirical study of the effects of using the leaky noisy-OR model in the CSF network, we exploited this simple function for multiple-valued cause variables.

4 Constructing a Leaky Noisy-OR Version

As described above, the leaky noisy-OR model assumes the property of exception independence to hold in an application domain, which may be hard to verify in practice. In this paper, we empirically study the effect of simply applying the leaky noisy-OR model, without proper verification of its underlying assumption, on a network's performance. For this purpose, we constructed a leaky noisy-OR version of the CSF network and compared the performances of the two networks.

From the original network, we selected all variables that could be considered effect variables in a causal mechanism. With this criterion, a total of 11 variables were selected, which are listed in Table 3. For the selected variables, we substituted new conditional probability tables which were constructed using the leaky noisy-OR model, for the original ones. For the model's parameters, we used the probabilities as originally assessed by our expert; we thus did not construct best-fitting approximations of the original tables. While for most of the selected variables very small Kullback-Leibler distances were found between the originally specified probability distributions and the leaky noisy-OR tables, for some of the variables these distances proved to be of considerable size; the fourth column of Table 3 records the maximum Kullback-Leibler distance found for each of the selected variables. The second and third columns of the table specify, for each variable separately, the numbers of probabilities which have to be assessed directly for the full probability table and for the leaky noisy-OR table, respectively. Table 3 shows that by using the leaky noisy-OR model, the number of probabilities to be specified explicitly by the domain expert, would be reduced from 220 to 98. Since for the entire network a total of 470 parameters

Table 3. The numbers of parameters of the original conditional probability tables (CPT) and their leaky noisy-OR versions, and the maximum Kullback-Leibler distance found between the original and noisy-OR distributions for selected variables of the network

Variable	*Number of CPT parameters needed*	*Number of noisy-OR parameters needed*	*max. KL distance*
Lung infection	64	8	0.0168
Wasting	40	25	0.0006
Skin haemorrhages	32	17	0.0009
Huddling	20	15	0.0050
Malaise	16	6	0.0019
Mucositis	16	6	0.2206
Body temperature	8	5	0.0052
Cyanosis	8	5	0.0000
Late intra-uterine inf.	8	5	0.0005
Appetite	4	3	0.0007
Activity	4	3	0.0138
Total	220	98	

had to be estimated, this number is reduced by more than 25% by exploiting the leaky noisy-OR model for one-third of the variables.

5 Comparing the Performances of the CSF Networks

The performance of a Bayesian network for diagnostic reasoning in biomedical applications is commonly expressed in terms of its sensitivity and specificity characteristics. The network's specificity then is the percentage of cases without the disease under study whom the network identifies as not having the disease; its sensitivity is the percentage of cases with the disease whom the network singles out as having the disease. Since the CSF network chosen for our empirical study was designed for diagnostic reasoning, we use these two characteristics for comparing the performances of the original network and its noisy-OR version.

For establishing the specificity of the two networks, field data from 375 pigs without Classical Swine Fever were used. For each of these pigs, the posterior probability of the observed clinical signs being caused by a CSF infection was computed from both networks. The computed probabilities were subsequently compared against a threshold probability α. If a computed probability exceeded this threshold probability, a suspicion of CSF was issued. Based upon the numbers of generated suspicions, the specificities of both networks could be readily established. Table 4 records these specificities for various realistic values of α.

The sensitivities of the two networks were estimated using experimental data. These data were collected from small groups of pigs in which some individuals had been inoculated with the CSF virus. A total of 91 animals were followed over a period of up to 35 days. Data were recorded at least every two or three days; the recording days were expressed in terms of the number of days after infection of the inoculated animals. For each recording day, for each pig, the posterior probability of the observed clinical signs being caused by a CSF infection was computed from both CSF networks and compared against a threshold probability α as before. Figure 2 shows, as an example, the cumulative number of animals for which a CSF suspicion was issued by the original network as a function of the number of days post infection, using $\alpha = 0.001$; with the leaky noisy-OR version of the network, the exact same results were found. With the threshold values of 0.0005, 0.005, and 0.01, we also found no differences in sensitivity between the

Table 4. The specificities of the CSF network and its leaky noisy-OR version, respectively, for different values of the threshold probability α for issuing a CSF suspicion

threshold α	*specificity CSF network*	*specificity noisy-OR version*
0.05	99%	99%
0.01	98%	98%
0.005	96%	96%
0.001	89%	88%
0.0005	84%	84%

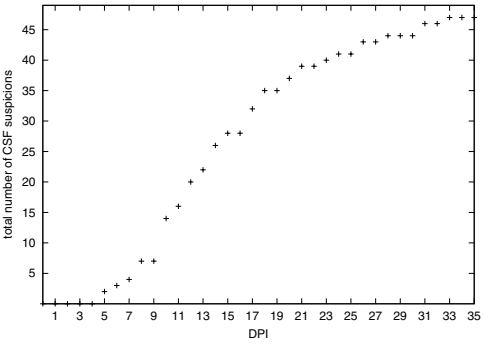


Fig. 2. The cumulative number of CSF suspicions issued by the original CSF network for 91 pigs, using $\alpha = 0.001$, as a function of the number of days post infection (DPI) of the inoculated animals

two networks. With $\alpha = 0.05$, the original network issued a CSF suspicion for one more animal than did the leaky noisy-OR version of the network.

From the sensitivities and specificities reviewed above, we may conclude that the overall performance of the original CSF network does not degrade when the conditional probability tables of one-third of the variables are replaced by leaky noisy-OR tables. These characteristics, however, might hide differences in performance of the two networks for individual cases, which could indicate a fundamental problem of the leaky noisy-OR version. We therefore also studied, using the same threshold values, the networks' performance on all pig cases individually. With $\alpha = 0.005$, the leaky noisy-OR version of the network issued a false suspicion for one of the pigs which did not receive such a suspicion from the original network; it further did not issue a suspicion for one of the pigs which did receive a false suspicion from the original network. Using probability thresholds of 0.001 and 0.005, the leaky noisy-OR version of the CSF network issued an additional false suspicion on the one hand, yet also resulted in three fewer false suspicions on the other hand. With respect to the networks' sensitivities, we found that, with $\alpha = 0.05$, the leaky noisy-OR version failed to detect one of the cases that was detected by the original network. No further differences in performance were found. From these observations, we conclude that also in view of the 466 individual pig cases, the performance of the original CSF network is hardly affected by substituting leaky noisy-OR tables for its expert-provided conditional probability tables.

6 Concluding Observations

The use of causal independency models, and of the noisy-OR and noisy-MAX models more specifically, is advocated throughout the literature as an approach to substantially lightening the task of obtaining all probabilities required for a Bayesian network. Since these models assume a disjunctive interaction of the causes of a common effect, a network engineer has to verify, before using such a

model, that the assumed interaction actually holds in reality. In practice, however, the assumption is hard to verify. Thus far, little evidence had been gathered about the effects of simply using these models without elaborate verification efforts, on the performance of a real-life network in practice. In this paper, we presented the results from an empirical study of these effects on a Bayesian network for the early detection of Classical Swine Fever in pigs. In this network, we substituted leaky noisy-OR tables for the expert-provided conditional probability tables for one-third of the variables. The performances of the original network and its leaky noisy-OR version on real-life data were investigated in terms of their sensitivity and specificity characteristics. The results of our study showed that using the leaky noisy-OR model had little effect on the performance of our CSF network. Although more research is required to corroborate our findings, a cautious conclusion from our study is that the (leaky) noisy-OR model can indeed be applied, without extensive knowledge elicitation efforts, to lighten the quantification task for Bayesian networks for diagnostic applications.

Acknowledgment. The research in this paper was supported by the Netherlands Organisation for Scientific Research.

References

1. Druzdzel, M.J., van der Gaag, L.C.: Building probabilistic networks: "Where do the numbers come from?" Guest editors' introduction. IEEE Transactions on Knowledge and Data Engineering 12, 481–486 (2000)
2. van der Gaag, L.C., Renooij, S., Witteman, C.L.M., Aleman, B., Taal, B.G.: How to elicit many probabilities. In: Laskey, K.B., Prade, H. (eds.) Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 647–654 (1999)
3. Díez, F.J., Druzdzel, M.J.: Canonical probabilistic models for knowledge engineering. Technical Report CISIAD-06-01 (2007)
4. Lucas, P.J.F.: Bayesian network modelling through qualitative patterns. Artificial Intelligence 163, 233–263 (2005)
5. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, Palo Alto (1988)
6. Díez, F.J.: Parameter adjustment in Bayes networks. The generalized noisy OR gate. In: Proceedings of the Nineth Annual Conference on Uncertainty in Artificial Intelligence, pp. 99–105 (1993)
7. Zagorecki, A., Druzdzel, M.J.: Knowledge engineering for Bayesian networks: How common are noisy-MAX distributions in practice? In: Brewka, G., Coradeschi, S., Perini, A., Traverso, P. (eds.) Proceedings of the Seventeenth European Conference on Artificial Intelligence, pp. 482–489 (2006)
8. Onisko, A., Druzdzel, M.J., Wasyluk, H.: Learning Bayesian network parameters from small data sets: Application of noisy-OR gates. International Journal of Approximate Reasoning 27(2), 165–182 (2001)
9. Henrion, M.: Some practical issues in constructing belief networks. In: Kanal, L.N., Levitt, T.S., Lemmer, J.F. (eds.) Uncertainty in Artificial Intelligence 3, pp. 161–173 (1989)
10. Zagorecki, A., Druzdzel, M.J.: An empirical study of probability elicitation under noisy-OR assumption. In: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, pp. 880–885 (2004)

Possibilistic Graphical Models and Compositional Models

Jiřina Vejnarová

Institute of Information Theory and Automation of the AS CR,
Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic
&
University of Economics, Prague
vejnar@utia.cas.cz

Abstract. We overview three kinds of possibilistic graphical models (based on directed acyclic graphs) and present, how they can be expressed by means of non-graphical approach to multidimensional models, so-called compositional models. We show that any of these graphical models can be transformed into a compositional model, but not vice versa. The only exception are directed possibilistic graphs, which are as general as so-called prefect sequences of low-dimensional distributions.

Keywords: Possibility distributions, graphical models, triangular norms.

1 Introduction

High dimensionality of problems usually solved in the field of artificial intelligence led in late 1980's to the emergence of new kind of models, usually called *graphical Markov models.* These models, sometimes characterized as a "marriage between probability and graph theories", utilize different types of graphs to express (in)dependences among variables.

Nevertheless, uncertainty can be modeled also by other calculi; among them we concentrated to possibility theory, which has in common with probability theory the advantage, that possibility measure can be expressed by means of possibility distribution. In this contribution we overview three kinds of possibilistic graphical models (based on directed acyclic graphs) and present, how they can be expressed by means of non-graphical approach to multidimensional possibilistic models, so-called *compositional models* — introduced already in [7] and further developed e.g. in [8,11].

The paper is organized as follows. After an overview of necessary notions form possibility theory in Section 2, in Section 3 we will present the most important results on compositional models. Section 4 will be devoted to the graphical models and their relationship to compositional models.

2 Basic Notions

The purpose of this section is to give, as briefly as possible, an overview of basic notions of De Cooman's measure-theoretical approach to possibility theory [3],

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 21–30, 2010.

necessary for understanding the paper. Special attention will be paid to conditioning, independence and conditional independence [9]. We will start with the notion of a triangular norm, since most notions in this paper are parameterized by it.

2.1 Triangular Norms

A *triangular norm* (or a *t-norm*) T is a nondecreasing, associative and commutative binary operator on $[0,1]$ satisfying the boundary condition: for any $a \in [0,1]$

$$T(1,a) = a.$$

A t-norm T is called *continuous* if T is a continuous function. Within this paper, we will only deal with continuous t-norms.

Let $x, y \in [0,1]$ and T be a t-norm. We will call an element $z \in [0,1]$ *T-inverse* of x w.r.t. y if

$$T(z,x) = T(x,z) = y. \tag{1}$$

It is obvious that if $x \leq y$ then the equation (1) admits no solution, i.e. there are no T-inverses of x w.r.t. y. On the other hand, if a T-inverse exists, it need not be unique. Nevertheless, we can obtain a unique representative (which is even maximal) using the notion of *T-residual* $y \triangle_T x$ of y by x defined for any $x, y \in [0,1]$ as

$$y \triangle_T x = \sup\{z \in [0,1] : T(z,x) \leq y\}.$$

From the viewpoint of this paper the following lemma proven in [5] is important, as it gives a hint, how to compute with residuals.

Lemma 1. *If $c \geq b$, then $T(a,b) \triangle_T c = T(a, b \triangle_T c)$.*

2.2 Possibility Measures and Distributions

Let $\mathbf{X}$ be a finite set called *universe of discourse* which is supposed to contain at least two elements. A *possibility measure* Π is a mapping from the power set $\mathcal{P}(\mathbf{X})$ of $\mathbf{X}$ to the real unit interval $[0,1]$ satisfying the following two requirements:

(i) $\Pi(\emptyset) = 0$;
(ii) for any family $\{A_j, j \in J\}$ of elements of $\mathcal{P}(\mathbf{X})$

$$\Pi(\bigcup_{j \in J} A_j) = \max_{j \in J} \Pi(A_j)^{[1]}.$$

Within this paper we will always assume that Π is *normal*, i.e. $\Pi(\mathbf{X}) = 1$.

For any Π there exists a mapping $\pi : \mathbf{X} \to [0,1]$, called a *distribution* of Π, such that for any $A \in \mathcal{P}(\mathbf{X})$, $\Pi(A) = \max_{x \in A} \pi(x)$. This function is a

[1] Max must be substituted by sup if $\mathbf{X}$ is not finite.

possibilistic counterpart of a density function in probability theory. It is evident that (in the finite case) Π is normal iff there exists at least one $x \in \mathbf{X}$ such that $\pi(x) = 1$. Throughout this paper we will use possibility distributions instead of possibility measures.

Let $\mathbf{X}_1$ and $\mathbf{X}_2$ denote two finite universes of discourse provided by possibility measures Π_1 and Π_2, respectively. The possibility measure Π on $\mathbf{X}_1 \times \mathbf{X}_2$ is called *T-product possibility measure* of Π_1 and Π_2 (denoted $\Pi_1 \times_T \Pi_2$) if for the corresponding possibility distributions for any $(x_1, x_2) \in \mathbf{X}_1 \times \mathbf{X}_2$

$$\pi(x_1, x_2) = T(\pi_1(x_1), \pi_2(x_2)). \tag{2}$$

Now, let us consider an arbitrary possibility distribution π defined on a product universe of discourse $\mathbf{X}_1 \times \mathbf{X}_2$. The *marginal possibility distribution* $\pi^{\downarrow 1}$ on $\mathbf{X}_1$ is defined by the expression

$$\pi^{\downarrow 1}(x_1) = \max_{x_2 \in \mathbf{X}_2} \pi(x_1, x_2) \tag{3}$$

for any $x_1 \in \mathbf{X}_1$.

2.3 Conditioning, Independence and Conditional Independence

Let T be a continuous t-norm on $[0, 1]$. The *conditional possibility distribution* $\pi_{X|_T Y}$ is defined (in accordance with [3]) as *any* solution of the equation

$$\pi_{XY}(x, y) = T(\pi_Y(y), \pi_{X|_T Y}(x|_T y)) \tag{4}$$

for any $(x, y) \in \mathbf{X} \times \mathbf{Y}$. Continuity of a t-norm T guarantees the existence of a solution of this equation. This solution is not unique (in general), but the ambiguity vanishes when almost-everywhere equality is considered (for more details see [3]). As mentioned in [3,9], this way of conditioning brings a unifying view on several conditioning rules and it also plays an important role in the definition of (conditional) independence, therefore its importance from the theoretical viewpoint is obvious. On the other hand, from the practical point of view, its expression by residual $\pi_{XY}(x, \cdot) \triangle_T \pi_Y(\cdot)$, i.e. the least specific (or maximal) solution of (4), is very useful (for more details see [11]).

Two variables X and Y (taking their values in $\mathbf{X}$ and $\mathbf{Y}$, respectively) are *possibilistically T-independent*[2] [3] if for any $x \in \mathbf{X}$ and $y \in \mathbf{Y}$

$$\pi_{XY}(x, y) = T(\pi_X(x), \pi_Y(y)). \tag{5}$$

In light of these facts, we defined the conditional possibilistic independence in the following way in [8]: Given a possibility measure Π on $\mathbf{X} \times \mathbf{Y} \times \mathbf{Z}$ with the respective distribution $\pi(x, y, z)$, variables X and Y are *possibilistically*

[2] Let us note that the definition presented in [3] is different and (5) is its equivalent characterization. Nevertheless, from the viewpoint of this paper (5) is more convenient.

conditionally T-independent[3] given Z (in symbols $I_T(X,Y|Z)$) if, for any pair $(x,y) \in \mathbf{X} \times \mathbf{Y}$,

$$\pi_{XYZ}(x,y,z) = T(T(\pi_{X|_T Z}(x|_T z), \pi_{Y|_T Z}(y|_T z)), \pi_Z(z)). \tag{6}$$

In [9] we proved its formal properties and studied its relationship with other definitions of conditional possibilistic independence, among others those introduced in [2].

3 Compositional Models

From now on, we will deal with joint possibility distributions π on Cartesian product of universes of discourse

$$\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n,$$

and their marginals $\pi^{\downarrow K}$ on its subspaces

$$\mathbf{X}_K = \bigtimes_{i \in K} \mathbf{X}_i.$$

3.1 Operators of Composition

Operators of composition of possibility distributions introduced in [7] are, in a way, a generalization of T-product possibility distributions defined by (2). Considering a continuous t-norm T, two subsets K_1, K_2 of $\{1,\ldots,n\}$ (not necessarily disjoint) and two normal possibility distributions $\pi_1(x_{K_1})$ and $\pi_2(x_{K_2})$ we define the *operator of right composition* of these possibilistic distributions by the expression

$$\pi_1\left(x_{K_1}\right) \triangleright_T \pi_2\left(x_{K_2}\right) = T\left(\pi_1\left(x_{K_1}\right), \pi_2\left(x_{K_2}\right) \triangle_T \pi_2^{\downarrow K_1 \cap K_2}\left(x_{K_1 \cap K_2}\right)\right),$$

and analogously the *operator of left composition* by the expression

$$\pi_1\left(x_{K_1}\right) \triangleleft_T \pi_2\left(x_{K_2}\right) = T\left(\pi_1\left(x_{K_1}\right) \triangle_T \pi_1^{\downarrow K_1 \cap K_2}\left(x_{K_1 \cap K_2}\right), \pi_2\left(x_{K_2}\right)\right).$$

It is evident that both $\pi_1 \triangleright_T \pi_2$ and $\pi_1 \triangleleft_T \pi_2$ are (generally different) possibility distributions of variables $\{X_i\}_{i \in K_1 \cup K_2}$.

Now, we will present two lemmata proven in [7], expressing basic properties of these operators.

Lemma 2. *Let T be a continuous t-norm and $\pi_1(x_{K_1})$ and $\pi_2(x_{K_2})$ be two distributions. Then*

$$(\pi_1 \triangleright_T \pi_2)^{\downarrow K_1}(x_{K_1}) = \pi_1(x_{K_1})$$

and

$$(\pi_1 \triangleleft_T \pi_2)^{\downarrow K_2}(x_{K_2}) = \pi_2(x_{K_2}).$$

[3] Let us note that a similar definition of conditional independence can be found in [4].

Lemma 3. *Consider two distributions* $\pi_1(x_{K_1})$ *and* $\pi_2(x_{K_2})$. *Then*

$$(\pi_1 \triangleright_T \pi_2)(x_{K_1 \cup K_2}) = (\pi_1 \triangleleft_T \pi_2)(x_{K_1 \cup K_2})$$

for any continuous t-norm T *iff* π_1 *and* π_2 *are projective, i.e.*

$$\pi_1^{\downarrow K_1 \cap K_2}(x_{K_1 \cap K_2}) = \pi_2^{\downarrow K_1 \cap K_2}(x_{K_2 \cap K_1}).$$

The following theorem proven in [8] reveals the relationship between conditional T-independence and operators of composition.

Theorem 1. *Let* T *be a continuous t-norm and* π *be a possibility distribution of* $X_{K_1 \cup K_2}$ *with marginals* π_1 *and* π_2 *of* X_{K_1} *and* X_{K_2}, *respectively. Then*

$$\begin{aligned} \pi(x_{K_1 \cup K_2}) &= (\pi_1 \triangleright_T \pi_2)(x_{K_1 \cup K_2}) \\ &= (\pi_1 \triangleleft_T \pi_2)(x_{K_1 \cup K_2}), \end{aligned} \tag{7}$$

if and only if $X_{K_1 \setminus K_2}$ *and* $X_{K_2 \setminus K_1}$ *are conditionally independent, given* $X_{K_1 \cap K_2}$.

3.2 Generating Sequences

In this section we will show how to apply the operators iteratively. Consider a sequence of possibility distributions $\pi_1(x_{K_1}), \pi_2(x_{K_2}), \ldots, \pi_m(x_{K_m})$ and the expression

$$\pi_1 \triangleright_T \pi_2 \triangleright_T \ldots \triangleright_T \pi_m.$$

Before beginning a discussion of its properties, we have to explain how to interpret it. Though we did not mention it explicitly, the operator $\triangleright_T$ (as well as $\triangleleft_T$) is neither commutative nor associative.[4] Therefore, generally

$$(\pi_1 \triangleright_T \pi_2) \triangleright_T \pi_3 \neq \pi_1 \triangleright_T (\pi_2 \triangleright_T \pi_3).$$

Nevertheless, under specific conditions this equality is satisfied. One of these situations, important from the viewpoint of this paper, is described by the following lemma.

Lemma 4. *Let* T *be a continuous t-norm and* π_1, π_2 *and* π_3 *be defined on* $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}$ *and* $\mathbf{X}_{K_3}$, *respectively, such that* K_1 *and* K_3 *are disjoint. Then*

$$(\pi_1 \triangleright_T \pi_2) \triangleright_T \pi_3 = \pi_1 \triangleright_T (\pi_2 \triangleright_T \pi_3). \tag{8}$$

Proof. Let $x \in \mathbf{X}_{K_1 \cup K_2 \cup K_3}$ then the right-hand side of (8) is by definition

$$\begin{aligned} &\pi_1 \triangleright_T (\pi_2 \triangleright_T \pi_3)(x) \\ &\quad = T(\pi_1(x_{K_1}), (\pi_2 \triangleright_T \pi_3)(x_{K_2 \cup K_3}) \triangle_T (\pi_2 \triangleright_T \pi_3)(x_{(K_2 \cup K_3) \cap K_1})) \\ &\quad = T(\pi_1(x_{K_1}), T(\pi_2(x_{K_2}), \pi_3(x_{K_3}) \triangle_T \pi_3(x_{K_3 \cap K_2})) \triangle_T \pi_2(x_{K_2 \cap K_1})) \\ &\quad = T(\pi_1(x_{K_1}), T(\pi_2(x_{K_2}) \triangle_T \pi_2(x_{K_2 \cap K_1}), \pi_3(x_{K_3}) \triangle_T \pi_3(x_{K_3 \cap K_2}))) \\ &\quad = T(T(\pi_1(x_{K_1}), \pi_2(x_{K_2}) \triangle_T \pi_2(x_{K_2 \cap K_1})), \pi_3(x_{K_3}) \triangle_T \pi_3(x_{K_3 \cap K_2})) \\ &\quad = (\pi_1 \triangleright_T \pi_2) \triangleright_T \pi_3(x), \end{aligned}$$

[4] Counterexamples can be found in [7].

where we used the fact that $K_1 \cap K_2 = \emptyset$, Lemma 2, Lemma 1 and associativity of a t-norm. □

For the above reason, let us note that in the part that follows, we always apply the operators from left to right, i. e.

$$\pi_1 \rhd_T \pi_2 \rhd_T \pi_3 \rhd_T \ldots \rhd_T \pi_m = (\ldots((\pi_1 \rhd_T \pi_2) \rhd_T \pi_3) \rhd_T \ldots \rhd_T \pi_m). \quad (9)$$

This expression defines a multidimensional distribution of $X_{K_1 \cup \ldots \cup K_m}$. Therefore, for any permutation $i_1, i_2, \ldots, i_m$ of indices $1, \ldots, m$ the expression

$$\pi_{i_1} \rhd_T \pi_{i_2} \rhd \ldots \rhd_T \pi_{i_m}$$

determines a distribution of the same family of variables, however, for different permutations these distributions can differ from one another. In the following paragraph we will deal with special generating sequences (or their special permutations), which seem to possess the most advantageous properties.

3.3 *T*-Perfect Sequences

An ordered sequence of possibility distributions $\pi_1, \pi_2, \ldots, \pi_m$ is said to be *T-perfect* if

$$\begin{aligned} \pi_1 \rhd_T \pi_2 &= \pi_1 \lhd_T \pi_2, \\ \pi_1 \rhd_T \pi_2 \rhd_T \pi_3 &= \pi_1 \lhd_T \pi_2 \lhd_T \pi_3, \\ &\vdots \\ \pi_1 \rhd_T \cdots \rhd_T \pi_m &= \pi_1 \lhd_T \cdots \lhd_T \pi_m. \end{aligned}$$

The notion of T-perfectness suggests that a sequence perfect with respect to one t-norm need not be perfect with respect to another t-norm, analogous to (conditional) T-independence. The following lemma, proven in [7], suggests that perfectness is a stronger property than pairwise projectivity (cf. Lemma 3).

Lemma 5. *Let T be a continuous t-norm. The sequence $\pi_1, \pi_2, \ldots, \pi_m$ is T-perfect, if and only if the pairs of distributions $(\pi_1 \rhd_T \cdots \rhd_T \pi_{k-1})$ and π_k are projective for all $k = 2, 3, \ldots, m$.*

Although T-perfect sequences may be defined for any continuous t-norm T, their semantics substantially differ from each other. For more details the reader is referred to [10].

The following characterization theorem proven in [11] expresses one of the most important results concerning T-perfect sequences. It says they compose into multidimensional distributions that are extensions of all the distributions from which the joint distribution is composed.

Theorem 2. *The sequence $\pi_1, \pi_2, \ldots, \pi_m$ is T-perfect iff all the distributions $\pi_1, \pi_2, \ldots, \pi_m$ are marginal to distribution $\pi_1 \rhd_T \pi_2 \rhd_T \ldots \rhd \pi_m$.*

If we translate this theorem to the language of artificial intelligence, its meaning is that the global knowledge expressed by the multidimensional distribution keeps all the local knowledge contained in the low-dimensional distributions, i.e. nothing was lost or changed.

4 Graphical Models

Probabilistic graphical models (well-known thanks e.g. to [6]) served as the inspiration for various authors e.g. [1,2] to introduce analogous models also in the framework of possibility theory.

4.1 Possibilistic Trees

Possibilistic trees suggested by de Campos and Huete in [2] for specific conditional independence concepts are based on the following simple idea. If $I_T(X, Y|Z)$, then the joint distribution $\pi(x, y, z)$ of X, Y, Z can be obtained from its marginals $\pi(x, z)$ and $\pi(y, z)$.

This idea can easily be generalized to n-dimensional case. Let us assume variables $X_1, \ldots, X_n$ such that $I_T(\{X_j\}_{j<i}\{X_j\}_{j>i}|i)$, then the joint possibility distribution of these variables can be obtained form the marginals $\pi(x_1, \ldots, x_i)$ and $\pi(x_i, \ldots, x_n)$. This idea can be recursively applied to both subsets of variables. Therefore to obtain the joint possibility distribution, it is enough to store low-dimensional distributions obtained by this process.

Resulting *possibilistic tree* $\mathcal{T}$ consists of two kinds of nodes — *leaf nodes* (which store marginal possibility distributions) and *internal nodes* (storing conditional independence statements).

De Campos and Huete presented two propositions concerning possibilistic trees (induced, in fact, by conditional independence concepts based on Gödel's and product t-norms), which can be generalized as suggested below.

Any possibilistic tree $\mathcal{T}$ can easily be transformed into a generating sequence of its leaves $\pi_{L_1}, \ldots \pi_{L_m}$. The joint possibiility distribution is then obtained in the following way: any fork of $\mathcal{T}$ is substituted by a composition operator connecting marginal distributions of corresponding branches. Let us note, that this transformation keeps according to Theorem 1 all the conditional independences expressed by the possibilistic tree.

Let us also note, that because of rather complicated system of brackets, the resulting model is not generally formed by a T-perfect sequence of possibility distributions. Nevertheless, it must exist, as the following lemma suggests.

Lemma 6. *Any possibilistic tree $\mathcal{T}$ defines a perfect sequence.*

Proof. It follows directly from Theorem 2, as distributions at leaves are marginals of the joint possibility distribution. □

Now, let us preset an example, which is a generalization of examples from [2].

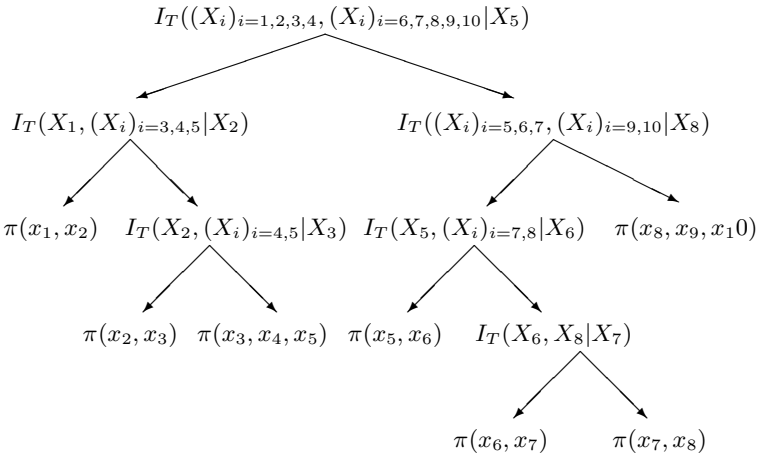


Fig. 1. Possibilistic tree from Example 1

Example 1 Let π be a joint possibility distribution of $X_i, i = 1,\dots,10$ with conditional independences (based on continuous t-norm T) expressed by possibilistic tree in Figure 1.

The generating sequence $\pi(x_1,x_2),\pi(x_2,x_3),\pi(x_3,x_4,x_5),\pi(x_5,x_6),\ \pi(x_6,x_7),\ \pi(x_7,x_8),\pi(x_8,x_9,x_{10})$ forms a joint distribution

$$(\pi(x_1,x_2)\rhd_T(\pi(x_2,x_3)\rhd_T\pi(x_3,x_4,x_5))) \rhd_T((\pi(x_5,x_6)\rhd_T(\pi(x_6,x_7)\rhd_T\pi(x_7,x_8)))\rhd_T\pi(x_8,x_9,x_{10})). \tag{10}$$

Although it is not obvious at the first sight it is also perfect as (10) can be transformed into

$$\pi(x_1,x_2)\rhd_T\pi(x_2,x_3)\rhd_T\pi(x_3,x_4,x_5) \rhd_T\pi(x_5,x_6)\rhd_T\pi(x_6,x_7)\rhd_T\pi(x_7,x_8)\rhd_T\pi(x_8,x_9,x_{10}).$$

due to Lemma 4 and convention (9). Therefore $\pi(x_1,x_2),\pi(x_2,x_3),\ \pi(x_3,x_4,x_5),\ \pi(x_5,x_6),\ \pi(x_6,x_7),\ \pi(x_7,x_8),\pi(x_8,x_9,x_{10})$ is a perfect sequence.

Let us note, that another ordering of the marginal possibility distributions, e.g. $\pi(x_1,x_2),\pi(x_3,x_4,x_5),\pi(x_2,x_3),\pi(x_5,x_6),\ \ \pi(x_6,x_7),\ \ \pi(x_7,x_8),\pi(x_8,x_9,x_{10})$, may lead to a different model than that expressed by a possibilistic tree $\mathcal{T}$, as the resulting model does not keep $\pi(x_2,x_3)$ unless $\pi(x_2,x_3)=\pi(x_2)\cdot\pi(x_3)$.

Let us also note that not every perfect sequence can be transformed into a possibilistic tree, e.g. if one variable appears in three (or more) marginals.

4.2 Dependence Trees

In dependence trees [2] nodes represent variables (or groups of variables) and edges represent direct dependence relationship among variables (or groups). Conditional independence statements can be obtained from the graph in an analogous way to Bayesian networks.

In [2] a simple algorithm for the construction of dependence tree of a possibility distribution in question is presented. In that paper it is also shown (by examples) how to a transform dependence tree to a possibilistic tree and vice versa. It is also mentioned that possibilistic tree is a more general structure than dependence tree. As we have shown that any possibilistic tree can be transformed to perfect sequence, it is obvious, that the same holds for dependence trees.

Nevertheless, we present a direct procedure of the transformation of dependence tree to a perfect sequence, which is extremely simple.

For each dependence tree one can construct a perfect sequence $\pi_1, \ldots, \pi_m$ of distributions of variables $X_{K_1}, X_{K_2}, \ldots, X_{K_m}$, respectively. These distributions are such that each $\{X_i\}_{i \in K_k}$ equals some $cl(X_j) = \{X_j\} \cup pa(X_j)$ and $\pi_1 \triangleright \ldots \triangleright \pi_m$ equals the distribution represented by the dependence tree.

This approach can be applied also to more general directed possibilistic graphs [1], which will be in the center of our attention in the next part.

4.3 Directed Possibilistic Graphs

Directed possibilistic graph (or *possibilistic belief network*) is a possibilistic counterpart of Bayesian network (and a generalization of dependence trees) and can be defined in the following way:

Relationships among variables in directed possibilistic graph are determined in two ways. Structural information describing the existence of a "direct" dependence of variables is given by a graph, while the quantitative information is given by a system of conditional possibility distributions. Thus, a possibilistic belief network is a couple: an *acyclic directed graph* and a *system of conditional probability distributions*. In this system there are as many distributions as variables, i.e. nodes of the graph (in contrary to dependence trees). For each variable there is a conditional distribution given all *parent* variables in the condition. Some of nodes (at least one because of acyclicity) are parentless and their distributions are in fact unconditional.

To transform a possibilistic belief network into a a perfect sequence the procedure described in the preceding section can be used. Here we present a reverse procedure for transformation of a perfect sequence into a possibilistic belief network.

Having a perfect sequence $\pi_1, \pi_2, \ldots, \pi_m$ (π_k being the distribution of X_{K_k}), we first order (in an arbitrary way) all the variables for which at least one of the distributions π_k is defined, i.e.

$$\{X_1, X_2, X_3, \ldots, X_n\} = \{X_i\}_{i \in K_1 \cup \ldots \cup K_m}.$$

Then we get a graph of the constructed possibilistic belief network in the following way:

1. the nodes are all the variables $X_1, X_2, X_3, \ldots, X_n$;
2. there is an edge $(X_i \to X_j)$ if there exists a distribution π_k such that both $i, j \in K_k$, $j \notin K_1 \cup \ldots \cup K_{k-1}$ and either $i \in K_1 \cup \ldots \cup K_{k-1}$ or $i < j$.

5 Conclusions

We overviewed the non-graphical approach to multidimensional possibilistic models based on operators of composition — so-called compositional models. We presented three types of graphical models and showed, how these models can be expressed by means of compositional models. Furthermore, we showed that any of these three models can be expressed by a perfect sequence of low-dimensional distributions. Finally, we presented a procedure by which any perfect sequence of low-dimensional distributions can be transformed into directed possibilistic graph (or possibilistic belief network).

Acknowledgments. The work of the author was supported by the grant GA ČR 201/09/1891, by the grant GA AV ČR A100750603 and by the grant MŠMT 2C06019.

References

1. Benferhat, S., Dubois, D., Gracia, L., Prade, H.: Directed possibilistic graphs and possibilistic logic. In: Bouchon-Meunier, B., Yager, R.R. (eds.) Proc. of the 7th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU 1998, Editions E.D.K. Paris, pp. 1470–1477 (1998)
2. de Campos, L.M., Huete, J.F.: Independence concepts in possibility theory: Part 1 and Part 2. Fuzzy sets and systems 103, 127–152, 487–505 (1999)
3. de Cooman, G.: Possibility theory I – III. Int. J. General Systems 25, 291–371 (1997)
4. Janssen, H., de Cooman, G., Kerre, E.E.: First results for a mathematical theory of possibilistic Markov processes. In: Proceedings of IPMU 1996. Information Processing and Management of Uncertainty in Knowledge-Based Systems, Granada, Spain, vol. III, pp. 1425–1431 (1996)
5. Jiroušek, R., Vejnarová, J.: General framework for multidimensional models. International Journal of Intelligent Systems 18, 107–127 (2003)
6. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
7. Vejnarová, J.: Composition of possibility measures on finite spaces: preliminary results. In: Proceedings of 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU 1998, Paris, pp. 25–30 (1998)
8. Vejnarová, J.: Possibilistic independence and operators of composition of possibility measures. In: Hušková, M., Víšek, J.Á., Lachout, P. (eds.) Prague Stochastics 1998, JČMF, pp. 575–580 (1998)
9. Vejnarová, J.: Conditional independence relations in possibility theory. Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems 8, 253–269 (2000)
10. Vejnarová, J.: On interpretation of T-product possibility distributions. In: Bouchon-Meunier, B., Foulloy, L., Yager, R.R. (eds.) 9th International Conference on Information Processing and Management of Uncertainty IPMU 2002, ESIA - Université de Savoie, France, pp. 1053–1060 (2002)
11. Vejnarová, J.: On possibilistic marfginal problem. Kybernetika 43, 657–674 (2007)

Bayesian Networks vs. Evidential Networks: An Application to Convoy Detection

Evangeline Pollard, Michèle Rombaut, and Benjamin Pannetier

ONERA - BP72 - 29 avenue de la Division Leclerc FR-92322 Châtillon Cedex, France
GIPSA-Lab 961 rue de la Houille Blanche - BP 46 - 38402 Grenoble Cedex, France
evangeline.pollard@onera.fr, michele.rombaut@gipsa-lab.inpg.fr,
benjamin.pannetier@onera.fr
http://www.onera.fr, http://www.gipsa-lab.inpg.fr/

Abstract. In this article, the evidential network is combined with a temporal credal filter in order to incorporate the time information and describe the information propagation from a node to another one. Then we describe an application in convoy detection and propose a complex simulated scenario. The results are compared with those of our previous approach with Bayesian networks.

Keywords: Bayesian networks, Evidential networks, Convoy detection.

1 Introduction

Graphical models, first formalized by Pearl [1], are commonly used for many applications like medical diagnosis, situation assessment [2] or biological applications [3]. Also called Bayesian networks, they are the merging of graph theory with probabilistic theory. They have a powerful formalism for reasoning under uncertainty because they compute the variable trends of a system, which is intuitively represented by a directed acyclic graph. This graph is composed of nodes and edges. The nodes correspond to the set of random variables representing the system evolution, and edges represent the dependencies between random variables, quantified by a set of conditional probability distributions. The dominant relevance is to limit the computational complexity by using the fundamental following formula:

$$p(x_1, \ldots, x_n) = \sum_{i=1}^{n} p\left(x_i | Pa(x_i)\right) \tag{1}$$

where $Pa(x_i)$ represents the parent node set of node x_i.

This kind of graphical model is specially effective when a very complete statistical knowledge description of the modeled system is available. If not, the use of *a priori* can strongly influence the final results.

Evidential networks, which are graphical models in an evidential context, are less known, but have similar applications in system analysis [4] or threat assessment [5]. They were for the first time formalized by Xu and Smets [6,7] by using a generalization of Bayes' theorem where all conditional probabilities are

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 31–39, 2010.

replaced by conditional belief functions. They limit the use of *a priori* and are consequently more flexible to model the knowledge.

On the same principle, evidential networks are the merging of graph theory with evidential theory, also called Dempster-Shafer theory. The main idea, in evidential theory, is to consider a larger frame of discernment as in probability, called power set. Indeed, instead of strictly considering probability distributions on $\Omega = \{\omega_1, \ldots, \omega_n\}$, where w_i represents a hypothesis by itself, the belief are also computed on the subset of Ω. In this paper, we adopt the Transferable Belief Model (TBM) representation which proposes to manage uncertainty in two levels: the credal level where beliefs are addressed and the pignistic level where beliefs are used to make decisions. The main difference between Dempster-Shafer theory and TBM representation is that with the latter, belief functions are unnormalized and the mass on conflict $m(\emptyset)$ can be non empty. In this case, it asks the question of the origin of this conflict (unreliable sources, missing hypothesis, open world...) [8].

In this paper, we first review transferable belief model formalism. In the second part, we describe the theoretical implementation of evidential networks, before we describe our specific application of convoy detection in the third part. Finally, we give some simulation results by using a complex simulated scenario.

2 The Transferable Belief Model (TBM) Framework

2.1 Background

The main idea with the Transferable Belief Model is to attribute a belief distribution on a variable to a larger state space as with probability. The power set of Ω, denoted 2^Ω, is a set composed of hypotheses and joined hypotheses and is of size $2^{|\Omega|}$. The basic belief assignment m^Ω (bba) is defined such that:

$$\begin{aligned} m^\Omega : 2^\Omega &\rightarrow [0,1] \\ B &\rightarrow m^\Omega(B) \end{aligned} \tag{2}$$

$$\sum_{B \in 2^\Omega} m(B) = 1 \tag{3}$$

The belief, given to a hypothesis or to a junction between hypotheses can be also expressed according to some other elementary functions called the plausibility function pl, the commonality function q and the the implicability function b, where $\forall A \subseteq 2^\Omega$:

$$bel^\Omega(A) = \sum_{B \subseteq A, B \neq \phi} m^\Omega(B) \tag{4}$$

$$pl^\Omega(A) = \sum_{A \cap B \neq \phi} m^\Omega(B) \tag{5}$$

$$b^\Omega(A) = bel^\Omega(A) + m^\Omega(\emptyset) \tag{6}$$

$$q^\Omega(A) = \sum_{B \supseteq A} m^\Omega(B) \tag{7}$$

These elementary functions bel, pl, q, b are in one-to-one correspondence with m. If bel and pl can be easily understood as the minimal and maximal likelihood

admitted to a proposition $A \in 2^{\Omega}$, q and b are less intuitive, but their formalisms are convenient for the propagation mechanism, as we will see in the next parts.

2.2 The Combination Rules

The Conjunctive Rule of Combination (CRC). It is an associative and commutative operation that combines belief functions coming from reliable and independent sources:

$$m_{1\textcircled{\cup}2}^{\Omega}(A) = \left(m_1^{\Omega} \textcircled{\cup} m_2^{\Omega}\right)(A) = \sum_{B \cup C = A} m_1^{\Omega}(B).m_2^{\Omega}(C) \tag{8}$$

The same equation can be more conveniently expressed with the commonality:

$$b_{1\textcircled{\cup}2}^{\Omega}(A) = \left(b_1^{\Omega} \textcircled{\cup} b_2^{\Omega}\right)(A) = b_1^{\Omega}(A).b_2^{\Omega}(A) \tag{9}$$

The Disjuntive Rule of Combination (DRC). The DRC is defined as the combination rule for unreliable sources. It can also be seen as a combination rule which can deal with conflict:

$$m_{1\textcircled{\cap}2}^{\Omega}(A) = \left(m_1^{\Omega} \textcircled{\cap} m_2^{\Omega}\right)(A) = \sum_{B \cap C = A} m_1^{\Omega}(B).m_2^{\Omega}(C) \tag{10}$$

The same equation can be more conveniently expressed with the implicability:

$$q_{1\textcircled{\cap}2}^{\Omega}(A) = \left(q_1^{\Omega} \textcircled{\cap} q_2^{\Omega}\right)(A) = q_1^{\Omega}(A).q_2^{\Omega}(A) \tag{11}$$

2.3 Generalized Bayes Theorem (GBT)

The GBT performs the same task as the Bayesian theorem, but within the TBM conflict. Given the set of conditional basic belief assignments $m^{\Omega}[\theta_i]$, $\forall \theta_i \in \Theta, \forall \omega \in \Omega$, if $\theta \subset \Theta$:

$$pl^{\Theta}[\omega](\theta) = 1 - \prod_{\theta_i \in \theta} \left(1 - pl^{\Omega}[\theta_i](\omega)\right) \tag{12}$$

2.4 Temporal Belief Filter

A temporal belief filter is proposed as in [9,10] in order to ensure a temporal consistency: the presence of objects of interest can be based on a long term detection. The predicted belief on X_i at time k can be written as:

$$\hat{m}^{\Omega_i}(X_i^k) = F^{\Omega_i}.m^{\Omega_i}(X_i^{k-1}) \tag{13}$$

where $m^{\Omega_i}(X^{k-1})$ at time $k-1$ is the belief function at time $k-1$, $\hat{m}^{\Omega_i}(X^k)$ is the predicted belief function at time k and F^{Ω_i} is the temporal evolution model.

Assuming that each node X_i is a binary node ($\Omega_i = \{X_i, \bar{X}_i\}$), the following vector notation for the belief distribution is used:

$$m^{\Omega_i} = [m^{\Omega_i}(\emptyset) \quad m^{\Omega_i}(X_i) \quad m^{\Omega_i}(\bar{X}_i) \quad m^{\Omega_i}(\Omega_i)]^T \tag{14}$$

Finally, the temporal evolution model F^{Ω_i} of size $2^{|\Omega_i|} \times 2^{|\Omega_i|}$ is written as:

$$F^{\Omega_i} = [F^{\Omega_i}(\emptyset) \quad F^{\Omega_i}(X_i) \quad F^{\Omega_i}(\bar{X}_i) \quad F^{\Omega_i}(\Omega_i)] \tag{15}$$

with $F^{\Omega_i}(\emptyset) = [1 \quad 0 \quad 0 \quad 0]^T$ and $F^{\Omega_i}(\Omega_i) = [0 \quad 0 \quad 0 \quad 1]^T$ because all conflict/doubt is transferred on itself. $F^{\Omega_i}(X_i)$ (*resp.* $F^{\Omega_i}(\bar{X}_i)$) represents the model evolution of the node X_i if its value is true (*resp.* false). In this case, the belief on X_i (*resp.* $\bar{X}_i$) is partly transferred on X_i (*resp.* $\bar{X}_i$) according to a certain confidence $\alpha_{\mathcal{T}}$ (*resp.* $\alpha_{\mathcal{F}}$) as:

$$F^{\Omega_i}(X_i) = [0 \quad \alpha_{\mathcal{T}} \quad 0 \quad 1-\alpha_{\mathcal{T}}]^T \qquad F^{\Omega_i}(\bar{X}_i) = [0 \quad 0 \quad \alpha_{\mathcal{F}} \quad 1-\alpha_{\mathcal{F}}]^T \tag{16}$$

Finally, the obtained belief at time k is combined with the measured belief distribution. This combination is made according to a CRC (*cf.* 2.2) which highlights conflict between the prediction $\tilde{m}^{\Omega}(X^k)$ and the measurement $\hat{m}^{\Omega}(X^k)$ and is written as:

$$m^{\Omega}(X^k) = \tilde{m}^{\Omega}(X^k) \textcircled{\cap} \hat{m}^{\Omega}(X^k) \tag{17}$$

2.5 Discounting a Belief Function

The discounting process is used to reduce the influence of a source of information. The new basic belief assignment ${}^{\alpha}m^{\Omega}$ is computed from m^{Ω} using the parameter $\alpha \in [0, 1]$:

$$\begin{aligned} {}^{\alpha}m^{\Omega}(A) &= (1-\alpha)\, m^{\Omega}(A) \qquad \forall A \subset \Omega \\ {}^{\alpha}m^{\Omega}(\Omega) &= (1-\alpha)\, m^{\Omega}(\Omega) + \alpha \end{aligned} \tag{18}$$

3 Dynamic Evidential Networks

A basic dynamic network illustrated in Figure 1, where X_1 and X_2 are parent nodes and the belief distribution on X_3 is computed at each iteration k illustrates Dynamic Evidential Network principle. But, before describing the inference mechanism, it is necessary to develop the two initialization steps:

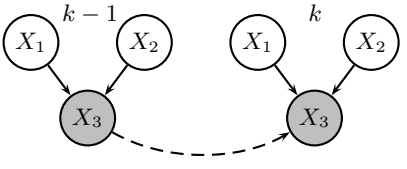


Fig. 1. A very simple example of evidential network

1. **Prior mass belief establishment:** With Bayesian approach, the first step would be to establish $p(X_3|X_1, X_2)$ of size 2×2, but whose size can quickly increase depending on the number of feasible states for X_1 and X_2 and more generally on the number of parent nodes. With evidential network, the conditional beliefs $m^{\Omega_3}[X_i]$ are established for each parent node $i, \forall i \in \{1, 2\}$ and independently of the others according to the knowledge on the system. However only conditional belief functions on X_3 knowing that X_i is in the state X_i or $\bar{X}_i$ can be established. The belief knowing that the node X_i is in the state Ω_i cannot be intuitively established but is computed $\forall i \in \{1, 2\}$ by using the DRC as in equation (11):

$$b^{\Omega_3}[\Omega_i](X_3) = b^{\Omega_3}[X_i](X_3).b^{\Omega_3}[\bar{X}_i](X_3) \tag{19}$$

2. **Discounting coefficient establishment:** When a node depends on many other nodes, it is possible to modify the importance of each node by using discounting coefficients. Another point of view could be that the parent nodes can be seen as independent sources which are strongly or weakly weighted, depending on their reliability.

The inference mechanism is now decomposed in simple operations for the basic dynamic network illustrated in Figure 1.

1. **Data transformation:** Data are transforming into belief distribution $\tilde{m}^{\Omega_i}$ for root nodes X_i. This transformation can be made by using fuzzy sets or Rayleigh distributions as done in [11].
2. **Propagation:** The information from parent nodes X_i are propagated to the node X_3. The obtained belief distributions are denoted $m^{\Omega}_{i\to 3}$. These are computed by using plausibility with the GBT equation (12) as, $\forall X_3 \subseteq \Omega_3$:

$$pl^{\Omega_3}_{i\to 3}(X_3) = \sum_{X_3 \subseteq \Omega_3} pl^{\Omega_3}[X_i](X_3).\tilde{m}^{\Omega_i}(X_i) \tag{20}$$

3. **Discounting:** If discounting coefficients are α_i for each node X_i, the formula (18) is applied on propagated belief distributions $m^{\Omega}_{i\to 3}$ to obtain ${}^{\alpha_i}m^{\Omega}_{i\to 3}$.
4. **Combination:** Discounted propagated belief distributions are finally combined by using the CRC with implicabilities as in equation (9):

$$q^{\Omega_3}(X_3) =^{\alpha_i} q^{\Omega_3}_{1\to 3}(X_3).{}^{\alpha_i}q^{\Omega_3}_{2\to 3}(X_3) \tag{21}$$

5. **Time propagation:** Assuming that the node X_3 evolves according to a model F^{Ω}_3, it is possible to predict the belief $\hat{m}^{\Omega_3}(X_3)$ and to combine it with the measured belief function $\tilde{m}^{\Omega_3}(X_3)$ according to the equation (17).

It is accepted that this propagation algorithms can only be applied on naive networks. With more complex networks, inference algorithms must be applied as in [12].

4 Application to the Convoy Detection

4.1 Application Description

The human expert describes a convoy as a vehicle set evolving approximately with the same dynamics over a long time. These vehicles are moving on the road at a limited velocity (<80m/s). They must stay within sight of each other with almost constant distances between them (mostly 100m). In some previous work, we developed a robust multitarget tracking algorithm in order to detect vehicle aggregates with precision in term of cardinality and state estimation [13] for battlefield surveillance with Ground Moving Target Indicator (GMTI) sensor [11]. Other information comes from Synthetic Aperture Radar (SAR) or video sensor in order to detect aggregates. In this application, the goal is to classify vehicle aggregates as convoys or not.

The criteria describing a convoy are manifold and of different nature, moreover the variables are discrete. That is why graphs represent an interesting formalism for this application. Figure 2 shows the convoy model represented by a graph. Gray nodes represent states depending on their previous state. For example, random variable X_5 is time dependent, because this type of information is provided infrequently from SAR or video sensor, unlike the other root node like X_1, or X_2 which are computed at the shorter GMTI intervals. Consequently, if the random variable X_5 is filled in at any moment, this information must be propagated for some time. Concerning the random variable X_9, it is time dependent because convoys are evolving as convoy for a long time by regarding GMTI sensor scanning time.

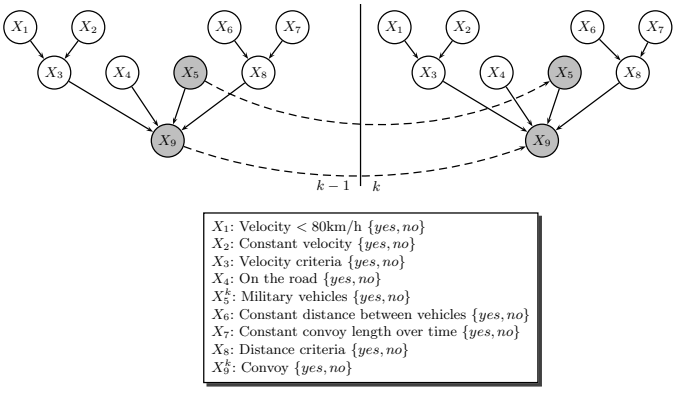


Fig. 2. Dynamic Evidential Network for convoy detection

4.2 Evidential Implementation

First of all, as described in part 3, prior belief functions and discounting coefficients must be established. For the first, they are determined by using logic model

like : *"If the aggregate velocity is higher than 80km/h, then the aggregate is not a convoy ; but if the aggregate velocity is lower than 80km/h, it can be a convoy or not."* Concerning discounting coefficient, we use, as previously with Bayesian networks, heuristic rules to represent relationships between variables as:

$$\alpha_3 = 0.3 \qquad \alpha_4 = 0.2 \qquad \alpha_5 = 0.2 \qquad \alpha_8 = 0.3 \tag{22}$$

This rule means that the criteria on distance and velocity are more important than criteria "on road" or "vehicle type".

Numerical values describing tracks are converted into belief functions concerning temporal velocity constancy or temporal length convoy constancy. In this way, a belief distribution is generated for node X_1, X_2, X_4, X_6 and X_7 at each GMTI scanning time. The node X_5^k is filled in when the type information is available. If not, the entire mass is attributed to the doubt. The goal is to process a belief distribution on node X_9^k at each time k.

5 Simulation Results

5.1 Scenario Description

The GMTI sensor has a linear trajectory, its velocity is 30m/s and its altitude is 4000m. The typical measurement error is 20m in range and 0.008rad in azimuth. The sensor scan time is $T = 10$s. Scenario time is limited to 500s. The false alarm density is $\beta_{FA} = 8.92.10^{-9}$ and the detection probability is $P_D = 0.9$. In the scenario, one 6 target convoy (Target 1-6) is moving on the main road with a constant velocity of 10m/s from South to North. An independent target (Target 7) is moving on the same road in the same direction but with a constant velocity of 15m/s and overtakes the convoy between time t=160s and t=360s approximately. A 8^{th} target is moving on another road completely independently to the others.

5.2 Results

The performances of tracking algorithms have been compared for 100 independent Monte Carlo runs. Figure 3 shows the combined belief functions $m^{\Omega_9}(X_9)$ over the time by considering different assumptions on the number of targets belonging to the convoy. This figure can be compared to the Figure 5 which represents the probability to have a convoy $p(X9)$ computed with Bayesian networks as described in previous work [11]. As expected, both curves have similar shapes, with decreasing values during the maneuver time (at time $t = 160$s et $t = 360$s which correspond to the entry and the output of the over-passing target). Belief of having a 5 or 6 target convoy are higher than the belief of having a 7 target convoy. Indeed, at these moments the 7 target convoy does not follow the convoy model specially because of the distance criteria shown on Figure 4. Finally, Figure 6 is the conflict belief on the combined belief functions $m^{\Omega_9}(X_9^k)$. We observe two conflict pics corresponding to the convoy maneuvers.

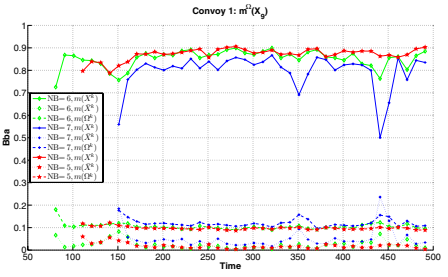


Fig. 3. Estimated belief function on $m^{\Omega_9}(X_9^k)$

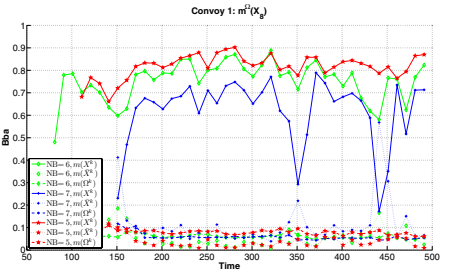


Fig. 4. Estimated belief function on $m^{\Omega_8}(X_8)$

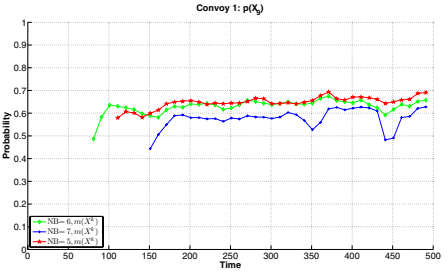


Fig. 5. Estimated belief function on $p(X_9^k)$

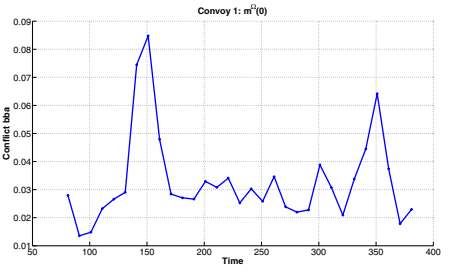


Fig. 6. Estimated belief function on $m^{\Omega_8}(X_8)$

Results obtained with evidential networks are coherent with previous results obtained with Bayesian networks. But the main advantage of using evidential theory is to add information on doubt and conflict. When the belief suddenly decreases, these information can be used to characterize the situation and the conflict mass indicates maneuvers. In future works, we can imagine to use doubt and conflict to characterize convoy maneuvers and to classify it.

6 Conclusion

From the study of evidential networks, a new dynamic approach is proposed based on the classical tools of the TBM framework. Then a simulated scenario of convoy detection is proposed to compare results with classical Bayesian networks, in order to demonstrate the relevance of using doubt and conflict.

References

1. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, San Francisco (1988)
2. Johansson, F., Falkman, G.: Implementation and integration of a Bayesian Network for prediction of tactical intention into a ground target simulator. In: 9th International Conference on Information Fusion, July 2006, pp. 1–7 (2006)

3. Murphy, K., Mian, S.: Modelling gene expression data using dynamic Bayesian networks. tech. rep., Lawrence Berkeley National Laboratory (1999)
4. Simon, C., Weber, P., Levrat, E.: Bayesian networks and evidence theory to model complex systems reliability. Journal of Computers 2(1), 33 (2007)
5. Benavoli, A., Ristic, B., Farina, A., Oxenham, M., Chisci, L.: An application of evidential networks to threat assessment. IEEE Transactions on Aerospace and Electronic Systems 45, 620–639 (2009)
6. Xu, H., Smets, P.: Reasoning in evidential networks with conditional belief functions. International Journal of Approximate Reasoning 14(2-3), 155–185 (1996)
7. Smets, P.: Belief functions: The disjunctive ule of Combination and the Generalized Bayesian theorem. International Journal of Approximate Reasoning 9(1), 1–35 (1993)
8. Smets, P.: Data fusion in the transferable belief model. In: Proc. Third Int. Conf. Information Fusion, Citeseer, pp. 10–13 (2000)
9. Ramasso, E., Rombaut, M., Pellerin, D.: A temporal belief filter improving human action recognition in videos. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006, May 2006, vol. 2, p. II (2006)
10. Ramasso, E., Rombaut, M., Pellerin, D.: State filtering and change detection using TBM conflict Application to human action recognition in athletics videos. IEEE Transactions on Circuits and Systems for Video Technology 17(7), 944 (2007)
11. Pollard, E., Pannetier, B., Rombaut, M.: Convoy detection processing by using the hybrid algorithm (GMCPHD/VS-IMMC-MHT) and dynamic bayesian networks. In: 12th International Conference on Information Fusion (2009)
12. Yaghlane, B., Mellouli, K.: Inference in directed evidential networks based on the transferable belief model. International Journal of Approximate Reasoning 48(2), 399–418 (2008)
13. Pollard, E., Pannetier, B., Rombaut, M.: Hybrid Algorithms for Multitarget Tracking using MHT and GMCPHD. IEEE Transactions on Aerospace and Electronic Systems (to appear)

Approximation of Data by Decomposable Belief Models

Radim Jiroušek

Faculty of Management, University of Economics
and
Institute of Information Theory and Automation, Acad. of Sciences
radim@utia.cas.cz

Abstract. It is well known that among all probabilistic graphical Markov models the class of decomposable models is the most advantageous in the sense that the respective distributions can be expressed with the help of their marginals and that the most efficient computational procedures are designed for their processing (for example professional software does not perform computations with Bayesian networks but with decomposable models into which the original Bayesian network is transformed). This paper introduces a definition of the counterpart of these models within Dempster-Shafer theory of evidence, makes a survey of their most important properties and illustrates their efficiency on the problem of approximation of a "sample distribution" for a data file with missing values.

1 Introduction

For data analysis, data preprocessing and management of missing values form an important step substantially influencing the expected results. This concerns in particular the analysis performed with the help of "classical" statistical procedures based on probability. The situation is changing fundamentally when one starts considering models within Dempster-Shafer theory of evidence [13]. In this theory (and it is the main difference with probability theory) one can easily model ignorance and therefore missing data may remain missing - unknown. Unfortunately, nothing is free and this advantage is paid by an increase of computational complexity. This is due to the fact that a basic assignment, in contrast to a probability distribution, cannot be represented by a point function. Therefore any idea decreasing computational complexity of the necessary procedures is desirable.

The goal of this paper is to show that within the framework of Dempster-Shafer theory one can construct decomposable models and that their representation is much less space-demanding than general Dempster-Shafer models. Moreover, by an example of data approximation with the help of a decomposable model we show that we gain not only an efficient representation of basic assignments but also possibility to design efficient ("local") computational procedures.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 40–49, 2010.

After introducing the necessary notation we will define an operator of composition, which plays a leading role in the definition of decomposable models. When introducing these models in Section 3 we will also deal with the concepts of independence and bring reasons in favor of a (relatively) new definition of conditional independence in Dempster-Shafer theory.

2 Basic Notion

2.1 Set Notation

In the whole paper we shall deal with a finite number of variables $X_1, X_2, \ldots, X_n$ each of which is specified by a finite set $\mathbf{X}_i$ of its values. So, we will consider multidimensional space of discernment

$$\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n,$$

and its *subspaces*. For $K \subset N = \{1, 2, \ldots, n\}$, $\mathbf{X}_K$ denotes a Cartesian product of those $\mathbf{X}_i$, for which $i \in K$:

$$\mathbf{X}_K = \bigtimes_{i \in K} \mathbf{X}_i.$$

A *projection* of $x = (x_1, x_2, \ldots, x_n) \in \mathbf{X}_N$ into $\mathbf{X}_K$ will be denoted $x^{\downarrow K}$, i.e. for $K = \{i_1, i_2, \ldots, i_\ell\}$

$$x^{\downarrow K} = (x_{i_1}, x_{i_2}, \ldots, x_{i_\ell}) \in \mathbf{X}_K.$$

Analogously, for $K \subset L \subseteq N$ and $A \subset \mathbf{X}_L$, $A^{\downarrow K}$ will denote a *projection* of A into $\mathbf{X}_K$:

$$A^{\downarrow K} = \{y \in \mathbf{X}_K : \exists x \in A \;\; (y = x^{\downarrow K})\}.$$

Let us remark that we do not exclude situations when $K = \emptyset$. In this case $A^{\downarrow \emptyset} = \emptyset$.

Set $A \subseteq \mathbf{X}_N$ is said to be a *point-cylinder* if it can be expressed as a Cartesian product of a singleton and a subspace $\mathbf{X}_L$. More precisely: a point-cylinder is a set $A \subseteq \mathbf{X}_N$ for which there exists an index set (possibly empty) $L \subseteq N$ such that $|C^{\downarrow L}| \leq 1$ and

$$C = C^{\downarrow L} \times \mathbf{X}_{N \setminus L}.$$

Let us stress that if $L = \emptyset$ then $C = \mathbf{X}_N$ (it is the only situation when $|C^{\downarrow L}| < 1$), and when $L = N$ then $|C| = 1$, C is a singleton.

In addition to the projection, in this text we will also need the opposite operation which is called a join. By a *join* of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ we understand a set

$$A \otimes B = \{x \in \mathbf{X}_{K \cup L} : x^{\downarrow K} \in A \;\; \& \;\; x^{\downarrow L} \in B\}.$$

Notice that if K and L are disjoint then the join of the corresponding sets is just their Cartesian product $A \otimes B = A \times B$. For $K = L$, $A \otimes B = A \cap B$. If $K \cap L \neq \emptyset$ and $A^{\downarrow K \cap L} \cap B^{\downarrow K \cap L} = \emptyset$ then also $A \otimes B = \emptyset$.

In view of this paper it is important to realize that if $x \in C \subseteq \mathbf{X}_{K\cup L}$, then $x^{\downarrow K} \in C^{\downarrow K}$ and $x^{\downarrow L} \in C^{\downarrow L}$, which means that always

$$C \subseteq C^{\downarrow K} \otimes C^{\downarrow L}.$$

However, it does not mean that $C = C^{\downarrow K} \otimes C^{\downarrow L}$.

2.2 Assignment Notation

The role of a probability distribution from a probability theory is in Dempster-Shafer theory played by any of the set functions: belief function, plausibility function or basic (*probability or belief*) assignment. Knowing one of them, one can deduce the two remaining. In this paper we shall use exclusively basic assignments.

A *basic assignment* m on $\mathbf{X}_K$ $(K \subseteq N)$ is a function

$$m : \mathcal{P}(\mathbf{X}_K) \longrightarrow [0, 1],$$

for which

$$\sum_{\emptyset \neq A \subseteq \mathbf{X}_K} m(A) = 1.$$

For the sake of this paper it is reasonable to consider only normalized basic assignments, for which $m(\emptyset)$ equals always 0. If $m(A) > 0$, then A is said to be a *focal element* of m.

Having a basic assignment m on $\mathbf{X}_K$ one can consider its *marginal assignment* on $\mathbf{X}_L$ (for $L \subseteq K$), which is defined (for each $\emptyset \neq B \subseteq \mathbf{X}_L$):

$$m^{\downarrow L}(B) = \sum_{A \subseteq \mathbf{X}_K : A^{\downarrow L} = B} m(A).$$

Basic assignment m is said to be *Bayesian* if all its focal elements are *singletons* (i.e. $m(A) > 0 \implies |A| = 1$). Basic assignment m is said to be *cylindrical* if all its focal elements are point-cylinders. Since each singleton is a point-cylinder, it is obvious that a Bayesian basic assignments is also cylindrical. An advantage of Bayesian and cylindrical basic assignments is that the number of possible focal elements does not grow up superexponentially (as it is for general basic assignments) with the number of dimensions but only exponentially.

2.3 Operator of Composition

Definition 1. For two arbitrary basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ $(K \neq \emptyset \neq L)$ a *composition* $m_1 \triangleright m_2$ is defined for each $C \subseteq \mathbf{X}_{K\cup L}$ by one of the following expressions:

[a] if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) > 0$ and $C = C^{\downarrow K} \otimes C^{\downarrow L}$ then

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})};$$

[b] if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) = 0$ and $C = C^{\downarrow K} \times \mathbf{X}_{L\setminus K}$ then

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow K});$$

[c] in all other cases $(m_1 \triangleright m_2)(C) = 0$.

First of all we want to stress that the operator of composition is something else than the famous Dempster's rule of combination [4]. For example it is (in contrary to Dempster's rule) neither commutative nor associative. In [9,8] we proved a number of properties concerning the operator of composition. In view of the forthcoming text the following ones are the most important(m_1 and m_2 are basic assignments defined on $\mathbf{X}_K, \mathbf{X}_L$, respectively):

(i) $m_1 \triangleright m_2$ is a basic assignment on $\mathbf{X}_{K\cup L}$;
(ii) $(m_1 \triangleright m_2)^{\downarrow K} = m_1$;
(iii) $m_1 \triangleright m_2 = m_2 \triangleright m_1 \iff m_1^{\downarrow K\cap L} = m_2^{\downarrow K\cap L}$;
(iv) If $A \subseteq \mathbf{X}_{K\cup L}$ is a focal element of $m_1 \triangleright m_2$ then $A = A^{\downarrow K} \otimes A^{\downarrow L}$;
(v) If m_1 and m_2 are cylindrical then $m_1 \triangleright m_2$ is also cylindrical.

3 Decomposable Models

3.1 Conditional Independence

First, let us present a generally accepted notion of unconditional independence[1] ([1,14,16]).

Definition 2. Let m be a basic assignment on $\mathbf{X}_N$ and $K, L \subset N$ be nonempty disjoint. We say that groups of variables X_K and X_L are *independent*[2] with respect to basic assignment m (in notation $K \perp\!\!\!\perp L\ [m]$) if for all $A \subseteq \mathbf{X}_{K\cup L}$

$$m^{\downarrow K\cup L}(A) = (m^{\downarrow K} \textcircled{\cap} m^{\downarrow L})(A^{\downarrow K\cup L}).$$

Symbol $\textcircled{\cap}$ denotes the famous *conjunctive combination rule* (non-normalized Dempster's rule of combination). It was proved in [8] that Definition 2 is equivalent to the following Definition 2a.

Definition 2a. Let m be a basic assignment on $\mathbf{X}_N$ and $K, L \subset N$ be nonempty disjoint. We say that groups of variables X_K and X_L are *independent* with respect to basic assignment m if for all $A \subseteq \mathbf{X}_{K\cup L}$

$$m^{\downarrow K\cup L}(A) = \begin{cases} m^{\downarrow K}(A^{\downarrow K}) \cdot m^{\downarrow L}(A^{\downarrow L}) & \text{if } A = A^{\downarrow K} \times A^{\downarrow L}, \\ 0 & \text{otherwise.} \end{cases}$$

[1] Some authors call it *marginal* independence.
[2] Couso et al. [3] call this independence *independence in random sets*, Klir [11] (*non-interactivity)).*

Though it it not obvious these definitions are equivalent to each other. However, when considering a generalization of these definitions to the conditional case we can get different notions. Most of the authors use the generalization based on Definition 2 (see for example papers [2,3,11,14,15,16]). In this text we will use a simple and straightforward generalization of Definition 2a, which was introduced in [6] and [8], and which can hardly be expressed with the help Dempster's rule of combination (or with the help of its non-normalized version: conjunctive combination rule). The resulting notion differs from the notion of conditional independence used, for example, by Shenoy [14] and Studený [16] (their notion of conditional independence is the same as the *conditional non-interactivity* used by Ben Yaghlane *et al.* in [2]).

Definition 3. Let m be a basic assignment on $\mathbf{X}_N$ and $K, L, M \subset N$ be disjoint, $K \neq \emptyset \neq L$. We say that groups of variables X_K and X_L are *conditionally independent given* X_M *with respect to* m (and denote it by $K \perp\!\!\!\perp L|M\ [m]$), if for any $A \subseteq \mathbf{X}_{K\cup L\cup M}$ such that $A = A^{\downarrow K\cup M} \otimes A^{\downarrow L\cup M}$ the equality

$$m^{\downarrow K\cup L\cup M}(A) \cdot m^{\downarrow M}(A^{\downarrow M}) = m^{\downarrow K\cup M}(A^{\downarrow K\cup M}) \cdot m^{\downarrow L\cup M}(A^{\downarrow L\cup M})$$

holds true, and $m^{\downarrow K\cup L\cup M}(A) = 0$ for all the remaining $A \subseteq \mathbf{X}_{K\cup L\cup M}$, for which $A \neq A^{\downarrow K\cup M} \otimes A^{\downarrow L\cup M}$.

Our definition (in the same way as the definition used in [2,14,16]) meets the following important properties:

- for $M = \emptyset$ the concept coincides with Definition 2;
- the notion meets all the properties required from the notion of conditional independence, so-called *semigraphoid properties* ([12,16,17]):

  (A1) $K \perp\!\!\!\perp L \,|\, M\ [m] \implies L \perp\!\!\!\perp K \,|\, M\ [m]$;
  (A2) $K \perp\!\!\!\perp L \cup M \,|\, J\ [m] \implies K \perp\!\!\!\perp M \,|\, J\ [m]$;
  (A3) $K \perp\!\!\!\perp L \cup M \,|\, J\ [m] \implies K \perp\!\!\!\perp L \,|\, M \cup J\ [m]$;
  (A4) $(K \perp\!\!\!\perp L \,|\, M \cup J\ [m])\ \&\ (K \perp\!\!\!\perp M \,|\, J\ [m]) \implies K \perp\!\!\!\perp L \cup M \,|\, J\ [m]$.

The main differences between our definition and that used in [2,14,16]) are the following

- our definition does not suffer from the *inconsistency with marginalization*[3];
- for our notion, the Dempster-Shafer counterpart to the probabilistic factorization lemma has been proved in [7].

[3] As it was showed by Studený, when the definition used in [2,14,16] is accepted, then it can happen that for two consistent overlapping basic assignments there does not exist their common extension with the required conditional independence property (for the Studený's example see [2,8]).

3.2 Decomposition

Consider a sequence $K_1, K_2, \ldots, K_r$ meeting the *running intersection property* (RIP), i.e. the sequence for which for all $i = 2, \ldots, r$ there exists $j, 1 \leq j < i$, such that

$$K_i \cap (K_1 \cup \ldots \cup K_{i-1}) \subseteq K_j.$$

Without a loss of generality we will assume that $K_1 \cup \ldots \cup K_r = N$.

Definition 4. We say that a basic assignment m is decomposable (with respect to a sequence $K_1, K_2, \ldots, K_r$ meeting RIP) if

$$m = (\ldots((m_1 \triangleright m_2) \triangleright m_3) \triangleright \ldots \triangleright m_{r-1}) \triangleright m_r.$$

In [7] we showed that, analogously to probabilistic decomposable models, also Dempster-Shafer decomposable models possess special independence structures described in the following assertion.

Theorem 1. *If a basic assignment m is decomposable with respect to a sequence $K_1, K_2, \ldots, K_r$ (meeting RIP) then for all $i = 2, \ldots, r$*

$$(K_i \setminus (K_1 \cup \ldots \cup K_{i-1})) \perp\!\!\!\perp ((K_1 \cup \ldots \cup K_{i-1}) \setminus K_i) \mid (K_i \cap (K_1 \cup \ldots \cup K_{i-1}))\ [m].$$

As showed in the following example, the dependence structure of decomposable models allows for their very efficient representation.

Example 1. Consider a 4-dimensional basic assignment on $\mathbf{X}_{\{1,2,3,4\}} = \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3 \times \mathbf{X}_4$, where $|\mathbf{X}_i| = 2$ for all $i = 1, 2, 3, 4$. Since there are

$$2^{(2^4)} - 1 = 2^{16} - 1 = 65\ 535$$

nonempty subsets of the considered frame of discernment, this number expresses also the maximum number of focal elements of a general basic assignment. However, the situation drastically simplifies when one considers a basic assignment decomposable with respect (let us say) $\{1, 2\}, \{2, 3\}, \{3, 4\}$ (it is obvious that this sequence meets RIP). The simplification follows immediately from the fact that, due to Theorem 1, there is a system of conditional independence relations valid for basic assignment m. From this one can deduce that for all the focal elements A of m

$$A = A^{\downarrow\{1,2\}} \otimes A^{\downarrow\{2,3\}} \otimes A^{\downarrow\{3,4\}}$$

holds true. This equality holds only for 657 out of 65 535 nonempty subsets of $\mathbf{X}_{\{1,2,3,4\}}$. Nevertheless, thanks to the fact that

$$m = m^{\downarrow\{1,2\}} \triangleright m^{\downarrow\{2,3\}} \triangleright m^{\downarrow\{3,4\}},$$

we do not need to store basic assignment m but only its three marginals $m^{\downarrow\{1,2\}}$, $m^{\downarrow\{2,3\}}$ and $m^{\downarrow\{3,4\}}$. Each of them has at most 15 focal elements and therefore one needs only 45 numbers to represent this 4-dimensional decomposable basic assignment.

4 Approximation of a Primitive Sample Assignment

In this section we will show that application of decomposable models may result not only in possibility to store multidimensional[4] basic assignments but also in possibility to compute with them using "local" computational procedures. Let us illustrate this possibility by the way of example of approximation of a data file with the help of a decomposable basic assignment. The reader is asked to keep in mind that it is just an illustration. We do not propose to realize the following primitive procedure for practical applications. Because of lack of space we cannot present here any more sophisticated process based on more complex ideas like the procedures studied in [5].

Having a data file with missing values one can quite naturally assign to each data record a point-cylinder $C = C^{\downarrow L} \times \mathbf{X}_{N\setminus L}$ from $\mathbf{X}_N$ expressing that the record contains $|L|$ specific data values corresponding to $C^{\downarrow L}$ and $|N \setminus L|$ missing values. By a *primitive sample assignment* we will understand a basic assignment m, where value $m(C)$ is computed as a relative frequency (within the data file) of records assigned with point cylinder C. It means that any primitive sample assignment is cylindrical.

The approximation task is to find a decomposable basic assignment m, which is in a sense best approximation of the primitive sample assignment for a given data file. For this one has to specify a criterion according to which a "goodness" of the approximation is evaluated. To do so one can consider a number of possible divergences proposed in literature (for a nice survey see [10]). However, not all of them are such that they make the "local" computations possible. As the simplest example of a suitable distance let us consider a "relative entropy" type of divergence defined

$$Div(m;\bar{m}) = \sum_{A\subseteq\mathcal{F}(m)} m(A)\log\frac{m(A)}{\bar{m}(A)},$$

where m is the primitive basic assignment to be approximated, $\bar{m}$ is an approximating decomposable basic assignment and $\mathcal{F}(m) \subset \mathbf{X}_N$ is the set of focal elements of m. It is well known that this divergence is always nonnegative, equals 0 if and only if $m = \bar{m}$ and if

$$m(A) > 0 \quad \Longrightarrow \quad \bar{m}(A) > 0 \tag{1}$$

then it is also finite.

Consider a sequence $K_1, K_2, \ldots, K_r$ meeting RIP such that $K_1 \cup \ldots \cup K_r = N$, and an arbitrary basic assignment $\bar{\bar{m}}$ decomposable with respect to $K_1, \ldots, K_r$. Further define a decomposable basic assignment constructed from the marginals of m

$$\bar{m} = (\ldots((m^{\downarrow K_1} \triangleright m^{\downarrow K_2}) \triangleright m^{\downarrow K_3}) \triangleright \ldots \triangleright m^{\downarrow K_{r-1}}) \triangleright m^{\downarrow K_r}. \tag{2}$$

[4] When speaking about multidimensionality in connection with Dempster-Shafer theory we have in mind several tens rather than hundreds of dimensions.

It is not difficult to show that for $\bar{m}$ defined by formula (2) implication (1) is valid (this is because we assume that m is cylindrical and for a cylinder A, $A = A^{\downarrow K} \otimes A^{\downarrow L}$ always holds true) and so we get that $Div(m; \bar{m})$ is finite. Moreover

$$Div(m; \bar{m}) \leq Div(m; \bar{\bar{m}}).$$

This is why it is enough to look for an approximation of m in the form of a *compositional model* (2)[5].

Let us now show that the search for the best approximation of a basic assignment m (i.e. for the most advantageous sequence $K_1, K_2, \ldots, K_r$ meeting RIP) can be based on "local" computations only, i.e. that the procedure stores only and computes with marginal basic assignments $m^{\downarrow K_i}$.

To make our consideration more lucid, consider first $r = 2$. For this

$$Div(m; \bar{m}) = \sum_{A \subseteq \mathcal{F}(m)} m(A) \log \frac{m(A)}{(m^{\downarrow K_1} \triangleright m^{\downarrow K_2})(A)}.$$

The following modifications are correct because for $A \subseteq \mathcal{F}(m)$

$$m^{\downarrow K_1 \cap K_2}(A^{\downarrow K_1 \cap K_2}) > 0$$

and therefore value of $(m^{\downarrow K_1} \triangleright m^{\downarrow K_2})(A)$ is positive and computed according to case [**a**] of Definition 1.

$$\begin{aligned}
Div(m; \bar{m}) &= \sum_{A \subseteq \mathcal{F}(m)} m(A) \log \frac{m(A)}{(m^{\downarrow K_1} \triangleright m^{\downarrow K_2})(A)} \\
&= \sum_{A \subseteq \mathcal{F}(m)} m(A) \log \frac{m(A) \cdot m^{\downarrow K_1 \cap K_2}(A^{\downarrow K_1 \cap K_2})}{m^{\downarrow K_1}(A^{\downarrow K_1}) \cdot m^{\downarrow K_2}(A^{\downarrow K_2})} \\
&= \sum_{A \subseteq \mathcal{F}(m)} m(A) \log m(A) + \sum_{A \subseteq \mathcal{F}(m)} m(A) \log m^{\downarrow K_1 \cap K_2}(A^{\downarrow K_1 \cap K_2}) \\
&\quad - \sum_{A \subseteq \mathcal{F}(m)} m(A) \log m^{\downarrow K_1}(A^{\downarrow K_1}) - \sum_{A \subseteq \mathcal{F}(m)} m(A) \log m^{\downarrow K_2}(A^{\downarrow K_2}).
\end{aligned}$$

The second term of the last expression can be simplified in the following way

$$\begin{aligned}
\sum_{A \subseteq \mathcal{F}(m)} m(A) \log m^{\downarrow K_1}(A^{\downarrow K_1}) &= \sum_{B \subseteq \mathcal{F}(m^{\downarrow K_1})} \sum_{\substack{A \subseteq \mathcal{F}(m) \\ A^{\downarrow K_1} = B}} m(A) \log m^{\downarrow K_1}(A^{\downarrow K_1}) \\
&= \sum_{B \subseteq \mathcal{F}(m^{\downarrow K_1})} \log m^{\downarrow K_1}(B) \sum_{\substack{A \subseteq \mathcal{F}(m) \\ A^{\downarrow K_1} = B}} m(A) \\
&= \sum_{B \subseteq \mathcal{F}(m^{\downarrow K_1})} m^{\downarrow K_1}(B) \log m^{\downarrow K_1}(B).
\end{aligned}$$

[5] Notice that in spite of the fact that the described approximation does not decrease the number of focal elements, it can be very efficiently represented.

Denoting

$$\mathfrak{H}(m^{\downarrow K_1}) = -\sum_{B \subseteq \mathcal{F}(m^{\downarrow K_1})} m^{\downarrow K_1}(B) \log m^{\downarrow K_1}(B),$$

and using analogous symbols also for the other marginals of m we get

$$Div(m; \bar{m}) = \mathfrak{H}(m^{\downarrow K_1}) + \mathfrak{H}(m^{\downarrow K_2}) - \mathfrak{H}(m^{\downarrow K_1 \cap K_2}) - \mathfrak{H}(m).$$

Repeating the above computations for a general r one gets

$$Div(m; \bar{m}) = \mathfrak{H}(m^{\downarrow K_1}) + \left(\sum_{i=2}^{r} \mathfrak{H}(m^{\downarrow K_i}) - \mathfrak{H}(m^{\downarrow K_i \cap (K_1 \cup \ldots \cup K_{i-1})}) \right) - \mathfrak{H}(m).$$

This formula shows that when searching for a suitable sequence $K_1, K_2, \ldots, K_r$ meeting RIP one can omit the term $\mathfrak{H}(m)$ because it appears in all compared expressions. Moreover, when modifying the sequence $K_1, K_2, \ldots, K_r$ only slightly one usually does not need to recompute all the terms

$$\left(\mathfrak{H}(m^{\downarrow K_i}) - \mathfrak{H}(m^{\downarrow K_i \cap (K_1 \cup \ldots \cup K_{i-1})}) \right)$$

but only some of them. These properties indicate that a quite efficient method searching for a suboptimal approximation can easily be designed.

5 Conclusions

In the paper we supported a relatively new notion of conditional independence for Dempster-Shafer theory of evidence. This notion was first introduced in [6] (in that paper under the name of *conditional irrelevance*, though) and later also in [8] and [7], where its theoretical properties were studied. It appears that our notion (in comparison with the notion usually used by other authors [2,3,11,14,15,17,16]) possesses more properties of the probabilistic notion of conditional independence: here we have in mind especially that it does not suffer from the *inconsistency with marginalization* [2] and that it enables us to prove the *factorization lemma*. And it is these very properties that enables us to define decomposable models within Dempster-Shafer theory. Perhaps we do not need to stress that we believe that the introduced decomposable models, just as the probabilistic decomposable models, will allow us to design efficient computational procedures for computation in Dempster-Shafer theory.

Acknowledgement

The research was financially supported by the Czech National Science Foundation under the grants no. ICC/08/E010, and 201/09/1891, and by the Ministry of Education of the Czech Republic under grants 1M0572 and 2C06019.

References

1. Ben Yaghlane, B., Smets, P., Mellouli, K.: Belief Function Independence: I. The Marginal Case. Int. J. of Approximate Reasoning 29(1), 47–70 (2002)
2. Ben Yaghlane, B., Smets, P., Mellouli, K.: Belief Function Independence: II. The Conditional Case. Int. J. of Approximate Reasoning 31(1-2), 31–75 (2002)
3. Couso, I., Moral, S., Walley, P.: Examples of independence for imprecise probabilities. In: de Cooman, G., Cozman, F.G., Moral, S., Walley, P. (eds.) Proceedings of ISIPTA 1999, pp. 121–130 (1999)
4. Dempster, A.: Upper and lower probabilities induced by a multi-valued mapping. Annals of Mathematical Statistics 38, 325–339 (1967)
5. Hutter, M., Zaffalon, M.: Bayesian treatment of incomplete discrete data applied to mutual information and feature selection. In: Günter, A., Kruse, R., Neumann, B. (eds.) KI 2003. LNCS (LNAI), vol. 2821, pp. 396–406. Springer, Heidelberg (2003)
6. Jiroušek, R.: On a conditional irrelevance relation for belief functions based on the operator of composition. In: Beierle, C., Kern-Isberner, G. (eds.) Dynamics of Knowledge and Belief, Proceedings of the Workshop at the 30th Annual German Conference on Artificial Intelligence, Fern Universität in Hagen, Osnabrück, pp. 28–41 (2007)
7. Jiroušek, R.: Factorization and Decomposable Models in Dempster-Shafer Theory of Evidence. In: Workshop on the Theory of Belief Functions, Brest, France (2010)
8. Jiroušek, R., Vejnarová, J.: Compositional models and conditional independence in evidence theory. Accepted to Int. J. Approx. Reasoning, doi: 10.1016/j.ijar.2010.02.005
9. Jiroušek, R., Vejnarová, J., Daniel, M.: Compositional models of belief functions. In: de Cooman, G., Vejnarová, J., Zaffalon, M. (eds.) Proc. of the 5th Symposium on Imprecise Probabilities and Their Applications, Praha, pp. 243–252 (2007)
10. Jousselme, A.L., Maupin, P.: On some properties of distances in evidence theory. In: Workshop on the Theory of Belief Functions, Brest, France (2010)
11. Klir, G.J.: Uncertainty and Information. Foundations of Generalized Information Theory. Wiley, Hoboken (2006)
12. Lauritzen, S.L.: Graphical models. Oxford University Press, Oxford (1996)
13. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
14. Shenoy, P.P.: Conditional independence in valuation-based systems. Int. J. of Approximate Reasoning 10(3), 203–234 (1994)
15. Shenoy, P.P.: Binary join trees for computing marginals in the Shenoy-Shafer architecture. Int. J. of Approximate Reasoning 17(2-3), 239–263 (1997)
16. Studený, M.: Formal properties of conditional independence in different calculi of AI. In: Moral, S., Kruse, R., Clarke, E. (eds.) ECSQARU 1993. LNCS, vol. 747, pp. 341–351. Springer, Heidelberg (1993)
17. Studený, M.: On stochastic conditional independence: the problems of characterization and description. Annals of Mathematics and Artificial Intelligence 35, 323–341 (2002)

A Gambler's Gain Prospects with Coherent Imprecise Previsions

Paolo Vicig

University of Trieste, Piazzale Europa 1, I-34127 Trieste, Italy
paolo.vicig@econ.units.it

Abstract. We explore some little investigated aspects of the well known betting scheme defining coherent lower or upper previsions in terms of admissible gains. A limiting situation (lose-or-draw) where the supremum of some gain is zero is discussed, deriving a gambler's gain evaluations and comparing the differences between the imprecise and precise prevision cases. Then, the correspondence of the betting scheme for imprecise previsions with real-world situations is analysed, showing how the gambler's profit objectives may compel him to select certain types of bets.

Keywords: Imprecise previsions, coherence, betting schemes, arbitrage.

1 Introduction

Imprecise (*lower* or *upper*) *previsions* are very general uncertainty measures for random variables, or for (indicators of) events, in which case we preferably speak of *imprecise probabilities*. Their most significant *consistency* criterion, discussed in [6], is that of *coherence*.

In this paper, some little explored facets of the notion of coherence are investigated. The key issue to introduce them is the well known *betting scheme* interpretation of the definitions of coherent lower or upper prevision (cf. Definitions 2, 3): for each *bounded random variable* or, following [6], *gamble* X, a gambler's lower prevision $\underline{P}(X)$ for X is his supremum buying price for X, while the upper prevision $\overline{P}(X)$ is his infimum selling price for X. Coherence considers the gambler's *gain*, $\underline{G}$ or $\overline{G}$ respectively, resulting from his exchanging (buying or selling) a finite number of gambles, with the rules of Definitions 2, 3. Each exchange of a single gamble is called a *bet*, and different gains $\underline{G}$ or $\overline{G}$ are obtained by varying the number and amount of the bets. Coherence for lower previsions (for upper previsions) requires that the supremum of every $\underline{G}$ (of every $\overline{G}$) is non-negative. Given this, we focus on the following issues:

(a) in extreme situations like $\sup \underline{G} = 0$ or $\sup \overline{G} = 0$ (*lose-or-draw* case), what are the gambler's beliefs about his gain prospects?
(b) A gambler will obviously try to avoid, whenever possible, case (a). How should he then select his bets, in order to increase his gain outlooks?

After supplying some preliminary material in Sect. 2, we tackle item (a) in Sect. 3, discussing it first in the special case of a de Finetti coherent, or briefly, dF-coherent *precise* prevision. In fact, the betting schemes for coherent imprecise

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 50–59, 2010.

previsions modify the earlier de Finetti's scheme for precise previsions (or probabilities), which also asks for non-negativity of the supremum of every gain G (Definition 1). Thus the situations in (a) concern dF-coherence too, and it is interesting to compare the different implications on the gambler's gain outlooks when previsions are either precise or imprecise. It is known [1] that when a dF-coherent probability P is such that $\max G = 0$ for some gain G, then necessarily the probability that just the gain G is negative is zero, $P(G < 0) = 0$. We extend this result to gains for dF-coherent previsions regarding *simple* gambles, i.e. gambles with finitely many distinct possible values (Proposition 3 (b)), while in the general case that $\sup G = 0$ we show that necessarily $P(G \leq -\varepsilon) = 0, \forall \varepsilon > 0$ (Proposition 3 (a)). Similar, but weaker results hold for coherent lower or upper previsions: only *lower* zero probabilities are induced on the corresponding gain evaluations (Propositions 4, 5). For instance, $\max \underline{G} = 0$ implies $\underline{P}(\underline{G} < 0) = 0$ with simple gambles, while $\overline{P}(\underline{G} < 0)$ might even be 1. See Example 1 for a justification of this fact, and the discussion concluding Sect. 3 for some comments on the relationship between coherence and arbitrage.

In real situations, a gambler clearly tries to avoid the lose-or-draw case, selecting rather those bets he believes to gain money from. While dF-coherent previsions do not fit well with this attitude, corresponding to a *fair* game ($P(G) = 0$ whatever is G), coherent upper/lower previsions may ensure prospects of positive gains. We study this problem in Sect. 4, showing that the gambler's bets should be one-sided: when assigning upper (lower) previsions, he should act only as a vendor (only as a buyer). This in fact happens in many practical situations: a bookie's prices are an instance of upper probabilities for "selling" events which he would certainly not buy at the same prices. Even among one-sided bets, some are better than others: betting on atomic events in a partition, say on ω_1 and separately on ω_2 (two bets), may be more profitable than allowing a single bet on $\omega_1 \vee \omega_2$. Measuring the minimum expected gain from such policies may be operationally difficult. This is done in Example 2 in the case of plausibility functions, while relationships with desirability concepts are outlined in the subsequent comments. Sect. 5 contains concluding remarks.

2 Preliminaries

Notation. Denote with $\mathcal{D}$ an *arbitrary* set of gambles.

Definition 1. *[2] Given $P : \mathcal{D} \to \mathbb{R}$, P is a* dF-coherent prevision *on $\mathcal{D}$ if and only if $\forall\, n \in \mathbb{N}^+$, $\forall\, s_1, \ldots, s_n \in \mathbb{R}$, $\forall\, X_1, \ldots, X_n \in \mathcal{D}$, defining $G = \sum_{i=1}^{n} s_i(X_i - P(X_i))$, it holds that $\sup G \geq 0$.*

Here $s_i(X_i - P(X_i))$ is the elementary gain from betting on X_i, with *stake* s_i. By introducing sign constraints on the stakes, we come to the definition of coherent lower prevision ([6], Sect. 2.5.4 (a)):

Definition 2. *Given $\underline{P} : \mathcal{D} \to \mathbb{R}$, $\underline{P}$ is a* coherent lower prevision *on $\mathcal{D}$ if and only if $\forall\, n \in \mathbb{N}$, $\forall\, s_0, \ldots, s_n \geq 0$, $\forall\, X_0, \ldots, X_n \in \mathcal{D}$, defining $\underline{G} = \sum_{i=1}^{n} s_i(X_i - \underline{P}(X_i)) - s_0(X_0 - \underline{P}(X_0))$, it holds that $\sup \underline{G} \geq 0$.*

As already recalled, $\underline{P}(X)$ is interpreted as a supremum buying price for X. The gain $\underline{G}$ is made up of $n \geq 0$ elementary gains $g_1, \ldots, g_n$ where $g_i = s_i(X_i - \underline{P}(X_i))$ arises from buying $s_i X_i$ at the price $s_i \underline{P}(X_i)$, and of the term $g_0 = -s_0(X_0 - \underline{P}(X_0)) = s_0\underline{P}(X_0) - s_0 X_0$, which may be viewed as the gambler's elementary gain for selling $s_0 X_0$ at the price $s_0 \underline{P}(X_0)$. This term, forcing the gambler to *sell* X_0 at his supremum *buying* price for it, ensures that coherent lower previsions have appropriate consistency properties not guaranteed by weaker consistency concepts, like that of previsions *avoiding sure loss* (ASL) which is in fact obtained from Definition 2 by putting $s_0 = 0$ ([6], Sect. 2.4.4 (a)). For instance, if $\mathcal{D} = \{X\}, \underline{P}(X) < \inf X$ avoids sure loss but is incoherent, since $\underline{G} = -s_0(X - \underline{P}(X))$ is such that $\sup \underline{G} < 0$ for $s_0 > 0$. On the contrary, $\underline{P}(X) = \inf X$ is a coherent assessment on X.

More generally, the *vacuous* lower prevision $\underline{P}(X) = \inf X, \forall X \in \mathcal{D}$ is coherent, and may express total lack of information about the gambles in $\mathcal{D}$.

Whenever both upper and lower previsions are assessed, they are customarily *conjugate*: $\overline{P}(X) = -\underline{P}(-X)$.

Conjugacy allows us to refer to lower or alternatively upper previsions only; results about one kind of imprecise prevision are easily reworded for the other one. The definition of coherent upper prevision is:

Definition 3. *Given* $\overline{P} : \mathcal{D} \to \mathbb{R}$, $\overline{P}$ *is a* coherent upper prevision *on* $\mathcal{D}$ *if and only if* $\forall\ n \in \mathbb{N}$, $\forall\ s_0, \ldots, s_n \geq 0$, $\forall\ X_0, \ldots, X_n \in \mathcal{D}$, *defining* $\overline{G} = \sum_{i=1}^{n} s_i(\overline{P}(X_i) - X_i) - s_0(\overline{P}(X_0) - X_0)$, *it holds that* $\sup \overline{G} \geq 0$.

Here the gambler acts as a vendor of $s_i X_i$ at the price $s_i \overline{P}(X_i)$, but may be obliged to buy $s_0 X_0$ at his selling price for it, $s_0 \overline{P}(X_0)$. The *vacuous* upper prevision $\overline{P}(X) = \sup X, \forall X \in \mathcal{D}$ is coherent.

Properties of coherent imprecise previsions are extensively discussed in [6]. We gather some of them, needed in the sequel, in the next proposition.

Proposition 1. *Let* $\underline{P}$ ($\overline{P}$) *be a coherent lower (upper) prevision on* $\mathcal{D}$. *It holds that (for properties (b) to (d), whenever the previsions are defined):*

(a) Whatever is $\mathcal{D}' \supset \mathcal{D}$, *there exists a coherent lower (upper) prevision which extends* $\underline{P}$ ($\overline{P}$) *on* $\mathcal{D}'$. *Such an extension is generally not unique, but there always exists a least-committal one, the* natural extension $\underline{E}$ ($\overline{E}$), *i.e.* $\underline{E} = \underline{P}$ *on* $\mathcal{D}$, *and* $\underline{E} \leq \underline{P}^*$ *on* $\mathcal{D}'$, *if* $\underline{P}^*$ *is any coherent extension of* $\underline{P}$ ($\overline{E} = \overline{P}$ *on* $\mathcal{D}$, *and* $\overline{E} \geq \overline{P}^*$ *on* $\mathcal{D}'$, *if* $\overline{P}^*$ *is a coherent extension of* $\overline{P}$)

(b) $\underline{P}(X + Y) \geq \underline{P}(X) + \underline{P}(Y)$, $\overline{P}(X + Y) \leq \overline{P}(X) + \overline{P}(Y)$

(c) $\underline{P}(aX + b) = a\underline{P}(X) + b$, $\overline{P}(aX + b) = a\overline{P}(X) + b, \forall a > 0, b \in \mathbb{R}$

(d) $\inf X \leq \underline{P}(X) \leq \overline{P}(X) \leq \sup X$.

Coherent imprecise previsions include several uncertainty models as special cases, for instance dF-coherent precise previsions. A dF-coherent prevision is simultaneously both a lower and an upper coherent prevision. Further,

Proposition 2. *If* P *is a dF-coherent prevision on* $\mathcal{D}$, *then:* (a) *there exists a dF-coherent extension of* P *on any* $\mathcal{D}' \supset \mathcal{D}$; (b) $P(aX + bY) = aP(X) + bP(Y)$.

When all gambles in $\mathcal{D}$ are (indicators of) events, previsions narrow down to probabilities, and the conjugacy relation between upper and lower probabilities is written as $\overline{P}(A) = 1 - \underline{P}(A^c)$, where A^c is the negation of event A.

The symbol I_A will denote the indicator of event A.

3 The Lose-or-Draw Case

Let us consider first a dF-coherent prevision P on $\mathcal{D}$. We prove that

Proposition 3. *Given a dF-coherent prevision P on $\mathcal{D}$, suppose that for some $n \in \mathbb{N}^+$, $s_1, \ldots, s_n \in \mathbb{R}$, $X_1, \ldots, X_n \in \mathcal{D}$ it holds for $G = \sum_{i=1}^{n} s_i(X_i - P(X_i))$ that $\sup G = 0$. Then (the dF-coherent extension of) P is uniquely determined on certain events concerning G, and precisely:*

(a) $P(G \leq -\varepsilon) = 0, \forall \varepsilon > 0$
(b) If in addition $X_1, \ldots, X_n$ are all simple, we also have that $P(G < 0) = 0$.[1]

Proof. Proof of (a). Suppose $\inf G < 0$ (if $\inf G = 0$, the proof is trivial). The thesis claims that if a dF-coherent P is assessed on $\mathcal{D}$ and there is a gain G such that $\sup G = 0$, then for any given $\varepsilon > 0$ the dF-coherent extension of P on $\mathcal{D}' = \mathcal{D} \cup \{G \leq -\varepsilon\}$ (there exists one by Proposition 2 (a)) is unique and such that $P(G \leq -\varepsilon) = 0$. Thus, it is sufficient to show by contradiction that when $P(G \leq -\varepsilon) > 0$ there is some gain G' on $\mathcal{D}'$ such that $\max G' < 0$.

Given $\varepsilon > 0$, assume then $P(G \leq -\varepsilon) > 0$ and consider the atoms of a partition $\mathbb{P}_c$ describing all jointly possible values of $X_1, \ldots, X_n$ and hence of G. These are of two types: atoms of type ω^+, if $G(\omega^+) > -\varepsilon$; atoms of type ω^-, if $G(\omega^-) \leq -\varepsilon$.

Define $G' = G + s(I_{G \leq -\varepsilon} - P(G \leq -\varepsilon))$ and choose k such that $0 < k < \varepsilon P(G \leq -\varepsilon)$.

At ω^+, $G'(\omega^+) = G(\omega^+) - sP(G \leq -\varepsilon) \leq -k$ iff $s \geq \frac{G(\omega^+)+k}{P(G \leq -\varepsilon)}$. Hence $G' \leq -k$ for *all* atoms of this type if

$$s > \sup_{\omega^+} \left\{ \frac{G(\omega^+) + k}{P(G \leq -\varepsilon)} \right\} = \frac{k}{P(G \leq -\varepsilon)} > 0 \ . \tag{1}$$

At ω^-, $G'(\omega^-) = G(\omega^-) + s(1 - P(G \leq -\varepsilon))$. Since $-G(\omega^-) \geq \varepsilon, \forall \omega^-$, we have that $G' \leq -k$ for *all* atoms of this type if

$$s(1 - P(G \leq -\varepsilon)) < -k + \varepsilon \ . \tag{2}$$

Note that $-k + \varepsilon > 0$, because $k < \varepsilon P(G \leq -\varepsilon)$. If $P(G \leq -\varepsilon) = 1$, then (2) trivially holds, while (1) can be satisfied by choosing s sufficiently large. This establishes (a).

[1] The dF-coherent extension of P is mentioned explicitly because (the indicators of) the events $(G \leq -\varepsilon)$ and $(G < 0)$ need not belong to $\mathcal{D}$. We shall omit similar specifications in the analogous Propositions 4 and 5.

On the other hand, if $P(G \leq -\varepsilon) < 1$, then $\sup_{\omega^-} G'(\omega^-) \leq -k$ if $s < \frac{-k+\varepsilon}{1-P(G\leq-\varepsilon)}$; hence, using also (1), $\sup G' \leq -k$ if $0 < \frac{k}{P(G\leq-\varepsilon)} < s < \frac{-k+\varepsilon}{1-P(G\leq-\varepsilon)}$.

It is immediate to check that $\frac{k}{P(G\leq-\varepsilon)} < \frac{-k+\varepsilon}{1-P(G\leq-\varepsilon)}$ iff $k < \varepsilon P(G \leq -\varepsilon)$, as already assumed. Hence also in this case (a) is established.

Proof of (b). Follows from (a), noting that when $X_1, \ldots, X_n$ are all simple G takes on finitely many values, hence there exists $\varepsilon > 0$ such that $(G < 0)$ and $(G \leq -\varepsilon)$ are equal. □

When the supremum of some gain, let it be G^*, is 0 for a dF-coherent prevision P, then $\inf G = 0$ for $G = -G^*$. Also G is an admissible gain, by Definition 1, for *the same* prevision P, and realises the opposite *win-or-draw* case for the gambler. However, the gambler is rather skeptical about his winning chances:

Corollary 1. *If for some dF-coherent P there is some gain G in Definition 1 such that* $\inf G = 0$*, then necessarily $P(G \geq \varepsilon) = 0, \forall \varepsilon > 0$.*

Proof. Since $\inf G = 0$ iff $\sup(-G) = 0$, we may apply Proposition 3 (a) to the admissible gain $-G$, getting $P(-G \leq -\varepsilon) = P(G \geq \varepsilon) = 0$. □

We turn now to coherent imprecise previsions. The next proposition is the version for coherent lower previsions of Proposition 3.

Proposition 4. *Given a coherent lower prevision $\underline{P}$ on $\mathcal{D}$, suppose that for some $n \in \mathbb{N}$, $s_0, \ldots, s_n \geq 0$, $X_0, \ldots, X_n \in \mathcal{D}$, the gain $\underline{G} = \sum_{i=1}^{n} s_i(X_i - \underline{P}(X_i)) - s_0(X_0 - \underline{P}(X_0))$ is such that* $\sup \underline{G} = 0$*. This implies that*

(a) $\underline{P}(\underline{G} \leq -\varepsilon) = 0, \forall \varepsilon > 0$.
(b) If $X_0, \ldots, X_n$ are all simple, $\underline{P}(\underline{G} < 0) = 0$.

Proof. The proof replicates that of Proposition 3, after replacing P, G, G' with $\underline{P}, \underline{G}, \underline{G}'$, $X_1, \ldots, X_n$ with $X_0, \ldots, X_n$, and the terms "dF-coherent", "Proposition 2 (a)" with "coherent", "Proposition 1 (a)".

Note that $\underline{G}'$ is an admissible gain by Definition 2, since it adds to $\underline{G}$ a term, $s(I_{\underline{G}\leq-\varepsilon} - \underline{P}(\underline{G} \leq -\varepsilon))$, whose stake s is positive. □

One might wonder whether stronger implications than those in Proposition 4 hold, in particular whether $\overline{P}(\underline{G} < 0) = 0$ in case (b) (or $\overline{P}(\underline{G} \leq -\varepsilon) = 0$ in case (a)). The answer is no, as shown by the next example.

Example 1. Let $\mathcal{D} = \{X\}$, where X is a simple gamble, $\underline{P}(X) = \min X$. If $n = 0, s_0 > 0$ in Definition 2, the gain $\underline{G} = -s_0(X - \min X)$ is such that $\max \underline{G} = 0$. It is coherent to assess $\overline{P}(\underline{G} < 0) = 1 - \underline{P}(\underline{G} \geq 0) = 1 - \underline{P}(\underline{G} = 0) = 1$. In fact this is equivalent to extending $\underline{P}$ on $\mathcal{D}' = \{X\} \cup \{\underline{G} = 0\}$ letting $\underline{P}(\underline{G} = 0) = 0$, and the extension is coherent, being the vacuous lower prevision on $\mathcal{D}'$.

Putting $\overline{P}(\underline{G} < 0) = 1$ in this example is also intuitively sound: the assignment $\underline{P}(X) = \min X$ is vacuous, hence the gambler has no significant information on X. Quite reasonably then, his opinion about the gain $\underline{G}$ may be vacuous too. A dF-coherent assessment $P(X) = \min X$ would be radically different, meaning that the gambler is almost sure that X is equal to its minimum. □

A result similar to Proposition 4 holds for upper previsions:

Proposition 5. *Given a coherent upper prevision $\overline{P}$ on $\mathcal{D}$, suppose there exist $n \in \mathbb{N}$, $s_0, \ldots, s_n \geq 0$, $X_0, \ldots, X_n \in \mathcal{D}$, such that for the corresponding gain $\overline{G} = \sum_{i=1}^{n} s_i(\overline{P}(X_i) - X_i) - s_0(\overline{P}(X_0) - X_0)$ it is $\sup \overline{G} = 0$. Then*

(a) $\underline{P}(\overline{G} \leq -\varepsilon) = 0, \forall \varepsilon > 0$.
(b) If $X_0, \ldots, X_n$ are all simple, $\underline{P}(\overline{G} < 0) = 0$.

Proof. Use conjugacy to write $\overline{G}$ as a gain concerning the conjugate $\underline{P}$ of $\overline{P}$, which is a lower prevision coherent on $\mathcal{D}^- = \{-X : X \in \mathcal{D}\}$: $\overline{G} = \sum_{i=1}^{n} s_i(-X_i - \underline{P}(-X_i)) - s_0(-X_0 - \underline{P}(-X_0))$. Then apply Proposition 4. □

Also $\overline{P}(\overline{G} < 0)$ may be 1 when $\max \overline{G} = 0$. We can see this in a modified version of Example 1, where again $\mathcal{D} = \{X\}$ and X is a simple gamble, but we assess now $\overline{P}(X) = \max X$. Then $\max \overline{G} = 0$ for $\overline{G} = -s_0(\max X - X)$, $s_0 > 0$, but $\overline{P}(\overline{G} < 0) = 1$ is a coherent extension of $\overline{P}$ on $\mathcal{D}' = \mathcal{D} \cup \{\overline{G} < 0\}$, being the vacuous upper prevision on $\mathcal{D}'$.

Finally, note that Propositions 3, 4 and 5 imply, respectively, that the *lower distribution functions* $\underline{F}_G(x) = \underline{P}(G \leq x)$, $\underline{F}_{\overline{G}}(x) = \underline{P}(\overline{G} \leq x)$ and the *distribution function* $F_G(x) = \overline{P}(X \leq x)$ are all piecewise constant and equal to 0 for $x < 0$, to 1 for $x \geq 0$, thus being left-discontinuous at 0.

Discussion. It is well known from the least general coherence concept, dF-coherence for precise probabilities, that replacing condition $\max G \geq 0$ with $\max G > 0$ would rule out lose-or-draw bets while introducing other significant constraints (every event $A \neq \emptyset, A \neq \Omega$ should be given probability *strictly* between 0 and 1, cf. [1], Sect. 9.3.4). More general coherence concepts are obviously concerned with similar problems. The existence of gains whose supremum is 0 is therefore something to cope with, in the coherence framework.

Those of such gains whose infimum is negative correspond to an *arbitrage* opportunity for the gambler's competitor, since his gain (the opposite of gambler's gain) is non-negative, and positive in some cases. More restrictive concepts of arbitrage may leave out some, but not all, of these instances, cf. [4]. Thus, while incoherence implies the existence of an arbitrage opportunity, coherence does not always exclude it. We saw that the gambler's fears of suffering from an arbitrage are however rather limited. A basic difference between precise and imprecise previsions is that these beliefs are expressed in terms of zero probabilities in the former case, of zero lower probabilities only in the latter.

There is another important difference: the dF-coherence condition $\sup G \geq 0$ is equivalent to $\inf G \cdot \sup G \leq 0$, hence $\inf G$ cannot be greater than 0, and even when it is 0, Corollary 1 shows that the gambler does not really expect to gain much. The equivalence depends on the fact that when G is an admissible gain for dF-coherence, so is also $-G$. In particular, when $\sup G = 0$, then $\inf(-G) = 0$.

With imprecise previsions, there exist gains from coherent evaluations whose infimum is positive. Take the vacuous upper probability on a finite partition $\mathbb{P} = \{\omega_1, \ldots, \omega_n\}$, i.e. $\overline{P}(\omega_i) = 1$, $i = 1, \ldots, n$. The admissible gain $\overline{G} = \sum_{i=1}^{n}(1 - I_{\omega_i})$

is constant and equal to $n-1$. These facts are possible because condition $\sup \overline{G} \geq 0$ is not equivalent to $\inf \overline{G} \cdot \sup \overline{G} \leq 0$: what precludes the equivalence is that if $\overline{G}$ is an admissible gain for $\overline{P}$, $-\overline{G}$ is generally not.

Thus, while dF-coherence treats symmetrically the lose-or-draw and win-or-draw cases, coherence does not, accepting even the "win only" case.

4 Bet Selection and Win Prospects

In the ideal schemes of Definitions 1, 2, 3, a gambler should accept all bets guaranteeing gains whose supremum is non-negative. In real-world, the organizer of a game, let us call him *House*, usually sells each gamble X to a counterpart, *Bettor*, at a price $\overline{P}(X)$ fixed by himself. We shall mainly focus on House's viewpoint, since Bettor is often the weaker party, having individually no power to change the rules or prices in the game, but being merely allowed to decide whether to play or not. Real situations, such as those where House is the organizer of some lottery, a bookie or an insurer, are rather close to this pattern. In such instances, House's prices can not be modelled by dF-coherent previsions: the gain G in Definition 1 is expected to be 0, i.e. $P(G) = 0$, as follows easily using Proposition 2 (b), while House aims at a positive (expected) gain. Coherent upper previsions, interpreted as House's selling prices, appear an adequate model. The counterpart, Bettor, should use his coherent lower previsions as buying prices.

The problem we tackle in this section is item (b) in the Introduction, i.e. how should House and Bettor select their bets? In House's eyes (the viewpoint for Bettor is specular), gains where $\sup \overline{G} = 0$ should be avoided. Those bets ensuring $\inf \overline{G} > 0$ are the ideal ones, but they might of course be hard to place.

In the generic case that $\overline{G}$ ($\underline{G}$) can be both positive and negative, it is convenient for House (Bettor) not to accept bets with $s_0 \neq 0$, thus acting exclusively as a sellor (buyer). This ensues from the following simple results.

Proposition 6. *Let $\underline{P}$ be a coherent lower prevision on $\mathcal{D}$ and $\overline{P}$ its conjugate. Consider a gain concerning $\underline{P}$.*

(a) For a gain $\underline{G}_0 = -s_0(X_0 - \underline{P}(X_0))$, it is necessary for coherence on $\mathcal{D} \cup \{\underline{G}_0\}$ that: $\underline{P}(\underline{G}_0) = s_0(\underline{P}(X_0) - \overline{P}(X_0))$ (≤ 0), $\overline{P}(\underline{G}_0) = 0$;
(b) For a gain $\underline{G} = \sum_{i=1}^{n} s_i(X_i - \underline{P}(X_i)) - s_0(X_0 - \underline{P}(X_0))$, the following bounds hold: $\underline{P}(\underline{G}) \geq s_0(\underline{P}(X_0) - \overline{P}(X_0))$, $\overline{P}(\underline{G}) \leq \sum_{i=1}^{n} s_i(\overline{P}(X_i) - \underline{P}(X_i))$.

Let $\overline{P}$ be a coherent upper prevision on $\mathcal{D}$ and $\underline{P}$ its conjugate. Consider a gain concerning $\overline{P}$.

(c) For a gain $\overline{G}_0 = -s_0(\overline{P}(X_0) - X_0)$, it is necessary for coherence on $\mathcal{D} \cup \{\overline{G}_0\}$ that: $\underline{P}(\overline{G}_0) = s_0(\underline{P}(X_0) - \overline{P}(X_0))$ (≤ 0), $\overline{P}(\overline{G}_0) = 0$;
(d) For a gain $\overline{G} = \sum_{i=1}^{n} s_i(\overline{P}(X_i) - X_i) - s_0(\overline{P}(X_0) - X_0)$, the following bounds hold: $\underline{P}(\overline{G}) \geq s_0(\underline{P}(X_0) - \overline{P}(X_0))$, $\overline{P}(\overline{G}) \leq \sum_{i=1}^{n} s_i(\overline{P}(X_i) - \underline{P}(X_i))$.

Proof. Proof of (a). Using Proposition 1 (c) and conjugacy, $\underline{P}(\underline{G}_0) = s_0\underline{P}(\underline{P}(X_0) - X_0) = s_0(\underline{P}(X_0) + \underline{P}(-X_0)) = s_0(\underline{P}(X_0) - \overline{P}(X_0))$ $(\leq 0$ by Proposition 1 (d)), and $\overline{P}(\underline{G}_0) = s_0\overline{P}(\underline{P}(X_0) - X_0) = s_0(\underline{P}(X_0) + \overline{P}(-X_0)) = 0$.

Proof of (b). Use Proposition 1 (b), (c), and (a) above: $\underline{P}(\underline{G}) \geq \sum_{i=1}^{n} s_i \underline{P}(X_i - \underline{P}(X_i)) + \underline{P}(-s_0(X_0 - \underline{P}(X_0))) = \underline{P}(\underline{G}_0) = s_0(\underline{P}(X_0) - \overline{P}(X_0))$, while $\overline{P}(\underline{G}) \leq \sum_{i=1}^{n} s_i \overline{P}(X_i - \underline{P}(X_i)) + \overline{P}(-s_0(X_0 - \underline{P}(X_0))) = \sum_{i=1}^{n} s_i(\overline{P}(X_i) - \underline{P}(X_i))$.

The proof of (c), (d) is obtained easily from (a), (b), using conjugacy. □

Proposition 7. *Let $\underline{P}$ be a coherent lower prevision on $\mathcal{D}$, and consider a gain $\underline{G} = \underline{G}_{ASL} + \underline{G}_0$,[2] where $\underline{G}_{ASL} = \sum_{i=1}^{n} s_i(X_i - \underline{P}(X_i))$, $\underline{G}_0 = -s_0(X_0 - \underline{P}(X_0))$. Then, for any coherent extension of $\underline{P}$ on $\mathcal{D} \cup \{\underline{G}, \underline{G}_{ASL}, \underline{G}_0\}$, and of its conjugate $\overline{P}$, it holds that $\underline{P}(\underline{G}_{ASL}) \geq \underline{P}(\underline{G})$, $\overline{P}(\underline{G}_{ASL}) \geq \overline{P}(\underline{G})$.*

For a coherent upper prevision $\overline{P}$ on $\mathcal{D}$, it holds (using corresponding definitions) that

$$\underline{P}(\overline{G}_{ASL}) \geq \underline{P}(\overline{G}), \quad \overline{P}(\overline{G}_{ASL}) \geq \overline{P}(\overline{G}) \ . \tag{3}$$

Proof. We prove the inequalities in (3) (the proof of those for $\underline{G}_{ASL}$ is analogue), using Proposition 1 (b), conjugacy and Proposition 6 (c):

$$\underline{P}(\overline{G}_{ASL}) = \underline{P}(\overline{G} - \overline{G}_0) \geq \underline{P}(\overline{G}) + \underline{P}(-\overline{G}_0) = \underline{P}(\overline{G}) - \overline{P}(\overline{G}_0) = \underline{P}(\overline{G}).$$
$$\overline{P}(\overline{G}) = \overline{P}(\overline{G}_{ASL} + \overline{G}_0) \leq \overline{P}(\overline{G}_{ASL}) + \overline{P}(\overline{G}_0) = \overline{P}(\overline{G}_{ASL}).$$ □

Propositions 6 and 7 have important implications in the bet selection strategy of House and Bettor. Considering House, he should not accept any one-bet gain $\overline{G}_0$ where he buys $s_0 X_0$ at his selling price $s_0 \overline{P}(X_0)$. In fact, from Proposition 6 (c) a negative or at best null gain is expected in this case.

More generally, House should not engage in any game $\mathcal{G}$ including a bet where he buys $s_0 X_0$ for $s_0 \overline{P}(X_0)$, that is, he should buy nothing at his selling prices. In fact, House's upper and lower previsions $\overline{P}(\overline{G}_{ASL}), \underline{P}(\overline{G}_{ASL})$ for the gain $\overline{G}_{ASL}$ obtained from $\mathcal{G}$ by cancelling the bet on $s_0 X_0$ are not smaller than those for the gain $\overline{G}$ from $\mathcal{G}$, by (3), and at any rate $\underline{P}(\overline{G}_{ASL}) \geq 0$, from the first inequality in Proposition 6 (d) with $s_0 = 0$.

Although House expects the gain $\overline{G}_{ASL}$ to be at least non-negative, it is possible that $\underline{P}(\overline{G}_{ASL}) = 0$, for instance when House makes a single bet ($n = 1$): $\underline{P}(s_1(\overline{P}(X_1) - X_1)) = 0$.

The lowest coherent value for $\underline{P}(\overline{G}_{ASL})$, i.e. the *natural extension* $\underline{E}(\overline{G}_{ASL})$, may also be strictly positive, even when $\inf \overline{G}_{ASL} < 0$, cf. Example 2.

Given that House will take up one-sided bets only, how can he try to maximise his expected gain? Clearly, restricting the subadditivity property in Proposition 1 (b), $\overline{P}(X+Y) \leq \overline{P}(X) + \overline{P}(Y)$: the strict inequality corresponds to a *discount* to Bettor when he makes a single bet on $X+Y$ (with stake s) rather than two bets (both with the same stake s), on X, Y respectively.

An interesting possibility for House is to employ coherent upper previsions which are *2-alternating*, because this ensures that they are additive for *comonotonic* gambles [3].

In the case that the potential bets regard events in the powerset $\mathcal{A}(\mathbb{P})$ of a *finite* partition $\mathbb{P} = \{\omega_1, \ldots, \omega_n\}$, House has the further option of restricting to

[2] We write $\underline{G}_{ASL}$ as this is the generic gain in the consistency condition of avoiding sure loss ([6], Sect. 2.4.4 (a)).

the atomic events in $\mathbb{P}$. This obliges a counterpart, willing to bet on $A \in \mathcal{A}(\mathbb{P})$ with stake s, to bet *separately* on some or all $\omega_i \in \mathbb{P}, \omega_i \Rightarrow A$. If he bets on all such ω_i with the same stake s, his total expense $s \sum_{\omega_i \Rightarrow A} \overline{P}(\omega_i)$ will be not smaller than $s\overline{P}(A)$, which means that House allows for no discount at all. The option is sometimes applied in practice: a bookie offering bets on "A wins against B", "A and B draw", "B wins", does not always offer a bet also on, say, "B does not lose". In the next example we consider the consequences of this strategy on House's gain outlooks when his upper probabilities are plausibility functions.

Example 2. Suppose House assigns a *plausibility function* $\overline{P}$ on $\mathcal{A}(\mathbb{P})$, $\overline{P}(A) = \sum_{B:B\wedge A\neq\emptyset} m(B)$, $\forall A \in \mathcal{A}(\mathbb{P})$, where m is a *mass function*, i.e. $m : \mathcal{A}(\mathbb{P}) \rightarrow [0,1]$, $m(\emptyset) = 0$ and $\sum_{B\in\mathcal{A}(\mathbb{P})} m(B) = 1$. This is equivalent to assigning the conjugate *belief function* $\underline{P}$ on $\mathcal{A}(\mathbb{P})$ using the same m, $\underline{P}(A) = \sum_{B\Rightarrow A} m(B)$. Even though they were introduced independently [5], belief functions are notable instances of coherent lower probabilities [7], and the *natural extension* $\underline{E}(X)$ of a belief function $\underline{P}$ on any gamble X defined on $\mathbb{P}$ is known ([6], Note 2 to Sect. 3.2) to be $\underline{E}(X) = \sum_{B\in\mathcal{A}(\mathbb{P})} m(B) \min_{\omega_i\Rightarrow B}\{X(\omega_i)\}$.

If House bets on some or all atomic events in $\mathbb{P} = \{\omega_1, \ldots, \omega_n\}$ only, $\overline{G} = \sum_{i=1}^{n} s_i(\overline{P}(\omega_i) - I_{\omega_i})$, with $s_i = 0$ if House does not bet on ω_i. Applying the natural extension formula above to $\overline{G}$, the property $\sum_{B\in\mathcal{A}(\mathbb{P})} m(B) = 1$, writing $\max_B\{s_i\}$ and $\sum_B s_i$ to mean that the maximum, respectively the summation is made over all s_i such that $\omega_i \Rightarrow B$, and noting (at the fourth equality) that $\{B : B \wedge \omega_i \neq \emptyset\} = \{B : \omega_i \Rightarrow B\}$,

$$
\begin{aligned}
&\underline{E}(\overline{G}) = \textstyle\sum_{B\in\mathcal{A}(\mathbb{P})} m(B) \cdot \min_{\omega_i\Rightarrow B}\{\sum_{i=1}^{n} s_i(\overline{P}(\omega_i) - I_{\omega_i})\} \\
&= \textstyle\sum_{i=1}^{n} s_i\overline{P}(\omega_i) - \sum_{B\in\mathcal{A}(\mathbb{P})} m(B) \cdot \max_{\omega_i\Rightarrow B} \{\sum_{i=1}^{n} s_i I_{\omega_i}\} \\
&= \textstyle\sum_{i=1}^{n} s_i \sum_{B:B\wedge\omega_i\neq\emptyset} m(B) - \sum_{B\in\mathcal{A}(\mathbb{P})} m(B) \cdot \max_B \{s_i\} \\
&= \textstyle\sum_{B\in\mathcal{A}(\mathbb{P})} m(B) \sum_B s_i - \sum_{B\in\mathcal{A}(\mathbb{P})} m(B) \cdot \max_B \{s_i\} \\
&= \textstyle\sum_{B\in\mathcal{A}(\mathbb{P})-\mathbb{P}} m(B) \cdot (\sum_B s_i - \max_B\{s_i\}) \geq 0.
\end{aligned}
$$

In accordance to previously pointed out facts, $\underline{E}(\overline{G}) = 0$ when $m(B) = 0, \forall B \in \mathcal{A}(\mathbb{P}) - \mathbb{P}$ (in which case the plausibility reduces to a precise probability), or when House bets on just one $\omega_i \in \mathbb{P}$. Otherwise, if House bets on $r \geq 2$ atomic events, say for notational ease on $\omega_1, \ldots, \omega_r$, $\underline{E}(\overline{G}) > 0$ if and only if at least one non-atomic event $B \in \mathcal{A}(\mathbb{P}) - \mathbb{P}$, such that $B \wedge (\omega_1 \vee \ldots \vee \omega_r) \neq 0$, is given positive mass, $m(B) > 0$. In other words, House expects a positive gain if and only if his uncertainty evaluations on the non-atomic events compatible with $\omega_1 \vee \ldots \vee \omega_r$ are imprecise for at least one of them. □

Comments. It seems reasonable to assume that in real-world situations House is inclined to propose those and only those bets guaranteeing a strictly positive expected gain $\overline{G}$, which may be (prudentially) evaluated by the natural extension $\underline{E}(\overline{G})$ of House's assessments. This criterion clearly enforces the following two of the three axioms for *strictly desirable* (SD, in short) gambles in [6], Sect. 3.7.8: avoiding partial loss (if $\overline{G} \leq 0$, $\overline{G}$ is not SD), openness (if $\overline{G}$ is SD, then either $\overline{G} \geq 0$ or $\overline{G} - \delta$ is SD for some $\delta > 0$), while it is slightly stronger than the third

axiom (accepting partial gains: $\overline{G} \geq 0$ and $\overline{G} \neq 0$ imply that $\overline{G}$ is SD). Thus the bet selection criterion based on $\underline{E}(\overline{G})$ may be also viewed as an operational way of requiring strict desirability for $\overline{G}$.

Note further that when $\sup \overline{G} = 0$, as in Section 3, $\overline{G}$ is not SD. However, $\overline{G}$ might possibly be almost desirable ([6], Sect. 3.7.3).

5 Conclusions

The investigation in this paper confirms the distinction between the theoretical and practical significance of the coherence definition. A game organizer (House) does not necessarily take up a bet just because it is allowed in the coherence theoretical framework, and precisely coherence arguments show how his aim for profit makes him restrict the range of accepted bets to (a subset of) those admissible with the weaker condition of avoiding sure loss. On the other hand, House could not propose selling prices that avoid sure loss only, as they might be inconsistent or simply unrealistic (for instance, $\overline{P}(X)$ might be higher than $\sup X$). In this sense, the bets in the coherence framework not operated by House, like that corresponding to $-s_0(\overline{P}(X_0) - X_0)$ or those ensuring only $\sup \overline{G} = 0$, have an also practical valiance, forcing House to apply coherent (and more acceptable) prices. Partial results not reported here show that many of these conclusions apply to other concepts of consistency, for instance to coherence for precise or imprecise conditional previsions.

Acknowledgements. I wish to thank the referees for their helpful comments, and to acknowledge financial support from the PRIN Project 'Metodi di valutazione di portafogli assicurativi danni per il controllo della solvibilità'.

References

1. Crisma, L.: Introduzione alla teoria delle probabilità coerenti. EUT, Trieste (2006)
2. de Finetti, B.: Theory of Probability, vol. I. Wiley, London (1974)
3. de Cooman, G., Troffaes, M.C.M., Miranda, E.: n-Monotone lower previsions. Journal of Intelligent & Fuzzy Systems 16, 253–263 (2005)
4. Schervish, M.J., Seidenfeld, T., Kadane, J.B.: The fundamental theorem of prevision and asset pricing. Internat. J. Approx. Reason. 49, 148–158 (2008)
5. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
6. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman & Hall, London (1991)
7. Walley, P.: Measures of uncertainty in expert systems. Artificial Intelligence 83, 1–58 (1996)

Infinite Exchangeability for Sets of Desirable Gambles

Gert de Cooman[1] and Erik Quaeghebeur[1,2]

[1] Ghent University, SYSTeMS Research Group
Technologiepark–Zwijnaarde 914, 9052 Zwijnaarde, Belgium
{Gert.deCooman,Erik.Quaeghebeur}@UGent.be
[2] Department of Philosophy, Baker Hall 135, Carnegie Mellon University
Pittsburgh, PA 15213-3890, United States
ErikQ@andrew.cmu.edu

Abstract. Sets of desirable gambles constitute a quite general type of uncertainty model with an interesting geometrical interpretation. We study infinite exchangeability assessments for them, and give a counterpart of de Finetti's infinite representation theorem. We show how the infinite representation in terms of frequency vectors is tied up with multivariate Bernstein (basis) polynomials. We also lay bare the relationships between the representations of updated exchangeable models, and discuss conservative inference (natural extension) under exchangeability.

Keywords: desirability, weak desirability, sets of desirable gambles, coherence, exchangeability, representation, natural extension, updating.

1 Introduction

In this paper, we bring together desirability,[1] an interesting approach to modelling uncertainty, with infinite exchangeability, a structural assessment for uncertainty models that is important for inference purposes.

Desirability, or the theory of (coherent) sets of desirable gambles, represents the uncertainty of a subject with a set of gambles[2] that she finds desirable. This theory is more expressive than the theory of (coherent) lower previsions [11], which itself is a generalization of the theory of linear previsions [5]. A complete model for a rational subject's uncertainty is a coherent set of desirable gambles, or mathematically, a cone satisfying some constraints. This geometric aspect is one of the points that make desirability appealing. Another is its generality [12]. We introduce the necessary desirability-related concepts in Sec. 2.

Here we study infinite exchangeability[1] for sets of desirable gambles. An exchangeability assessment expresses that the order of the samples in a sequence of them is irrelevant for inference purposes. This study builds on earlier work on exchangeability for coherent lower previsions [2] and finite exchangeability for

[1] For a brief historical overview, see some of our earlier work [1, Sec. 1].
[2] Gambles are also called bets or random rewards.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 60–69, 2010.

sets of desirable gambles [1]. We need to recall the definition and representation results for finite exchangeability of sets of desirable gambles (Sec. 3). After that, we can give our definition and representation result of infinite exchangeability for sets of desirable gambles and give some results about updating and natural extension under exchangeability (Sec. 4).

We end with some conclusions (Sec. 5).

2 Desirability

Consider an experiment with a non-empty set Ω describing its mutually exclusive possible outcomes, and a subject who is uncertain about its outcome.

Sets of desirable gambles. A *gamble* f is a bounded real-valued map on Ω, and it is interpreted as an uncertain reward. When the actual outcome of the experiment is ω, then the corresponding (possibly negative) reward is $f(\omega)$. The set of all gambles is $\mathcal{G}(\Omega)$.

We say that a non-zero gamble f is *desirable* to a subject if she accepts to engage in the following transaction, where: (i) the actual outcome ω of the experiment is determined, and (ii) she receives the reward $f(\omega)$. The zero gamble is not considered to be desirable.[3]

We try to model the subject's beliefs about the outcome of the experiment by considering which gambles are desirable for her. We suppose the subject has some set $\mathcal{R} \subseteq \mathcal{G}(\Omega)$ of desirable gambles.

Coherence. Not every such set should be considered as a reasonable model, and in what follows, we give an abstract and fairly general treatment of ways to impose 'rationality' constraints on sets of desirable gambles.

The set $\mathcal{G}(\Omega)$ of all gambles on Ω is a linear space with respect to the (point-wise) addition of gambles, and the (point-wise) scalar multiplication of gambles with real numbers. The *positive hull operator* posi generates the set of finite *strictly* positive linear combinations of elements of its argument set. A subset $\mathcal{C}$ of $\mathcal{G}(\Omega)$ is a *convex cone* if $\text{posi}(\mathcal{C}) = \mathcal{C}$.

Consider a linear subspace $\mathcal{K}$ of the linear space $\mathcal{G}(\Omega)$. With any convex cone $\mathcal{C} \subset \mathcal{K}$ such that $0 \in \mathcal{C}$ we can always associate a *vector ordering* $\succeq$ on $\mathcal{K}$, defined by $f \succeq g \Leftrightarrow f - g \in \mathcal{C} \Leftrightarrow f - g \succeq 0$. The partial ordering $\succeq$ turns $\mathcal{K}$ into an ordered linear space [10, Section 11.44]. We also write $f \succ g$ if $f \succeq g$ and $f \neq g$. Finally, we let $\mathcal{K}_{\preceq 0} := \{f \in \mathcal{K} \colon f \preceq 0\} = -\mathcal{C}$ and $\mathcal{K}_{\succ 0} := \{f \in \mathcal{K} \colon f \succ 0\} = \mathcal{C}_0$.[4]

Definition 1 (Avoiding non-positivity and coherence). *Let $\mathcal{K}$ be a linear subspace of $\mathcal{G}(\Omega)$ and let $\mathcal{C} \subset \mathcal{K}$ be a convex cone containing the zero gamble 0.*

A set of desirable gambles $\mathcal{R} \subseteq \mathcal{K}$ avoids non-positivity relative to $(\mathcal{K}, \mathcal{C})$ *if $f \not\preceq 0$ for all gambles f in* $\text{posi}(\mathcal{R})$*, or in other words if $\mathcal{K}_{\preceq 0} \cap \text{posi}(\mathcal{R}) = \emptyset$.*

A set of desirable gambles $\mathcal{R} \subseteq \mathcal{K}$ is coherent relative to $(\mathcal{K}, \mathcal{C})$ *if it satisfies the following requirements, for all gambles f, f_1, and f_2 in $\mathcal{K}$ and all real $\lambda > 0$:*

[3] For clarification on the confusing nomenclature in the literature, see [1, footnote 2].

[4] Subscripting a set of gambles with zero *removes* the zero gamble, if present.

D1. *if* $f = 0$ *then* $f \notin \mathcal{R}$*;*
D2. *if* $f \succ 0$ *then* $f \in \mathcal{R}$*, or equivalently* $\mathcal{K}_{\succ 0} \subseteq \mathcal{R}$*;*
D3. *if* $f \in \mathcal{R}$ *then* $\lambda f \in \mathcal{R}$ *[scaling];*
D4. *if* $f_1, f_2 \in \mathcal{R}$ *then* $f_1 + f_2 \in \mathcal{R}$ *[combination].*
We denote by $\mathbb{D}_{(\mathcal{K},\mathcal{C})}(\Omega)$ *the set of sets of desirable gambles that are coherent relative to* $(\mathcal{K},\mathcal{C})$*.*

These requirements make $\mathcal{R}$ a *cone* that excludes $\mathcal{K}_{\preceq 0}$:
D5. if $f \preceq 0$ then $f \notin \mathcal{R}$, or equivalently $\mathcal{K}_{\preceq 0} \cap \mathcal{R} = \emptyset$.
We see that $\mathcal{K}$ is *never*, and $\mathcal{K}_{\succ 0}$ is *always* coherent relative to $(\mathcal{K},\mathcal{C})$.

Natural extension. If we consider an arbitrary non-empty family of sets of desirable gambles that are coherent relative to $(\mathcal{K},\mathcal{C})$, then so is their intersection. If a subject gives us an *assessment*, a set $\mathcal{A} \subseteq \mathcal{K}$ of gambles on Ω that she finds desirable, then the following theorem tells us exactly when this assessment can be extended to a coherent set, and how to construct the smallest such set.

Theorem 1 (Natural extension). *Let* $\mathcal{K}$ *be a linear subspace of* $\mathcal{G}(\Omega)$ *and let* $\mathcal{C} \subset \mathcal{K}$ *be a convex cone containing the zero gamble* 0*.*

Consider an assessment $\mathcal{A} \subseteq \mathcal{K}$*, and define its* $(\mathcal{K},\mathcal{C})$-natural extension*:* $\mathcal{E}_{(\mathcal{K},\mathcal{C})}(\mathcal{A}) := \bigcap \{\mathcal{R} \in \mathbb{D}_{(\mathcal{K},\mathcal{C})}(\Omega) \colon \mathcal{A} \subseteq \mathcal{R}\}$*, with* $\bigcap \emptyset = \mathcal{K}$*. Then the following statements are equivalent:*

(i) $\mathcal{A}$ *avoids non-positivity relative to* $(\mathcal{K},\mathcal{C})$*;*
(ii) $\mathcal{A}$ *is included in some set of desirable gambles that is coherent relative to* $(\mathcal{K},\mathcal{C})$*;*
(iii) $\mathcal{E}_{(\mathcal{K},\mathcal{C})}(\mathcal{A}) \neq \mathcal{K}$*;*
(iv) *the set of desirable gambles* $\mathcal{E}_{(\mathcal{K},\mathcal{C})}(\mathcal{A})$ *is coherent relative to* $(\mathcal{K},\mathcal{C})$*;*
(v) $\mathcal{E}_{(\mathcal{K},\mathcal{C})}(\mathcal{A})$ *is the smallest set of desirable gambles that is coherent relative to* $(\mathcal{K},\mathcal{C})$ *and includes* $\mathcal{A}$*.*

When any of these equivalent statements holds, then $\mathcal{E}_{(\mathcal{K},\mathcal{C})}(\mathcal{A}) = \operatorname{posi}\big(\mathcal{K}_{\succ 0} \cup \mathcal{A}\big)$*.*

Point-wise comparison coherence. We now turn to the important special case, commonly considered in the literature [12], where $\mathcal{K} := \mathcal{G}(\Omega)$ and the partial order $\succeq$ is the point-wise ordering $\geq$.[5] This partial order is associated to $\mathcal{C} := \mathcal{G}_0^+(\Omega) := \mathcal{G}(\Omega)_{>0}$, the cone of all non-negative gambles.

If $\mathcal{R}$ avoids non-positivity relative to $\big(\mathcal{G}(\Omega), \mathcal{G}_0^+(\Omega)\big)$, we simply say that $\mathcal{R}$ *avoids non-positivity*: $\mathcal{G}^-(\Omega) \cap \operatorname{posi}(\mathcal{R}) = \emptyset$, where $\mathcal{G}^-(\Omega) := \mathcal{G}(\Omega)_{\leq 0}$ is the set of all non-positive gambles. Similarly, if $\mathcal{R}$ is coherent relative to $\big(\mathcal{G}(\Omega), \mathcal{G}_0^+(\Omega)\big)$, we simply say that $\mathcal{R}$ is *coherent*, and we denote the set of coherent sets of desirable gambles by $\mathbb{D}(\Omega)$. In this case, the coherence conditions D1–D5 are to be seen as *rationality criteria*. The $\big(\mathcal{G}(\Omega), \mathcal{G}_0^+(\Omega)\big)$-natural extension of an assessment $\mathcal{A} \subseteq \mathcal{G}(\Omega)$ is simply denoted by $\mathcal{E}(\mathcal{A})$, and is called the *natural extension* of $\mathcal{A}$.

Weakly desirable gambles. We now define *weak desirability*, a concept that lies at the basis of our definition of exchangeability.

[5] $f \geq g$ iff $f(\omega) \geq g(\omega)$ for all ω in Ω; $f > g$ iff $f \geq g$ and $f \neq g$.

Definition 2 (Weak desirability). *Consider a coherent set $\mathcal{R}$ of desirable gambles. Then a gamble f is called* weakly desirable *if $f + f'$ is desirable for all desirable f', i.e., if $f + f' \in \mathcal{R}$ for all f' in $\mathcal{R}$. So the set of weakly desirable gambles is $\mathcal{D}_{\mathcal{R}} := \{f \in \mathcal{G}(\Omega) \colon f + \mathcal{R} \subseteq \mathcal{R}\}$.*

Every desirable gamble is also weakly desirable, so $\mathcal{R} \subseteq \mathcal{D}_{\mathcal{R}}$. Moreover, $\mathcal{D}_{\mathcal{R}}$ also satisfies the scaling and combination requirements D3–D4, so it is a cone as well.

Updating sets of desirable gambles. Consider a set of desirable gambles $\mathcal{R}$ on Ω. With a non-empty subset B of Ω, we associate an *updated* set of desirable gambles $\mathcal{R}\rfloor B := \{f_B \colon f = I_B f \in \mathcal{R}\} \subseteq \mathcal{G}(B)$,[6] where f_B is the restriction of f to B and I_B is the indicator function of B, i.e., 1 on B and 0 elsewhere. $\mathcal{R}\rfloor B$ is our subject's set of desirable gambles contingent on observing the event B.

Proposition 1. *If $\mathcal{R}$ is a coherent set of desirable gambles on Ω, then $\mathcal{R}\rfloor B$ is a coherent set of desirable gambles on B.*

3 Finite Exchangeable Sequences

Now that we have familiarised ourselves with sets of desirable gambles, we turn to exchangeability. In this section, we recall the basic definitions and results about finite exchangeable sequences from our earlier work [1], and add some new material related to frequency vector representations.

Consider random variables $X_1, \ldots, X_N$ taking values in a non-empty finite set $\mathcal{X}$, where $N \in \mathbb{N}_0$, i.e., a positive integer. The possibility space is $\Omega = \mathcal{X}^N$.

Let $x = (x_1, \ldots, x_N)$ be an arbitrary element of $\mathcal{X}^N$. $\mathcal{P}_N$ is the set of all permutations of the index set $\{1, \ldots, N\}$. With any such permutation π, we associate a permutation of $\mathcal{X}^N$, also denoted by π, and defined by $(\pi x)_k = x_{\pi(k)}$. Similarly, we lift π to a permutation π^t of $\mathcal{G}(\mathcal{X}^N)$ by letting $\pi^t f = f \circ \pi$.

The *counting map* T^N maps a sequence x to its count vector, an $\mathcal{X}$-tuple with a z-component $T^N_z(x) := |\{k \in \{1, \ldots, N\} \colon x_k = z\}|$ for all z in $\mathcal{X}$. The set of possible *count vectors* is given by $\mathcal{N}^N := \left\{m \in \mathbb{N}^{\mathcal{X}} \colon \sum_{x \in \mathcal{X}} m_x = N\right\}$. The permutation invariant atoms $[x] := \{\pi x \colon \pi \in \mathcal{P}_N\}$ are the smallest permutation invariant subsets of $\mathcal{X}^N$. If $m = T^N(x)$, then $[x] = \left\{y \in \mathcal{X}^N \colon T^N(y) = m\right\}$, so the atom $[x]$ is completely determined by the count vector m of all its the elements, and is therefore also denoted by $[m]$.

Defining exchangeability. If a subject assesses that $X_1, \ldots, X_N$ are exchangeable, this means that for any gamble f and any permutation π, she finds exchanging $\pi^t f$ for f weakly desirable, because she is indifferent between them [11, Section 4.1.1]. Taking into account that $\mathcal{D}_{\mathcal{R}}$ is a cone, we introduce the linear[7] space $\mathcal{D}_{\mathcal{U}_N} := \text{posi}\left\{f - \pi^t f \colon f \in \mathcal{G}(\mathcal{X}^N) \text{ and } \pi \in \mathcal{P}_N\right\}$. It holds that $\mathcal{R} \cap \mathcal{D}_{\mathcal{U}_N} = \emptyset$.

[6] Our definition is different from, but equivalent to the usual one [12].

[7] Due to the negation invariance of posi's argument.

Definition 3 (Exchangeability). *A coherent set $\mathcal{R}$ of desirable gambles on $\mathcal{X}^N$ is called* exchangeable *if one (and hence both) of the following equivalent conditions is (are) satisfied: (i) all gambles in $\mathcal{D}_{\mathcal{U}_N}$ are weakly desirable: $\mathcal{D}_{\mathcal{U}_N} \subseteq \mathcal{D}_{\mathcal{R}}$; and (ii) $\mathcal{D}_{\mathcal{U}_N} + \mathcal{R} \subseteq \mathcal{R}$.*

Updating exchangeable models. Consider an exchangeable and coherent set of desirable gambles $\mathcal{R}$ on $\mathcal{X}^N$. Assume that the subject has observed the values $\check{x} = (\check{x}_1, \check{x}_2, \ldots, \check{x}_{\check{n}})$ or a count vector $\check{m} \in \mathcal{N}^{\check{n}}$ of the variables $X_1, \ldots, X_{\check{n}}$. She wants to make inferences about the remaining $\hat{n} := N - \check{n}$ variables. Consider the updated models $\mathcal{R}\rfloor\{\check{x}\} \times \mathcal{X}^{\hat{n}}$ and $\mathcal{R}\rfloor[\check{m}] \times \mathcal{X}^{\hat{n}}$, and their restrictions to these $\hat{n}$ variables, $\mathcal{R}\rfloor\check{x}$ and $\mathcal{R}\rfloor\check{m}$.

Proposition 2. *Consider $\check{x}$ in $\mathcal{X}^{\check{n}}$, $\check{m}$ in $\mathcal{N}^{\check{n}}$, and a coherent and exchangeable set of desirable gambles $\mathcal{R}$ on $\mathcal{X}^N$. Then $\mathcal{R}\rfloor\check{x}$ and $\mathcal{R}\rfloor\check{m}$ are coherent and exchangeable sets of desirable gambles on $\mathcal{X}^{\hat{n}}$. If $\check{m} = T^{\check{n}}(\check{x})$, then $\mathcal{R}\rfloor\check{x} = \mathcal{R}\rfloor\check{m}$.*

Finite representation in terms of count vectors. In earlier work [1], we have proved that a coherent and exchangeable set of desirable gambles on sequences can be represented by a coherent set of desirable gambles on count vectors. To move between both spaces of gambles, we were led to the linear map MuHy^N that maps a gambler f on $\mathcal{X}^N$ to the gamble $\mathrm{MuHy}^N(f) := \mathrm{MuHy}^N(f|\cdot)$ on $\mathcal{N}^N$. Here, for every m in $\mathcal{N}^N$, $\mathrm{MuHy}^N(f|m) := \sum_{y\in[m]} f(y)/|[m]|$ is the expectation of f under the *multivariate hyper-geometric distribution* [7, Section 39.2] associated with random sampling without replacement from an urn, whose composition is characterised by the count vector m. For the other direction, we use the the linear map T^N that maps a gamble g on $\mathcal{N}^N$ to the permutation invariant gamble $\mathrm{T}^N(g) := g \circ T^N$ on $\mathcal{X}^N$ assuming the value $g(m)$ on the invariant atom $[m]$.

Theorem 2 (Finite Representation). *A set of desirable gambles $\mathcal{R}$ on $\mathcal{X}^N$ is coherent and exchangeable iff there is some coherent set $\mathcal{S}$ of desirable gambles on $\mathcal{N}^N$ such that $\mathcal{R} = (\mathrm{MuHy}^N)^{-1}(\mathcal{S})$, and in that case this $\mathcal{S}$ is uniquely determined by $\mathcal{S} = \left\{g \in \mathcal{G}(\mathcal{N}^N)\colon \mathrm{T}^N(g) \in \mathcal{R}\right\} = \mathrm{MuHy}^N(\mathcal{R})$. We call $\mathcal{S}$ the* count representation *of the exchangeable set $\mathcal{R}$.*

Multinomial processes. Next, we turn to a number of important ideas related to multinomial processes. They are useful for comparisons with the existing literature [6, for example], and essential for our treatment of countable exchangeable sequences in Section 4.

Consider the *$\mathcal{X}$-simplex* $\Sigma_{\mathcal{X}} := \left\{\theta \in \mathbb{R}^{\mathcal{X}}\colon \theta \geq 0 \text{ and } \sum_{x\in\mathcal{X}} \theta_x = 1\right\}$, and, for $N \in \mathbb{N}_0$, the linear map CoMn^N from $\mathcal{G}(\mathcal{N}^N)$ to $\mathcal{G}(\Sigma_{\mathcal{X}})$ defined by $\mathrm{CoMn}^N(g) = \mathrm{CoMn}^N(g|\cdot)$, where for all θ in $\Sigma_{\mathcal{X}}$, $\mathrm{CoMn}^N(g|\theta) := \sum_{m\in\mathcal{N}^N} g(m)B_m(\theta)$ is the expectation associated with the *count multinomial* distribution with parameters N and θ, and where B_m is the multivariate *Bernstein (basis) polynomial of degree* N given by $B_m(\theta) := \binom{N}{m}\prod_{x\in\mathcal{X}} \theta_x^{m_x} = |[m]|\prod_{x\in\mathcal{X}} \theta_x^{m_x}$. We also consider the

related linear map Mn^N from $\mathcal{G}(\mathcal{X}^N)$ to $\mathcal{G}(\Sigma_{\mathcal{X}})$ defined by $\mathrm{Mn}^N(f) = \mathrm{Mn}^N(f|\cdot)$, where for all θ in $\Sigma_{\mathcal{X}}$, $\mathrm{Mn}^N(f|\theta) := \sum_{m\in\mathcal{N}^N} \mathrm{MuHy}^N(f|m)B_m(\theta)$ is the expectation associated with the *multinomial* distribution with parameters N and θ. We then have that $\mathrm{CoMn}^N = \mathrm{Mn}^N \circ \mathrm{T}^N$ and $\mathrm{Mn}^N = \mathrm{CoMn}^N \circ \mathrm{MuHy}^N$.

The Bernstein basis polynomials B_m, $m \in \mathcal{N}^N$ form a basis for the linear space $\mathcal{V}_N(\Sigma_{\mathcal{X}})$ of all polynomials on $\Sigma_{\mathcal{X}}$ of degree up to N [9]. This means that for each polynomial p whose degree $\deg(p)$ does not exceed N, there is a unique gamble b_p^N on $\mathcal{N}^N$ such that $p = \mathrm{CoMn}^N(b_p^N)$. We denote by $\mathcal{V}(\Sigma_{\mathcal{X}})$ the linear space of all polynomials on $\Sigma_{\mathcal{X}}$.

Finite representation in terms of polynomials. We see that the range of the linear maps CoMn^N and Mn^N is the linear space $\mathcal{V}_N(\Sigma_{\mathcal{X}})$. Moreover, since for every polynomial p of degree up to N, i.e., for every p in $\mathcal{V}_N(\Sigma_{\mathcal{X}})$, there is a *unique* count gamble $b_p^N \in \mathcal{G}(\mathcal{N}^N)$ such that $p = \mathrm{CoMn}^N(b_p^N)$, CoMn^N is a *linear isomorphism* between the linear spaces $\mathcal{G}(\mathcal{N}^N)$ and $\mathcal{V}_N(\Sigma_{\mathcal{X}})$.

In summary, everything that can be expressed using the language of gambles on $\mathcal{N}^N$, can also be expressed using the language of polynomial gambles on $\Sigma_{\mathcal{X}}$ of degree up to N, and *vice versa*. The map CoMn^N and its inverse are the tools that take care of the translation between the two languages. This is essentially what is behind the representation theorem for countable exchangeable sequences that we will turn to in Section 4. In order to lay the proper foundations for this work, we now present a version of the finite representation theorem in terms of polynomial gambles of degree N on $\Sigma_{\mathcal{X}}$, rather than count gambles on $\mathcal{N}^N$.

We first introduce a concept of coherence for Bernstein polynomials:

Definition 4 (Bernstein coherence). *We call a set $\mathcal{H}$ of polynomials in $\mathcal{V}_N(\Sigma_{\mathcal{X}})$* Bernstein coherent at degree N *if it satisfies the following properties: for all p, p_1, and p_2 in $\mathcal{V}_N(\Sigma_{\mathcal{X}})$ and all real $\lambda > 0$,*
B$_N$1. *if $p = 0$ then $p \notin \mathcal{H}$;*
B$_N$2. *if p is such that $b_p^N > 0$ then $p \in \mathcal{H}$;*
B$_N$3. *if $p \in \mathcal{H}$ then $\lambda p \in \mathcal{H}$;*
B$_N$4. *if $p_1, p_2 \in \mathcal{H}$ then $p_1 + p_2 \in \mathcal{H}$.*

Bernstein coherence at degree N is very closely related to coherence, the only difference being that we do not consider whether a polynomial p is positive, but whether its Bernstein expansion b_p^N is. Interpretation-wise, this means that models in terms of sequences or count vectors are authoritative over those in terms of frequency vectors. Bernstein coherence at degree N is a special case of the general concept of coherence relative to $(\mathcal{K}, \mathcal{C})$, discussed in Section 2, where $\mathcal{K} := \mathcal{V}_N(\Sigma_{\mathcal{X}})$ and $\mathcal{C} := \{p \in \mathcal{V}_N(\Sigma_{\mathcal{X}})\colon b_p^N \geq 0\}$ is the convex cone of all polynomials of degree at most N with a non-negative expansion b_p^N in the Bernstein basis of degree N.

Theorem 3 (Finite Representation). *A set of desirable gambles $\mathcal{R}$ on $\mathcal{X}^N$, with count representation $\mathcal{S} := \mathrm{MuHy}^N(\mathcal{R})$, is coherent and exchangeable iff there is some subset $\mathcal{H}$ of $\mathcal{V}_N(\Sigma_{\mathcal{X}})$, Bernstein coherent at degree N, such that $\mathcal{R} = (\mathrm{Mn}^N)^{-1}(\mathcal{H})$ or equivalently $\mathcal{S} = (\mathrm{CoMn}^N)^{-1}(\mathcal{H})$, and in that case this $\mathcal{H}$ is*

uniquely determined by $\mathcal{H} = \mathrm{Mn}^N(\mathcal{R}) = \mathrm{CoMn}^N(\mathcal{S})$. *We call* $\mathcal{H}$ *the* frequency representation *of the coherent and exchangeable set of desirable gambles* $\mathcal{R}$.

4 Countable Exchangeable Sequences

With the experience gained in investigating finite exchangeable sequences, we are now ready to address reasoning about countably infinite exchangeable sequences. The first step is to use finite frequency representation to find a Representation Theorem for infinite exchangeable sequences. We can then show what updating and natural extension look like in terms of this frequency representation.

Infinite representation. We consider a countable sequence $X_1, \ldots, X_N, \ldots$ of random variables assuming values in the same finite set $\mathcal{X}$. We call this sequence *exchangeable* if each of its finite subsequences is, or equivalently, if for all n in $\mathbb{N}_0$, the random variables $X_1, \ldots, X_n$ are exchangeable.

How can we model this? First of all, this means that for each n in $\mathbb{N}_0$, there is a coherent and exchangeable set of desirable gambles $\mathcal{R}^n$ on $\mathcal{X}^n$. Equivalently, we have a coherent set of desirable gambles (count representation) $\mathcal{S}^n := \mathrm{MuHy}^n(\mathcal{R}^n)$ on $\mathcal{N}^n$, or a set (frequency representation) $\mathcal{H}^n := \mathrm{Mn}^n(\mathcal{R}^n) = \mathrm{CoMn}^n(\mathcal{S}^n)$ of polynomials in $\mathcal{V}_n(\Sigma_{\mathcal{X}})$, Bernstein coherent at degree n.

In addition, there is a time-consistency constraint. Consider the following linear *projection operators* $\mathrm{proj}_{n_2}^{n_1} \colon \mathcal{X}^{n_2} \to \mathcal{X}^{n_1}$ defined by $\mathrm{proj}_{n_2}^{n_1}(x_1, \ldots, x_{n_2}) := (x_1, \ldots, x_{n_1})$, where $n_1 \leq n_2$. With each such operator there corresponds a linear map $\mathrm{ext}_{n_1}^{n_2}$ between the linear spaces $\mathcal{G}(\mathcal{X}^{n_1})$ and $\mathcal{G}(\mathcal{X}^{n_2})$, defined as follows: $\mathrm{ext}_{n_1}^{n_2}(f) = f \circ \mathrm{proj}_{n_2}^{n_1}$. In other words, $\mathrm{ext}_{n_1}^{n_2}(f)$ is the *cylindrical extension* of the gamble f on $\mathcal{X}^{n_1}$ to a gamble on $\mathcal{X}^{n_2}$.

Time-consistency now means that if we consider a gamble on $\mathcal{X}^{n_2}$ that really only depends on the first n_1 variables, it should not matter, as far as its desirability is concerned, whether we consider it to be a gamble on $\mathcal{X}^{n_1}$ or a gamble on $\mathcal{X}^{n_2}$. More formally, we require that $(\forall n_1 \leq n_2)\, \mathrm{ext}_{n_1}^{n_2}(\mathcal{R}^{n_1}) = \mathcal{R}^{n_2} \cap \mathrm{ext}_{n_1}^{n_2}\big(\mathcal{G}(\mathcal{X}^{n_1})\big)$.

How can we translate this constraint in terms of the count representations $\mathcal{S}^n$ or the frequency representations $\mathcal{H}^n$? If we introduce the linear *extension map* $\mathrm{enl}_{n_1}^{n_2}$ from the linear space $\mathcal{G}(\mathcal{N}^{n_1})$ to the linear space $\mathcal{G}(\mathcal{N}^{n_2})$ by $\mathrm{enl}_{n_1}^{n_2}(g) := \sum_{m \in \mathcal{N}^{n_1}} g(m) |[\cdot - m]|\,|[m]|/|[\cdot]|$, then $\mathrm{MuHy}^{n_2} \circ \mathrm{ext}_{n_1}^{n_2} = \mathrm{enl}_{n_1}^{n_2} \circ \mathrm{MuHy}^{n_1}$, and time-consistency is equivalent to $(\forall n_1 \leq n_2)\, \mathrm{enl}_{n_1}^{n_2}(\mathcal{S}^{n_1}) = \mathcal{S}^{n_2} \cap \mathrm{enl}_{n_1}^{n_2}\big(\mathcal{G}(\mathcal{N}^{n_1})\big)$, which is in turn equivalent to $(\forall n_1 \leq n_2)\mathcal{H}^{n_1} = \mathcal{H}^{n_2} \cap \mathcal{V}_{n_1}(\Sigma_{\mathcal{X}})$. The time-consistency condition is most elegantly expressed using frequency representations.

We call the family $\mathcal{R}^n$, $n \in \mathbb{N}_0$ *time-consistent, coherent and exchangeable* when each member $\mathcal{R}^n$ is coherent and exchangeable, and when the family $\mathcal{R}^n$, $n \in \mathbb{N}_0$ satisfies the time-consistency requirement.

We can generalise the concept of Bernstein coherence given in Definition 4 to sets of polynomials of arbitrary degree:

Definition 5 (Bernstein coherence). *We call a set $\mathcal{H}$ of polynomials in $\mathcal{V}(\Sigma_{\mathcal{X}})$* Bernstein coherent *if it satisfies the following properties: for all p, p_1, and p_2 in $\mathcal{V}(\Sigma_{\mathcal{X}})$ and all real $\lambda > 0$,*
B1. *if $p = 0$ then $p \notin \mathcal{H}$;*
B2. *if p is such that $b_p^n > 0$ for some $n \geq \deg(p)$, then $p \in \mathcal{H}$;*
B3. *if $p \in \mathcal{H}$ then $\lambda p \in \mathcal{H}$;*
B4. *if $p_1, p_2 \in \mathcal{H}$ then $p_1 + p_2 \in \mathcal{H}$.*

We can replace B1 by the following requirement, equivalent to it under B2–B4:
B5. If p is such that $b_p^n \leq 0$ for some $n \geq \deg(p)$, then $p \notin \mathcal{H}$.

This type of Bernstein coherence is again very closely related to coherence, the only difference being that not all positive polynomials, but rather all polynomials with some positive Bernstein expansion are required to belong to a Bernstein coherent set. Bernstein coherence is a special case of the general concept of coherence relative to $(\mathcal{K}, \mathcal{C})$, discussed in Section 2, where $\mathcal{K} := \mathcal{V}(\Sigma_{\mathcal{X}})$ and $\mathcal{C} := \mathcal{V}^+(\Sigma_{\mathcal{X}}) := \{p \in \mathcal{V}(\Sigma_{\mathcal{X}}) \colon (\exists n \geq 0) b_p^n \geq 0\}$ is the convex cone of all polynomials with some non-negative Bernstein expansion. We also denote the set $\mathbb{D}_{(\mathcal{V}(\Sigma_{\mathcal{X}}), \mathcal{V}^+(\Sigma_{\mathcal{X}}))}(\Sigma_{\mathcal{X}})$ of all Bernstein coherent subsets of $\mathcal{V}(\Sigma_{\mathcal{X}})$ by $\mathbb{D}_{\mathrm{Be}}(\Sigma_{\mathcal{X}})$.

We are now ready to formulate our Infinite Representation Theorem, which is a significant generalisation of de Finetti's representation result for countable sequences [3]. A similar result can be proved for coherent lower previsions [2].

Theorem 4 (Infinite Representation). *A family $\mathcal{R}^n$, $n \in \mathbb{N}_0$ of sets of desirable gambles on $\mathcal{X}^n$, with associated count representations $\mathcal{S}^n := \mathrm{MuHy}^n(\mathcal{R}^n)$ and frequency representations $\mathcal{H}^n := \mathrm{Mn}^n(\mathcal{R}^n) = \mathrm{CoMn}^n(\mathcal{S}^n)$, is time-consistent, coherent and exchangeable iff there is some Bernstein coherent set $\mathcal{H}$ of polynomials in $\mathcal{V}(\Sigma_{\mathcal{X}})$ such that, for all n in $\mathbb{N}_0$, both $\mathcal{S}^n = (\mathrm{CoMn}^n)^{-1}(\mathcal{H})$ and $\mathcal{R}^n = (\mathrm{Mn}^n)^{-1}(\mathcal{H})$, and in that case this $\mathcal{H}$ is uniquely given by $\mathcal{H} = \bigcup_{n \in \mathbb{N}_0} \mathcal{H}^n$. We call $\mathcal{H}$ the* frequency representation *of the coherent, exchangeable and time-consistent family of sets of desirable gambles $\mathcal{R}^n$, $n \in \mathbb{N}_0$.*

Updating and infinite representation. Suppose we have a coherent, exchangeable and time-consistent family of sets of desirable gambles $\mathcal{R}^n$, $n \in \mathbb{N}_0$, with associated count representations $\mathcal{S}^n := \mathrm{MuHy}^n(\mathcal{R}^n)$ and frequency representation $\mathcal{H} := \bigcup_{n \in \mathbb{N}} \mathcal{H}^n$ with $\mathcal{H}^n := \mathrm{Mn}^n(\mathcal{R}^n)$. Now suppose we observe the values $\check{x}$ of the first $\check{n}$ variables. It turns out that updating becomes especially easy in terms of the frequency representation.

Theorem 5. *Consider a coherent, exchangeable and time-consistent family of sets of desirable gambles $\mathcal{R}^n$, $n \in \mathbb{N}_0$, with associated frequency representation $\mathcal{H}$. After updating with a sample with count vector $\check{m} \in \mathcal{N}_{\mathcal{X}}^{\check{n}}$, the family $\mathcal{R}^{\hat{n}} \rfloor \check{m}$, $\hat{n} \in \mathbb{N}_0$ is still coherent, exchangeable and time-consistent, and has frequency representation $\mathcal{H} \rfloor \check{m} := \{p \in \mathcal{V}(\Sigma_{\mathcal{X}}) \colon B_{\check{m}} p \in \mathcal{H}\}$.*

Independence: iid sequences. We can use Theorem 5 to find an intriguing characterisation of a sequence of *independent identically distributed* (iid) random variables $X_1, \ldots, X_N, \ldots$ assuming values in a finite set $\mathcal{X}$. This is an

exchangeable sequence where learning the value of any finite number of variables does not change our subject's beliefs about the remaining, unobserved ones. This is the case iff the frequency representation $\mathcal{H}$ of the sequence satisfies $(\forall \check{n} \in \mathbb{N}_0)(\forall \check{m} \in \mathcal{N}_{\mathcal{X}}^{\check{n}})\mathcal{H}\rfloor\check{m} = \mathcal{H}$. This is equivalent to $(\forall \check{n} \in \mathbb{N}_0)(\forall \check{m} \in \mathcal{N}_{\mathcal{X}}^{\check{n}})$ $\big(\forall p \in \mathcal{V}(\Sigma_{\mathcal{X}})\big)(p \in \mathcal{H} \Leftrightarrow B_{\check{m}}p \in \mathcal{H})$, and also to $\mathcal{H} = \mathcal{V}^+(\Sigma_{\mathcal{X}})\mathcal{H}$, which is shorthand for $\big(\forall p \in \mathcal{V}(\Sigma_{\mathcal{X}})\big)\big(\forall p^+ \in \mathcal{V}^+(\Sigma_{\mathcal{X}})\big)(p \in \mathcal{H} \Leftrightarrow p^+p \in \mathcal{H})$.

Bernstein natural extension. The intersection of an arbitrary non-empty family of Bernstein coherent sets of polynomials is still Bernstein coherent. This is the idea behind the following theorem, which is a special instance of Theorem 1 with $\mathcal{K} := \mathcal{V}(\Sigma_{\mathcal{X}})$ and $\mathcal{C} := \mathcal{V}^+(\Sigma_{\mathcal{X}})$.

We denote by $\mathcal{V}_0^+(\Sigma_{\mathcal{X}})$ the set of all polynomials on $\Sigma_{\mathcal{X}}$ with some positive Bernstein expansion: $\mathcal{V}_0^+(\Sigma_{\mathcal{X}}) = \big\{p \in \mathcal{V}(\Sigma_{\mathcal{X}})\colon \big(\exists n \geq \deg(p)\big)b_p^n > 0\big\}$ and by $\mathcal{V}^-(\Sigma_{\mathcal{X}}) = \big\{p \in \mathcal{V}(\Sigma_{\mathcal{X}})\colon \big(\exists n \geq \deg(p)\big)b_p^n \leq 0\big\}$ the set of all polynomials on $\Sigma_{\mathcal{X}}$ with some non-positive Bernstein expansion. Moreover, we say that a set $\mathcal{A}$ of polynomials *avoids Bernstein non-positivity* if no polynomial in its positive hull $\operatorname{posi}(\mathcal{A})$ has any non-positive Bernstein expansion, i.e. $\operatorname{posi}(\mathcal{A}) \cap \mathcal{V}^-(\Sigma_{\mathcal{X}}) = \emptyset$; clearly, this is the case iff $\mathcal{A}$ avoids non-positivity relative to $\big(\mathcal{V}(\Sigma_{\mathcal{X}}), \mathcal{V}^+(\Sigma_{\mathcal{X}})\big)$. We also call the $\big(\mathcal{V}(\Sigma_{\mathcal{X}}), \mathcal{V}^+(\Sigma_{\mathcal{X}})\big)$-natural extension $\mathcal{E}_{(\mathcal{V}(\Sigma_{\mathcal{X}}),\mathcal{V}^+(\Sigma_{\mathcal{X}}))}(\mathcal{A})$ of $\mathcal{A}$ its Bernstein natural extension, and denote it by $\mathcal{E}_{\mathrm{Be}}(\mathcal{A})$.

Theorem 6 (Bernstein natural extension). *Consider $\mathcal{A} \subseteq \mathcal{V}(\Sigma_{\mathcal{X}})$ and its* Bernstein natural extension $\mathcal{E}_{\mathrm{Be}}(\mathcal{A}) := \bigcap\{\mathcal{H} \in \mathbb{D}_{\mathrm{Be}}(\Sigma_{\mathcal{X}})\colon \mathcal{A} \subseteq \mathcal{H}\}$. *The following statements are then equivalent:*

(i) *$\mathcal{A}$ avoids Bernstein non-positivity;*
(ii) *$\mathcal{A}$ is included in some Bernstein coherent set of polynomials;*
(iii) *$\mathcal{E}_{\mathrm{Be}}(\mathcal{A}) \neq \mathcal{V}(\Sigma_{\mathcal{X}})$;*
(iv) *$\mathcal{E}_{\mathrm{Be}}(\mathcal{A})$ is a Bernstein coherent set of polynomials;*
(v) *$\mathcal{E}_{\mathrm{Be}}(\mathcal{A})$ is the smallest Bernstein coherent set of polynomials including $\mathcal{A}$.*

When any these equivalent statements holds, then $\mathcal{E}_{\mathrm{Be}}(\mathcal{A}) = \operatorname{posi}\big(\mathcal{V}_0^+(\Sigma_{\mathcal{X}}) \cup \mathcal{A}\big)$.

Exchangeable natural extension for infinite sequences. To finish this discussion of exchangeability for infinite sequences of random variables, we take up the issue of deductive inference, and introduce exchangeable natural extension.

Suppose we have an assessment consisting of a set $\mathcal{A}^n$ of desirable gambles on $\mathcal{X}^n$ for each n in $\mathbb{N}_0$. We are looking for the (element-wise) smallest coherent, exchangeable and time-consistent family $\mathcal{R}^n$, $n \in \mathbb{N}_0$ that includes this assessment in the sense that $\mathcal{A}^n \subseteq \mathcal{R}^n$ for all n in $\mathbb{N}_0$, or equivalently $\bigcup_{n\in\mathbb{N}_0} \operatorname{Mn}^n(\mathcal{A}^n) \subseteq \bigcup_{n\in\mathbb{N}_0} \operatorname{Mn}^n(\mathcal{R}^n) =: \mathcal{H}$, a condition formulated in terms of the frequency representation $\mathcal{H}$ of the family $\mathcal{R}^n$, $n \in \mathbb{N}_0$.

Theorem 7. *Consider an assessment consisting of the sets of desirable gambles $\mathcal{A}^n$ on $\mathcal{X}^n$ for each n in $\mathbb{N}_0$, and the corresponding set of 'desirable' polynomials $\mathcal{A} := \bigcup_{n\in\mathbb{N}_0} \operatorname{Mn}^n(\mathcal{A}^n)$. There is a coherent, exchangeable and time-consistent family $\mathcal{R}^n$, $n \in \mathbb{N}_0$ that includes this assessment iff $\mathcal{A}$ avoids Bernstein non-positivity. In that case $\mathcal{E}_{\mathrm{Be}}(\mathcal{A})$ is the frequency representation of the (element-wise) smallest coherent, exchangeable and time-consistent family that includes this assessment.*

5 Conclusions

Modelling an infinite exchangeability assessment using sets of desirable gambles is not only possible, but also quite elegant. Our Infinite Representation Theorem reduces reasoning about infinite exchangeable sequences to reasoning about (polynomials of) frequency vectors. This automatically guarantees that, next to the exchangeability of finite subsequences, time-consistency of these subsequences is satisfied. The representation for the natural extension and for updated models can be derived directly from the representation of the original model.

Acknowledgements. Erik Quaeghebeur was supported by a Fellowship of the Belgian American Educational Foundation. We wish to thank the reviewer.

References

1. De Cooman, G., Quaeghebeur, E.: Exchangeability for sets of desirable gambles. In: Augustin, T., Coolen, F.P.A., Moral, S., Troffaes, M.C.M. (eds.) ISIPTA 2009: proceedings of the Sixth International Symposium on Imprecise Probabilities: Theories and Applications. pp. 159–168. SIPTA, Durham, United Kingdom (2009), http://hdl.handle.net/1854/LU-718913
2. De Cooman, G., Quaeghebeur, E., Miranda, E.: Exchangeable lower previsions. Bernoulli 15(3), 721–735 (2009), http://hdl.handle.net/1854/LU-498518, an expanded version can be found at http://arxiv.org/abs/0801.1265
3. de Finetti, B.: La prévision: ses lois logiques, ses sources subjectives. Annales de l'Institut Henri Poincaré 7, 1–68 (1937), English translation in [8]
4. de Finetti, B.: Teoria delle Probabilità. Einaudi, Turin (1970)
5. de Finetti, B.: Theory of Probability: A Critical Introductory Treatment. John Wiley & Sons, Chichester (1974–1975), English translation of [4], two volumes
6. Heath, D.C., Sudderth, W.D.: De Finetti's theorem on exchangeable variables. The American Statistician 30, 188–189 (1976)
7. Johnson, N.L., Kotz, S., Balakrishnan, N.: Discrete Multivariate Distributions. Wiley Series in Probability and Statistics, John Wiley and Sons, New York (1997)
8. Kyburg Jr., H.E., Smokler, H.E. (eds.): Studies in Subjective Probability. Wiley, New York (1964), second edition (with new material) 1980
9. Prautzsch, H., Boehm, W., Paluszny, M.: Bézier and B-Spline Techniques. Springer, Berlin (2002)
10. Schechter, E.: Handbook of Analysis and Its Foundations. Academic Press, San Diego, CA (1997)
11. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)
12. Walley, P.: Towards a unified theory of imprecise probability. International Journal of Approximate Reasoning 24, 125–148 (2000)

Ergodicity Conditions for Upper Transition Operators

Filip Hermans and Gert de Cooman

SYSTeMS Research Group, Ghent University
Technologiepark–Zwijnaarde 914, 9052 Zwijnaarde, Belgium
{filip.hermans,gert.decooman}@UGent.be

Abstract. We study ergodicity for upper transition operators: bounded, sub-additive and non-negatively homogeneous transformations of finite-dimensional linear spaces. Ergodicity provides a necessary and sufficient condition for Perron–Frobenius-like convergence behaviour for upper transition operators. It can also be characterised alternatively using accessibility relations: ergodicity is equivalent with there being a single maximal communication (or top) class that is moreover regular and absorbing. We present efficient algorithms for checking these conditions.

Keywords: upper transition operator, ergodicity, imprecise probability.

1 Introduction

Throughout the paper, $\mathcal{X}$ denotes a finite non-empty set of elements that we also refer to as *states*, and $\mathcal{L}(\mathcal{X})$ is the set of all real-valued maps on $\mathcal{X}$. We provide the finite-dimensional linear space $\mathcal{L}(\mathcal{X})$ with the supremum norm $\|\cdot\|_\infty$, or with the topology of uniform convergence, so the result is a Banach space. Uniform and point-wise convergence coincide on this finite-dimensional space.

Definition 1. *An* upper transition operator *on* $\mathcal{L}(\mathcal{X})$ *is a transformation* T *of* $\mathcal{L}(\mathcal{X})$ *that has the following properties: for arbitrary* f, g *in* $\mathcal{L}(\mathcal{X})$ *and real* $\lambda \geq 0$,

T1. $\min f \leq \mathrm{T}f \leq \max f$	T *is* bounded*;*
T2. $\mathrm{T}(f+g) \leq \mathrm{T}f + \mathrm{T}g$	T *is* sub-additive*;*
T3. $\mathrm{T}(\lambda f) = \lambda \mathrm{T}g$	T *is* non-negatively homogeneous.

Where do such upper transition operators come from? We recall finite-state and discrete-time Markov chains. At any time k, such a Markov chain can be described by a stochastic transition matrix $M^{(k)}$ whose x-th row $M^{(k)}_{x,\cdot}$ is a probability mass function over the states at time $k+1$, conditional on the chain being in state x at time k. If we now assume that these conditional mass functions $M^{(k)}_{x,\cdot}$ can be picked from a convex closed set $\mathcal{M}_x$ depending on the state x, then any such transition matrix has to belong to the set $\mathcal{T} := \left\{M \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}} : (\forall x \in \mathcal{X})(M_{x,\cdot} \in \mathcal{M}_x)\right\}$. It can be shown [1] that there corresponds exactly one upper transition operator T with $\mathcal{T}$ in the following sense: if

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 70–79, 2010.

we start out in state x, the maximum of the possible expectations—the so-called *upper expectation*—of a map $f \in \mathcal{L}(\mathcal{X})$ after k time steps can be calculated as

$$\max\left\{M^{(1)}_{x,\cdot} M^{(2)} \dots M^{(k)} f \colon M^{(j)} \in \mathcal{T}\right\} = \mathrm{T}^k f(x).$$

As a result, $\mathrm{T}^k \mathbb{I}_A(x)$ can be interpreted as the upper probability and $1 - \mathrm{T}\mathbb{I}_{A^c}(x)$ as the lower probability to go from state x in k steps to a state in the set A; here $\mathbb{I}_A$ is the indicator of A, which assumes the value one on A and zero elsewhere, and A^c is the complement of A. Generally speaking, an upper transition operator can be seen as summarising robust inference for a set of not necessarily stationary Markov chains [1,2,7].

Any upper transition operator T automatically also satisfies the following interesting properties: for arbitrary f, g, f_n in $\mathcal{L}(\mathcal{X})$ and real μ,

T4. $\mathrm{T}(f + \mu) = \mathrm{T}f + \mu$	T is *constant-additive*;
T5. if $f \leq g$ then $\mathrm{T}f \leq \mathrm{T}g$	T is *order-preserving*;
T6. if $f_n \to f$ then $\mathrm{T}f_n \to \mathrm{T}f$	T is *continuous*;
T7. $\mathrm{T}f + \mathrm{T}(-f) \geq 0$	T is *upper–lower consistent*.

Clearly, for any n in the set of natural numbers (with zero) $\mathbb{N}_0$, T^n is an upper transition operator as well.

Properties T4 and T5 define a so-called *topical map* [4]. It is easy to see [4] that every topical map is also non-expansive under the supremum norm: for every f and g in $\mathcal{L}(\mathcal{X})$,

T8. $\|\mathrm{T}f - \mathrm{T}g\|_\infty \leq \|f - g\|_\infty$ T is *non-expansive*.

A very useful result for non-expansive maps by Sine [6] states that for every element f of the finite-dimensional domain of a non-expansive transformation T, there is some natural number p such that the sequence $\mathrm{T}^{np} f$ converges. More importantly, Sine proves that we can find a finite 'period' p *common* to all maps f in the domain $\mathcal{L}(\mathcal{X})$. This means that for any f, the set $\omega_{\mathrm{T}}(f)$ of limit points of the set of iterates $\{\mathrm{T}^n f \colon n \in \mathbb{N}\}$ has a number of elements $|\omega_{\mathrm{T}}(f)|$ that divides this p.[1] T is cyclic on $\omega_{\mathrm{T}}(f)$, with period $|\omega_{\mathrm{T}}(f)|$ (and therefore also with period p). Lemmens and Scheutzow [4] managed to prove that an upper bound for the common periods of all topical functions $\mathrm{T} \colon \mathbb{R}^n \to \mathbb{R}^n$ is $\binom{n}{\lfloor n/2 \rfloor}$. This upper bound is tight in the sense that there is always at least one topical function that has this bound as its smallest common period.

In Sec. 2 we exploit these ideas to introduce ergodicity for upper transition operators, and to explain its link with Perron–Frobenius conditions. In Sec. 3 we develop methods for checking ergodicity in practise. We order the state space by means of an accessibility relation. Using the equivalence classes induced by this ordering, we show that ergodicity is equivalent to the combination of top class regularity and top class absorption, and we work out an efficient test for each condition.

[1] $|A|$ denotes the cardinality of a set A and $\mathbb{N}$ is the set of natural numbers (without zero).

2 Perron–Frobenius Condition

We introduce the notion of ergodicity for upper transition operators. We allow ourselves to be inspired by corresponding notions for non-stationary Markov chains [5, p. 136] and Markov set chains [2].

Definition 2 (Ergodicity). *An upper transition operator* T *on* $\mathcal{L}(\mathcal{X})$ *is called* ergodic *if for all* $f \in \mathcal{L}(\mathcal{X})$, $\lim_{n\to\infty} \mathrm{T}^n f$ *exists and is a constant function.*

Consider any $f \in \mathcal{L}(\mathcal{X})$. Ergodicity of an upper transition operator T not only means that the sequence $\mathrm{T}^n f$ converges, so $\omega_{\mathrm{T}}(f)$ is a singleton $\{\xi_f\}$, but also that this limit ξ_f is a constant function. Observe that by T6, ξ_f is a fixed point for all T^k: $T^k\xi_f = \xi_f$ and therefore $\xi_{T^k f} = \xi_f$ for all $k \in \mathbb{N}$. If we denote the constant value of ξ_f by $\overline{E}_{\mathrm{T}}(f)$, meaning that $\xi_f = \overline{E}_{\mathrm{T}}(f)\mathbb{I}_{\mathcal{X}}$, then this defines a real functional $\overline{E}_{\mathrm{T}}$ on $\mathcal{L}(\mathcal{X})$. This functional is an *upper expectation*: it is bounded, sub-additive and non-negatively homogeneous [compare with T1–T3]. It is T-invariant in the sense that $\overline{E}_{\mathrm{T}} \circ \mathrm{T} = \overline{E}_{\mathrm{T}}$, and it is the only such upper expectation.

Definition 3. *We call an upper transition operator* T *on* $\mathcal{L}(\mathcal{X})$ Perron–Frobenius-like *if there is some real functional* $\overline{E}_\infty$ *on* $\mathcal{L}(\mathcal{X})$ *such that* $\lim_{n\to\infty} \overline{E}(\mathrm{T}^n f) = \overline{E}_\infty(f)$ *for all upper expectations* $\overline{E}$ *on* $\mathcal{L}(\mathcal{X})$ *and all* $f \in \mathcal{L}(\mathcal{X})$*, or in other words, if the sequence of upper expectations* $\overline{E} \circ \mathrm{T}^n$ *converges to some limit* $\overline{E}_\infty$ *that does not depend on the initial value* $\overline{E}$*.*

As an immediate result, conditions for ergodicity of upper transition operators are conditions for a Perron–Frobenius-like theorem to hold.

Theorem 1 (Perron–Frobenius). *An upper transition operator* T *is Perron–Frobenius-like if and only if it is ergodic, and in that case* $\overline{E}_\infty = \overline{E}_{\mathrm{T}}$*.*

Proof. Sufficiency. Suppose T is ergodic. Then using the notations established above, $\mathrm{T}^n f \to \xi_f$ and therefore $\overline{E}(\mathrm{T}^n f) \to \overline{E}(\xi_f)$ because any upper expectation $\overline{E}$ is continuous [compare with T6]. Observe that, since any upper expectation $\overline{E}$ is constant-additive [compare with T4 and T1], $\overline{E}(\xi_f) = \overline{E}_{\mathrm{T}}(f)$. Hence $\overline{E} \circ \mathrm{T}^n \to \overline{E}_{\mathrm{T}}$, and therefore T is Perron–Frobenius-like, with $\overline{E}_\infty = \overline{E}_{\mathrm{T}}$.

Necessity. Suppose that T is Perron–Frobenius-like, with limit upper expectation $\overline{E}_\infty$. Fix any $x \in \mathcal{X}$, and consider the upper expectation $\overline{E}_x$ defined by $\overline{E}_x(f) := f(x)$ for all $f \in \mathcal{L}(\mathcal{X})$. Then by assumption $T^n f(x) = \overline{E}_x(\mathrm{T}^n f) \to \overline{E}_\infty(f)$. Since this holds for all $x \in \mathcal{X}$, we see that T is ergodic with $\overline{E}_{\mathrm{T}} = \overline{E}_\infty$.

It follows from the discussion in Sec. 1 that $\bigcup_{f\in\mathcal{L}(\mathcal{X})} \omega_{\mathrm{T}}(f)$ is the set of all periodic points of T—*a periodic point* being an element $f \in \mathcal{L}(\mathcal{X})$ for which there is some $n \in \mathbb{N}$ for which $\mathrm{T}^n f = f$. By T4, this set contains all constant maps. We now see that for T to be ergodic, this set cannot contain any other maps.

Proposition 1. *An upper transition operator* T *is ergodic if and only if all of its periodic points are constant maps.*

3 Ergodicity in Practise

We now turn to the issue of determining in actual practise whether an upper transition operator is ergodic. In the case of finite-state, discrete-time Markov chains, a nice approach to deciding upon ergodicity was given by Kemeny and Snell [3, Sec. 1.4]. It is based on the notion of an accessibility relation. This is a binary (weak order) relation on set of states $\mathcal{X}$ that captures whether it is possible to go from one state to another in a finite number of steps. We intend to show that it is possible to associate an accessibility relation with an upper transition operator, and that this relation provides an intuitive interpretation of ergodicity in terms of accessibility. We refer to [1] for a detailed discussion of accessibility relations and their connections with upper transition operators.

Definition 4. *Consider an upper transition operator* T *on* $\mathcal{L}(\mathcal{X})$*, and two states* x *and* y *in* $\mathcal{X}$*. We say that* y *is* accessible *from* x in n steps *(notation:* $x \xrightarrow{n} y$*) if* $\mathrm{T}^n\mathbb{I}_{\{y\}}(x) > 0$*. We say that state* y *is* accessible *from state* x *(notation:* $x \to y$*) if* $\mathrm{T}^n\mathbb{I}_{\{y\}}(x) > 0$ *for some* $n \in \mathbb{N}_0$*. We say that* x *and* y communicate *(notation:* $x \leftrightarrow y$*) if both* $x \to y$ *and* $y \to x$*.*

The relation $\to$ is a weak order (reflexive and transitive), and consequently $\leftrightarrow$ is an equivalence relation. The equivalence classes for this relation are called *communication classes*: maximal subsets of $\mathcal{X}$ for which every element has access to any other element. The accessibility relation induces a partial order on these communication classes.

In the case of finite-state, discrete-time Markov chains, this partial order gives us clues about the ergodicity of the Markov chain. For such a Markov chain to be ergodic, it is necessary and sufficient [1] that it should be *top class regular*, meaning that: (i) there should be only one *maximal* or *undominated* communication class—elements of a maximal communication class have no access to states not in that class—, in which case we call this unique maximal class $\mathcal{R}$ the *top class*; and (ii) the top class $\mathcal{R}$ should be *regular*, meaning that after some time k, all elements of this class become accessible to each other in any number of steps: for all x and y in $\mathcal{R}$ and for all $n \geq k$, $x \xrightarrow{n} y$.

For upper transition operators, it turns out that top class regularity is a necessary condition for ergodicity. However, top class regularity is by itself not a sufficient condition: we need some guarantee that the top class will eventually be reached—a requirement that is automatically fulfilled in finite-state discrete-time Markov chains.

Proposition 2. *An upper transition operator* T *is ergodic if and only if it is* regularly absorbing*, meaning that it satisfies the following properties:*

(i) *it is* top class regular*:*

$$\mathcal{R} := \left\{x \in \mathcal{X} \colon (\exists n \in \mathbb{N})(\forall k \geq n) \min \mathrm{T}^k\mathbb{I}_{\{x\}} > 0\right\} \neq \emptyset, \tag{TCR}$$

(ii) *it is* top class absorbing*: with* $\mathcal{R}^c := \mathcal{X} \setminus \mathcal{R}$*,*

$$(\forall y \in \mathcal{R}^c)(\exists n \in \mathbb{N})\, \mathrm{T}^n\mathbb{I}_{\mathcal{R}^c}(y) < 1. \tag{TCA}$$

For a proof that (TCR) is equivalent to $\mathcal{R} \neq \emptyset$, we refer to [1, Prop. 4.3]. (TCA) means that for every element y not in the top class, there is some finite number of steps n after which the top class can be reached with a strictly positive *lower* probability $1 - \mathrm{T}^n \mathbb{I}_{\mathcal{R}^c}(y)$.

Proof. (TCR) $\wedge$ (TCA) $\Rightarrow$ (ER). Consider any periodic point ξ of T with period $p \in \mathbb{N}$. By Prop. 1, we have to show that ξ is constant. Using T1, T5, T3 and the periodic character of ξ, we infer that $\min \mathrm{T}^k \xi = \min \xi$ and $\max \mathrm{T}^k \xi = \max \xi$ for all $k \in \mathbb{N}_0$. Using T5, T4 and T3 we construct from $\mathrm{T}^k \xi \geq \min \mathrm{T}^k \xi + [\mathrm{T}^k \xi(x) - \min \mathrm{T}^k \xi] \mathbb{I}_{\{x\}} = \min \xi + [\mathrm{T}^\xi(x) - \min \xi] \mathbb{I}_{\{x\}}$ the following inequality, which holds for all $n, k \in \mathbb{N}_0$ and all $x \in \mathcal{X}$: $\mathrm{T}^n \xi \geq \min \xi + \left[\mathrm{T}^k \xi(x) - \min \xi\right] \mathrm{T}^n \mathbb{I}_{\{x\}}$. Taking the minimum on both sides of this inequality, we find that $0 \geq \left[\mathrm{T}^k \xi(x) - \min \xi\right] \min \mathrm{T}^n \mathbb{I}_{\{x\}}$. We infer from (TCR) that by taking n large enough, we can ensure that $\min \mathrm{T}^n \mathbb{I}_{\{x\}} > 0$ if $x \in \mathcal{R}$. So we already find that $\mathrm{T}^k \xi(x) = \min \xi$ for all $k \in \mathbb{N}_0$ and $x \in \mathcal{R}$.

Using T5, T4 and T3 we construct from $\xi \leq \max \xi - [\max \xi - \max_{x \in \mathcal{R}} \xi(x)] \mathbb{I}_{\mathcal{R}}$ and $-\mathbb{I}_{\mathcal{R}} = \mathbb{I}_{\mathcal{R}^c} - 1$ the following inequality, which holds for all $k \in \mathbb{N}$:

$$\mathrm{T}^k \xi \leq \max \xi + \left[\max \xi - \max_{x \in \mathcal{R}} \xi(x)\right] (\mathrm{T}^k \mathbb{I}_{\mathcal{R}^c} - 1).$$

By taking the maximum over $\mathcal{R}^c$ on both sides of this inequality, we get

$$\max_{y \in \mathcal{R}^c} \mathrm{T}^k \xi(y) - \max \xi \leq \left[\max \xi - \max_{x \in \mathcal{R}} \xi(x)\right] \left(\max_{y \in \mathcal{R}^c} \mathrm{T}^k \mathbb{I}_{\mathcal{R}^c}(y) - 1\right). \tag{1}$$

It follows from (TCA) that for each $y \in \mathcal{R}^c$, we can consider some $n_y \in \mathbb{N}$ such that $\mathrm{T}^{n_y} \mathbb{I}_{\mathcal{R}^c}(y) < 1$. Let $n := \max_{y \in \mathcal{R}^c} n_y$. Then we see that for every $y \in \mathcal{R}^c$:

$$\mathrm{T}^n \mathbb{I}_{\mathcal{R}^c}(y) = \mathrm{T}^{n_y}[(\mathbb{I}_{\mathcal{R}} + \mathbb{I}_{\mathcal{R}^c}) \mathrm{T}^{n-n_y} \mathbb{I}_{\mathcal{R}^c}](y) = \mathrm{T}^{n_y}[\mathbb{I}_{\mathcal{R}^c} \mathrm{T}^{n-n_y} \mathbb{I}_{\mathcal{R}^c}](y) \leq \mathrm{T}^{n_y} \mathbb{I}_{\mathcal{R}^c}(y).$$

The second equality follows from the fact that $\mathbb{I}_{\mathcal{R}} \mathrm{T}^{n-n_y} \mathbb{I}_{\mathcal{R}^c} = 0$: an element in the top class $\mathcal{R}$ has no access to any element outside it. The inequality follows from $\mathbb{I}_{\mathcal{R}^c} \leq 1$ and T5. We conclude that $\max_{y \in \mathcal{R}^c} \mathrm{T}^n \mathbb{I}_{\mathcal{R}^c}(y) - 1 < 0$. If $\mathrm{T}^n \xi$ reaches its maximum on $\mathcal{R}$, then $\max \xi = \max \mathrm{T}^n \xi = \min \xi$. If the maximum of $\mathrm{T}^n \xi$ is not reached on $\mathcal{R}$, then $\max \xi = \max_{y \in \mathcal{R}^c} \mathrm{T}^n \xi(y)$, and Eq. (1) for $k = n$ tells us that $\max \xi = \max_{x \in \mathcal{R}} \xi(x) = \min \xi$. In both cases, ξ is indeed constant.

(ER) $\Rightarrow$ (TCR) $\wedge$ (TCA). We use contraposition and show first that $\neg$(TCR) $\Rightarrow \neg$(ER). Then we show that $\neg$(TCA) $\wedge$ (TCR) $\Rightarrow \neg$(ER).

$\neg$(TCR) $\Rightarrow \neg$(ER). Not being top class regular means that $\mathcal{R} = \emptyset$, which is equivalent to $(\forall x \in \mathcal{X})(\forall n \in \mathbb{N})(\exists k \geq n)(\exists z \in \mathcal{X}) \mathrm{T}^k \mathbb{I}_{\{x\}}(z) = 0$. Since we infer from $\mathbb{I}_{\{x\}} \geq 0$ and T1 that $\mathrm{T}^k \mathbb{I}_{\{x\}} \geq 0$, this implies that $\liminf_{n \to \infty} \min \mathrm{T}^n \mathbb{I}_{\{x\}} = 0$. But for any $n \in \mathbb{N}$, $\mathrm{T}^{n+1} \mathbb{I}_{\{x\}} = \mathrm{T}(\mathrm{T}^n \mathbb{I}_{\{x\}}) \geq \min \mathrm{T}^n \mathbb{I}_{\{x\}}$ by T1, and therefore also $\min \mathrm{T}^{n+1} \mathbb{I}_{\{x\}} \geq \min \mathrm{T}^n \mathbb{I}_{\{x\}}$. This implies that the sequence $\min \mathrm{T}^n \mathbb{I}_{\{x\}}$ is non-decreasing, and bounded above [by 1], and therefore convergent. This leads to the conclusion that $(\forall x \in \mathcal{X}) \lim_{n \to \infty} \min \mathrm{T}^n \mathbb{I}_{\{x\}} = 0$.

We also infer from T1 and T2 that $1 = \mathrm{T}^k \mathbb{I}_{\mathcal{X}} \leq \sum_{x \in \mathcal{X}} \mathrm{T}^k \mathbb{I}_{\{x\}}$. Since the cardinality $|\mathcal{X}|$ of the state space is finite, this means that for all $z \in \mathcal{X}$ and all $n \in \mathbb{N}$

there is some $x \in \mathcal{X}$ such that $\mathrm{T}^n\mathbb{I}_{\{x\}}(z) \geq 1/|\mathcal{X}|$. This tells us that $\max \mathrm{T}^n\mathbb{I}_{\{x\}} \geq 1/|\mathcal{X}|$. Since we can infer from a similar argument as before that the sequence $\max \mathrm{T}^n\mathbb{I}_{\{x\}}$ converges, this tells us that $(\forall x \in \mathcal{X}) \lim_{n\to\infty} \max \mathrm{T}^n\mathbb{I}_{\{x\}} \geq 1/|\mathcal{X}|$. Combining both limit results tells us that $\lim_{n\to\infty}(\max \mathrm{T}^n\mathbb{I}_{\{x\}} - \min \mathrm{T}^n\mathbb{I}_{\{x\}}) > 0$, so T cannot be ergodic.

$\neg(\mathrm{TCA}) \wedge (\mathrm{TCR}) \Rightarrow \neg(\mathrm{ER})$. Since T is not top class absorbing, we know that there is some $y \in \mathcal{R}^c$ such that $\mathrm{T}^n\mathbb{I}_{\mathcal{R}^c}(y) = 1$ for all $n \in \mathbb{N}$. As the top class $\mathcal{R}$ is non-empty, we know that there is some $x \in \mathcal{R}$, and this x has no access to any state outside the maximal communication class $\mathcal{R}$: $\mathrm{T}^n\mathbb{I}_{\mathcal{R}^c}(x) = 0$ for all $n \in \mathbb{N}$. Consequently, $\lim_{n\to\infty}(\max \mathrm{T}^n\mathbb{I}_{\mathcal{R}^c} - \min \mathrm{T}^n\mathbb{I}_{\mathcal{R}^c}) = 1 - 0 > 0$, so T cannot be ergodic.

3.1 Checking for Top Class Regularity

Checking for top class regularity directly using the definition would involve calculating for every state x the maps $\mathrm{T}\mathbb{I}_{\{x\}}, \mathrm{T}^2\mathbb{I}_{\{x\}}, \ldots, \mathrm{T}^n\mathbb{I}_{\{x\}}$ until a first number $n = n_x$ is found for which $\min \mathrm{T}^{n_x}\mathbb{I}_{\{x\}} > 0$. Unfortunately, it is not clear whether this procedure is guaranteed to terminate after a certain number of iterations, or whether we can stop checking after a fixed number of iterations. Here, we want to take a closer look at this problem.

The next proposition shows that all the information we need in order to check top class regularity is incorporated in a single application of T to the atoms of $\mathcal{X}$.

Proposition 3. *Let* T *be an upper transition operator on* $\mathcal{L}(\mathcal{X})$, $n \in \mathbb{N}$ *and* $x, y \in \mathcal{X}$. *Then* $\mathrm{T}^n\mathbb{I}_{\{y\}}(x) > 0$ *if and only if there is some sequence* x_0, x_1, x_2, $\ldots$, x_{n-1}, x_n *in* $\mathcal{X}$ *with* $x_0 = x$ *and* $x_n = y$ *such that* $\mathrm{T}\mathbb{I}_{\{x_{k+1}\}}(x_k) > 0$ *for all* $k \in \{0, 1, \ldots, n-1\}$.

Proof. Sufficiency. Fix arbitrary k and ℓ in $\mathbb{N}$, and u and v in $\mathcal{X}$. Since $\mathrm{T}^\ell\mathbb{I}_{\{y\}} = \sum_{z\in\mathcal{X}} \mathbb{I}_{\{z\}}\mathrm{T}^\ell\mathbb{I}_{\{y\}}(z) \geq \mathbb{I}_{\{v\}}\mathrm{T}^\ell\mathbb{I}_{\{y\}}(v)$, it follows from T5 and T3 that $\mathrm{T}^{k+\ell}\mathbb{I}_{\{y\}} \geq \mathrm{T}^k\mathbb{I}_{\{v\}}\mathrm{T}^\ell\mathbb{I}_{\{y\}}(v)$ and therefore $\mathrm{T}^{k+\ell}\mathbb{I}_{\{y\}}(x) \geq \mathrm{T}^k\mathbb{I}_{\{v\}}(x)\mathrm{T}^\ell\mathbb{I}_{\{y\}}(v)$. Applying this inequality repeatedly, we get:

$$\mathrm{T}^n\mathbb{I}_{\{y\}}(x) \geq \prod_{k=0}^{n-1} \mathrm{T}\mathbb{I}_{\{x_{k+1}\}}(x_k)$$

for any sequence x_0, x_1, x_2, $\ldots$, x_{n-1}, x_n in $\mathcal{X}$ with $x_0 = x$ and $x_n = y$. It follows that the left-hand side is positive as soon as all factors on the right-hand side are.

Necessity. We infer using T2 and T3 that

$$\mathrm{T}^n\mathbb{I}_{\{y\}}(x) = \mathrm{T}\left(\sum_{x_1\in\mathcal{X}} \mathbb{I}_{\{x_1\}}\mathrm{T}^{n-1}\mathbb{I}_{\{y\}}(x_1)\right)(x) \leq \sum_{x_1\in\mathcal{X}} \mathrm{T}^{n-1}\mathbb{I}_{\{y\}}(x_1)\mathrm{T}\mathbb{I}_{\{x_1\}}(x),$$

and repeating the same argument recursively leads to

$$\mathrm{T}^n\mathbb{I}_{\{y\}}(x) \leq \sum_{\substack{x_0,x_1,\ldots,x_{n-1},x_n\in\mathcal{X} \\ x_0=x, x_n=y}} \prod_{k=0}^{n-1} \mathrm{T}\mathbb{I}_{\{x_{k+1}\}}(x_k).$$

Since all the factors (and therefore all the terms) on the right-hand side are non-negative by T1 and T5, the positivity of the left-hand side implies that there must be at least one positive term on the right-hand side, all of whose factors must therefore be positive.

This proposition not only implies that the set $\left\{\mathrm{T}\mathbb{I}_{\{x\}} : x \in \mathcal{X}\right\}$ completely determines the accessibility relation $\rightarrow$, but also that it determines the 'accessibility in n steps' relation $\xrightarrow{n}$. In other words, not only the communication and maximal classes can be determined from $\left\{\mathrm{T}\mathbb{I}_{\{x\}} : x \in \mathcal{X}\right\}$, but also their regularity.

Definition 5. *A stochastic matrix* $M \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ represents *an upper transition operator* T *on* $\mathcal{L}(\mathcal{X})$ *if* $M_{x,y} > 0 \Leftrightarrow \mathrm{T}\mathbb{I}_{\{y\}}(x) > 0$ *for all* x *and* y *in* $\mathcal{X}$.

Clearly, any upper transition operator T has an infinity of representing stochastic matrices. By Prop. 3, accessibility in n steps is completely determined by $\left\{\mathrm{T}\mathbb{I}_{\{x\}} : x \in \mathcal{X}\right\}$, so we may conclude that the finite-state, discrete-time Markov chain with transition matrix M on the one hand and the upper transition operator T on the other hand will invoke exactly the same 'accessibility in n steps' relations $\xrightarrow{n}$. They will have the same communication classes, the same maximal classes and the same regular class. This allows us to use the entire machinery of finite-state, discrete-time Markov chains to decide upon top class regularity for upper transition operators. We are led to the following immediate conclusion, see [2, Theorem 1.7] for an explanation of what (ii) means, and the equivalence between (ii) and (iii).

Proposition 4 (Top class regularity). *Consider an upper transition operator* T *and any stochastic matrix* M *that represents it. Then the following statements are equivalent:* (i) T *is top class regular;* (ii) M *is regular; and* (iii) M *has exactly one eigenvalue with modulus 1.*

Clearly this single eigenvalue must be equal to 1, because M is a stochastic matrix.

Example 1. Let $\mathcal{X} := \{x, y\}$ and $\mathrm{T}f := f(x)\mathbb{I}_{\{x\}} + \max\{f(x), f(y)\}\mathbb{I}_{\{y\}}$ for all $f \in \mathcal{L}(\mathcal{X})$. Then $\mathrm{T}\mathbb{I}_{\{x\}} = \mathbb{I}_{\mathcal{X}}$ and $\mathrm{T}\mathbb{I}_{\{y\}} = \mathbb{I}_{\{y\}}$. This means that the stochastic matrix $M = \left(\begin{smallmatrix} 1 & 0 \\ 1/2 & 1/2 \end{smallmatrix}\right)$ represents T. Since M has eigenvalues 1 and $1/2$, we conclude that T is top class regular.

3.2 Checking for Top Class Absorption

We now present an efficient procedure to check for top class absorption.

Proposition 5 (Top class absorption). *Let* T *be an upper transition operator with regular top class* $\mathcal{R}$. *Consider the nested sequence of subsets of* $\mathcal{R}^c$ *defined by the iterative scheme:*

$$A_0 := \mathcal{R}^c \text{ and } A_{n+1} := \{a \in A_n : \mathrm{T}\mathbb{I}_{A_n}(a) = 1\}, \quad n \geq 0.$$

After $k \leq |\mathcal{R}^c|$ *iterations, we reach* $A_k = A_{k+1}$. *Then* T *is top class absorbing if and only if* $A_k = \emptyset$.

Proof. We start by showing inductively that under the given assumptions, the statement

$$H_n: \quad \mathbb{I}_{A_n}\mathrm{T}^n\mathbb{I}_{\mathcal{R}^c} = \mathbb{I}_{A_n} \wedge (\forall a \in A_{n+1}^c)\mathrm{T}\mathbb{I}_{A_n}(a) < 1 \wedge (\forall a \in A_n^c)\mathrm{T}^n\mathbb{I}_{\mathcal{R}^c}(a) < 1$$

holds for all $n \geq 0$. We first prove that the statement H_n holds for $n = 0$. The first and third statements of H_0 hold trivially. For the second statement, we have to prove that $\mathrm{T}\mathbb{I}_{A_0}(a) < 1$ for all $a \in A_1^c = A_0^c \cup A_0 \setminus A_1$. On $A_0 \setminus A_1$, the desired inequality holds by definition. On $A_0^c = \mathcal{R}$ it holds because there $\mathrm{T}\mathbb{I}_{A_0}$ is zero: no state in the top class $\mathcal{R}$ has access to any state outside it. Next, we prove that $H_n \Rightarrow H_{n+1}$. First of all,

$$\mathrm{T}^{n+1}\mathbb{I}_{A_0} = \mathrm{T}(\mathrm{T}^n\mathbb{I}_{A_0}) = \mathrm{T}[\mathbb{I}_{A_n}\mathrm{T}^n\mathbb{I}_{A_0} + \mathbb{I}_{A_n^c}\mathrm{T}^n\mathbb{I}_{A_0}] = \mathrm{T}[\mathbb{I}_{A_n} + \mathbb{I}_{A_n^c}\mathrm{T}^n\mathbb{I}_{A_0}], \quad (2)$$

where the last equality follows from the induction hypothesis H_n. It follows from the definition of A_{n+1} that $\mathbb{I}_{A_{n+1}}\mathrm{T}\mathbb{I}_{A_n} = \mathbb{I}_{A_{n+1}}$, and therefore

$$\begin{aligned}
\mathbb{I}_{A_{n+1}} &= \mathbb{I}_{A_{n+1}}\mathrm{T}[\mathbb{I}_{A_n} + \mathbb{I}_{A_n^c}\mathrm{T}^n\mathbb{I}_{A_0} - \mathbb{I}_{A_n^c}\mathrm{T}^n\mathbb{I}_{A_0}] \\
&\leq \mathbb{I}_{A_{n+1}}\mathrm{T}[\mathbb{I}_{A_n} + \mathbb{I}_{A_n^c}\mathrm{T}^n\mathbb{I}_{A_0}] + \mathbb{I}_{A_{n+1}}\mathrm{T}[-\mathbb{I}_{A_n^c}\mathrm{T}^n\mathbb{I}_{A_0}] \\
&= \mathbb{I}_{A_{n+1}}\mathrm{T}^{n+1}\mathbb{I}_{A_0} + \mathbb{I}_{A_{n+1}}\mathrm{T}[-\mathbb{I}_{A_n^c}\mathrm{T}^n\mathbb{I}_{A_0}] \leq \mathbb{I}_{A_{n+1}}\mathrm{T}^{n+1}\mathbb{I}_{A_0} \leq \mathbb{I}_{A_{n+1}},
\end{aligned}$$

where the first inequality follows from T2, the second from $-\mathbb{I}_{A_n^c}\mathrm{T}^n\mathbb{I}_{A_0} \leq 0$ and therefore $\mathbb{I}_{A_{n+1}}\mathrm{T}[-\mathbb{I}_{A_n^c}\mathrm{T}^n\mathbb{I}_{A_0}] \leq 0$ [use T1 and T5], and the third from $\mathrm{T}^{n+1}\mathbb{I}_{A_0} \leq 1$ [use T5]. The second equality follows from Eq. (2). Hence indeed $\mathbb{I}_{A_{n+1}} = \mathbb{I}_{A_{n+1}}\mathrm{T}^{n+1}\mathbb{I}_{A_0}$. Next, observe that $A_{n+2}^c = A_{n+1}^c \cup A_{n+1} \setminus A_{n+2}$. By definition, $\mathrm{T}\mathbb{I}_{A_{n+1}}(a) < 1$ for all $a \in A_{n+1} \setminus A_{n+2}$. It also follows from the induction hypothesis H_n that $\mathrm{T}\mathbb{I}_{A_n}(a) < 1$ for all $a \in A_{n+1}^c$. But since $A_{n+1} \subseteq A_n$, it follows from T5 that $\mathrm{T}\mathbb{I}_{A_{n+1}} \leq \mathrm{T}\mathbb{I}_{A_n}$, and therefore also $\mathrm{T}\mathbb{I}_{A_{n+1}}(a) < 1$ for all $a \in A_{n+1}^c$. Hence indeed $\mathrm{T}\mathbb{I}_{A_{n+1}}(a) < 1$ for all $a \in A_{n+2}^c$. To finish the induction proof, let $\beta := \max_{b \in A_n^c} \mathrm{T}^n\mathbb{I}_{\mathcal{R}^c}(a)$, then $\beta < 1$ by the induction hypothesis H_n. We then infer from Eq. (2) that

$$\mathrm{T}^{n+1}\mathbb{I}_{A_0} \leq \mathrm{T}[\mathbb{I}_{A_n} + \beta\mathbb{I}_{A_n^c}] = \mathrm{T}[\beta + (1-\beta)\mathbb{I}_{A_n}] \leq \beta + (1-\beta)\mathrm{T}\mathbb{I}_{A_n}.$$

Consider any $a \in A_{n+1}^c$, then $\mathrm{T}\mathbb{I}_{A_n}(a) < 1$ by the induction hypothesis H_n, and therefore $\mathrm{T}^{n+1}\mathbb{I}_{A_0}(a) \leq \beta + (1-\beta)\mathrm{T}\mathbb{I}_{A_n}(a) < 1$ since also $\beta < 1$. We conclude that H_{n+1} holds too.

To continue the proof: since the sequence $A_0, A_1, \ldots, A_n, \ldots$ is non-increasing and A_0 is finite, there is some first $k \in \mathbb{N}$ such that $A_{k+1} = A_k$. Clearly, $k \leq |A_0|$. We now prove by induction that the statement $G_n: \mathbb{I}_{A_k}\mathrm{T}^{n+k}\mathbb{I}_{A_0} = \mathbb{I}_{A_k}$ holds for all $n \geq 0$. The statement G_n clearly holds for $n = 0$: it follows directly from H_k. We show that $G_n \Rightarrow G_{n+1}$. First of all, as above

$$\mathrm{T}^{n+k+1}\mathbb{I}_{A_0} = \mathrm{T}[\mathbb{I}_{A_k}\mathrm{T}^{n+k}\mathbb{I}_{A_0} + \mathbb{I}_{A_k^c}\mathrm{T}^{n+k}\mathbb{I}_{A_0}] = \mathrm{T}[\mathbb{I}_{A_k} + \mathbb{I}_{A_k^c}\mathrm{T}^{n+k}\mathbb{I}_{A_0}],$$

where the last equality follows from the induction hypothesis G_n. As above, it follows from the definition of A_{k+1} that $\mathbb{I}_{A_{k+1}}\mathrm{T}\mathbb{I}_{A_k} = \mathbb{I}_{A_{k+1}}$, and therefore $\mathbb{I}_{A_k}\mathrm{T}\mathbb{I}_{A_k} = \mathbb{I}_{A_k}$, so

$$\mathbb{I}_{A_k} = \mathbb{I}_{A_k}\mathrm{T}[\mathbb{I}_{A_k} + \mathbb{I}_{A_k^c}\mathrm{T}^{n+k}\mathbb{I}_{A_0} - \mathbb{I}_{A_k^c}\mathrm{T}^{n+k}\mathbb{I}_{A_0}]$$

$$\leq \mathbb{I}_{A_k} \mathrm{T}[\mathbb{I}_{A_n k} + \mathbb{I}_{A_k^c} \mathrm{T}^{n+k} \mathbb{I}_{A_0}] + \mathbb{I}_{A_k} \mathrm{T}[-\mathbb{I}_{A_k^c} \mathrm{T}^{n+k} \mathbb{I}_{A_0}]$$
$$= \mathbb{I}_{A_k} \mathrm{T}^{n+k+1} \mathbb{I}_{A_0} + \mathbb{I}_{A_k} \mathrm{T}[-\mathbb{I}_{A_k^c} \mathrm{T}^{n+k} \mathbb{I}_{A_0}] \leq \mathbb{I}_{A_k} \mathrm{T}^{n+k+1} \mathbb{I}_{A_0} \leq \mathbb{I}_{A_k}.$$

The first inequality follows from T2, the second from $-\mathbb{I}_{A_k^c} \mathrm{T}^{n+k} \mathbb{I}_{A_0} \leq 0$ and therefore $\mathbb{I}_{A_k} \mathrm{T}[-\mathbb{I}_{A_k^c} \mathrm{T}^{n+k} \mathbb{I}_{A_0}] \leq 0$ [use T1 and T5], and the third from $\mathrm{T}^{n+k+1} \mathbb{I}_{A_0} \leq 1$ [use T5]. Hence indeed $\mathbb{I}_{A_k} = \mathbb{I}_{A_k} \mathrm{T}^{n+k+1} \mathbb{I}_{A_0}$.

There are now two possibilities. The first is that $A_k \neq \emptyset$. It follows from the arguments above [namely $H_1, \ldots, H_k$ and all G_n] that for any element a of A_k, $\mathrm{T}^\ell \mathbb{I}_{\mathcal{R}^c}(a) = 1$ for all $\ell \in \mathbb{N}$, which implies that T cannot be top class absorbing. The second possibility is that $A_k = \emptyset$. It follows from the arguments above [namely H_k] that $\mathrm{T}^k \mathbb{I}_{\mathcal{R}^c}(a) < 1$ for all $a \in A_k^c = \mathcal{X}$, which implies that T is top class absorbing.

Example 2. Define $\mathrm{T}f = \max\{Mf : L \leq M \leq U \text{ and } M \text{ stochastic}\}$ where L and U are given by

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 0 & 0 \\ 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/4 \end{pmatrix} \text{ and } U = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/2 & 3/4 & 1/2 & 0 & 0 \\ 3/4 & 1/2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1/4 & 3/4 & 0 & 0 & 1/4 \end{pmatrix}.$$

Here, any representing matrix's non-zero elements correspond to those of U. One representing matrix is

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 0 & 1/3 \end{pmatrix}.$$

with characteristic polynomial $\chi_M(s) = (s-1)(s-1/3)^2(s-(1+\sqrt{7})/6)(s-(1-\sqrt{7})/6)$. So T is top class regular. From the form of M we infer that the 'first' state is absorbing which implies that $\mathbb{I}_{\mathcal{R}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix}^T$.

To check for top class absorbtion, we start iterating:

(step 1) $\mathrm{T}\mathbb{I}_{\mathcal{R}^c} = \begin{pmatrix} 0 & 1 & 1/2 & 1 & 1 \end{pmatrix}^T$ whence $\mathbb{I}_{A_1} = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \end{pmatrix}^T$,
(step 2) $\mathrm{T}\mathbb{I}_{A_1} = \begin{pmatrix} 0 & 3/4 & 1/2 & 1 & 1 \end{pmatrix}^T$ whence $\mathbb{I}_{A_2} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \end{pmatrix}^T$,
(step 3) $\mathrm{T}\mathbb{I}_{A_2} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1/4 \end{pmatrix}^T$ whence $\mathbb{I}_{A_3} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \end{pmatrix}^T$,
(step 4) $\mathrm{T}\mathbb{I}_{A_3} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \end{pmatrix}^T$ whence $\mathbb{I}_{A_4} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \end{pmatrix}^T$.

Because $A_4 = A_3 \neq \emptyset$ we conclude that T is not top class absorbing and therefore not ergodic.

4 Conclusion

We have discussed a number of equivalent characterisations of the ergodicity of an upper transition operator—which corresponds to a set of Markov chains.

We have shown that ergodicity is completely determined by the eigenvalues and eigenfunctions of the upper transition operator, as is the case in Markov chains. This could open the door to a spectral theorem for upper transition operators. Unfortunately, at this point it is not known how to calculate such eigenvalues, and for this reason we have developed an alternative test for ergodicity, which needs at most $2|\mathcal{X}| - 1$ evaluations of the upper transition operator. The test consists of two steps: the first checks for top class regularity by building a representing stochastic matrix and solving a linear eigenvalue problem; the second checks for top class absorption.

We have not yet investigated the conditions under which $\{\mathrm{T}^n f\}$ will converge in general. Extrapolating from the conditions for Perron–Frobenius-like convergence given above, it might be conjectured that there is convergence if and only if all classes that are not regular are absorbed by some union of regular classes. Also worth investigating is whether we could find stochastic matrices that represent the upper transition operator in a more quantitative way than the purely qualitative one we introduced above. Such representing matrices might then for instance be used to make inference about the rate or speed of transition from individual states to an absorbing class.

Acknowledgements

This paper presents research results of the Belgian programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office, and SBO project 060043 of the IWT-Vlaanderen. The scientific responsibility rests with its authors.

References

1. De Cooman, G., Hermans, F., Quaeghebeur, E.: Imprecise Markov chains and their limit behavior. Probability in the Engineering and Informational Sciences 23(4), 597–635 (2009)
2. Hartfiel, D.J.: Markov Set-Chains. Lecture Notes in Mathematics, vol. 1695. Springer, Berlin (1998)
3. Kemeny, J.G., Snell, J.L.: Finite Markov Chains. Undergraduate Text in Mathematics. Springer, New York (1976); reprint of the 1960 Edition
4. Lemmens, B., Scheutzow, M.: On the dynamics of sup-norm non-expansive maps. Ergod. Theor. Dyn. Syst. 25, 861–871 (2005)
5. Seneta, E.: Non-negative Matrices and Markov Chains. Springer Series in Statistics. Springer, Heidelberg (2006)
6. Sine, R.: A nonlinear Perron–Frobenius theorem. Proceedings of the American Mathematical Society 109(2), 331–336 (1990)
7. Škulj, D., Hable, R.: Coefficients of ergodicity for imprecise markov chains. In: Augustin, T., Coolen, F.P., Serafin, M., Troffaes, M.C. (eds.) ISIPTA 2009 – Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications, pp. 377–386. SIPTA (2009)

An Empirical Comparison of Bayesian and Credal Set Theory for Discrete State Estimation

Alexander Karlsson, Ronnie Johansson, and Sten F. Andler

Informatics Research Centre, University of Skövde, Sweden
{alexander.karlsson,ronnie.johansson,sten.f.andler}@his.se

Abstract. We are interested in whether or not there exist any advantages of utilizing credal set theory for the discrete state estimation problem. We present an experiment where we compare in total six different methods, three based on Bayesian theory and three on credal set theory. The results show that Bayesian updating performed on centroids of operand credal sets significantly outperforms the other methods. We analyze the result based on degree of imprecision, position of extreme points, and second-order distributions.

Keywords: Bayesian theory, credal sets, imprecise probability.

1 Introduction

There exist many theories for *reasoning under uncertainty*. Two categories of such theories can be formulated based on the mathematical structure used for representing ones *belief*: (1) *a single probability function*, and (2) *a set of probability functions*. One of the most commonly utilized theories, belonging to the first category, is *Bayesian theory* [1]. However, one criticism of Bayesian theory is that its *belief structure*, i.e., a probability function, seriously limits the types of belief that one might want to express [2], and that such limitation implies a risk of drawing erroneous conclusions from the *posterior*. As an example, the use of a single probability function prohibits one of expressing a *probability interval* as belief, something that seems natural in many situations where limited information is available. In order to overcome this deficiency of Bayesian theory, it has been proposed that a *closed convex set of probability functions*, also known as a *credal set* [3,4], should be utilized as a belief structure. Credal sets are in fact a part of what is a quite natural extension to Bayesian theory, referred to as *credal set theory*[1] [4], where one applies Bayes theorem *point-wise* on a *prior credal set* and a *closed convex set of likelihood functions*.

So far, research concerning empirical evaluation of the performance between Bayesian and credal set theory has been scarce, leaving the importance to applications of the latter theory undetermined. Moreover, we argue that existing such evaluations (e.g., [6,7]) have been implemented with a bias towards credal set theory, since one most often has, in some way, compared a *set of decisions*,

[1] Also known as "theory of credal sets" [4] and "quasi-Bayesian theory" [5].

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 80–89, 2010.

as a result of a credal approach, to a *single decision*, as a result of a Bayesian approach. We argue that if such comparison should be regarded as fair then the decision output from the approaches must be of the same type. This can be done in two ways: (1) one can construct decision set-output from the Bayesian approach, or (2) one can force a single decision from the credal approach. In this paper, we will explore alternative (2) with respect to the *discrete estimation problem*, i.e., to determine the true state of a discrete random variable by forcing such decision based on a credal set.

The paper is organized as follows: in Sect. 2, we will present the foundations of Bayesian and credal set theory. In Sect. 3, we present an experiment that compares Bayesian theory with credal set theory with respect to discrete state estimation. In Sect. 4 and 5 we present and analyze the result of the experiment. Lastly, in Sect. 6, we give a summary of the paper and the main conclusions.

2 Preliminaries

We elaborate on Bayesian and credal set theory, on their main differences, and how these theories can be utilized for the discrete state estimation problem.

2.1 Bayesian Theory

Bayesian theory is based on the assumption that *belief* can be adequately represented by a *single probability function* and that *Bayes' theorem* is utilized for *updating* belief whenever new observations from the environment of interest become available. Let X and Y be random variables with discrete state spaces Ω_X and Ω_Y, respectively, and let $x \in \Omega_X$ and $y \in \Omega_Y$. The Bayesian update operator (Bayes' theorem) can now be defined as [1]:

Definition 1. *The Bayesian update operator $\Phi_{\mathcal{B}}$ is defined as:*

$$\Phi_{\mathcal{B}}(p(X), p(y|X)) \triangleq \frac{p(y|X)p(X)}{\sum\limits_{x \in \Omega_X} p(y|x)p(x)}\,, \tag{1}$$

where $p(X)$ is the prior probability function and $p(y|X)$ the likelihood function. The operator is undefined if $\sum_{x \in \Omega_X} p(y|x)p(x) = 0$.

The Bayesian approach to the discrete state estimation problem is simply to select a state from the following set:

$$\mathcal{D}_{\mathcal{B}}(p(X)) \triangleq \{x_i \in \Omega_X : p(x_i) \geq p(x_j),\, \forall x_j \in \Omega_X\}\,, \tag{2}$$

where $\mathcal{D}_{\mathcal{B}}(p(X))$ is an operator that returns the set of all optimal states with respect to $p(X)$. Although $\mathcal{D}_{\mathcal{B}}(p(X))$ can in general be non-singleton, that is not the common case.

2.2 Credal Set Theory

As already mentioned in the introduction credal set theory [5,3] is an extension to Bayesian theory where one allows for closed convex sets of probability and likelihood functions. Let $\mathcal{P}_X$ denote a credal set that contains functions of the form $p(X)$, $\mathcal{P}_{X|y}$ functions of the form $p(X|y)$, and let $\mathcal{P}_{x|Y}$ denote a closed convex set of likelihood functions $p(x|Y)$. By utilizing this notation the credal update operator, corresponding to a "point-wise" version of the Bayesian operator in Definition 1, can now be defined as, cf [5,8,9]:

Definition 2. *The credal update operator $\Phi_{\mathcal{C}}$ is defined as:*

$$\Phi_{\mathcal{C}}\left(\mathcal{P}_X, \mathcal{P}_{y|X}\right) \triangleq CH\left\{\Phi_{\mathcal{B}}(p(X), p(y|X)) : p(X) \in \mathcal{P}_X, p(y|X) \in \mathcal{P}_{y|X}\right\}, \quad (3)$$

where CH denotes the convex-hull operator, $\mathcal{P}_X$ is a credal set of prior probability functions $p(X)$ and $\mathcal{P}_{x|Y}$ is a closed convex set of likelihood functions $p(x|Y)$. The operator is undefined if there exists a pair $(p(X), p(x|Y)) \in \mathcal{P}_X \times \mathcal{P}_{x|Y}$ where $\Phi_{\mathcal{B}}(p(X), p(y|X))$ is undefined.

In order to be able to compute the credal update operator one adopts operand sets in the form of *polytopes*, i.e., the *convex hull* of a finite number of points (here functions), since it enables one to calculate the credal update operator in an exact way by utilizing the sets' *extreme points* [10, Theorem 1]. One key feature of credal sets is their *imprecision*, which can loosely be thought of as the volume of the sets. We will use the following measure for *degree of imprecision* [2,10]:

Definition 3. *The degree of imprecision for a credal set $\mathcal{P}_X$ is defined as:*

$$\mathcal{I}(\mathcal{P}_X) \triangleq \frac{1}{|\Omega_X|} \sum_{x \in \Omega_X} \left(\max_{p(x) \in \mathcal{P}_X} p(x) - \min_{p(x) \in \mathcal{P}_X} p(x) \right),$$

where Ω_X is a discrete state space for X.

In contrast to Bayesian theory, there is often no straightforward way of determining a single true state for the discrete state estimation problem based on a credal set [5]. The reason for this is the caution regarding conclusions induced by the imprecision. If we generalize the Bayesian way of decision making, taking into account that there now exists a set of possibilities for a probability function, we obtain:

$$\mathcal{D}_{\mathcal{C}}(\mathcal{P}_X) \triangleq \bigcup_{p(X) \in \mathcal{P}_X} \mathcal{D}_{\mathcal{B}}(p(X)), \quad (4)$$

where $\mathcal{D}_{\mathcal{B}}(p(X))$ is defined by (2). We see that unless all probability functions $p(X) \in \mathcal{P}_X$ agrees on some state as being the most probable, additional imprecision is likely to be introduced compared to the Bayesian counterpart. In order to still be able to decide for a single state in such cases, one can select a single representative probability function $p(X) \in \mathcal{P}_X$, by using some *selector function* $\Upsilon(\mathcal{P}_X)$, which then can be used in the same way as in the Bayesian approach

[9,5], i.e., $\mathcal{D}_{\mathcal{B}}(\Upsilon(\mathcal{P}_X))$. Such approach is actually very similar to what has previously been stipulated in a variant of *evidence theory* [11], called the *transferable belief model* (TBM) [12]. The idea in TBM is that one should maintain as much representational ability from the belief structure as long as a decision is not required to be implemented (i.e., for posterior calculations). In this paper, we focus on three selector functions that often can be found in the literature (see e.g., [9,5]): (1) *the centroid* $\Upsilon_c(\mathcal{P}_X)$, defined as:

$$\Upsilon_c(\mathcal{P}_X) \triangleq \mathrm{E}_{u(\mathcal{P}_X)}[p(X)], \tag{5}$$

where $u(\mathcal{P}_X)$ denotes the uniform distribution over $\mathcal{P}_X$, (2) *the maximum entropy function* $\Upsilon_e(\mathcal{P}_X)$, defined as:

$$\Upsilon_e(\mathcal{P}_X) \triangleq \arg \sup_{p(X)\in\mathcal{P}_X} \left(-\sum_{x\in\Omega_X} p(x)\ln(p(x)) \right), \tag{6}$$

and (3) sampling a function from the uniform distribution over the credal set, denoted $\Upsilon_u(\mathcal{P}_X)$:

$$\Upsilon_u(\mathcal{P}_X) \sim u(\mathcal{P}_X). \tag{7}$$

3 Experiment Design

Let us use a simple state space, i.e., $|\Omega_X| = 2$. The reason for choosing such low dimension is that we then can sample both credal sets and functions without using a number of approximations methods needed in higher dimensions. Suppose that we need to formulate a prior and a likelihood function, but where we have insufficient information in order to do so in a precise way, i.e., we report operand sets $\mathcal{P}_X$ and $\mathcal{P}_{y|X}$. Since the posterior is invariant to normalization of likelihoods we will use a normalized version of $\mathcal{P}_{y|X}$ throughout the paper, denoted by $\hat{\mathcal{P}}_{y|X}$ (i.e., $\hat{\mathcal{P}}_{y|X}$ is a credal set). Let us assume that the semantics behind the reported sets is that one is equally willing to perform reasoning based on any function within the sets, i.e., there implicitly exists a *uniform second-order distribution* over each of the sets. Consider utilizing Bayesian theory for this case. Since the Bayesian update operator dictates single functions as operands, it is necessary to select representative functions from the sets to utilize for updating. Such selection can be implemented by anyone of the selection functions defined by (5), (6), and (7), hence let:

$$\begin{aligned} p^u_{\mathcal{B}}(X|y) &\triangleq \Phi_{\mathcal{B}}(\Upsilon_u(\mathcal{P}_X), \Upsilon_u(\hat{\mathcal{P}}_{y|X})) \\ p^c_{\mathcal{B}}(X|y) &\triangleq \Phi_{\mathcal{B}}(\Upsilon_c(\mathcal{P}_X), \Upsilon_c(\hat{\mathcal{P}}_{y|X})) \\ p^e_{\mathcal{B}}(X|y) &\triangleq \Phi_{\mathcal{B}}(\Upsilon_e(\mathcal{P}_X), \Upsilon_e(\hat{\mathcal{P}}_{y|X})). \end{aligned} \tag{8}$$

Note that since we have assumed a uniform distribution over the operand sets, $p^c_{\mathcal{B}}(X|y)$ represents a method that utilizes the expected values of the operand sets, i.e., the centroids.

Now, in contrast to Bayesian theory, updating based on the sets $\mathcal{P}_X$ and $\hat{\mathcal{P}}_{y|X}$ is straightforward in credal set theory; simply use the $\Phi_{\mathcal{C}}$ operator. However, here we are likely to encounter the problem of determining a single state due to the posterior imprecision (see (4)). Thus, we will here utilize the selector functions on the posterior credal set instead, i.e.:

$$\begin{aligned} p^u_{\mathcal{C}}(X|y) &\triangleq \Upsilon_u(\Phi_{\mathcal{C}}(\mathcal{P}_X, \hat{\mathcal{P}}_{y|X})) \\ p^c_{\mathcal{C}}(X|y) &\triangleq \Upsilon_c(\Phi_{\mathcal{C}}(\mathcal{P}_X, \hat{\mathcal{P}}_{y|X})) \\ p^e_{\mathcal{C}}(X|y) &\triangleq \Upsilon_e(\Phi_{\mathcal{C}}(\mathcal{P}_X, \hat{\mathcal{P}}_{y|X}))\,. \end{aligned} \tag{9}$$

Now, let the "true" prior and likelihood functions, i.e., functions formulated under a condition of a large amount of information, exist in $\mathcal{P}_X$ and $\hat{\mathcal{P}}_{y|X}$. This means that the posterior truth, denoted by $p_t(X|y)$, must exists in $\Phi_{\mathcal{C}}(\mathcal{P}_X, \hat{\mathcal{P}}_{y|X})$. Furthermore, let us assume that the truth in each operand set is distributed in the same way as our semantics behind such reported sets, i.e., the true function is uniformly distributed over each set. We will utilize two *score functions* that measure the performance of the methods with respect to $p_t(X|y)$. First consider a score function that measures the *accuracy*[2]:

$$s_a(p(X|y), p_t(X|y)) \triangleq \begin{cases} 1 & \text{if } \mathcal{D}_{\mathcal{B}}(p(X|y)) = \mathcal{D}_{\mathcal{B}}(p_t(X|y)) \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

where $p(X|y)$ is any of the methods in (8) or (9). In order to be able to discriminate the performance of the different methods at a finer level, we also use a *loss function* known as *the Brier loss* [13]:

$$s_l(p(X|y), p_t(X|y)) \triangleq \sum_{x \in \Omega_X} (p(x|y) - p_t(x|y))^2\,. \tag{11}$$

Let us sample n true posteriors, denoted $\{p^i_t(X|y)\}^n_{i=1}$, by sampling from the uniform second-order distributions over the operand sets, and then applying the Bayesian update operator, i.e., each posterior sample $p^i_t(X|y)$ has been obtained by the following expression:

$$p^i_t(X|y) \triangleq \Phi_{\mathcal{B}}(\Upsilon_u(\mathcal{P}_X), \Upsilon_u(\hat{\mathcal{P}}_{y|X}))\,. \tag{12}$$

By drawing a large number of samples we can obtain an approximation of the *expected score* of the methods in (8) and (9):

$$\mathrm{E}[s_\bullet(p(X|y), p_t(X|y))|\mathcal{P}_X, \hat{\mathcal{P}}_{y|X}] \approx \frac{1}{n} \sum_{i=1}^{n} s_\bullet(p(X|y), p^i_t(X|y))\,, \tag{13}$$

[2] Accuracy in this setting can be thought of as an performance measure with respect to the "best guess". If $p_t(X|y)$ or $p(X|y)$ is uniform, we evaluate $s_a(p(X|y), p_t(X|y))$ by sampling.

where $s_\bullet$ denotes any of the function in (10) or (11). Now, for each problem instance $(\mathcal{P}_X, \hat{\mathcal{P}}_{y|X})$ we can *rank*[3] each method in (8) and (9) with respect to the expected scores in (13) (i.e., the best method receives rank 1, the second best rank two etc.). Let $\mathcal{R}_a(p(X|y))$ and $\mathcal{R}_l(p(X|y))$ denote the rank for any of the methods in (8) and (9), with respect to (10) and (11), respectively, given a specific problem instance $(\mathcal{P}_X, \hat{\mathcal{P}}_{y|X})$. By uniformly sampling a large number of credal operands[4], i.e., $\{\mathcal{P}_X^i\}_{i=1}^m$ and $\{\hat{\mathcal{P}}_{y|X}^i\}_{i=1}^m$, we can obtain an approximation of the expected rank over all problem instances with respect to each of the score functions, i.e.:

$$\mathrm{E}_{u(\Omega_{\mathcal{P}_X})}[\mathcal{R}_\bullet(p(X|y))] \approx \frac{1}{m}\sum_{i=1}^{m} \mathcal{R}_\bullet^i(p(X|y))\,, \tag{14}$$

where $\mathcal{R}_\bullet^i(p(X|y))$ denotes the rank, with respect to (10) and (11), for sampled operand credal sets $\mathcal{P}_X^i$ and $\hat{\mathcal{P}}_{y|X}^i$, and where the subscript $u(\Omega_{\mathcal{P}_X})$ indicates that the sampling of credal sets has been performed with respect to the uniform distribution over the state space of all credal sets for X, denoted by $\Omega_{\mathcal{P}_X}$. However, sampling operand credal sets uniformly from $\Omega_{\mathcal{P}_X}$ might not do the credal operator justice, since there exists considerably more credal sets with a low degree of imprecision than high, resulting in a bias to the former. Therefore we also sample operand credal sets with respect to the uniform distribution over the degree of imprecision, denoted $u_{\mathcal{I}}(\Omega_{\mathcal{P}_X})$. The corresponding expected ranks with respect to this form of sampling is denoted by $\mathrm{E}_{u_{\mathcal{I}}(\Omega_{\mathcal{P}_X})}[\mathcal{R}_\bullet(p(X|y))]$.

4 Results

The average rank of each method, where $m = 10^6$ (see (14)) and $n = 10^2$ (see (13)), can be seen in Table 1. A Friedman test [14] on the 5%-level is passed for both of the sampling methods and score functions, hence the differences in ranks are significant. A post-hoc Nemenyi test [14] on the same level is passed

Table 1. Expected rank for the methods defined in (8) and (9)

Expected rank	$p_{\mathcal{B}}^u(X\|y)$	$p_{\mathcal{B}}^c(X\|y)$	$p_{\mathcal{B}}^e(X\|y)$	$p_{\mathcal{C}}^u(X\|y)$	$p_{\mathcal{C}}^c(X\|y)$	$p_{\mathcal{C}}^e(X\|y)$
$\mathrm{E}_{u(\Omega_{\mathcal{P}_X})}[\mathcal{R}_a(\cdot)]$	3.489	**3.086**	3.511	3.715	3.190	4.010
$\mathrm{E}_{u(\Omega_{\mathcal{P}_X})}[\mathcal{R}_l(\cdot)]$	3.853	**1.449**	3.925	4.288	2.574	4.912
$\mathrm{E}_{u_{\mathcal{I}}(\Omega_{\mathcal{P}_X})}[\mathcal{R}_a(\cdot)]$	3.529	**2.931**	3.555	3.831	3.104	4.050
$\mathrm{E}_{u_{\mathcal{I}}(\Omega_{\mathcal{P}_X})}[\mathcal{R}_l(\cdot)]$	4.239	**1.597**	3.497	4.573	2.697	4.398

[3] The average rank is used in case of draw.

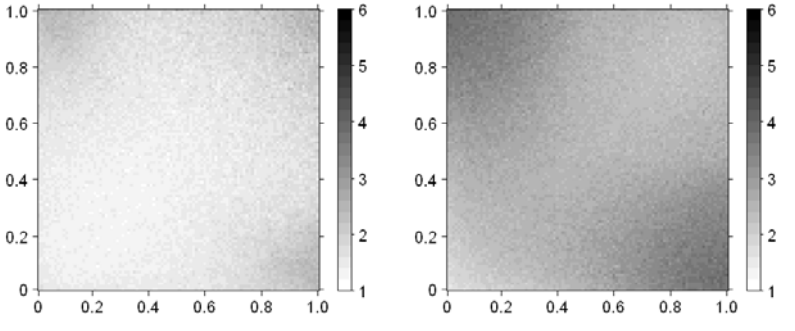
[4] Since $|\Omega_X| = 2$, uniformly sampling a credal set can be implemented by sampling from a triangle where the base represents the centroid and the height the imprecision.

for all comparisons. Therefore we conclude that the Bayesian method utilizing the centroid of the credal operands, i.e., $p^c_{\mathcal{B}}(X|y)$, significantly outperforms the other methods. Note that the corresponding credal method, i.e., $p^c_{\mathcal{C}}(X|y)$, is the second best method.

5 Analysis of Results

Consider the level plot in Fig. 1 where the degree of imprecision of the operands are shown on the x and y-axis and where the intensity indicates the expected rank $\mathrm{E}_{u_{\mathcal{I}}(\Omega_{\mathcal{P}_X})}[\mathcal{R}_l(\cdot)|\mathcal{I}(\mathcal{P}_X), \mathcal{I}(\hat{\mathcal{P}}_{y|X})]$ for $p^c_{\mathcal{B}}(X|y)$ and $p^c_{\mathcal{C}}(X|y)$. From the level plots we see that the performance of $p^c_{\mathcal{B}}(X|y)$ and $p^c_{\mathcal{C}}(X|y)$ is to a large extent uniformly distributed with respect to imprecision (the corresponding level plots for the accuracy has a similar appearance with the exception that the intensity is lower). A degradation of the performance of $p^c_{\mathcal{C}}(X|y)$ can be seen when one of the operands has a high degree of imprecision. In such cases the extreme points of one operand credal set has a tremendous effect on the other operand through the credal update operator, since the points are close to the boundaries of the *probability simplex* (the set of probability functions over a given state space). Consider Fig. 2 where we have plotted the expected Brier loss given the minimum distance from an operand to the boundary of the probability simplex, i.e., $\mathrm{E}_{u_{\mathcal{I}}(\Omega_{\mathcal{P}_X})}[s_l(\cdot, p_t(X|y))|\gamma(\mathcal{P}_X, \hat{\mathcal{P}}_{y|X})]$ for $p^c_{\mathcal{B}}(X|y)$ and $p^c_{\mathcal{C}}(X|y)$, where:

$$\gamma(\mathcal{P}_X, \hat{\mathcal{P}}_{y|X}) \triangleq \min_{p_i(X) \in ext(\mathcal{P}_X) \cup ext(\hat{\mathcal{P}}_{y|X})} \min_{p_j(X) \in ext(\mathcal{P}^*_X)} ||p_i(X) - p_j(X)||\,, \quad (15)$$



(a) $p^c_{\mathcal{B}}(X|y)$ (b) $p^c_{\mathcal{C}}(X|y)$

Fig. 1. Level plot with intensity according to expected rank conditional on the degree of imprecision of the operands where the operand credal sets have been sampled from the uniform distribution over the degree of of imprecision, i.e., $\mathrm{E}_{u_{\mathcal{I}}(\Omega_{\mathcal{P}_X})}[\mathcal{R}_l(\cdot)|\mathcal{I}(\mathcal{P}_X), \mathcal{I}(\hat{\mathcal{P}}_{y|X})]$ for $p^c_{\mathcal{B}}(X|y)$ and $p^c_{\mathcal{C}}(X|y)$. The degree of imprecision on each axis has been discretized into one hundred bins.

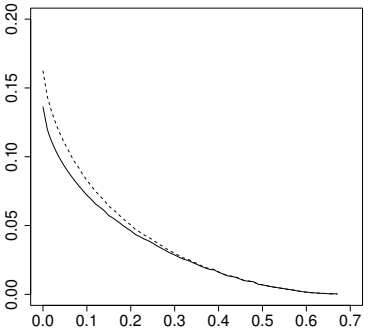


Fig. 2. $\mathrm{E}_{u_{\mathcal{I}(\Omega_{\mathcal{P}_X})}}[s_l(\cdot, p_t(X|y))|\gamma(\mathcal{P}_X, \hat{\mathcal{P}}_{y|X})]$ (y-axis) and $\gamma(\mathcal{P}_X, \hat{\mathcal{P}}_{y|X})$ (x-axis) for $p^c_{\mathcal{B}}(X|y)$ (solid line) and $p^c_{\mathcal{C}}(X|y)$ (dashed line)

where $\mathcal{P}^*_X$ denotes the probability simplex for Ω_X ($\mathcal{P}^*_X \triangleq \{\lambda(0,1)^T + (1-\lambda)(1,0)^T : 0 \leq \lambda \leq 1\}$) and where $||\cdot||$ is the *Euclidean distance*. From the figure we see that in cases where there exists an operand with an extreme point close to the boundary of the simplex, it is clearly beneficial to use the Bayesian method $p^c_{\mathcal{B}}(X|y)$. The reason for this is that $p^c_{\mathcal{B}}(X|y)$ suppresses the effect of such operand on the posterior, since it uses the centroid and such point is closer to the uniform distribution than the extreme point. The credal update operator, on the other hand, performs posterior calculations on the operand's extreme points which then significantly affects the posterior extreme points and thus also the posterior's centroid.

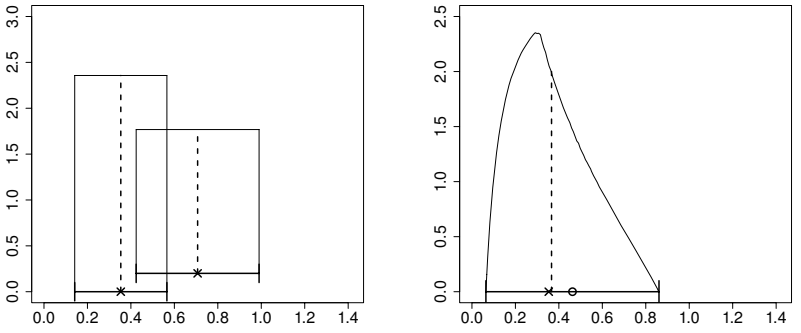
Let us further analyze the result of the experiment by exploring the methods $p^c_{\mathcal{B}}(X|y)$ and $p^c_{\mathcal{C}}(X|y)$ for a single problem instance $\mathcal{P}_X$ and $\hat{\mathcal{P}}_{y|X}$:

$$\begin{aligned} \mathcal{P}_X &= \left\{\lambda(0.1, 0.9)^T + (1-\lambda)(0.4, 0.6)^T : 0 \leq \lambda \leq 1\right\} \\ \hat{\mathcal{P}}_{y|X} &= \left\{\lambda(0.3, 0.7)^T + (1-\lambda)(0.7, 0.3)^T : 0 \leq \lambda \leq 1\right\} . \end{aligned} \tag{16}$$

The operands and the result after updating are seen in Fig. 3 where also an approximation of the *second-order distribution* over the truth has been plotted (remember that we sampled the truth uniformly from the operands). From the figure, it is evident that the second-order distribution over the posterior credal set can be considerably skewed. In such cases the centroid is clearly not a good approximation of the expected value of such distribution, i.e., $\mathrm{E}[p_t(X|y)|\mathcal{P}_X, \hat{\mathcal{P}}_{y|X}]$. We also see that $p^c_{\mathcal{B}}(X|y)$ is a better estimate of the expected value of the second-order distribution over the posterior credal set.

6 Summary and Conclusions

In this paper, we have studied the performance of six different methods with respect to the discrete state estimation problem; three which are Bayesian and



(a) $\mathcal{P}_X$ and $\hat{\mathcal{P}}_{y|X}$ (b) $\mathcal{P}_{X|y}$

Fig. 3. Second-order distribution (approximated with 10^7 samples discretized into 100 bins) over $\mathcal{P}_X$ and $\hat{\mathcal{P}}_{y|X}$ where the x-axis is the probability simplex $\mathcal{P}_X^*$. The dashed line shows the expected value (approximated) with respect to the second-order distribution. The cross shows the selected operand functions and result of $p^c_{\mathcal{B}}(X|y)$ and the circle indicates the result of $p^c_{\mathcal{C}}(X|y)$.

three based on credal set theory. We performed an experiment using expected ranks with respect to accuracy and Brier loss and it was found that Bayesian updating performed on centroids of operand credal sets significantly outperforms the other methods. We have analyzed the result based on degree of imprecision, position of extreme points, and second-order distributions.

In principle, our results suggest that if sources choose to express imprecision for the prior and likelihood and such imprecision can be interpreted as a uniform distribution over the operands then such information can in general be sufficiently summarized by centroids of credal operands. At first such result may appear somewhat provocative, since the credal update operator has been specifically designed to represent and maintain imprecision. However, the operator has not been designed to preserve any information about the second-order distribution.

Even though the credal centroid method is not a good approximation of the second-order distribution, it can still be an interesting choice for an application as a cautious decision policy. Our future work therefore concerns to further explore whether or not using such method can be beneficial with respect to an application.

Acknowledgements. We wish to thank Professor Stefan Arnborg for his feedback during the work on this paper. Computations have been performed with *R* [15]. This work was supported by the Information Fusion Research Program (University of Skövde, Sweden) in partnership with the Swedish Knowledge Foundation under grant 2003/0104 (URL: http://www.infofusion.se).

References

1. Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. John Wiley and Sons, Chichester (2000)
2. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, Boca Raton (1991)
3. Levi, I.: The enterprise of knowledge. The MIT press, Cambridge (1983)
4. Cozman, F.G.: Credal networks. Artificial Intelligence 120, 199–233 (2000)
5. Cozman, F.: Decision Making Based on Convex Sets of Probability Distributions: Quasi-Bayesian Networks and Outdoor Visual Position Estimation. PhD thesis, The Robotics Institute, Carnegie Mellon University (1997)
6. Zaffalon, M., Fagiuoli, E.: Tree-based credal networks for classification. Reliable Computing 9, 487–509 (2003)
7. Corani, G., Zaffalon, M.: Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. Journal of Machine Learning Research 9, 581–621 (2008)
8. Arnborg, S.: Robust Bayesianism: Imprecise and paradoxical reasoning. In: Proceedings of the 7th International Conference on Information fusion, pp. 407–414 (2004)
9. Arnborg, S.: Robust Bayesianism: Relation to evidence theory. Journal of Advances in Information Fusion 1, 63–74 (2006)
10. Karlsson, A., Johansson, R., Andler, S.F.: On the behavior of the robust Bayesian combination operator and the significance of discounting. In: 6th International Symposium on Imprecise Probability: Theories and Applications, pp. 259–268 (2009)
11. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
12. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66, 191–234 (1994)
13. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1–3 (1950)
14. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30 (2006)
15. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009)

On the Complexity of Non-reversible Betting Games on Many-Valued Events

Martina Fedel[1] and Tommaso Flaminio[2]

[1] Department of Mathematics and Computer Science, University of Siena, Pian dei Mantellini 44, 53100 Siena, Italy
fedel@mail.dm.unipi.it
[2] Articial Intelligence Research Institute (IIIA - CSIC), Campus UAB, Bellaterra 08193, Spain
tommaso@iiia.csic.es

Abstract. A bookmaker is said to be *non-reversible* if she does not accept negative stakes on the events. A rational criterion to choose stakes on a book Γ arranged by a non-reversible bookmaker, avoids *bad bets*. A bad bet is a stake δ on an event ϕ, that gives the bettor a strictly better payoff independently of the truth values of the events involved. In this paper we study the computational complexity for the problem of deciding whether a book arranged on many-valued events, admits bad bets, or it does not. In this short paper we show our problem to be NP-complete.

Keywords: Computational complexity, Non-reversible bookmakers, Many-valued events, NP-completeness.

1 Introduction

De Finetti's probability theory relies on two different key concepts. The first one is the assumption that for each event of interest we can find some betting rate that we regard as fair, in the sense that we are willing to accept a bet both on and against the event at that rate. The second one is the coherence criterion of betting odds: Let $\phi_1, \ldots, \phi_n$ be classical events and let $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to [0, 1]$ be an assessment of $\phi_1, \ldots, \phi_n$. Then $\mathbf{a}$ is said to be coherent if and only if there is no system of reversible bets on the events leading to a win independently of the truth of $\phi_1, \ldots, \phi_n$. Formally, $\mathbf{a}$ is coherent if and only if, for every $\mathbf{b} : \{\phi_1, \ldots, \phi_n\} \to \mathbb{R}$, there exists a boolean valuation $\mathbf{v} : \{\phi_1, \ldots, \phi_n\} \to \{0, 1\}$ such that

$$\sum_{i=1}^{n} \mathbf{b}(\phi_i)(\mathbf{a}(\phi_i) - \mathbf{v}(\phi_i)) \geq 0$$

The celebrated de Finetti's theorem states that an assessment $\mathbf{a}$ is coherent if and only if it can be extended to a finitely additive measure on the boolean algebra of formulas. In [13] Paris shows that the problem of checking whether or not a rational-valued assessment $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to \mathbb{Q} \cap [0, 1]$ is coherent is NP-complete.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 90–97, 2010.

A natural generalization of de Finetti's coherence criterion is obtained allowing an infinite-valued interpretation of events $\phi_1, \ldots, \phi_n$, instead of their classical two-valued interpretation.

MV-algebras has been introduced by Chang in [2] to provide an algebraic semantics to the infinite valued Łukasiewicz logic, and in [11] Mundici introduces states on MV-algebras as averaging processes for formulas in Łukasiewicz logic. Moreover, states constitute measures on their associated MV-algebras which generalize the usual probability measures on boolean algebras and they are related to de Finetti's coherence criterion as follows:

Theorem 1 (Mundici [12]). *Let $\phi_1, \ldots, \phi_n$ formulas of Łukasiewicz logic and $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to [0,1]$ an assessment. Then the following are equivalent:*

(i) The assessment $\mathbf{a}$ is coherent in the sense explained above, that is for every $\mathbf{b} : \{\phi_1, \ldots, \phi_n\} \to \mathbb{R}$, there exists a valuation $\mathbf{v} : \{\phi_1, \ldots, \phi_n\} \to [0,1]$ such that $\sum_{i=1}^{n} \mathbf{b}(\phi_i)(\mathbf{a}(\phi_i) - \mathbf{v}(\phi_i)) \geq 0$

(ii) There is a state s on the Lindenbaum algebra of Łukasiewicz logic generated by the propositional variables occurring in $\phi_1, \ldots, \phi_n$, such that $s([\phi_i]) = \mathbf{a}(\phi_i)$ for all $i = 1, \ldots, n$, where $[\phi_i]$ denotes the equivalence class of ϕ_i.

In [12], Mundici shows that the coherence of rational-valued assessments of formulas of infinite-valued Łukasiewicz logic is decidable. As regards to the computational complexity of the problem, in [6] Flaminio and Montagna show that the problem is in PSPACE and in [1] Bova and Flaminio settle the computational complexity issue, showing that the problem is NP-complete.

As explained in [4] and [5], if we consider many-valued events under a condition of ignorance both partial and total, the existence of a fair betting rate for each event seems not reasonable. There are several reasons which suggest the use of non-reversible games in which it is not possible for the bettor to reverse his role with the bookmaker at the same conditions. In this case the non existence of a winning strategy for the bettor is a necessary but not sufficient condition for coherence. Whereas a coherence criterion à la de Finetti becomes the non existence of a bad bet, that is of a bet for which there is an alternative system of bets ensuring to the bettor a strictly better payoff. More formally a rational-valued Łukasiewicz assessment $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to [0,1]$ admits a bad bet if there is a $1 \leq j \leq n$, and a real number $\delta \in [0,1]$ such that for every valuation $\mathbf{v} : \{\phi_1, \ldots, \phi_n\} \to [0,1]$

$$\sum_{i=1}^{n} \mathbf{b}(\phi_i)(\mathbf{a}(\phi_i) - \mathbf{v}(\phi_i)) > \delta(\mathbf{h}(\phi_j) - \mathbf{a}(\phi_j)).$$

and the following theorem holds:

Theorem 2 (Fedel, Montagna, Keimel, Roth [4]). *Let $\phi_1, \ldots, \phi_n$ formulas of Łukasiewicz logic and $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to [0,1]$ an assessment. Then the following are equivalent:*

(i) The assessment $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to [0,1]$ is admissible that is, it does not admit a bad bet.

(ii) There is a $t \leq n$, and a set of states $\{s_1, \ldots, s_t\}$ over the Lindenbaum algebra of Łukasiewicz logic generated by the propositional variables occurring in $\phi_1, \ldots, \phi_k$, such that, for every $i = 1, \ldots, n$,

$$\mathbf{a}(\phi_i) = \max\{s_j([\phi_i]) \mid j = 1, \ldots, t\}.$$

In this paper we will show that the problem of checking whether or not a rational-valued Łukasiewicz assessment on a non-reversible betting game is admissible (we will denote by Luk-Adm the set of admissible Łukasiewicz assessments) is NP-complete. In particular the paper is organized as follows: after this short non-technical introduction, in Section 2 we will present some preliminary notions and results fundamental in order to understand the rest of this paper. The essential technical tool is Theorem 5 that is a sharpening of [1, Lemma 2.7]. In Section 3 firstly we will prove that Luk-Adm is in NP using the fact the feasibility problem of linear systems is decidable in polynomial time, in the size of the binary encoding of the linear system. Then, since the set of 1-satisfiable formulas is NP-complete, we will prove that Luk-Adm is NP-hard.

2 Preliminaries

Let MV be the signature $(\oplus, \neg, \bot, \top)$ of type $(2, 1, 0, 0)$. The set T of formulas over MV is the smallest set containing a countable set of variables $X_1, X_2, \ldots$ and satisfying: $\bot, \top \in T$, and $\varphi \oplus \psi \in T$, $\neg\varphi \in T$, whenever $\varphi, \psi \in T$. We denote by T_k the set of formulas defined over the finite set of variables $X_1, \ldots, X_k$. Further binary operation symbols are definable over the signature MV: $\varphi \to \psi$ is $\neg\varphi \oplus \psi$, $\varphi \leftrightarrow \psi$ is $(\varphi \to \psi) \odot (\psi \to \varphi)$, $\varphi \ominus \psi$ is $\neg(\varphi \to \psi)$, $\varphi \vee \psi$ is $(\varphi \to \psi) \to \psi$, and $\varphi \wedge \psi$ is $\neg(\neg\varphi \vee \neg\psi)$. The system

$$[0,1]_{MV} = ([0,1], \oplus^{[0,1]}, \neg^{[0,1]}, \bot^{[0,1]}, \top^{[0,1]}), \tag{1}$$

where, for every $x, y \in [0,1]$, $x \oplus^{[0,1]} y = \min\{1, x+y\}$, $\neg^{[0,1]} x = 1 - x$, $\bot^{[0,1]} = 0$, and $\top^{[0,1]} = 1$, is called the *standard MV-algebra*, and the variety $\mathbb{MV}$ of MV-algebra is generated, as a quasivariety, by $[0,1]_{MV}$ [2].

Let $k \geq 1$ be a natural number. For every $\varphi \in T_k$, denote by f_φ the function from $[0,1]^k$ into $[0,1]$ so inductively defined: for every $\mathbf{x} = (x_1, \ldots, x_k) \in [0,1]^k$, $\bot(\mathbf{x}) = 0$, $\top(\mathbf{x}) = 1$, $X_i(\mathbf{x}) = x_i$, $(f_\varphi \oplus f_\gamma)(\mathbf{x}) = f_\varphi(\mathbf{x}) \oplus^{[0,1]} f_\gamma(\mathbf{x})$, and $(\neg f_\varphi)(\mathbf{x}) = 1 - f_\varphi(\mathbf{x})$. Let $F(k) = \{f_\varphi \mid \varphi \in T_k\}$. Then the free k-generated MV-algebra is the algebra

$$\mathcal{F}(k) = (F(k), \oplus^F, \neg^F, \bot^F, \top^F), \tag{2}$$

where the operations are the operations of $[0,1]_{MV}$ defined pointwise. McNaughton theorem (cf. [10], [3, Theorem 9.1.5]) states that a function $f \in F(k)$ iff f is continuous, piecewise linear, each piece having integer coefficient. In other words $f \in F(k)$ iff there exist linear polynomial $p_1, \ldots, p_l$ with integer coefficients, such that, for every $\mathbf{x} \in [0,1]^k$, there exists a $j \in \{1, \ldots, l\}$ such that $f(\mathbf{x}) = p_j(\mathbf{x})$.

We will henceforth denote by $\mathcal{F}(\omega)$ the free MV-algebra generated by countably many free generators.

In every MV-algebra A, the order relation $\leq$ can be defined by the following stipulation:

$$x \leq y \text{ iff } x \to y = \top.$$

If the order $\leq$ is linear, then the MV-algebra A is said to be linearly ordered, or an *MV-chain*.

2.1 Coherent Books

Let $\phi_1, \ldots, \phi_n$ be formulas in T_k. A *Łukasiewicz assessment* is a map $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to [0, 1]$. A Łukasiewicz assessment $\mathbf{a}$ is said to be:

- *Rational-valued* if, for all $i = 1, \ldots, n$, $\mathbf{a}(\phi_i) \in \mathbb{Q}$.
- *Coherent for reversible bets* if, for every $\mathbf{b} : \{\phi_1, \ldots, \phi_n\} \to \mathbb{R}$, there exists a homomorphism $\mathbf{h} : \mathcal{F}(k) \to [0, 1]_{MV}$ such that

$$\sum_{i=1}^{n} \mathbf{b}(\phi_i)(\mathbf{a}(\phi_i) - \mathbf{h}(\phi_i)) \geq 0$$

  holds

The notion of *state* (cf. [11]) is key for the investigation of coherent books.

Definition 1. *Let A be an MV-algebra. A state on A is a map $s : A \to [0, 1]$ satisfying:*

(s1) $s(\top) = 1$ *(*normality*),*
*(s2) for all $x, y \in A$, whenever $x \odot y = \bot$, then $s(x \oplus y) = s(x) + s(y)$ (*additivity*).*

Mundici (cf. [12]), Mundici and Kühr (cf. [9]), and Bova and Flaminio (cf. [1]) provided characterizations of coherent books. We recall the result by Bova and Flaminio, being a sharpness of [9, Theorem 3.2].

First we introduce the notation we are going to use. We assume a reasonably compact binary encoding of $\phi \in T$, such that the number $\text{size}(\phi)$ of bits in the encoding of ϕ is bounded above by a polynomial $e_1 : \mathbb{N} \to \mathbb{N}$ of the number $c(\phi)$ of symbols $\odot$, $\to$ occurring in ϕ, that is,

$$\text{size}(\phi) \leq e_1(c(\phi)).$$

We similarly assume that the length in bits of the encoding of a finite set of formulas $\{\phi_1, \ldots, \phi_k\} \subseteq T$, in symbols $\text{size}(\{\phi_1, \ldots, \phi_k\})$ satisfies

$$\text{size}(\{\phi_1, \ldots, \phi_k\}) \leq e_2(\text{size}(\phi_1) + \ldots + \text{size}(\phi_k)),$$

for some polynomial $e_2 : \mathbb{N} \to \mathbb{N}$. Also, letting $\mathbf{a} : \{\phi_1, \ldots, \phi_k\} \to [0, 1]$ be a rational-valued assessment such that $\mathbf{a}(\phi_i) = k_i/z_i$ with k_i and z_i relatively

prime integers for all $i = 1, \ldots, n$, we assume a binary encoding of $\mathbf{a}$ such that the number of bits in the encoding of $\mathbf{a}$, in symbols, size($\mathbf{a}$), satisfies

$$\text{size}(\mathbf{a}) \leq e_3 \left(\text{size}(\{\phi_1, \ldots, \phi_n\}) + k \cdot \log_2 \max\{k_1, \ldots, z_k\}\right),$$

for some polynomial $e_3 \colon \mathbb{N} \to \mathbb{N}$.

Then the characterization theorem reads as follows:

Theorem 3. *Let $\phi_1, \ldots, \phi_n$ be formulas in T_k and let $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to \mathbb{Q} \cap [0, 1]$ be a rational-valued Łukasiewicz assessment. The following are equivalent:*

(i) $\mathbf{a}$ *is coherent.*

(ii) There exists a state s over $\mathcal{F}(k)$ such that for all $i = 1, \ldots, n$

$$s(\phi_i) = \mathbf{a}(\phi_i).$$

(iii) There exists a unary polynomial $p : \mathbb{N} \to \mathbb{N}$ and $m \leq n + 1$ homomorphisms $\mathbf{h}_1, \ldots, \mathbf{h}_m$ from $\mathcal{F}(k)$ to $[0, 1]_{MV}$ satisfying the following. For all $i = 1, \ldots, m$, $\mathbf{h}_i$ ranges over $\{0, 1/d_i, \ldots, (d_i - 1)/d_i, 1\}$ where

$$\log_2 d_i \leq p(\text{size}(\mathbf{a})),$$

and for all $i = 1, \ldots, n$, $\mathbf{a}(\phi_i)$ is a convex combination of $\mathbf{h}_1, \ldots, \mathbf{h}_m$.

2.2 Rationality Criterion for Non-reversible Books

Let $\phi_1, \ldots, \phi_n$ be formulas in T_k. Consider a Łukasiewicz assessment $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to [0, 1]$. We say that $\mathbf{a}$ admits a *bad bet* if there is a $1 \leq j \leq n$, and a real number $\delta \in \mathbb{R}^+$ such that there exists a system of bets $\mathbf{b} : \{\phi_1, \ldots, \phi_n\} \to \mathbb{R}^+$ such that

$$\sum_{i=1}^{n} \mathbf{b}(\phi_i)(\mathbf{h}(\phi_i) - \mathbf{a}(\phi_i)) > \delta(\mathbf{h}(\phi_j) - \mathbf{a}(\phi_j)),$$

independently on the homomorphism $\mathbf{h} : \mathcal{F}(k) \to [0, 1]_{MV}$.

Similarly we say that $\mathbf{a}$ admits a *good bet* if there is a $1 \leq j \leq n$, and a real number $\delta \in \mathbb{R}^+$ such that there exists a system of bets $\mathbf{b} : \{\phi_1, \ldots, \phi_n\} \to \mathbb{R}^+$ such that

$$\sum_{i=1}^{n} \mathbf{b}(\phi_i)(\mathbf{h}(\phi_i) - \mathbf{a}(\phi_i)) < \delta(\mathbf{h}(\phi_j) - \mathbf{a}(\phi_j)),$$

independently on the homomorphism $\mathbf{h} : \mathcal{F}(k) \to [0, 1]_{MV}$.

A Łukasiewicz assessment $\mathbf{a}$ is said to be

- *admissible* iff there is no bad bet based on $\mathbf{a}$.
- *appropriate* iff there is no good bet based on $\mathbf{a}$.

As explained in [4, Remark 7.1] an assessment $\mathbf{a}$ on $\{\phi_1, \ldots, \phi_n\}$ is admissible in a game where only positive bets are allowed, then the assessment $-\mathbf{a}$ is appropriate for non reversible bets.

The following result characterizes admissible and appropriate assessments for non reversible bets, and generalizes [9, Theorem 3.2].

Theorem 4. *Let $\phi_1, \ldots, \phi_n \in T_k$, and let $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to [0,1]$ be a Łukasiewicz assessment. Then:*

(i) **a** *is admissible iff there is a $t \leq n$, and a set of states $\{s_1, \ldots, s_t\}$ over $\mathcal{F}(k)$ such that, for every $i = 1, \ldots, n$, $\mathbf{a}(\phi_i) = \max\{s_j([\phi_i]) \mid j = 1, \ldots, t\}$.*

(ii) **a** *is appropriate for non-reversible bets iff there is a $t \leq n$, and a set of states $\{s_1, \ldots, s_t\}$ over $\mathcal{F}(k)$ such that, for every $i = 1, \ldots, n$, $\mathbf{a}(\phi_i) = \min\{s_j([\phi_i]) \mid j = 1, \ldots, t\}$.*

(iii) **a** *is admissible and appropriate for non-reversible bets iff* **a** *is coherent.*

Now we are ready to prove an analogous of Theorem 3 for the case of non-reversible betting games.

Theorem 5. *Let $\phi_1, \ldots, \phi_n$ be formulas in T_k and $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to [0,1]$ be a Łukasiewicz assessment. The following are equivalent:*

(i) **a** *is admissible.*

(ii) There exists a $t \leq n$ such that, for every $i = 1, \ldots, t$, there exists a unary polynomial $p_i : \mathbb{N} \to \mathbb{N}$, a natural number m_i, and $m_i \leq n+1$ homomorphisms $\mathbf{h}^i_1, \ldots, \mathbf{h}^i_{m_i}$ from $\mathcal{F}(k)$ to $[0,1]_{MV}$ satisfying the following. For all $l = 1, \ldots, m_i$, there are natural numbers $r_{i,l} \in \mathbb{N}$ such that:

1. $\mathrm{range}(\mathbf{h}^i_l) \subseteq \{0, 1/r_{i,l}, \ldots, (r_{i,l}-1)/r_{i,l}, 1\}$
2. $\log_2 r_{i,l} \leq p_i(\mathrm{size}(\mathbf{a}))$,
3. *There are positive real numbers $\lambda^i_1, \ldots, \lambda^i_{m_i}$ such that $\sum_{l=1}^{m_i} \lambda^i_l = 1$, and, for every $j = 1, \ldots, n$,*

$$\mathbf{a}(\phi_j) = \max\left\{ \sum_{l=1}^{m_i} \lambda^i_l \cdot \mathbf{h}^i_l(\phi_j) \mid i = 1, \ldots, t \right\}.$$

Proof. $(i) \Rightarrow (ii)$. Since **a** is admissible, Theorem 4 (i) ensures the existence of $t(\leq n)$ states $s_1, \ldots, s_t$ on the free k-generated MV-algebra $\mathcal{F}(k)$ such that, for every $i = 1, \ldots, n$, $\mathbf{a}(\phi_i) = \max\{s_j([\phi_i]) \mid j = 1, \ldots, t\}$. For every $j = 1, \ldots, t$, the assessment $\mathbf{c} : \phi_i \mapsto s_j([\phi_i])$ is clearly coherent. Therefore, from Theorem 3 $(i) \Rightarrow (iii)$, for every $j = 1, \ldots, t$, there exist a unary polynomial $p_j : \mathbb{N} \to \mathbb{N}$, $m_j \leq n+1$ homomorphisms $\mathbf{h}^j_1, \ldots, \mathbf{h}^j_{m_j}$ such that each $\mathbf{h}^j_l$ ranging on $\{0, 1/r_{j,l}, \ldots, 1\}$, with $r_{j,l} \leq 2^{p_j(\mathrm{size}(\mathbf{a}))}$, and s_j is a convex combination of $\mathbf{h}^j_1, \ldots, \mathbf{h}^j_{m_j}$. Therefore our claim follows.

$(ii) \Rightarrow (i)$. The reader can easily verify that, for every $j = 1, \ldots, t$ the map $s_j : \mathcal{F}(k) \to [0,1]$ defined by $s_j([\psi]) = \sum_{l=1}^{m_j} \lambda^j_l \cdot \mathbf{h}^j_l([\psi])$ is a state on $\mathcal{F}(k)$. Therefore the claim trivially follows from Theorem 4 (i).

A direct inspection on the above proof, together with the characterization Theorem 4(ii) shows that an analogous of Theorem 3 can be easily proved even for the case of assessments that are *appropriate* for non reversible bets.

3 Upper and Lower Bound

We denote by Luk-Adm the set of all (the binary encoding of) Łukasiewicz assessments that are admissible.

In this section we settle the complexity Luk-Adm to be NP-complete. First of all remind that the feasibility problem of linear systems is decidable in polynomial time, in the size of the binary encoding of the linear system (cf. [14]). Therefore Theorem 5 provides a clear NP-algorithm deciding if either a Łukasiewicz assessment $\mathbf{a}$ belongs to Luk-Adm or it does not.

Lemma 1. Luk-Adm *is in NP.*

Proof. Let $\phi_1, \ldots, \phi_n \in T_k$, and let $\mathbf{a} : \{\phi_1, \ldots, \phi_n\} \to [0,1] \cap \mathbb{Q}$ be a rational-valued Łukasiewicz assessment. Following Theorem 5:

(1) The algorithm guesses a natural number $t \leq n$, and for all $i = 1, \ldots, t$, it also guesses a natural number $m_i \leq n+1$. Clearly $n \leq \text{size}(\mathbf{a})$, whence the guesses are performed in polynomial time in size($\mathbf{a}$).

(2) For all $i = 1, \ldots, t$, and for all $l = 1, \ldots, m_i$, the algorithm guesses a natural number $r_{i,l}$ and guesses the values $\mathbf{h}_l^i(X_1), \ldots, \mathbf{h}_l^i(X_k)$ in $\{0, 1/r_{i,l}, \ldots (r_{i,l} - 1)/r_{i,l}, 1\}$. Again $k \leq \text{size}(\mathbf{a})$, then these values still are guesses in polynomial time. Then, for all $i = 1, \ldots, t$, and $l = 1, \ldots, m_i$, it computes the values $\mathbf{h}_l^i(\phi_1), \ldots, \mathbf{h}_l^i(\phi_n)$. [8, Lemma 2] ensures this computation to be performed in polynomial time.

(3) Now the algorithm checks the feasibility of the following linear system in the unknowns $x_1^1, \ldots, x_{m_1}^1, x_1^2, \ldots, x_1^t, \ldots, x_{m_t}^t$:

$$\begin{aligned}
& x_1^1 + \ldots + x_{m_1}^1 && = 1 \\
& \vdots \\
& x_1^t + \ldots + x_{m_t}^t && = 1 \\
& \max\left\{ \sum_{l=1}^{m_i} x_l^i \cdot \mathbf{h}_l^i(\phi_1) \mid i = 1, \ldots, t \right\} && = \mathbf{a}(\phi_1) \\
& \vdots \\
& \max\left\{ \sum_{l=1}^{m_i} x_l^i \cdot \mathbf{h}_l^i(\phi_n) \mid i = 1, \ldots, t \right\} && = \mathbf{a}(\phi_n)
\end{aligned}$$

As we previously recalled, every guess needed to define the above system can be performed in polynomial time in size($\mathbf{a}$). Moreover, checking the system feasibility requires a polynomial time procedure. Therefore our algorithm is in NP, and from Theorem 5 it provides a solution to the above system iff the assessment $\mathbf{a}$ is admissible.

Call Luk-Sat the set of all (the binary encodings of) formulas in T that are satisfied by a valuation in the standard MV-algebra $[0,1]_{MV}$. In [7] Luk-Sat is proved to be NP-complete.

Lemma 2. Luk-Adm *is NP-hard.*

Proof. We provide a logarithmic space reduction from Luk-Sat to Luk-Adm. As a matter of fact, let Ψ be a formula in T_k, and define the assessment $\mathbf{a} : \{\Psi\} \to \{1\}$. Then $\Psi \in$ Luk-Sat iff $\mathbf{a} \in$ Luk-Adm.

($\Rightarrow$) Assume that $\Psi \in$ Luk-Sat. Then there is a homomorphism $\mathbf{h} : \mathcal{F}(k) \to [0,1]_{MV}$ such that $\mathbf{h}([\Psi]) = 1$. Moreover $\mathbf{h}$ is a state, therefore $\mathbf{a}$ is extended by a state, whence is coherent, and therefore admissible.

($\Leftarrow$) Conversely, assume that $\Psi \notin$ Luk-Adm. Therefore, for every homomorphism $\mathbf{h} : \mathcal{F}(k) \to [0,1]_{MV}$, one has $\mathbf{h}([\Psi]) < 1$. Then there is a bad bet for $\mathbf{a}$. In fact, letting $\lambda = 0$, for every homomorphism $\mathbf{h} : \mathcal{F}(k) \to [0,1]_{MV}$, $\lambda(\mathbf{h}([\Psi]) - \mathbf{a}(\Psi)) = 0 > 1 \cdot (\mathbf{h}([\Psi]) - \mathbf{a}(\Psi)) = \mathbf{h}([\Psi]) - 1$. Therefore $\mathbf{a} \notin$ Luk-Adm, and our claim is settled.

Lemma 1 and Lemma 2 give us the main result of this paper:

Theorem 6. Luk-Adm *is NP-complete.*

References

1. Bova, S., Flaminio, T.: The coherence of Łukasiewicz assessments is NP-complete. International Journal of Approximate Reasoning 51, 294–301 (2010)
2. Chang, C.C.: Algebraic Analysis of Many-valued Logics. Trans. Am. Math. Soc. 88, 467–490 (1958)
3. Cignoli, R., D'Ottaviano, I.M.L., Mundici, D.: Algebraic Foundations of Many-valued Reasoning. Kluwer, Dordrecht (2000)
4. Fedel, M., Keimel, K., Montagna, F., Roth, W.: Imprecise probabilities, bets and functional analytic methods in Łukasiewicz logic (submitted)
5. Fedel, M.: Uncertainty, Indeterminacy and Fuzziness: A Probabilistic Approach. In: Hosni, H., Montagna, F. (eds.) Probability, Uncertainty and Rationality, Publications of the Scuola Normale Superiore. CRM series. Springer, Heidelberg (In print)
6. Flaminio, T., Montagna, F.: Models for Many-valued Probabilistic Reasoning. Journal of Logic and Computation, doi:10.1093/logcom/exp013
7. Hájek, P.: Metamathematics of Fuzzy Logic. Kluwer, Dordrecht (1998)
8. Hájek, P.: Complexity of Fuzzy Probabilistic Logics II. Fuzzy Sets and Systems 158(23), 2605–2611 (2007)
9. Kühr, J., Mundici, D.: De Finetti Theorem and Borel states in [0, 1]-valued Algebraic Logic. International Journal of Approximate Reasoning 46(3), 605–616 (2007)
10. McNaughton, R.: A Theorem about Infinite-valued Sentential Logic. J. Symb. Log. 16, 1–13 (1951)
11. Mundici, D.: Averaging the Truth-value in Łukasiewicz Logic. Studia Logica 55(1), 113–127 (1995)
12. Mundici, D.: Bookmaking over Infinite-valued Events. International Journal of Approximate Reasoning 43, 223–240 (2006)
13. Paris, J.B.: The uncertain reasoner's companion, - a mathematical perspective. Cambridge University Press, Cambridge (1994)
14. Schrijver, A.: Theory of Linear and Integer Programming. Wiley, Chichester (1998)

Sequential Decision Processes under Act-State Independence with Arbitrary Choice Functions

Matthias C.M. Troffaes[1], Nathan Huntley[1], and Ricardo Shirota Filho[2]

[1] Durham University, UK
[2] University of São Paulo, Brazil

Abstract. We investigate how sequential decision processes can be solved, under act-state independence. We first identify a reasonable requirement that such solutions ought to satisfy under act-state independence, which we call locality. We then identify a simple necessary and sufficient algebraic condition on choice functions for locality to be satisfied. As an example, we study locality for some choice functions used in imprecise probability theory, and find that marginal extension plays a crucial role.

1 Introduction

The traditional approach to decision making is to model uncertainty via probability, and to model preferences via utility. However, it has been argued that classical (or precise) probability assessments cannot represent the kind of uncertainty usually involved in complex decision making problems. For this reason, more general models have been advocated, such as Dempster-Shafer belief functions [2,9], imprecise probability [15,13], game-theoretic probability [10], info-gap theory [1], and many more. In this paper we simply consider arbitrary choice functions. Choice functions originate from social choice theory [8], and provide an elegant and unifying framework to study most existing decision theories.

The aim of this paper is to investigate how sequential decision processes can be solved, under act-state independence,[1] and using arbitrary choice functions. This builds further on work by Huntley and Troffaes [4] on subtree perfectness in decision trees. Subtree perfectness roughly says that the restriction of a normal form solution to a subtree coincides with the normal form solution of that subtree, and is similar to subgame perfectness [7] introduced in game theory.

We first identify a reasonable requirement that such solutions ought to satisfy under act-state independence, which we call *locality*, and which essentially represents a very strong form of subtree perfectness for particular decision problems: locality means that the global sequential problem can be solved by solving a local problem at each stage, for each stage taking only into account rewards incurred locally at that stage, and any events observed from previous stages. An important special case of a sequential decision problem which ought to satisfy

[1] In this paper we have no ambition for solving the complex problem of act-state dependence under arbitrary choice functions.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 98–107, 2010.

locality is a sequence of unrelated (in the sense of rewards and information) decision problems. The latter example is philosophically particularly important, because it means that the intuitively logical answer—solving each decision problem separately—coincides with the answer where one models the sequential problem in full and solves them jointly in the normal form. Locality implies this not only for unrelated decision problems, but also for slightly more general situations where information and rewards at each stage depend on previous stages.

We identify necessary and sufficient conditions on choice functions for locality to be satisfied. As an example, we study locality for some choice functions employed in imprecise probability theory (Γ-maximin, maximality, E-admissibility, and interval dominance), and find that marginal extension plays a crucial role.

Before we start, consider a classical Markov decision process with discount rate γ, and transition probabilities p^d_{st} and rewards r^d_{st} (in utiles) for going from state s to t under decision d. The optimal expected utility of an $n+1$-stage Markov decision process starting in state s can be recursively calculated:

$$V_0(s) = 0, \qquad V_{n+1}(s) = \max_d \textstyle\sum_t p^d_{st}(r^d_{st} + \gamma V_n(t)). \tag{1}$$

This corresponds to the usual *value iteration* algorithm [11, Sec. 4.4] for finding optimal policies in infinite horizon Markov decision processes, with precisely one policy evaluation step and one policy improvement step at each stage.

Under act-state independence, transition probabilities do not depend on the decisions, and the solution turns out to be extremely simple:

$$V_{n+1}(s) = \left(\max_d \textstyle\sum_t p_{st} r^d_{st}\right) + \gamma \textstyle\sum_t p_{st} V_n(t). \tag{2}$$

In other words, at every stage, the optimal decision d^* can be obtained by solving a simple one-stage problem, namely $\max_d \sum_t p_{st} r^d_{st}$, and the sequential decision problem reduces to a sequence of static decision problems. In essence, this property is what we will call locality. Of course, under act-state dependence, locality is usually violated, since in such case Markov decision processes cannot be solved just locally, and backward induction techniques are required.

As stated before, we will not be concerned with the act-state dependent case, however, we will generalise the above result to the case where

- we use arbitrary choice functions (no probabilities or utilities assumed),
- rewards can depend on the full state history,
- state spaces and decision spaces can depend on the stage.

We clarify the concepts presented by solving a toy problem: an agent sequentially bets on a coin toss, aiming to maximise profit over the full sequence.

Section 2 introduces basic notions. Section 3 provides precise definitions of sequential decision processes and optimal policies. Section 4 defines locality, and presents our main result: a simple necessary and sufficient condition for locality to be satisfied. Section 5 explores four specific choice functions induced by coherent lower previsions. Section 6 concludes. Proofs are long and omitted.

2 Choice Functions on Gambles

Let Ω be the *possibility space*—we only consider finite ones, no probabilities over Ω are assumed, and subsets of Ω are called *events*. Let $\mathcal{R}$ be a set of rewards—no utility function over $\mathcal{R}$ is assumed. A *gamble* is a function $X\colon \Omega \to \mathcal{R}$, and is interpreted as an uncertain reward: if $\omega \in \Omega$ is the true state, $X(\omega)$ is incurred.

When solving a decision problem, it may not be possible to identify a single action the subject considers the best. The subject might, however, still identify some decisions that he would never consider choosing, leaving an *optimal* subset he is willing to choose from. Various criteria will be explored in Section 5.

Commonly, optimal decisions are determined by comparison of gambles. We therefore suppose that the subject has some way of determining an optimal subset of any set of gambles, conditional upon an event A:

Definition 1. *A choice function* opt *maps, for any event $A \neq \emptyset$, each non-empty finite set $\mathcal{X}$ of gambles to a non-empty subset of $\mathcal{X}$: $\emptyset \neq \mathrm{opt}(\mathcal{X}|A) \subseteq \mathcal{X}$.*

3 Sequential Decision Processes

3.1 Definition and Notation

Consider for Ω a Cartesian product of *state spaces*: $\Omega = S_0 \times S_1 \times \cdots \times S_n$, where S_0 is the set of possible states of the system at time 0, and so on. Particular elements of these spaces are denoted by s_0, s_1, …, s_n. We identify any such element s_k also with an event $E_{s_k} = \{(s'_0, \ldots, s'_n)\colon s'_k = s_k\}$. For brevity, we will sometimes write s_k instead of E_{s_k} when no confusion is possible.

The states are observed sequentially, and after each observation, we can take a decision from some set, and receive a reward from a set $\mathcal{R}$. With $h_k = (s_0, \ldots, s_k)$, we can describe the process as:

- observe $s_0 \in S_0$,
- choose $d_1 \in D_1$, observe $s_1 \in S_1$, receive $r_1(s_0 d_1 s_1) = r_1(h_0 d_1 s_1)$,
- choose $d_2 \in D_2$, observe $s_2 \in S_2$, receive $r_2(s_0 s_1 d_2 s_2) = r_2(h_1 d_2 s_2)$,
- …
- choose $d_n \in D_n$, observe $s_n \in S_n$, receive $r_n(h_{n-1} d_n s_n)$.

The total reward is again assumed to be in $\mathcal{R}$. More precisely, we assume an operator $+$ on $\mathcal{R}$ which maps every two elements r and r' of $\mathcal{R}$ to another element $r + r'$ of $\mathcal{R}$. We assume that $+$ has an identity element 0; no further properties of $+$ are assumed. To avoid many brackets, we always let $+$ evaluate from right to left (this simplifies proofs). The total reward is assumed to be:

$$r_1(h_0 d_1 s_1) + r_2(h_1 d_2 s_2) + \cdots + r_n(h_{n-1} d_n s_n).$$

Finally, we assume that the subject has a choice function opt over any finite set $\mathcal{X}$ of gambles on $F_{k+1} = S_{k+1} \times \cdots \times S_n$, given any $h_k = s_0 \ldots s_k$.

This process is a special case of a sequential decision problem, in that:

- chance and decision nodes follow each other consecutively, hence the problem consists of clearly defined *stages*,
- a variable is observed regardless of the history (for instance, you cannot decide to observe a different variable),
- at each stage, a decision incurs a reward, which may depend on the state history, the current decision, and the next state, but not on anything else,
- the final reward is obtained through combining local rewards.

In our toy example, at each stage, the agent must bet on the outcome—hence, heads or tails—and he loses one utile if wrong, but gains one utile if right. So,

$$S_1 = S_2 = \cdots = S_n = \{H, T\}, \qquad D_1 = D_2 = \cdots = D_n = \{d_H, d_T\},$$
$$r_k(h_{k-1} d_k s_k) = \begin{cases} 1 & \text{if } d_k \text{ matches } s_k, \\ -1 & \text{otherwise.} \end{cases}$$

The initial state s_0 is not of relevance in this problem (S_0 can be any singleton).

There might be insufficient information to affirm the coin's fairness. Therefore, we allow the agent to learn about the possible bias of the coin whilst still performing optimally given his existing knowledge—we come back to this later.

3.2 Normal Form Solution

Policies. After $s_0 d_1 s_1 \dots d_{k-1} s_{k-1}$ has already occurred, a *policy* specifies:

- $d_k \in D_k$,
- $d_{k+1}(s_k) \in D_{k+1}$ for all s_k,
- $\dots$
- $d_n(s_k \dots s_{n-1}) \in D_n$ for all $s_k \dots s_{n-1}$.

A policy is also called a *normal form decision*. Denote the set of all policies by:

$$\Pi_k^n = D_k \times D_{k+1}^{S_k} \times \cdots \times D_n^{S_k \times \cdots \times S_{n-1}}.$$

In our toy example, to mention a few, a policy could be 'always bet tails', or 'bet tails if we had more tails than heads in the past, otherwise bet heads'.

Gambles. Each state history $s_0 \dots s_{k-1} = h_{k-1}$ in H_{k-1} and each policy π_k^n in Π_k^n incurs a *gamble* $X_k^n(h_{k-1}, \pi_k^n)$, that is, a mapping from $S_k \times \cdots \times S_n$ to $\mathcal{R}$:

$$\bigoplus_{s_k} E_{s_k} \bigoplus_{s_{k+1}} E_{s_{k+1}} \dots \bigoplus_{s_n} E_{s_n} \Big(r_k(h_{k-1} d_k s_k) + r_{k+1}\big(h_k d_{k+1}(h_k) s_{k+1}\big) + \dots \\ \cdots + r_n\big(h_{n-1} d_n(h_{n-1}) s_n\big). \Big)$$

We abused notation: $d_\ell(h_{\ell-1})$ denotes of course simply $d_\ell(s_k \dots s_{\ell-1})$.

The gamble $X_k^n(h_{k-1}, \pi_k^n)$ describes the reward $X_k^n(h_{k-1}, \pi_k^n)(f_k)$ that we receive for each possible f_k when we follow π_k^n after having observed h_{k-1}.

In our toy example, 'always bet tails' would induce the gamble:

$$X_k^n(h_{k-1}, \pi_k^n)(f_k) = n_T - n_H,$$

where n_T is the number of tails in f_k, and n_H the number of heads in f_k. In a single stage, the policy 'bet heads' has gamble:

$$X_k^k(h_{k-1}, d_H)(s_k) = \begin{cases} 1 & \text{if } s_k = H, \\ -1 & \text{if } s_k = T. \end{cases} \tag{3}$$

Optimal Policies. A policy π_k^n is *optimal* for a given state history h_{k-1} if

$$X_k^n(h_{k-1}, \pi_k^n) \in \operatorname{opt}(\mathcal{X}_k^n(h_{k-1})|h_{k-1}),$$

where $\mathcal{X}_k^n(h_{k-1})$ is the set of gambles of all policies starting from h_{k-1}. The set of all optimal policies, for a given state history h_{k-1}, is:

$$\Pi_k^n(h_{k-1}) = \{\pi_k^n \in \Pi_k^n \colon X_k^n(h_{k-1}, \pi_k^n) \in \operatorname{opt}(\mathcal{X}_k^n(h_{k-1})|h_{k-1})\}.$$

4 Locality

As seen previously, under act-state independence, we can solve an n-stage Markov decision process simply by solving n one-stage ones. We now generalise this idea.

Let $\Pi_k^k(\cdot)$ denote all locally optimal decision functions:

$$\Pi_k^k(\cdot) = \{d_k(\cdot) \in (D_k)^{H_{k-1}} \colon d_k(h_{k-1}) \in \Pi_k^k(h_{k-1})\}.$$

(It may be useful at this point to recall that $\Pi_k^k = D_k$.) More generally,

$$\begin{aligned} \Pi_k^n(\cdot) = \{(d_k(\cdot), d_{k+1}(\cdot), \ldots, d_n(\cdot)) \in (\Pi_k^n)^{H_{k-1}} \colon \\ \big(d_k(h_{k-1}), d_{k+1}(h_{k-1}\cdot), \ldots, d_n(h_{k-1}\cdot)\big) \in \Pi_k^n(h_{k-1})\}, \end{aligned}$$

where we used the identity

$$\begin{aligned} (\Pi_k^n)^{H_{k-1}} &= \left(D_k \times D_{k+1}^{S_k} \times \cdots \times D_n^{S_k \times \cdots \times S_{n-1}}\right)^{H_{k-1}} \\ &= D_k^{H_{k-1}} \times D_{k+1}^{H_{k-1} \times S_k} \times \cdots \times D_n^{H_{k-1} \times S_k \times \cdots \times S_{n-1}} \\ &= D_k^{H_{k-1}} \times D_{k+1}^{H_k} \times \cdots \times D_n^{H_{n-1}}. \end{aligned}$$

Property 1 (Locality). We say that opt satisfies *locality* (on $S_0, \ldots, S_n$) whenever, for each sequential decision process on $S_0, \ldots, S_n$ and each $1 \leq k < n$,

$$\Pi_k^n(\cdot) = \Pi_k^k(\cdot) \times \Pi_{k+1}^{k+1}(\cdot) \times \cdots \times \Pi_n^n(\cdot).$$

With any opt on $S_0, \ldots, S_n$, we can associate the following property:

Property 2 (Sequential Distributivity). For any $1 \leq k < n$, any value h_{k-1} of H_{k-1}, all finite sets of gambles $\mathcal{X}$ on S_k, all finite sets of gambles $\mathcal{Y}(s_k)$ on F_{k+1} (one such set for each $s_k \in S_k$), and all $X \in \mathcal{X}$ and $Y(s_k) \in \mathcal{Y}(s_k)$:

$$X + \bigoplus_{s_k \in S_k} E_{s_k} Y(s_k) \in \text{opt}\left(\mathcal{X} + \bigoplus_{s_k \in S_k} E_{s_k} \mathcal{Y}(s_k) \middle| h_{k-1}\right)$$
$$\iff X \in \text{opt}(\mathcal{X}|h_{k-1}) \text{ and } Y(s_k) \in \text{opt}(\mathcal{Y}(s_k)|h_{k-1}s_k) \text{ for all } s_k.$$

Intuitively, locality clearly implies sequential distributivity, once seen that sequential distributivity essentially amounts to locality for (a particular set of) two-stage problems. It is a purely algebraic property of the optimality operator, and is relatively easy to verify. Our main result (whose proof is omitted due to lack of space) is that the converse implication holds too:

Theorem 1. opt *satisfies locality if & only if it satisfies sequential distributivity.*

So, sequential distributivity yields a necessary and sufficient test for locality.

5 Locality for Choice Functions Induced by Coherent Lower Previsions

Coherent lower previsions generalize the classical theory of probability by bounding probabilities, and are useful when information is scarce or conflicting [13,6].

First, we must suppose that rewards are in utiles, hence $\mathcal{R} = \mathbb{R}$, and that $+$ represents addition. This is merely a practical assumption in order to work with lower previsions, and will be assumed throughout the remainder.

A *full conditional probability* [3] yields a probability $p(A|B)$ for every $A \subseteq \Omega$ and $\emptyset \neq B \subseteq \Omega$,[2] and induces for every gamble X and every $\emptyset \neq B \subseteq \Omega$,

$$E_p(X|B) = \textstyle\sum_{\omega \in B} p(\omega|B) X(\omega).$$

E_p is coherent in the sense of Williams [15, §1.2.1], or Walley [13, §7.1.4].

A *coherent conditional lower (upper) prevision* is then simply the lower (upper) envelope of some set $\mathcal{M}$ of coherent conditional previsions E_p:

$$\underline{P}(X|B) = \inf_{E_p \in \mathcal{M}} E_p(X|B), \qquad \overline{P}(X|B) = \sup_{E_p \in \mathcal{M}} E_p(X|B).$$

There are different ways of obtaining and interpreting lower previsions [6].

Many properties of coherent conditional lower and upper previsions are well known [15]; for instance, $\underline{P}$ completely determines $\overline{P}$.

The next property will reoccur frequently in our study.

Definition 2. *Let the possibility space be $\Omega = S_0 \times \cdots \times S_n$. A coherent lower prevision $\underline{P}$ is then said to* satisfy marginal extension *(with respect to $S_0, \ldots, S_n$) whenever, for all $1 \leq k < n$, all gambles Z on F_k, and all $h_{k-1} \in H_{k-1}$,*

$$\underline{P}(Z|h_{k-1}) = \underline{P}(\underline{P}(Z|h_{k-1}S_k)|h_{k-1}).$$

[2] We cannot simply start from an unconditional probability, and use Kolmogorov's approach [5], because zero probability plays an essential role in our results, and Kolmogorov's approach excludes conditioning on events of probability zero.

In the above definition, $\underline{P}(Z|h_{k-1}S_k)\colon s_k \mapsto \underline{P}(Z|h_{k-1}s_k)$.

The order of the state spaces is relevant for marginal extension. For conditional previsions E_p, marginal extension corresponds to disintegrability [3, p. 90, Eq. (3)], and is always satisfied in our case as we consider finite state spaces only.

5.1 Maximality

Suppose we are given $\underline{P}(\cdot|\cdot)$—by natural (or regular) extension [13, Sec. 8.1], we can assume that $\underline{P}(X|A)$ is defined for all gambles X on Ω and non-empty $A \subseteq \Omega$. Then, a policy $\pi_k^{n*} \in \Pi_k^n$ is optimal, in the sense of maximality, whenever:

$$\overline{P}(X_k^n(h_{k-1}, \pi_k^{n*}) - X_k^n(h_{k-1}, \pi_k^n)|h_{k-1}) \geq 0, \text{ for all policies } \pi_k^n \in \Pi_k^n. \quad (4)$$

Proposition 1. *Maximality satisfies locality on $S_0, \ldots, S_n$, if and only if*

(i) $\underline{P}(E_{s_k}|h_{k-1}) > 0$ for all $1 \leq k < n$, $h_{k-1} \in H_{k-1}$, and $s_k \in S_k$, and
(ii) $\underline{P}$ satisfies marginal extension with respect to $S_0, \ldots, S_n$.

A first observation is that locality provides a behavioural argument for marginal extension: if you violate marginal extension with respect to some sequence of states $S_0, \ldots, S_n$, then you must violate locality for some act-state independent sequential decision problems on $S_0, \ldots, S_n$. Although marginal extension is a convenient assumption to make, for instance due to computational reasons [13, §6.7.5, p. 316], we are not aware of any other behavioural motivation.

Secondly, we note that the condition of strictly positive lower probability can be relaxed through perturbation. Using results from [12], one can show that, locally, you do not need to be concerned about zero lower probabilities, if you are willing to accept an (arbitrarily close) approximate version of maximality.

In our toy example, first, our agent must assess a coherent lower prevision $\underline{P}(\cdot|\cdot)$ reflecting his beliefs about the coin. For instance, he could use the *imprecise Dirichlet model* [14], which states that, for any gamble X on $S_k = \{H, T\}$:

$$\underline{P}(X|h_{k-1}) = \frac{n_H X(H) + n_T X(T) + s \min\{X(H), X(T)\}}{n_H + n_T + s}, \quad (5)$$

where n_H is the number of heads observed in h_{k-1}, n_T is the number of tails in h_{k-1}, and s is a hyper-parameter, usually taken to be 1 or 2. Eq. (5) is called the *predictive lower prevision*. It models a completely vacuous state of knowledge if $n_H = n_T = 0$, and converges to the empirical expectation as $n_H + n_T$ grows.

The predictive lower previsions yield marginals on gambles on S_k conditional on h_{k-1}. The most conservative joint lower prevision that is compatible with these marginals is given by repeatedly applying the marginal extension theorem [13, p. 314–315, §6.7.2]: for any gamble Z on $S_k \times \cdots \times S_n$, define

$$\underline{P}(Z|h_{k-1}) = \underline{P}(\underline{P}(Z|h_{k-1}S_k)|h_{k-1})$$

(where $\underline{P}(Z|h_{k-1}S_k)$ is considered as a gamble on S_k).

Applying maximality to our toy example is now straightforward. By definition of maximality (Eq. (4)), betting on heads is locally maximal if

$$\overline{P}(X_k^k(h_{k-1}, d_H) - X_k^k(h_{k-1}, d_T)|h_{k-1}) \geq 0.$$

By $X_k^k(h_{k-1}, d_H) = -X_k^k(h_{k-1}, d_T)$, and Eqs. (3) and (5), we conclude that, locally, betting on heads is optimal whenever $n_H \geq n_T + s$, and similarly, betting on tails is optimal whenever $n_T \geq n_H + s$. By construction, $\underline{P}(\cdot|\cdot)$ satisfies marginal extension, so applying Proposition 1, we conclude that this is also the global solution (after small perturbation to get rid of any zero lower probabilities).

5.2 E-Admissibility

Given a full conditional expectation $Q(\cdot|\cdot)$, let $\text{opt}_Q(\cdot|A)$ be the choice function corresponding to maximising expected utility conditionally on A, and let $\mathcal{M}(\cdot|A)$ denote the set of full conditional expectations that dominate a given coherent lower prevision $\underline{P}(\cdot|A)$. Then E-admissibility with respect to $\underline{P}(\cdot|A)$ is:

$$\text{opt}(\mathcal{X}|A) = \bigcup_{Q(\cdot|A)\in\mathcal{M}(\cdot|A)} \text{opt}_Q(\mathcal{X}|A).$$

Proposition 2. *E-admissibility satisfies locality on $S_0, \ldots, S_n$ if and only if*

(i) $\underline{P}(E_{s_k}|h_{k-1}) > 0$ for all $1 \leq k < n$, $h_{k-1} \in H_{k-1}$, and $s_k \in S_k$, and
(ii) $\underline{P}$ satisfies marginal extension with respect to $S_0, \ldots, S_n$.

Perhaps surprisingly, the conditions are identical to those for maximality. Regarding the example, note that, in general, for binary choice, E-admissibility is equivalent to maximality. So, in the example, E-admissibility and maximality coincide locally, and therefore, by Proposition 2, also globally.

5.3 Γ-Maximin

Γ-maximin with respect to a coherent lower prevision $\underline{P}$ is the choice function

$$\text{opt}(\mathcal{X}|A) = \arg\max_{X\in\mathcal{X}} \underline{P}(X|A).$$

Proposition 3. *Γ-maximin satisfies locality on $S_0, \ldots, S_n$ if and only if*

(i) $\underline{P}(E_{s_k}|h_{k-1}) > 0$ for all $1 \leq k < n$, $h_{k-1} \in H_{k-1}$, and $s_k \in S_k$,
(ii) $\underline{P}$ satisfies marginal extension with respect to $S_0, \ldots, S_n$, and
(iii) $\underline{P}$ is locally linear in the sense that, for all $1 \leq k < n$ and all gambles X and Y on S_k, $\underline{P}(X + Y|h_{k-1}) = \underline{P}(X|h_{k-1}) + \underline{P}(Y|h_{k-1})$.

These conditions imply full linearity on all gambles on $S_0 \times \cdots \times S_{n-1}$. Interestingly, they do *not* imply linearity on gambles on S_n. Of course, in cases where such strong form of linearity is satisfied, usually full linearity will actually be satisfied. In other words, one cannot really endorse locality for Γ-maximin and at the same time use imprecise probabilities, except in some unusual cases.

However, a locally Γ-maximin policy is always locally maximal, and so if marginal extension holds it is also globally maximal, by Proposition 1. So, using a locally Γ-maximin policy may still be a reasonable choice.

In our example, betting on heads is optimal under local Γ-maximin whenever

$$\underline{P}(X_k^k(h_{k-1}, d_H)|h_{k-1}) \geq \underline{P}(X_k^k(h_{k-1}, d_T)|h_{k-1}).$$

By Eqs. (3) and (5), we obtain $n_H \geq n_T$. So, betting on heads is locally optimal if $n_H \geq n_T$ and similarly on tails when $n_T \geq n_H$. However, $\underline{P}$ does *not* satisfy the linearity condition, so the policy is not necessarily globally Γ-maximin. Still, the local Γ-maximin policy is an interesting alternative for the reasons above.

5.4 Interval Dominance

Interval dominance with respect to a coherent lower prevision $\underline{P}$ is:

$$\mathrm{opt}(\mathcal{X}|A) = \{X \in \mathcal{X} \colon (\forall Y \in \mathcal{X})(\overline{P}(X|A) \geq \underline{P}(Y|A))\}.$$

Proposition 4. *Interval dominance satisfies locality on $S_0, \ldots, S_n$ if & only if*

(i) $\underline{P}(E_{s_k}|h_{k-1}) > 0$ *for all* $1 \leq k < n$, $h_{k-1} \in H_{k-1}$, *and* $s_k \in S_k$,
(ii) $\underline{P}$ *satisfies marginal extension with respect to* $S_0, \ldots, S_n$, *and*
(iii) $\underline{P}(\cdot|h_{k-1})$ *is locally linear, in the sense that for all* $1 \leq k \leq n$, $h_{k-1} \in H_{k-1}$, *and gambles* X *on* S_k, $\underline{P}(X|h_{k-1}) = \overline{P}(X|h_{k-1})$.

Interval dominance requires even stronger conditions than Γ-maximin. Indeed, for interval dominance to satisfy locality, $\underline{P}$ must essentially correspond to a coherent prevision E_p for some full conditional probability p: you cannot be imprecise, and at the same time endorse locality for interval dominance.

6 Conclusion

We identified a reasonable and convenient condition that solutions to a particular class of sequential decision problems ought to satisfy under act-state independence: locality. Essentially, locality means that, under fairly general circumstances, solving a sequential decision problem can be reduced to solving a sequence of static decision problems, each contingent on the state history of the process. Locality seems quite compelling for both practical and philosophical reasons. For example, any sequence of unrelated (in the sense of rewards and information) decision problems clearly ought to be solvable by solving each problem in the sequence independently of the others.

For those theories that can be modelled via choice functions on gambles (which includes many), we identified a necessary and sufficient condition for locality to be satisfied: sequential distributivity. This property provides a test for any decision criterion whether locality is satisfied or not. We applied it to a number of well known choice functions induced by coherent lower previsions: maximality, E-admissibility, Γ-maximin, and interval dominance.

For maximality and E-admissibility, the conditions for locality interestingly coincide, and amount to (i) having strictly positive lower transition probabilities, (ii) satisfying marginal extension. The first condition can be relaxed, if we are willing to adopt an (arbitrarily close) approximate version of maximality. The second condition, marginal extension, is satisfied if lower previsions are specified locally at each stage, conditional on the full history at that stage. This is the natural way of specifying a joint model for a sequential problem, hence, usually, marginal extension will be satisfied. Interestingly, we can also turn the argument around, and interpret locality as a behavioural motivation for marginal extension: you must satisfy marginal extension if you wish to satisfy locality.

For Γ-maximin, locality will usually be violated, unless a precise conditional prevision is specified, in which case it reduces to maximising expected utility.

However, local Γ-maximin solutions will still be globally maximal (i.e. undominated), hence for this reason, Γ-maximin is perhaps still not that unreasonable.

For interval dominance, locality will also usually be violated unless a precise conditional prevision is specified.

References

1. Ben-Haim, Y.: Info-Gap Decision Theory. Academic Press, London (2001)
2. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. The Annals of Mathematical Statistics 38, 325–339 (1967)
3. Dubins, L.E.: Finitely additive conditional probabilities, conglomerability and disintegrations. The Annals of Probability 3(1), 89–99 (1975)
4. Huntley, N., Troffaes, M.C.M.: Subtree perfectness, backward induction, and normal-extensive form equivalence for single agent sequential decision making under arbitrary choice functions (submitted)
5. Kolmogorov, A.N.: Foundations of the Theory of Probability. Chelsea Publishing Company, New York (1950)
6. Miranda, E.: A survey of the theory of coherent lower previsions. International Journal of Approximate Reasoning 48(2), 628–658 (2008)
7. Selten, R.: Reexamination of the perfectness concept for equilibrium points in extensive games. International Journal of Game Theory 4(1), 25–55 (1975)
8. Sen, A.: Social choice theory: A re-examination. Econometrica 45(1), 53–89 (1977)
9. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
10. Shafer, G., Vovk, V.: Probability and Finance: It's Only a Game! Wiley, New York (2001)
11. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
12. Troffaes, M.C.M.: Finite approximations to coherent choice. International Journal of Approximate Reasoning 50(4), 655–665 (2009)
13. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman & Hall, Boca Raton (1991)
14. Walley, P.: Inferences from multinomial data: Learning about a bag of marbles. Journal of the Royal Statistical Society. Series B 58(1), 3–57 (1996)
15. Williams, P.M.: Notes on conditional previsions. Technical report, School of Math. and Phys. Sci., Univ. of Sussex (1975); Reprinted in International Journal of Approximate Reasoning 44(3), 366–383 (2007)

Similarity-Based Equality with Lazy Evaluation

Ginés Moreno

Department of Computing Systems
U. Castilla-La Mancha, 02071, Albacete, Spain
gmoreno@dsi.uclm.es

Abstract. Thanks to the high expressive power and the rule-based nature of declarative languages, their influences are growing in the fields of AI, knowledge representation, and so on. On the other hand, since the notion of "equality" plays a crucial role on such languages, in this paper we focus in the design of a flexible (fuzzy) but efficient (lazy) notion of equality for hybrid declarative languages amalgamating functional-fuzzy-logic features. Here, we show that, by extending at a very low cost the notion of "strict equality" typically used in lazy functional-logic languages (Curry, Toy), and by relaxing it to the more flexible one of similar equality used in fuzzy-logic programming languages (Likelog, Bousi∼Prolog), similarity relations can be successfully treated while mathematical functions are lazily evaluated in a given program. Our method represents a very easy, low-cost way, for fuzzifying lazy functional-logic languages and it can be implemented at a very high abstraction level by simply performing a static pre-process at compilation time which only manipulates the program at a syntactic level (i.e., the underlying operational mechanism based on rewriting/narrowing remains untouched).

Keywords: Declarative Programming, Equality, Similarity, Laziness.

1 Introduction

Due to the essential vagueness of human thinking, the logical treatment of uncertainty has an increasing importance in specification/verification/development software tasks involving Artificial Intelligence, Soft-Computing, etc. Nowadays, a considerable number of logical systems focus in the formalization of vague concepts and approximate reasoning in a rule-based way by making use of theoretical results coming up from the mathematical background of fuzzy logic [18].

As it is well-known, *Logic Programming* [11] has been widely used for problem solving and knowledge representation in the past. Nevertheless, traditional logic programming languages are not able to treat with partial truth. *Fuzzy Logic Programming* is an interesting and still growing research area that agglutinates the efforts for introducing Fuzzy Logic into Logic Programming, in order to provide these traditional languages with techniques or constructs to deal with uncertainty. During the last decades, several fuzzy logic programming systems have been developed, where the classical inference mechanism SLD–Resolution (based on syntactic unification) is replaced with a fuzzy variant which is able to handle

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 108–117, 2010.

partial truth [13,19,9]. This is also the case of Likelog [1] and Bousi~Prolog [7], two fuzzy extensions of Prolog which cope with similarity/proximity relations.

On the other hand, during the last three decades of investigation in the field of the integration of declarative programming paradigms (functional, fuzzy and logic), the scientific community of the area has produced important and advanced contributions related to both theoretical and practical aspects. However, whereas the functional and logic programming styles have been successfully integrated in the past and, as said before, more recently fuzzy logic has also been introduced into the logic programming paradigm, there is not precedent (apart from our preliminary approach presented in [16]), for a total integration of all these frameworks. In this paper, we plan to go a step beyond in this last sense, by proposing the fusion of the different equality models traditionally supported by each one of these declarative paradigms. It is important to take into account that an appropriate notion of equality has a capital importance when designing the repertoire of expressive resources for a particular declarative language. In general, when we use the term "equality" in declarative programming, there are several different meanings depending of the concrete paradigm being considered. A representative (not exhaustive) list of some cases could be:

- **Syntactic equality.** It is the simplest equality model used in the context of classical pure logic programming (Prolog) which refers to syntactic identity. In this sense, two element are considered "equal" if they have exactly the same syntax. For instance, $f(a)$ is equal to $f(a)$ but not to $g(b)$.
- **Strict equality.** When considering lazy languages, both pure functional and integrated functional-logic languages (Haskell and Curry, respectively), this new equality notion is the only applicable one mainly due to the possible presence of non terminating functions. For instance, if the evaluation of $f(a)$ does not finish then we can not say that $f(a)$ is strictly equal to itself. And, on the contrary, two terms with different syntax, such as $g(b)$ and $h(c)$, could be proved equal if they produce the same final value (for example 0) after being evaluated by rewriting or narrowing.
- **Similar equality.** This equality model, typically used in fuzzy logic languages such as Likelog and Bousi~Prolog, is a direct consequence of several attempts for fuzzifying the original notion of syntactic equality. In this case, the idea is to allow the presence of a set of the so called "similarity/proximity equations" between symbols of a given program. So, if we have a program with the equations $eq(a, b) = 0.5$ and $eq(f, g) = 0.3$ then, it could be proved that expressions $f(a)$ and $g(b)$ are similar with a concrete truth degree.

In the present work, we are looking for a sophisticated equality model fussing the two last equality versions above, in order to take into account the intrinsic particularities that laziness and fuzziness introduce into the maximally integrated functional-fuzzy-logic paradigm. The clever idea of our method is to simply add to a given functional-logic program (written in Curry, for instance) a set of rewriting rules defining the new symbol $\approx$:$\approx$ which captures similarities and implements at a very low cost the powerful notion of "strict similar equality".

2 Similarity Relations and Fuzzy Logic Programming

In the last two decades, several fuzzy logic programming languages have been developed where, in essence, the classical SLD resolution principle of Prolog has been replaced by a fuzzy variant of itself, with the aim of dealing with partial truth and reasoning with uncertainty. Most of these languages implement the resolution principle introduced by Lee[9], such as Elf-Prolog [5], Fril [2], F-Prolog [10], Bousi∼Prolog [7][1], Likelog [1] and the multi-adjoint logic approach of [13]. In this work we are interested in fuzzy languages like Likelog and Bousi∼Prolog, which are based in the mathematical notions of similarity and proximity respectively, since we think that they can be easily extended with mathematical functions in a natural way. Moreover, we will show that it is also possible to incorporate an important computational resource coming from the functional world, as it is the case of laziness.

A similarity relation is a mathematical notion able to manipulate alternative instances of a given entity that can be considered equals with concrete truth degrees. Similarity relations are closely related with equivalence relations (and, then, to closure operators) [20]. Let us recall that a T-norm $\wedge$ in $[0,1]$ is a binary operation $\wedge : [0,1] \times [0,1] \rightarrow [0,1]$ associative, commutative, non-decreasing in both the variables, and such that $x \wedge 1 = 1 \wedge x = x$ for any $x \in [0,1]$. Formally, a *similarity relation* $\Re$ on a domain $\mathcal{U}$ is a fuzzy subset $\Re : \mathcal{U} \times \mathcal{U} \rightarrow [0,1]$ of $\mathcal{U} \times \mathcal{U}$ such that, $\forall x, y, z \in \mathcal{U}$, the following properties hold: reflexivity $\Re(x,x) = 1$, symmetry $\Re(x,y) = \Re(y,x)$ and transitivity $\Re(x,z) \geq \Re(x,y) \wedge \Re(y,z)$. It is important to note that this last property is not required when considering *proximity relations*. In order to simplify our developments, as in [19], we assume that $x \wedge y$ is the minimum between the two elements $x, y \in [0,1]$.

A very simple, but effective way, to introduce similarity relations into pure logic programming, generating one of the most promising ways for the integrated paradigm of fuzzy logic programming, consists of modeling them by a set of the so-called *similarity equations* of the form $eq(s1, s2) = \alpha$, with the intended meaning that $s1$ and $s2$ are predicate/function symbols of the same arity with a similarity degree α. As in [16], we assume here that the intended similarity relation $\Re$ associated to a given program $\mathcal{R}$, is induced from the (safe) set of similarity equations of $\mathcal{R}$, verifying that the similarity degree of two symbols s_1 and s_2 is 1 if $s_1 = s_2$ or, otherwise, it is recursively defined as the transitive closure of the similarity equations.

This approach is followed, for instance, in the fuzzy logic languages Likelog [1] and Bousi∼Prolog [7], where a set of usual Prolog clauses are accompanied by a set of similarity equations playing an important role at (fuzzy) unification time. Instead of classical *syntactic unification*, we speak now about *weak unification* [7]. Of course, the set of similarity equations is assumed to be safe in the sense that each equation connects two symbols of the same arity and nature (both predicates or both functions) and the properties of the definition of similarity

[1] As it can be seen in `http://www.inf-cr.uclm.es/www/pjulian/bousi.html`, this language is being developed in our research group, at the U. of Castilla-La Mancha.

relation are not violated, as occurs, for instance, with the wrong set $\{eq(a,b) = 0.5,\ eq(b,a) = 0.9\}$ which, apart for introducing risks of infinite loops when treated computationally, in particular, it does not satisfy the symmetric property.

Example 1. Following [1], if we consider a database of books of different kinds containing the fact "`horror(drakula)`", then the goal "`?-adventurous(Book)`" would not have classical solution in the case that there were no rule in the database unifying with atom "`adventurous(X)`". Nevertheless, it seems reasonable that the user considers the constants "`adventurous`" and "`horror`" *similar* to a certain degree. More precisely, if the user introduces a similarity equation like "`eq(adventurous, horror) = 0.9`" into a Likelog/Bousi~Prolog interpreter, the system would successfully respond with a computed answer incorporating the corresponding truth degree "`0.9`" (i.e, something like the 90 % of credibility) to substitution "`Book` $\mapsto$ `drakula`", as obviously expected.

Before finishing this section, it is noteworthy to take in mind the following points:

- Likelog is oriented to manipulate inductive databases, where no function symbols of arity greater than 0 are allowed. So, only similarities between two predicates or constants (that is, constructor function symbols with no parameters) can be considered. Fortunately, both Bousi~Prolog and our present approach drop out this last limitation by also allowing similarity equations between any pair of constructor function symbols which need not be constants. Moreover, since our current language does not treat with proper predicate symbols, we use the notion of *constraint* which is a quite natural way to model predicates in (pure and/or integrated) functional languages.
- On the other hand, since Likelog and Bousi~Prolog can be seen as fuzzy logic languages extending the resolution principle used in pure logic languages like Prolog, they are "eager" languages, where no lazy computations are allowed (in contrast with the agile evaluation we will see in Section 5). In our approach, instead of mirroring these inference systems, we are much more oriented to the syntax (instead of Horn clauses, we prefer rewrite rules) and operational principles (rewriting/narrowing) of declarative languages with a "functional taste" (Haskell, Curry) in order to cope with laziness plus their combinations with similarity equations.

3 Rewriting, Narrowing and Laziness

The theory of Term Rewriting Systems (TRS) has been largely used in declarative programming to develop pure functional and integrated, functional-logic, languages, such as Haskell and Curry, respectively. A Haskell or a Curry program is no more than a TRS, that is, a set of rewrite rules (instead of a set of clauses, as occurs with logic languages) that can not be distinguished under a syntactic point of view: the differences appear only at the operational level, depending whether rewriting or narrowing is used to execute programs. This section is devoted to explain such concepts.

We consider a *signature* Σ partitioned into a set $\mathcal{C}$ of *constructors* and a set $\mathcal{F}$ of *defined* functions. The set of *constructor terms* (with *variables*) is obtained by using symbols from $\mathcal{C}$ (and a set of variables $\mathcal{X}$). The set of variables occurring in a term t is denoted by $\mathcal{V}ar(t)$. A *pattern* is a term of the form $f(d_1, \ldots, d_n)$ where $f/n \in \mathcal{F}$ and $d_1, \ldots, d_n$ are constructor terms (with variables). A term is *linear* if it does not contain multiple occurrences of one variable. A *position* p in a term t is represented by a sequence of natural numbers (Λ denotes the empty sequence, i.e., the root position). Positions are ordered by the *prefix* ordering: $p \leq q$, if $\exists w \mid p.w = q$. Moreover, $t|_p$ denotes the *subterm* of t at a given position p, and $t[s]_p$ denotes the result of *replacing the subterm* $t|_p$ by the term s. We denote by $\{x_1 \mapsto t_1, \ldots, x_n \mapsto t_n\}$ the *substitution* σ with $\sigma(x_i) = t_i$ for $i \in \{1, \ldots, n\}$ (with $x_i \neq x_j$ if $i \neq j$), and $\sigma(x) = x$ for all other variables x. The application of a substitution σ to a term t is denoted by $\sigma(t)$.

A rewrite rule is an expression of the form $l \to r$ such that $l \notin \mathcal{X}$, and $\mathcal{V}ar(r) \subseteq \mathcal{V}ar(l)$. The terms l and r are called the *left-hand side* (lhs) and the *right-hand side* (rhs) of the rule, respectively. A set of rewrite rules is called a *term rewriting system* (TRS). A TRS $\mathcal{R}$ is left-linear if l is linear for all $l \to r \in \mathcal{R}$. A TRS is *constructor–based* (CB) if each left-hand side is a pattern. In the remainder of this paper, a *program* is a left-linear CB-TRS where the lhs's of two different rewrite rules do not unify. A *rewrite step* is an application of a rewrite rule to a term, i.e., $t \to_{p,R} s$ if there exists a position p in t, a rewrite rule $R = (l \to r)$ and a substitution σ with $t|_p = \sigma(l)$ and $s = t[\sigma(r)]_p$. Rewriting is the operational principle of pure functional languages such as Haskell. However, the operational semantics of modern integrated functional-logic languages, such as Curry, is usually based on *(needed) narrowing*, a combination of variable instantiation and reduction. Formally, $s \leadsto_{p,R,\sigma} t$ is a *narrowing step* if p is a non-variable position in s and $\sigma(s) \to_{p,R} t$.

Despite of the rewriting/narrowing mechanism used to solve goals w.r.t. a given program, we distinguish between "eager (call-by-value)" and "lazy (call-by-name)" evaluation models. Intuitively, when we have an expression with several nested subterms to be evaluated, any eager strategy only should evaluate innermost sub-expressions in each computation step, whereas a lazy strategy should give priority to outermost subterms for being processed. Although a lazy strategy is always more difficult to implement than an eager one, the benefits of laziness have been largely reported in the specialized literature. For instance, in programs defining infinite data structures, computations with guarantees of termination can be only performed in a lazy way.

In the following section, we focus on another important element strongly related with the operational semantics of lazy (pure functional and functional logic) languages, called *Strict Equality*.

4 Strict Equality *versus* Similar Equality

As said in the previous sections, it is usual in functional logic programming to simulate typical (crisp) predicates of pure logic programming by means of

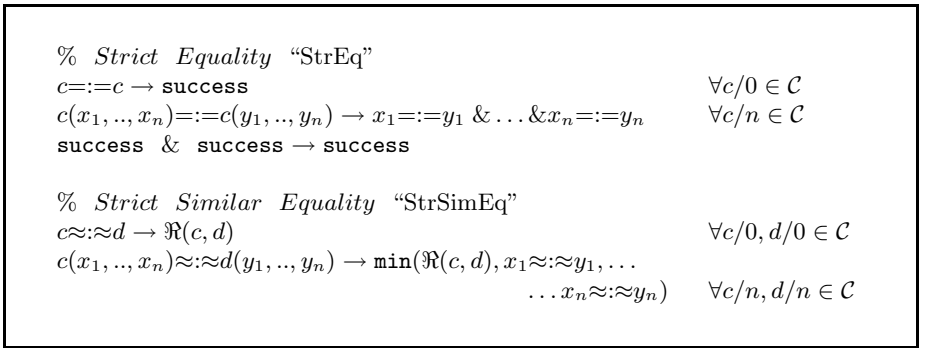


Fig. 1. Rewrite rules defining =:= (strict equality) and ≈:≈ (strict similar equality)

boolean functions. However, a second much more interesting way to face this problem is by using *constraints*. An elementary constraint is an *equational constraint* $\mathtt{e_1} =:= \mathtt{e_2}$ between two expressions (of base type). Then, $\mathtt{e_1} =:= \mathtt{e_2}$ is satisfied if both sides are reducible to a same ground data term. This notion of equality, which is the only sensible notion of equality in the presence of non-terminating functions [17] and also used in (lazy) functional languages, it is also called *strict equality*. As a consequence, if one side is undefined (non-terminating), then the strict equality does not hold (so, it is not reflexive).

Equational constraints should be distinguished from standard boolean functions since constraints are checked for satisfiability. For instance, the equational constraint `X =:= 0` is satisfied if variable `X` is bound to 0. However, the evaluation of `X =:= 0` does not deliver a boolean value `True` or `False`, since the latter value would require a binding for `X` to all values different from `0`. This is sufficient since, similarly to predicates in logic programming, constraints are only activated in conditions of equations which must be checked for satisfiability. Operationally, an equational constraint $\mathtt{e_1} =:= \mathtt{e_2}$ is solved by evaluating $\mathtt{e_1}$ and $\mathtt{e_2}$ to unifiable data terms. The equational constraint could also be solved in an incremental way by an interleaved lazy evaluation of the original expressions and binding variables to constructor terms [12]. Constraints can be also combined into a *conjunction* (which can be interpreted concurrently), written as $\mathtt{c_1}\&\mathtt{c_2}$. This evaluation mechanism can be implemented at a very high abstraction level by assuming that each program implicitly incorporates the standard set of rewrite rules shown in Figure 1, defining the semantics of the primitive "strict equality" relation symbols "=:=" and "&" [4,17].

On the other hand, since =:= represents a natural way to deal with strict equality and constraints simulating "crisp predicates", our next task consists of introducing a new operator, say ≈:≈, for modeling "fuzzy predicates" by means of the new notions of *similar equality* and *f-constraints*. Given an f-constraint $\mathtt{e_1}{\approx:\approx}\mathtt{e_2}$, the goal now is to reduce both expressions $\mathtt{e_1}$ and $\mathtt{e_2}$ to ground values, and then comparing the resulting data terms $\mathtt{v_1}$ and $\mathtt{v_2}$, having into account the similarity relation $\Re$ induced by the set of the similarity equations of the

corresponding program as shown in (see Figure 1). Now, instead of `success`, we are looking for a real number in the interval $[0, 1]$ representing the similarity degree between outputs $\mathtt{v}_1$ and $\mathtt{v}_2$.

Basically, the set of rewrite rules defining "$\approx$:$\approx$" in Figure 1 proceeds as follows. The similarity degree between two constructor symbols of arity 0 is the one returned by the induced similarity relation $\Re$. On the other hand, when comparing two data terms (obtained after reducing the original parameters of a f-constraint) with arguments, it is necessary to recursively compute the similarity degree between the corresponding pairs of arguments of the data terms, together with the similarity relation between the constructor symbols heading each data term.

5 A Running Example

In the following program $\mathcal{P}$, we consider that data terms are built with constants `a`, `b` and `c`, constructor symbols (of arity 1) `r` and `s`, and the typical binary constructor ":" for modeling lists, whereas `f`, `g` and `h` are function symbols defined by the the following three rewrite rules (note that we use capital letters to denote variable symbols) $R_1:\ \mathtt{f(X)} \rightarrow \mathtt{r(X)} : \mathtt{f(X)}$, $R_2:\ \mathtt{g(b)} \rightarrow \mathtt{s(c)}$, and $R_3:\ \mathtt{h(X : Y)} \rightarrow \mathtt{X}$. Intuitively, function `f` generates an infinite list of the form $\mathtt{r(X)} : \mathtt{r(X)} : ...$ (hence introducing the risk of non-termination in a non-lazy, i.e. eager, setting), function `g` produces the data term `s(c)` when it is invoked with argument `b`, and finally, function `h` returns the (first) element heading a given list. Assume now that program $\mathcal{P}$ also contains the following set of similarity equations between constructor symbols of the same arity $E_1:\ \mathtt{eq(a, b)} = 0.8$, $E_2:\ \mathtt{eq(b, c)} = 0.6$ and $E_3:\ \mathtt{eq(r, s)} = 0.5$. In order to generate a similarity relation based on the previous set of similarity equations, we must firstly apply the "transitivization" algorithm of [8], in order to obtain the following similarity relation $\Re$ induced from E_1, E_2 and E_3:

$$\Re = \begin{pmatrix} & \mathtt{a} & \mathtt{b} & \mathtt{c} & \mathtt{r} & \mathtt{s} \\ \mathtt{a} & 1 & 0.8 & 0.6 & 0 & 0 \\ \mathtt{b} & 0.8 & 1 & 0.6 & 0 & 0 \\ \mathtt{c} & 0.6 & 0.6 & 1 & 0 & 0 \\ \mathtt{r} & 0 & 0 & 0 & 1 & 0.5 \\ \mathtt{s} & 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}$$

Inspired by this matrix and following the method explained in the previous section, we can now construct the following set $StrSimEq$ of "strict similar equality rewrite rules" shown in Figure 2.

Now, we can safely replace the original set of similarity equations (i.e., E_1, E_2 and E_3) by the previous thirteen rewrite rules (collecting all the information on similarities among constructor symbols), which are added to the original program $\mathcal{P}$ in order to obtain the extended program $\mathcal{P}^+ = (\mathcal{P} - \{E_1, E_2, E_3\}) \cup StrSimEq = \{R_1, R_2, R_3\} \cup \{R_4, .., R_{16}\} = \{R_1, .., R_{16}\}$. It is important to note that, in contrast to $\mathcal{P}$, no similarity equations appear in $\mathcal{P}^+$, since it only contains rewrite rules which can be directly used by rewriting/narrowing to solve goals.

R_4 : a≈:≈a → 1 R_5 : a≈:≈b → 0.8 R_6 : a≈:≈c → 0.6
R_7 : b≈:≈a → 0.8 R_8 : b≈:≈b → 1 R_9 : b≈:≈c → 0.6
R_{10} : c≈:≈a → 0.6 R_{11} : c≈:≈b → 0.6 R_{12} : c≈:≈c → 1

R_{13} : r(X)≈:≈r(Y) → min(1, X≈:≈Y)
R_{14} : r(X)≈:≈s(Y) → min(0.5, X≈:≈Y)
R_{15} : s(X)≈:≈r(Y) → min(0.5, X≈:≈Y)
R_{16} : s(X)≈:≈s(Y) → min(1, X≈:≈Y)

Fig. 2. Rewrite rules for "StrSimEq" induced from $\Re$

h(f(a)) ≈:≈ g(X)
⇝ 1.1, R_1 h(r(a) : f(a)) ≈:≈ g(X)
⇝ 1, R_3 r(a) ≈:≈ g(X)
⇝ 2, R_2 {X↦b} r(a) ≈:≈ s(c)
⇝ Λ, R_{14} min(0.5, a ≈:≈ c)
⇝ 2, R_6 min(0.5, 0.6)
⇝ 0.5

Fig. 3. Derivation for goal h(f(a))≈:≈g(X) in the extended program $\mathcal{P}^+$

Let us explain in detail our method following the derivation of Figure 3:

• **Step 1.** This is properly a *pure functional* evaluation step, since it is based on *rewriting*. Note that, by using rule R_1, it simply exploits the unique redex of the original goal, that is, the (underlined) subterm located at position 1.1 (observe that f(a) is the first argument of h(f(a)), which, once again, is the first argument of the whole expression h(f(a))≈:≈g(X)).

• **Step 2.** This step is also based on rewriting (by using this time rule R_3), but in contrast with the previous one, here we can clearly observe the *lazy behaviour* of our operational principle. Since the exploited redex (subterm h(r(a) : f(a)) at position 1), also contains an inner redex (f(a)), an eager strategy would have (infinitely) exploited such innermost redex, thus producing a non-ending evaluation sequence with rule R_1 (remember that this rewrite rule defines the generation of an infinite list!). Fortunately, our lazy strategy avoids this risk of loop in a very easy and clever way: by simply giving priority to *outermost* redexes when reducing a term.

• **Step 3.** It is the first time that we perform a proper *narrowing* step (which is not only based on rewriting, but also in *variable instantiation*) in our example, thus introducing the *logical* dimension of the derivation. In fact, before reducing

subterm g(X) at position 2 of the original goal, narrowing firstly generates the binding X ↦ b, and then, after applying it to that subterm, it is able to reduce the resulting redex g(b) with rule R_2.

• **Step 4.** If in the previous steps we have evidenced the functional, lazy, and logic features, respectively, of our approach, in what follows, we focus our attention on its *fuzzy* properties. The fourth step, exploits the *similarity* between constructors r and s by simply performing a rewriting step on the whole goal (position Λ) with the synthesized rule R_{14}. Remember that this "strict similar equality rewrite rule" is not present in the original program $\mathcal{P}$, but only in the extended program $\mathcal{P}^+$. This rule collects the similarity degree 0.5 between r and s expressed in the original similarity equation E_3 of $\mathcal{P}$. Observe also in the resulting expression the recursive call to ≈:≈ with the actual parameters of subterms r(a) and s(c), which will be considered in the following step.

• **Step 5.** Once again, we are describing a new *fuzzy* step, which exploits now the similarity degree 0.6 between constants a and c by means of the new rewrite rule $R_6 \in \mathcal{P}^+$. It is important to note that, although there is no similarity equation in the original program $\mathcal{P}$ directly relating both constants, the corresponding similarity degree can be obtained by transitivity from E_1 and E_2 after applying the method of [8].

• **Step 6.** The last derivation step simply evaluates the min T-norm (with concrete similarity degrees), which can be considered a primitive operator of the language. So, the computed answer for the original goal can be interpreted as: "expressions h(f(a)) and g(X) can be reduced to similar output values with degree 0.5 when the input value for X is b".

6 Conclusions and Further Research

In this paper we have proposed a novel, powerful notion of equality based on similarity, which is especially well suited for hybrid declarative languages promoting a functional-fuzzy-logic programming style. Our approach prefers a compilation process instead of an interpretation way for managing rules at a very low computational cost and in a complete transparent way for the final user. Moreover, it captures similarities among constructor symbols of any arity when comparing data terms obtained as the output of more complex expressions, also enjoying an efficient execution via lazy evaluation. Moreover, it can be implemented at a very high abstraction level, without manipulating the semantics and the operational principle of the original language, by simply performing two purely syntactic pre-processing stages (from similarity equations to a complete similarity relation, and then to a set of rewriting rules defining symbol ≈:≈).

For the near future we plan to combine this refined notion of equality with the extension of *needed narrowing* proposed in [16], as well as to adapt to the new functional-fuzzy-logic language, the implementation, transformation and optimization techniques developed in our research group regarding the fuzzy logic programming field (see same representative works in [14,6,15,3,7,8]).

Acknowledgments. This work was supported by the EU (FEDER), and the Spanish Science and Innovation Ministry (MICIN) under grant TIN 2007-65749 and by the Castilla-La Mancha Administration under grant PII1I09-0117-4481.

References

1. Arcelli, F., Formato, F.: Likelog: A logic programming language for flexible data retrieval. In: Proc. of the ACM Symposium on Applied Computing (SAC 1999). Artificial Intelligence and Computational Logic, pp. 260–267. ACM, New York (1999)
2. Baldwin, J.F., Martin, T.P., Pilsworth, B.W.: Fril- Fuzzy and Evidential Reasoning in Artificial Intelligence. John Wiley & Sons, Inc., Chichester (1995)
3. Guerrero, J.A., Moreno, G.: Optimizing fuzzy logic programs by unfolding, aggregation and folding. ENTCS (Elsevier Science) 219, 19–34 (2008)
4. Hanus, M. (ed.): Curry: An Integrated Functional Logic Language (2003), http://www.informatik.uni-kiel.de/~mh/curry
5. Ishizuka, M., Kanai, N.: Prolog-ELF Incorporating Fuzzy Logic. In: Joshi, A.K. (ed.) Proc. of IJCAI 1985, pp. 701–703. Morgan Kaufmann, San Francisco (1985)
6. Julián, P., Moreno, G., Penabad, J.: On Fuzzy Unfolding. A Multi-Adjoint Approach. Fuzzy Sets and Systems 154, 16–33 (2005)
7. Julián, P., Rubio, C., Gallardo, J.: Bousi∼prolog: a prolog extension language for flexible query answering. ENTCS (Elsevier Science) 248, 131–147 (2009)
8. Julián-Iranzo, P.: A procedure for the construction of a similarity relation. In: Verdegay, J.L., Magdalena, L., Ojeda-Aciego, M. (eds.) Proc. of the 12th Int. Conf. on Information Processing and Management of Uncertainty in Knwoledge-based Systems, IPMU 2008, U. Málaga, pp. 489–496 (2008)
9. Lee, R.C.T.: Fuzzy Logic and the Resolution Principle. Journal of the ACM 19(1), 119–129 (1972)
10. Li, D., Liu, D.: A fuzzy Prolog database system. John Wiley & Sons, Inc., Chichester (1990)
11. Lloyd, J.W.: Foundations of Logic Programming. Springer, Berlin (1987)
12. Loogen, R., López-Fraguas, F., Rodríguez-Artalejo, M.: A Demand Driven Computation Strategy for Lazy Narrowing. In: Penjam, J., Bruynooghe, M. (eds.) PLILP 1993. LNCS, vol. 714, pp. 184–200. Springer, Heidelberg (1993)
13. Medina, J., Ojeda-Aciego, M., Vojtáš, P.: Similarity-based unification: a multi-adjoint approach. Fuzzy Sets and Systems 146(1), 43–62 (2004)
14. Morcillo, P.J., Moreno, G.: Programming with Fuzzy Logic Rules by using the FLOPER Tool. In: Bassiliades, N., Governatori, G., Paschke, A. (eds.) RuleML 2008. LNCS, vol. 5321, pp. 119–126. Springer, Heidelberg (2008)
15. Moreno, G.: Building a Fuzzy Transformation System. In: Wiedermann, J., Tel, G., Pokorný, J., Bieliková, M., Štuller, J. (eds.) SOFSEM 2006. LNCS, vol. 3831, pp. 409–418. Springer, Heidelberg (2006)
16. Moreno, G., Pascual, V.: A hybrid programming scheme combining fuzzy-logic and functional-logic resources. Fuzzy Sets and Systems 160(10), 1402–1419 (2009)
17. Moreno-Navarro, J.J., Rodríguez-Artalejo, M.: Logic Programming with Functions and Predicates: The language Babel. Journal of Logic Programming 12(3), 191–224 (1992)
18. Nguyen, H.T., Walker, E.A.: A First Course in Fuzzy Logic. Chapman & Hall/CRC, Boca Ratón (2000)
19. Sessa, M.I.: Approximate reasoning by similarity-based SLD resolution. Fuzzy Sets and Systems 275, 389–426 (2002)
20. Zadeh, L.A.: Similarity relations and fuzzy orderings. Informatica Science 3, 177–200 (1971)

Progressive Reasoning for Complex Dialogues among Agents

Josep Puyol-Gruart and Mariela Morveli-Espinoza

Artificial Intelligence Research Institute (IIIA)
Spanish Scientific Research Council (CSIC)
Campus UAB, 08193 Bellaterra, Spain

Abstract. In this paper we propose an architecture for deliberative agents based on progressive reasoning. When an agent receives a query, it tries to satisfy it by building an answer based on its current knowledge. Depending on the available time or the urgency of the requirement the agent can produce answers with different levels of quality. Agents could build progressively their answers with the information they receive from perception or during the dialogue with other agents. We assume that in the real world normally it is better to receive an answer with poor quality than no answer. The answer can be good enough for the receiver or the receiver can spend more time to wait for a better answer. Autonomy implies taking the best decision with the available information, avoiding blocking situations and no action.

1 Introduction

Dean and Boddy first used the term *anytime algorithm* in the late 1980's [2]. The main characteristic of these algorithms is that the quality of its results can be measured and that it improves gradually as computation time increases. This kind of algorithms are normally related to real time, where the time granularity is thinner than the long time needed to calculate a complete solution. They are able to communicate the best result obtained when interrupted or they can establish a compromise to deliver it in a given time. In the context of logics and knowledge-based systems some authors talk about *progressive (or anytime) reasoning* or *deduction* [5,4].

In multiagent systems, agents have particular goals. The conversations among deliberative agents aim to obtain information in order to produce solutions to those goals. In [3] we described how these conversational agents could be modeled.

Consider a rule-based agent. Classical inference mechanisms could take a long time to generate definitive results, depending on the availability of the external information provided by other agents. The model of specialization reasoning that will be described below always can generate—more or less useful—partial results in a much shorter period of time. It is very important to notice that rules are weighted with intervals of truth-values and that we can use negation in the facts and in the conditions and conclusions of rules. Then we can talk about

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 118–127, 2010.

different certainty levels and precision. In our language [6] the value of a fact a is an interval of truth-values $[\alpha, \beta]$. Rules concluding a are responsible of α (the minimum of the interval) and rules concluding $\neg a$ of β (the maximum).

We can introduce *quality* and precision measures. We will consider values are *provisional* when it is possible to improve their precision later using more information. Otherwise they are *definitive*. More certain rules produce more precision for the conclusion. Provisional values for facts are those less precise that can be used also to produce provisional deduction and so provisional values for other facts.

Another important issue is time. It may be reasonable to think in different strategies of specialization using provisional values, i.e. when a concrete timeout has been reached or when we need a value, we can use a less precise but useful result. The passing of time gives an opportunity to increase the accuracy, and then the agent's goals can persist until it is not possible to obtain more precise values.

In this paper we will introduce how reasoning based on specialization of rule-based systems can be the central mechanism to deliberate and also to produce *reasonable* dialogues among conversational agents [1,3,7]. Agents are anytime reasoners and can produce answers with different levels of quality: containing the best, a provisional or a conditional answer. We assume that in the real world normally it is better to receive an answer with poor quality than no answer. The answer can be good enough for the receiver or the receiver can spend more time waiting for a better answer.

In Section 2 we formally describe the deduction by specialization. Section 3 is devoted to quality measures and the impact of specialization over quality and precision. We present the description of the agent and its pragmatics in Section 4. In Section 5 we describe a very simple example of communication among agents. Finally, some conclusions and future work are developed in Section 6.

2 Deduction

The main component of the mental state of agents [8] is the knowledge base containing beliefs (facts) and knowledge (rules) for deliberation. In our model, both facts and rules are weighted with intervals of truth-values.

Specialization [6] can be considered as an anytime algorithm because it allows to obtain information before the completion of the inference process. It can be considered also a mechanism for progressive reasoning because it is a technique that successively refines a solution while making available intermediate solutions. In the following we introduce briefly a simplified version of the language and inference mechanism:

Language. $\mathcal{L} = \langle A, \Sigma, \mathcal{S} \rangle$ is defined by:

- Truth-values $A \in [0, 1]$ where 1 and 0 are the booleans *True* and *False*. $Int(A) = \{[i, j] \mid i, j \in A, 0 \leq i \leq j \leq 1\}$ are intervals of A.
- Σ is a set of propositional variables (atoms or facts).

– Sentences S composed by: literals (a, V), $(\neg a, V)$, with $a \in \Sigma$ and $V \in Int(A)$ and rules of the form $(p_1 \wedge p_2 \wedge \cdots \wedge p_n \rightarrow q, [i, 1])$, where $i \in A$, p_i and q are literals, and $\forall i, j(p_i \neq p_j, p_i \neq \neg p_j, q \neq p_j, q \neq \neg p_j)$

Inference Rules. We will use the following inference rules, where T is a T-norm:

– Not-introduction: from $(a, [i, j])$ infer $(\neg a, [1 - j, 1 - i])$
– Not-elimination: from $(\neg a, [i, j])$ infer $(a, [1 - j, 1 - i])$
– Parallel composition: from (a, V_1) and (a, V_2) infer $(a, V_1 \cap V_2)$
– Specialization: from $(p_i, [i, j])$ and $(p_1 \wedge \cdots \wedge p_n \rightarrow q, [k, 1])$ infer $(p_1 \wedge \cdots \wedge p_{i-1} \wedge p_{i+1} \wedge \cdots \wedge p_n \rightarrow q, [T(i, k), 1])$

Consider $R_q = R_q^+ \cup R_q^-$ the set of rules deducing the fact q. We can distinguish between the positive rules R_q^+ deducing q and the negative rules R_q^- deducing $\neg q$. Positive rules contribute to the minimum of the interval (positive evidences) and negative ones to the maximum (negative evidences).

The *specialization rule* above is the core of the progressive reasoning algorithm. When a rule is specialized it produces a new rule with less conditions and a new updated value. When a rule is totally specialized (there are no more conditions) it produces a value for the literal of the conclusion.

Given a rule $(P \rightarrow q, [k, 1]) \in R_q^+$, the most precise value for that literal will be $[k, 1]$ when P is true, because of the specialization rule and the T-norm property $T(1, a) = a$. Similarly, given $(P \rightarrow \neg q, [k, 1]) \in R_q^-$, the most precise value for that literal will be $[0, 1 - k]$.

A fact q is initially *unknown*, that is, its value is the most imprecise interval $[0, 1]$. Using the *parallel composition rule* and the values obtained from totally—positive and negative—specialized rules we will obtain a more precise interval for q, or a contradiction when the intersection is empty[1]. Given a set of r rules R_q^+ with truth-values $\{[\alpha_1, 1], \ldots, [\alpha_r, 1]\}$ the most precise interval will be $[max_{i=1}^{r}(\alpha_i), 1]$. Given a set of s rules R_q^- with truth-values $\{[\beta_1, 1], \ldots, [\beta_s, 1]\}$ the most precise interval will be $[0, 1 - max_{i=1}^{s}(\beta_i)]$. Finally we can say that the expected most precise interval for q will be $[max_{i=1}^{r}(\alpha_i), 1 - max_{i=1}^{s}(\beta_i)]$. We have to take into account that each specialization step produces a new knowledge base and then the expected most precise interval will be changed.

The new rules are provisional if they are deduced with provisional information, otherwise they are definitive. Facts are definitive if they are deduced with definitive information and there are no more rules that can improve its value. In this case rules can be deleted.

3 Quality Measures

Quality measures and their properties are important for anytime algorithms [9]. It has to be (i) Measurable and recognizable: the quality of an approximate

[1] A contradiction detected during the deduction of q means that the knowledge base is not useful to deduce q in that context.

result has to be determined precisely and easily at run time, (ii) Monotonic: the quality of the result is a non-decreasing function of time and input quality, and (iii) Consistent: the quality of the result is correlated with time and input quality.

Quality is evaluated based on a three-dimensional criterion that measures the level of certainty, precision and completeness of a given value, an interval of truth-values. The quality is determined based on the following characteristics:

Certainty: In an approximate reasoning context we want to know the degree of truth or falsity of propositions. Then, given a set of knowledge deducing a fact we are interested in using those relations that provides values close to *true* or *false*. We assume a uniform distribution of the certainty in the interval, the mean is then the expected value of the interval and it can be representative of its certainty:

$$C[i,j] = \frac{i+j}{2}$$

Precision: Values of facts are intervals. The most precise interval is when the difference between the maximum and the minimum is 0, and the least precise is when that difference is 1, that is, the only case $[0,1]$, or *unknown*.

$$P[i,j] = 1 - (j-i)$$

Completeness: To determine the value of a fact we need to know the values of other related facts contained in the rules that deduce that fact. Given two facts, with the same level of certainty and precision, we will consider of more quality that with less number of dependencies that could improve the result.

We have to distinguish between the quality of a certainty value in itself and that related to the quality given by a concrete *KB*. For instance it is obvious that a fact with value 1 is better than 0.8, but if the *KB* deducing that fact can not produce in any circumstance a better value than 0.8 then we have to consider 0.8 is the best quality value for that fact.

3.1 Absolute and Relative Quality

Precision and certainty are directly related because good precision is interesting only when the value of the fact is close to true or false. We can use the expression $f(p,c) = p \cdot |(2c-1)|$ to calculate a quality measure between 0 and 1. The first term p corresponds to the precision of the interval, better when more close to 1. The second term c corresponds to the value represented by the middle point of the interval, better when more close to 0 or 1, that is, true or false. It is a symmetric function with respect to the plane $i + j = 1$. The absolute quality is:

$$Q_a[i,j] = |P([i,j]) \cdot (2C([i,j]) - 1)| = \left|i^2 - (1-j)^2\right|$$

The *KB* of an agent determines the quality degree and the precision that can be obtained for a given fact with all the available information. We can consider

a KB is *good* deducing a fact f when it can deduce that fact with high precision and with values close to true or false. This depends on the designer of the KB. Finally the result for f will be a combination of the KB and also of the values of facts used to make the deduction.

We have to consider relative measures with respect to the maximum quality level that is possible to obtain with the current knowledge. In the next Section we will see how this value change with the time, when new information is known by the agent. After each execution cycle—to include the new information—quality of the results can be evaluated. We can know in each moment which is the most precise value the agent A can obtain for the fact a, $P_m^A(a)$. We will use precision to evaluate the relative quality.

$$Q_r^A(a, [\alpha, \beta]) = \frac{P[\alpha, \beta]}{P_m^A(a)}$$

If all the facts used to deduce the goal would have a definitive value then the completeness will be of 100%. If all those facts would have values—true of false with the maximum precision—such that the premises of rules are true, then we will obtain the maximum quality degree.

3.2 Specialization and Quality

The main goal of the inference engine is to find the more precise values for the facts. The nature of the inference rules in Section 2 always produce more precise values. This is not so obvious with the quality of facts. For instance, an agent can produce a provisional value quality $Q_a[0.4, 1] = 0.16$. Consider that a new information arrives and $Q_a([0.4, 1] \cap [0, 0.6]) = Q_a[0.4, 0.6] = 0$. This kind of behavior only depends on the design of the KB.

Consider a provisional result $[\alpha, \beta]$. The behavior of the specialization with respect to the quality index produces the following properties:

Positive rules. They produce results in the form $[\gamma, 1]$, with the middle point $C[\gamma, 1] \geq 0.5$. We have to consider the following cases:

- When $C[\alpha, \beta] \geq 0.5$, $\forall \gamma, Q_a([\alpha, \beta] \cap [\gamma, 1]) \geq Q_a[\alpha, \beta]$
- When $C[\alpha, \beta] \leq 0.5$, and $\beta \geq 0.5$, $Q_a([\alpha, \beta] \cap [\gamma, 1]) \geq Q_a[\alpha, \beta]$ only for $\sqrt{2(1-\beta)^2 - \alpha^2} \leq \gamma \leq \beta$
- Otherwise $Q_a([\alpha, \beta] \cap [\gamma, 1]) < Q_a[\alpha, \beta]$

Negative rules. They produce results in the form $[0, \delta]$, with the middle point $C[0, \delta] \leq 0.5$. We have to consider the following cases:

- When $C[\alpha, \beta] \leq 0.5$, $\forall \delta, Q_a([\alpha, \beta] \cap [0, \delta]) \geq Q_a[\alpha, \beta]$
- When $C[\alpha, \beta] \geq 0.5$ and $\alpha \leq 0.5$, $Q_a([\alpha, \beta] \cap [0, \delta]) \geq Q_a[\alpha, \beta]$ only for $\alpha \leq \delta \leq 1 - \sqrt{2\alpha^2 - (1-\beta)^2}$
- Otherwise $Q_a([\alpha, \beta] \cap [0, \delta]) < Q_a[\alpha, \beta]$

These results are obvious in the sense that an interval with a middle point greater that 0.5 is reinforced by positive rules producing a more precise result more close to 1. The same occurs with intervals close to false and negative rules. In the other cases when we combine two intervals with middle points in opposite parts—greater and less than 0.5—it is necessary to compensate with the precision the circumstance that the middle point of the resulting value now is farther from true or false.

At the meta-level we will be interested in exploring the best rules first to avoid unnecessary communication. It seems reasonable to try first the rules that can contribute increasing the absolute quality of facts. Notice that the values of rules change in each specialization step and that it can not be guaranteed a monotonic behavior of absolute quality.

It is easy to see that precision of facts is a monotonic function with respect to the time, new information will produce more precise facts by firing rules and applying parallel composition with old values of facts

$$P([\alpha, \beta] \cap [\gamma, \delta]) \geq max\{P[\alpha, \beta], P[\gamma, \delta]\}$$

Taking into account rules we can notice that their precision decreases with new information because of the specialization rule and the well-known property of t-norms: $T(a, b) \leq min\{a, b\}$

$$\forall \alpha \leq k, P([\alpha, 1]) \leq P([k, 1])$$

We can conclude that new information will produce less or equal precision for the set of rules and equal or more precision for the set of facts. This implies that we can use provisional values of facts to deduce more provisional values. Relative quality is the external quality of an agent that deals with the more important property of an anytime algorithm, monotonicity.

4 Deliberative Agents

The model of reasoning described above could take a long time to generate definitive results. This is not a consequence of the complexity of the deductive process. We consider that specialization time is irrelevant, we have to look at other things like communication time, availability of agents, collaborative behavior, etc. Agents have a deadline to answer a question. When an agent accepts a query, if necessary, it starts by asking other agents for information. But it cannot be waiting forever. When it is not possible to obtain a definitive value for a query and the deadline has been reached, it answers with less precision. Answers can be the best one, a provisional one because it can be improved later, or a conditional answer because the agent ignores some information.

Agents mainly contain facts, rules and goals. These goals can represent commitments with other agents or self-commitments. Agents proactively try to satisfy these goals by asking other agents and making deduction. When an agent receives a question, it assume a new goal and a new commitment of answering in a giving time.

Consider $\mathcal{A}_m = \{A_1, \ldots, A_m\}$ a multi-agent system with m agents. A deliberative **agent** $A_i = \langle KB_i, G_i, I_i, O_i, t_i \rangle$ is a tuple where: KB_i is the knowledge base of agent A_i, literals and rules; G_i is the set of goals of agent A_i, a tuple $\langle x, A_j, t_b \rangle$, where $x \in \Sigma$, $A_j \in \mathcal{A}_m$ and t_b is the remaining time for deadline; I_i is the input interface of agent A_i, the set of external facts that can be obtained by querying other agents, tuples $\langle x, A_j \rangle$, where $x \in \Sigma$, $A_j \in \mathcal{A}_m$ and $A_j \neq A_i$; O_i is the output interface of agent A_i, the set of facts agent A_i can answer to other agents; and t_i is the default deadline for giving an answer.

One of the most important topics in our model is the different variety of answers agents can express producing complex dialogues. A **response** is a tuple $R = \langle f, V, S, KB \rangle$ where: f is the fact which is been answered, V is the value of fact f (an interval of truth-values), S is the state of the fact f value, i.e. *provisional*, *definitive* or *pending* (a fact that is provisionally unknown), and KB is a knowledge base useful to improve the value of f. Let's define now the kinds of responses the agent can give:

a) Definitive value $R = \langle f, V, \text{def}, \emptyset \rangle$: this is the most useful result because it means that there is no more information that can improve the result, this is the most precise. After the specialization with definitive values we can substitute a rule using it by its specialized version.

b) Provisional value $R = \langle f, V, \text{prov}, \emptyset \rangle$: this is not a definitive value, it can be improved later. We can use it to produce only more provisional values. We can not delete rules that use it because they will be useful in the future to produce more precise values.

c) Provisional value and a set of knowledge related to it, $R = \langle f, V, \text{prov}, KB_f \rangle$: this is similar to the case above but the answer includes all the information needed for improving the value. We can use this provisional value and start the mechanism to find more information.

d) The same that the case above but without a provisional value. A set of rules related to the question $R = \langle f, [0, 1], \text{pending}, KB_f \rangle$

When an *agent's life* begins and receives a simple query, the agent starts a goal-driven—backward chaining style—work. This task will produce new goals that have to be solved and it judges the impact of these new goals in the quality of the original one. Some of them can be internal and others have to be obtained from other agents. Internal goals are considered a self-commitment and start a search process in order to find which are the new goals. When new facts are known, maybe from other agents answers, it is started a data-driven task of specialization—forward chaining style. The transition from one solution to a more precise one happens in this specialization step.

In Fig. 1 we can see a summary of the cycle. Agents are continuously checking the queue of messages and the set of goals or commitments. When a query is received it generates subgoals if necessary; if the message is an answer from another agent it integrates and specializes the knowledge base. It checks the goal base to decide if it is necessary to make requests, insists to obtain better values, or sends answers because it has obtained the value or the contract is over.

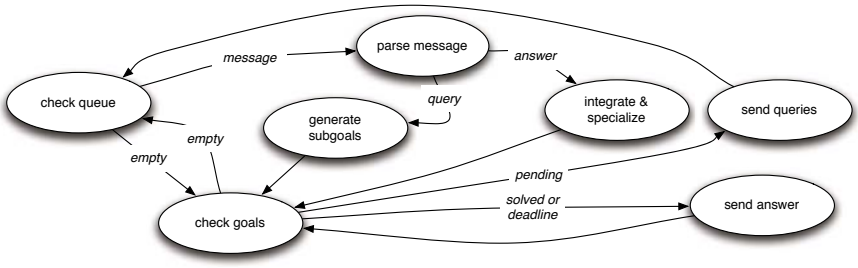


Fig. 1. Evaluation cycle

An incomplete answer to a query is generated when there is not enough time to complete the query processing. Each agent goal could achieve a definitive or a provisional value. If this value is not enough, further reasoning is required and new questions are sent to the corresponding agents. Agents can send and receive facts and rules as conditional answers or knowledge communication. When the deadline of a goal ends and it has a provisional value, the agent can send rules as part of the answer.

5 An Example

Is this section we will present a very simple example in order to understand the general mechanism for running this multiagent system.

Consider a system with four agents. The code in Fig. 2 corresponds to our implemented syntax for the two more important agents. For the sake of simplicity we use as default t-norm the minimum. The definition of agents contains the type of logic, in this case a many-valued logic based on numerical intervals (INMV), the predicates, the rules, interfaces and a default deadline for the agent. Considering that the form $a@A$ means that the fact a belongs to agent A, it is easy to see that the knowledge bases for agents corresponding to the two modules above are:

```
(defAgent AgentA
  :Logic (make-instance 'INMV-Logic)
  :Predicates ((fact i1 :name "inputA1")
               (fact i2 :name "inputA2")
               (fact i3 :name "inputA3")
               (fact o1 :name "outputA"))
  :Rules ((rule R001 (i1 i2) -> o1 is 1)
          (rule R002 (i3) -> not o1 is 0.8))
  :Inputs ((input i1 :fact o2 :from agentB)
           (input i2 :fact o3 :from agentC)
           (input i3 :fact o4 :from agentC))
  :Outputs ((output o1))
  :Deadline 30)
```

```
(defAgent AgentD
  :Logic (make-instance 'INMV-Logic)
  :Predicates ((fact i4 :name "inputD")
               (fact o2 :name "outputD"))
  :Rules ((rule R003 (not i4) -> o2 is 0.9))
  :Inputs ((input i4 :fact o1 :from agentA)
  :Outputs ((output o2))
  :Deadline 100)
```

Fig. 2. Code of the example

$$KB_A = \{(o_2@B \wedge o_3@C \rightarrow o_1, [1,1]),\ (o_4@C \rightarrow \neg o_1, [0.8,1])\}$$
$$KB_D = \{(\neg o_1@A \rightarrow o_2, [0.9,1])\}$$

Consider that an agent makes a question to agent D about the fact o_2. D contains only one rule R003 then it makes a question to A about o_1. Now we have to take into account that agent A has two rules, R001 and R002. We can calculate the absolute quality provided by these rules and they are 1 and 0.64, respectively. The best rule is R001 and considering the conditions in the writing order, finally the agent A makes the question o_2 to agent B.

Suppose that the answer of B is a definitive value for the fact: $\langle o_2, [0.7,1]$,def, $\emptyset\rangle$. The knowledge base is specialized deleting the first rule:

$$KB_A = \{(o_2@B, [0.7,1]),\ (o_3@C \rightarrow o_1, [0.7,1]),\ (o_4@C \rightarrow \neg o_1, [0.8,1])\}$$

Now we have to take into account that the truth-values of the rules has changed. Then the quality for rules R001 and R002 is now 0.49 and 0.64, respectively. Then it is reasonable to make the question o_4 to C. Suppose the answer of C is a provisional value: $\langle o_4, [0.4,1]$,prov, $\emptyset\rangle$. We have to maintain all the rules and we deduce a provisional value for o_1. The new knowledge base is:

$$KB_A = \{(o_2@B, [0.7,1]),\ (o_4@C, [0.4,1]),\ (o_1, [0,0.6]),$$
$$(o_3@C \rightarrow o_1, [0.7,1]),\ (o_4@C \rightarrow \neg o_1, [0.8,1])\}$$

Imagine that the is no more time for obtaining a better value and agent A has to obligation to answer. The answer of agent A to the initial question of agent D will be a conditioned answer:

$$\langle o_1, [0,0.6]\text{,prov, } \{(o_3@C \rightarrow o_1, [0.7,1]),\ (o_4@C \rightarrow \neg o_1, [0.8,1])\}\rangle$$

The new knowledge base for D after the integration:

$$KB_D = \{(o_1@A, [0,0.6]),\ (o_3@C \rightarrow o_1@A, [0.7,1]),$$
$$(o_4@C \rightarrow \neg o_1@A, [0.8,1])\},\ (\neg o_1@A \rightarrow o_2, [0.9,1]),\ (o_2, [0.4,1])\}$$

Notice that in this case it is possible to obtain a provisional value for o_2 because the specialization rule produces a value $[T(0.4,1),1] = [min(0.4,1),1] = [0.4,1]$. Now agent D has all the necessary elements to determine the value of o_1 by asking directly to agent C. Suppose that the answers are $\langle o_3, [0.2,1]$,def, $\emptyset\rangle$ and $\langle o_4, [0.6,1]$,def, $\emptyset\rangle$. It is easy to see that the final value for o_i is: $[0,0.6] \cap [0.2,1] \cap [0,0.4] = [0.2,0.4]$ and $[0.6,0.8] \cap [0.4,1] = [0.4,0.8]$ for o_2. The absolute quality of o_2 has decreased from 0.16 to 0.12 but the relative quality has arrived to 100%, that maximum precision for the concrete knowledge base.

6 Conclusions and Future Work

In this paper we have presented an architecture for multiagent systems and an anytime mechanism for deliberative agents based on a monotonous reasoning

over intervals of truth values. Both absolute and relative quality measures have been defined. After the respective analysis, relative measures fit better with anytime quality measures properties.

Deadline is considered fix for the sake of simplicity, but it could be variable and be calculated to improve agent's performance. Criteria like communication channels cost, confidence and agent's capacity can be considered for its estimation and effects on performance profiles. We have said that when an agent receives a provisional value it can be used to produce more provisional values, but we can think in a timeout or other rational subjective criteria to consider that a provisional value becomes definitive.

We are also designing a protocol to deal with provisional values and the knowledge received. It is reasonable to think that when a provisional value is received, agents can insist later in order to improve the value or use their own means to obtain that information.

Acknowledgements. Authors acknowledge partial support by the Spanish projects ARINF (TIN2009-14704-C03-03), MULOG2 (TIN2007-68005-C04-01) and Agreement Technologies (CONSOLIDER CSD2007-0022, INGENIO 2010) and the Government of Catalonia grant 2009-SGR-1434. Mariela Morveli-Espinoza is supported by the Programme Alβan, the European Union Programme of High Level Scholarships for Latin America, scholarship No.(E06D101440PE). We would like to thank the referees for their valuable suggestions and comments.

References

1. Barbuceanu, M., Lo, W.: Conversation oriented programming for agent interaction. In: Dignum, F.P.M., Greaves, M. (eds.) Issues in Agent Communication. LNCS (LNAI), vol. 1916, pp. 220–234. Springer, Heidelberg (2000)
2. Dean, T.L., Boddy, M.: An analysis of time-dependent planning. In: Proceedings of the Seventh National Conference on Artificial Intelligence, AAAI 1988, pp. 49–54 (1988)
3. Morveli-Espinoza, M., Puyol-Gruart, J.: On partial deduction and conversational agents. In: Alsinet, T., Puyol-Gruart, J., Torras, C. (eds.) Artificial Intelligence Research and Development. Frontiers in Artificial Intelligence and Applications, vol. 184, pp. 60–69. IOS Press, Amsterdam (2008)
4. Mouaddib, A.-I.: A study of a dynamic progressive reasoning system. Journal of Experimental and Theoretical Artificial Intelligence, 101–122 (2000)
5. Mouaddib, A.-I.M., Zilberstein, S.: Knowledge-based anytime computation. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 1995, pp. 775–783 (1995)
6. Puyol-Gruart, J., Godo, L., Sierra, C.: Specialisation calculus and communication. International Journal of Approximate Reasoning (IJAR) 18(1/2), 107–130 (1998)
7. Rago, F.: Conversational agent model in intelligent user interface. In: Di Gesú, V., Masulli, F., Petrosino, A. (eds.) Fuzzy Logic and Applications. LNCS (LNAI), vol. 2955, pp. 46–54. Springer, Heidelberg (2006)
8. Shoham, Y.: Agent-oriented programming. Artificial Intelligence 60, 51–92 (1993)
9. Zilberstein, S.: Using anytime algorithms in intelligent systems. The AI magazine 17(3), 73–83 (1996)

Measuring Instability in Normal Residuated Logic Programs: Discarding Information*

Nicolás Madrid and Manuel Ojeda-Aciego

Dept. Matemática Aplicada
Univ. Málaga, Spain

Abstract. Inconsistency in the framework of general residuated logic programs can be, somehow, decomposed in two notions: incoherence and instability. In this work, we focus on the measure of instability of normal residuated programs. Some measures are provided and initial results are obtained in terms of the amount of information that have to be discarded in order to recover stability.

1 Introduction

In many fields of automated information processing it becomes crucial to consider together imprecise, uncertain or inconsistent information. Although inconsistency is an undesirable property, it arises naturally in many real-world problems (for instance, consider the integration of information coming from different sources). Anyway, the analysis of inconsistent knowledge-bases can lead us to obtain useful information: for instance, a big number of contradictions in the statements of a suspect of a crime with respect to the forensic evidences may lead us to increase our confidence on his/her being the culprit; a sensor which send data which contradict other sensors may indicate a possible malfunction. In both cases, a good estimation of the degree of inconsistency of the data can help us to estimate the truth-degree up to which this new information can be safely considered.

There are several papers dealing with inconsistency in a classical logic programming framework. For instance, [1] uses consistency restoring rules as a means to recover whenever possible the consistency of a normal logic program; this approach has been used in [13] to formalize negotiations dealing with incomplete information, preferences, and changing goals. The Answer Set Programming (ASP) framework has been used to detect inconsistencies in large biological networks [2]. Argumentation theory is a suitable framework for inconsistency to arise. There are several non-classical approaches to ASP argumentation, some based on possibility theory, some other based on, for instance, fuzzy set theory [12,7].

The problem of measuring the degree of inconsistency contained in a knowledgebase has been already considered in the literature [5,4,6]. This approach

* Partially supported by the Spanish Ministry of Science projects TIN06-15455-C03-01 and TIN09-14562-C05-01 & Junta de Andalucía projects FQM-2049 and FQM-5233.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 128–137, 2010.

shows that measuring the inconsistency of a knowledgebase is useful to allow for the comparison of the inconsistency of various knowledgebases. On the other hand, Lozinskii provided a method [8] for defining the quantity of information of a knowledgebase in propositional logic. However, that method is not suitable when the knowledgebase is inconsistent. Furthermore, it is certainly false that all inconsistent knowledgebases contain the same (null) amount of information, this is especially relevant when considering fuzzy extensions of the theory.

This work is based on the Fuzzy Answer Set Programming for residuated logic programs defined in [10,9], in which we consider a fuzzy answer set attending to two dimensions: coherence and stability, the former is related to strong negation, whereas the latter is related to default negation and the GL-reduct [3]. An inconsistent fuzzy program is a program without fuzzy answer sets, and this can be due to the lack of stable models (instability) or, perhaps, to the inconsistency of every stable model (incoherence). This is why we talk about the two dimensions of inconsistency. In [11] some measures of inconsistency were defined in terms of incoherence; in this work, we aim at providing an initial step towards the measuring the degree of instability in normal residuated logic programs.

The structure of the paper is described as follows. In Section 2 we recall the definition of stable model. Section 3 describes the possible causes of the instability of a residuated logic program and defines the notion of information measure, which assigns a degree of information to any value in the truth space. In Section 4 we define the measure of instability which establish how many information has to be deleted from a set of rules in order to recovering the stability in the residuated logic program.

2 Preliminaries

Let us start this section recalling the definition of residuated lattice, which fixes the set of truth values and the relationship between the conjunction and the implication (the adjoint condition) occurring in our logic programs.

Definition 1. *A* residuated lattice *is a tuple* $(L, \leq, *, \leftarrow)$ *such that:*

1. $(L, \leq)$ *is a complete bounded lattice, with top and bottom elements 1 and 0.*
2. $(L, *, 1)$ *is a commutative monoid with unit element 1.*
3. $(*, \leftarrow)$ *forms an adjoint pair, i.e.* $z \leq (x \leftarrow y)$ *iff* $y * z \leq x \quad \forall x, y, z \in L.$

In the rest of the paper we will consider a residuated lattice enriched with a negation operator, $(L, *, \leftarrow, \neg)$. The negation $\neg$ will model the notion of default negation often used in logic programming. As usual, a negation operator, over L, is any decreasing mapping $n\colon L \to L$ satisfying $n(0) = 1$ and $n(1) = 0$. In the examples, we will use the following familiy of negation operators:

$$n_\alpha(x) = \begin{cases} 1 & \text{if } x \leq \alpha \\ 0 & \text{if } x > \alpha \end{cases} \qquad\qquad n(x) = 1 - x$$

Definition 2. *Given a residuated lattice with negation* $(L, \leq, *, \leftarrow, \neg)$*, a* normal residuated logic program $\mathbb{P}$ *is a set of weighted rules of the form*

$$\langle p \leftarrow p_1 * \cdots * p_m * \neg p_{m+1} * \cdots * \neg p_n; \quad \vartheta \rangle$$

where ϑ *is an element of* L *and* $p, p_1, \ldots, p_n$ *are propositional symbols.*

It is usual to denote the rules as $\langle p \leftarrow \mathcal{B}; \vartheta \rangle$. The formula $\mathcal{B}$ is usually called the *body* of the rule whereas p is called its *head*. A *fact* is a rule with empty body, i.e facts are rules with this form $\langle p \leftarrow \quad ; \vartheta \rangle$. The set of propositional symbols appearing in $\mathbb{P}$ is denoted by $\Pi_{\mathbb{P}}$.

Definition 3. *A fuzzy* L-interpretation *is a mapping* $I \colon \Pi_{\mathbb{P}} \to L$*; note that the domain of the interpretation can be lifted to any rule by homomorphic extension.*

We say that I satisfies *a rule* $\langle \ell \leftarrow \mathcal{B}; \quad \vartheta \rangle$ *if and only if* $I(\mathcal{B}) * \vartheta \leq I(\ell)$ *or, equivalently,* $\vartheta \leq I(\ell \leftarrow \mathcal{B})$.

Finally, I *is a* model *of* $\mathbb{P}$ *if it satisfies all rules (and facts) in* $\mathbb{P}$.

Note that the order relation in the residuated lattice $(L, \leq)$ can be extended over the set of all L-interpretations as follows: *Let* I *and* J *be two* L*-interpretations, then* $I \leq J$ *if and only if* $I(p) \leq J(p)$ *for all propositional symbol* $p \in \Pi_{\mathbb{P}}$.

Stable Models

Our aim in this section is to adapt the approach given in [3] to the normal residuated logic programs just defined in the section above.

Let us consider a normal residuated logic program $\mathbb{P}$ together with a fuzzy L-interpretation I. To begin with, we will construct a new normal program $\mathbb{P}_I$ by substituting each rule in $\mathbb{P}$ such as

$$\langle p \leftarrow p_1 * \cdots * p_m * \neg p_{m+1} * \cdots * \neg p_n; \quad \vartheta \rangle$$

by the rule[1]

$$\langle p \leftarrow p_1 * \cdots * p_m; \quad \neg I(p_{m+1}) * \cdots * \neg I(p_n) * \vartheta \rangle$$

Notice that the new program $\mathbb{P}_I$ is positive , that is, does not contain any negation; in fact, the construction closely resembles that of a reduct in the classical case, this is why we introduce the following:

Definition 4. *The program* $\mathbb{P}_I$ *is called the* reduct *of* $\mathbb{P}$ *wrt the interpretation* I.

As a result of the definition, note that given two fuzzy L-interpretations I and J, then the reducts $\mathbb{P}_I$ and $\mathbb{P}_J$ have the same rules, and might only differ in the values of the weights. By the monotonicity properties of $*$ and $\neg$, we have that if $I \leq J$ then the weight of a rule in $\mathbb{P}_I$ is greater or equal than its weight in $\mathbb{P}_J$.

[1] Note the overloaded use of the negation symbol, as a syntactic function in the formulas and as the algebraic negation in the truth-values.

It is not difficult to prove that every model M of the program $\mathbb{P}$ is a model of the reduct $\mathbb{P}_M$.

Recall that *a fuzzy interpretation* can be interpreted as a L-fuzzy subset. Now, as usual, the notion of reduct allows for defining a *stable set* for a program.

Definition 5. *Let $\mathbb{P}$ be a normal residuated logic program and let I be a fuzzy L-interpretation; I is said to be a* stable set *of $\mathbb{P}$ iff I is the least model of $\mathbb{P}_I$.*

Theorem 1. *Any stable set of $\mathbb{P}$ is a minimal model of $\mathbb{P}$.*

Thanks to Theorem 1 we know that every stable set is a model, therefore we will be able to use the term stable model to refer to a stable set. Obviously, this approach is a conservative extension of the classical approach. Note, as well, that a residuated logic program can have infinitely many stable models.

In the following example we use a simple normal logic program with just one rule in order to clarify the definition of stable set (stable model).

Example 1. Consider the program $\langle p \leftarrow \neg q \quad ; \vartheta \rangle$. Given a fuzzy L-interpretation $I \colon \Pi \to L$, the reduct $\mathbb{P}_I$ is the rule (actually, the fact) $\langle p \quad ; \vartheta * \neg I(q) \rangle$ for which the least model is $M(p) = \vartheta * \neg I(q)$, and $M(q) = 0$. As a result, I is a stable model of $\mathbb{P}$ if and only if $I(p) = \vartheta * \neg I(q) = \vartheta * \neg(0) = \vartheta * 1 = \vartheta$ and $I(q) = 0$. □

The following example shows that stable models for a normal residuated logic program need not exist.

Example 2. Consider the the following normal residuated logic program on the product logic

$$\langle p \leftarrow \neg p \quad ; 1 \rangle$$

defined over the residuated lattice $([0,1], \leq, *_P, \leftarrow_P, n_\alpha)$ (for any $\alpha \in [0,1)$). This normal residuated logic program does not have stable models. Let I be an interpretation. The reduct w.r.t. I is either the fact $\langle p \leftarrow \quad ; 1 \rangle$ if $I(p) \leq \alpha$ or the fact $\langle p \leftarrow \quad ; 0 \rangle$ if $I(p) > \alpha$. In any case, if I is a stable model then $I(p)$ is equal either 1 or 0. However, none of the interpretations is stable model of this normal residuated logic program. □

The aim of this work is to study normal residuated logic programs without any stable model by means of measures which determine how much information one has to add or delete in order to recover at least one stable model. We start by proposing the following definition:

Definition 6. *A normal residuated logic program $\mathbb{P}$ is stable if and only if there is an L-interpretation I that is a stable model of $\mathbb{P}$; i.e I is the least model of $\mathbb{P}_I$. Otherwise, $\mathbb{P}$ is called unstable.*

3 Causes of Instability: Measures of Information

Instability is an undesirable feature of a logic program. When representing knowledge as a (residuated) logic program it is usual to implement rules according to

a set of external data (obtained either from sensors or from suggestion of an expert); this data is subject to mistake and/or imprecisions, and may lead to the following shortcomings:

- Not to include relevant information. (Missing information)
- Include information which is either false or leading to contradiction. (Excess of information)

Any of the situations above might lead to instability. Let us further discuss this by means of an example: the following program tries to simulate a procedure to deduce which sports are practised by a person give some data.

$$\begin{aligned} r_1 &: \langle Football \leftarrow n_{0.4}(Basketball) *_G LivesInSuburb *_G AthleticBody \quad ; 0.6\rangle \\ r_2 &: \langle Basketball \leftarrow n_{0.4}(Cycling) *_G Tall *_G AthleticBody \quad ; 0.6\rangle \\ r_3 &: \langle Cycling \leftarrow n_{0.4}(Football) *_G Slim *_G AthleticBody \quad ; 0.6\rangle \end{aligned}$$

The first rule determines that if a person with an athletic body, which lives in a suburb and we do not know whether he practices regularly basketball, then this person practices football frequently (the interpretation of the other two rules is similar). These three rules do not imply any contradiction, in fact, the program consisting of the three rules has just one stable model[2] $I_\perp$. However, if we add the following facts

$$\begin{aligned} r_4 &: \langle AthleticBody \leftarrow \quad ; 0.8\rangle \qquad & r_5 &: \langle LivesInSuburb \leftarrow \quad ; 1\rangle \\ r_6 &: \langle Tall \leftarrow \quad ; 0.7\rangle & r_7 &: \langle Slim \leftarrow \quad ; 0.8\rangle \end{aligned}$$

the program turns out to be unstable. What are the reasons for this behaviour?

As we said above, it may be because of excess or lack of information. For the former, excess of information can reside in any of the seven rules, it might be that too much information is obtained by default from r_1, r_2 and r_3. Notice that if the weights are changed to 0.39, therefore reducing the amount of information provided by those rules, the program would remain stable. Lack of information is more difficult to handle, in that it is not possible to know which rules are needed; it might be just a fact (if we include the fact $\langle Football \leftarrow \quad ; 0.5\rangle$, the program gets stable again), or a more complex rule or set of rules.

In this work we focus on the treatment of excess of information, and we propose a framework to measure the instability of a program by means of the minimum amount of information which we have to delete in order to obtain a stable program. Our approach to reducing the amount of information provided by a program is based of the values of the weights, since the smaller they are the less information is produced. The key point is how to measure the amount of information which is eliminated.

We propose to fix an operator $m\colon L \to \mathbb{R}^+$ such that:

- $m(x) = 0$ if and only if $x = \perp$
- m is monotonic

[2] $I_\perp$ denotes the bottom element of the complete lattice of L-interpretations.

Such an operator will be called an *information measure*.

It is not difficult to provide examples of these operators in the unit interval or in any finite lattice:

Example 3. Any norm $||\cdot||$ on the lattice $([0,1],\leq)$ is an information measure, since $||x|| = 0$ if and only if $x = 0$; and if $x \leq y$ then

$$||x|| = ||\frac{x}{y} \cdot y|| = |\frac{x}{y}| \cdot ||y|| \leq ||y||$$

□

Example 4. Let $(L,\leq)$ be a finite lattice. An information measure can be defined as follows:

$$m(x) = \max\{n\colon \bot < x_1 < \cdots < x_n = x\}$$

Let us check that, in fact, it is an information measure: if $x \neq \bot$, then $\bot < x$, and this implies $m(x) \geq 1$. On the other hand, if $x < y$, then for all chain $\bot < x_1 < \cdots < x_n = x$ we have the chain $\bot < x_1 < \cdots < x_n = x < x_{n+1} = y$ which has a greater length, and this implies $m(x) < m(y)$. □

Information measures will be used to determine the amount of information inherently contained in any element of the lattice. From now on, we will consider that any lattice has an associated information measure.

4 Measuring Instability of Normal Residuated Programs

In this section we define an instability measure based on the amount of information deleted from a unsatble program so that it gets stable. Contrariwise to the classical case, in which the only form to delete information is by deleting rules completely, in our framework we can just reduce their weights by some amount. A specific operator will be defined for this task.

For that purpose, we need to fix a t-norm t to handle the values of $\mathcal{L}$ (recall that a t-norm is a commutative and monotonic map $L \times L \to L$ satisfying $t(\bot, x) = \bot$ and $t(\top, x) = x$). Fixed such a t-norm, we can define an operator to modify the weights of rules.

Given a normal residuated logic program $\mathbb{P}$, a set $\{\langle r_i; \vartheta_i\rangle\}_i$ of rules in $\mathbb{P}$ and a set of values $\{\varphi_i\}_i$ we define a new general residuated logic program $O_{\mathbb{P}}(\{\langle r_i; \vartheta_i\rangle\}_i, \{\varphi\}_i)$ as follows:

$$O_{\mathbb{P}}(\{\langle r_i; \vartheta_i\rangle\}_i, \{\varphi\}_i) = (\mathbb{P} \smallsetminus \{\langle r_i; \vartheta_i\rangle\}_i) \cup \{\langle r_i; t(\vartheta_i, \varphi_i)\rangle\}_i$$

In other words, the operator $O_{\mathbb{P}}$ substitutes the weight of any rule $\langle r_j; \vartheta_j\rangle$ in the given set by $t(\vartheta_j, \varphi_j)$.

It is not difficult to note that the resulting program has smaller weights than the original one. The following example illustrates this fact.

Example 5. Consider the residuated lattice with negation $([0,1], \leq, *_P, \leftarrow_P, n)$, and the following residuated program

$$\begin{array}{ll} r_1 \colon \langle p \leftarrow q * t * \neg t \quad ; 0.7 \rangle & r_2 \colon \langle p \leftarrow t * \neg s \quad ; 0.8 \rangle \\ r_3 \colon \langle q \leftarrow \neg v \quad ; 0.2 \rangle & r_4 \colon \langle t \leftarrow s * u * \neg v \quad ; 0.9 \rangle \end{array}$$

Assume the product t-norm ($t(x,y) = x \cdot y$) as the t-norm associated to the operator $O_{\mathbb{P}}$. Then, the program $O_{\mathbb{P}}(\{r_1, r_4\}, \{0.5, 0.9\})$ is shown below:

$$\begin{array}{ll} r_1 \colon \langle p \leftarrow q * t * \neg t \quad ; 0.35 \rangle & r_2 \colon \langle p \leftarrow t * \neg s \quad ; 0.8 \rangle \\ r_3 \colon \langle q \leftarrow \neg v \quad ; 0.2 \rangle & r_4 \colon \langle t \leftarrow s * u * v \quad ; 0.81 \rangle \end{array}$$

Notice that the weights of rules r_1 and r_4 are reduced by a factor 0.5 and 0.9 respectively. □

The instability measure will be defined in terms of the amount of discarded information needed to get stability, and this will be computed by means of an information measure, as those introduced in Section 3, and the formula

$$\sum_{i \in \mathbb{I}} \big(m(\top) - m(\varphi_i) \big)$$

The sum above, in some sense, measures the amount of information discarded from the program; the lesser the values of φ_i the more information discarded, and greater the sum. Notice as well that $O_{\mathbb{P}}$ does not reduce the weights of the program if and only if φ_i are $\top$ for all i, and the previous sum reduces to 0.

Example 6. Continuing with Example 5, if we consider in $[0,1]$ the information measure induced by the Euclidean norm, then the amount of discarded information by the use of $O_{\mathbb{P}}(\{r_1, r_4\}, \{0.5, 0.9\})$ would be $(1 - 0.5) + (1 - 0.9) = 0.6$. □

Now, we can define the following instability measure, given a general residuated logic program $\mathbb{P}$ and a set of rules $\{\langle r_i, \vartheta_i \rangle\}_i \subseteq \mathbb{P}$ (w.r.t. the respective residuated logic program) as:

$$\textsc{Instab}_{\mathbb{P}}(\{\langle r_i, \vartheta_i \rangle\}_i\}) = \inf\{\sum_{i \in \mathbb{I}} m(\top) - m(\varphi_i) \colon O_{\mathbb{P}}(\{\langle r_i, \vartheta_i \rangle\}_i, \{\varphi\}_i) \textit{ is stable } \}$$

It is important to note that this operator needs not be defined for any set of rules (the sum could be infinite). This is not a big problem, as that would indicate that it is not possible to recover stability by not even discarding completely all the rules in the set.

Example 7. On the residuated lattice with negation $([0,1], \leq, \wedge_P, \leftarrow_P, n_{0.4})$, let us consider the following unstable logic program:[3]

$$\begin{array}{rr} r_1 \colon & \langle p \leftarrow s \wedge \neg q \quad ; 0.8 \rangle \\ r_2 \colon & \langle q \leftarrow \neg r \wedge \neg u \quad ; 0.8 \rangle \\ r_3 \colon & \langle r \leftarrow \neg p \quad ; 0.5 \rangle \\ r_4 \colon & \langle s \leftarrow \quad ; 0.8 \rangle \\ r_5 \colon & \langle t \leftarrow \neg p \wedge \neg s \quad ; 0.5 \rangle \\ r_6 \colon & \langle v \leftarrow u \wedge \neg r \quad ; 0.7 \rangle \end{array}$$

[3] To increase readability, the subscripts P have been removed.

It is not difficult to check that this program does not have stable models. We will use the product t-norm and the Euclidean norm in the formulas above to measure the instability of the rules of the program. For the case of r_1, one can see that if its weight would be a value $\alpha \leq 0.5$, then the program would have a stable model; specifically, $M \equiv \{(p, 0.8 \cdot \alpha); (q, 0); (r, 0.5); (s, 0.8); (t, 0.4); (v, 0)\}$.

On the other hand, it is possible to set the weight of r_1 to 0.5 using the factor $\varphi = 0.625$. Therefore, the least amount of information to be discarded from r_1 has to be $1 - 0.625 = 0.375$. In other words, $\textsc{Instab}_{\mathbb{P}}(\{r_1\}) = 0.375$. Similarly, we can obtain the instability measures for the rest of rules:

x	r_1	r_2	r_3	r_4	r_5	r_6
$\textsc{Instab}_{\mathbb{P}}(\{x\})$	0.375	0.5	0.2	0.375	$\star$	$\star$

The symbol $\star$ for rules r_5 and r_6 denotes that it is impossible to get a stable program by reducing the weights of these rules. Notice that these results state that, in recovering stability by modifying just one rule, we need to discard much more information from r_2 than in r_3. □

A couple of straightforward results about the instability measure $\textsc{Instab}_{\mathbb{P}}$ are presented below. The first one establishes a relationship between stable programs and zero measure.

Proposition 1. *Let $\mathbb{P}$ be a normal residuated logic program:*

- *If $\mathbb{P}$ is stable then $\textsc{Instab}_{\mathbb{P}}(\mathbb{P}) = 0$*
- *If $\textsc{Instab}_{\mathbb{P}}(\mathbb{P}) = 0$ then for all $\varepsilon > 0$ there exists a set $\{\varphi_i\} \subseteq L$ such that $O_{\mathbb{P}}(\{\langle r_i; \vartheta_i\rangle\}_i, \{\varphi\}_i)$ is stable and $\sum_{i \in \mathbb{I}} \big(m(\top) - m(\varphi_i)\big) < \varepsilon$.*

The following proposition states the antitonicity of the measure $\textsc{Instab}_{\mathbb{P}}$:

Proposition 2. *Let $\mathbb{P}$ be a normal residuated logic program and let $\{\langle r_i; \vartheta_i\rangle\} \subseteq \{\langle \overline{r_i}; \overline{\vartheta_i}\rangle\}$ be two sets of rules of $\mathbb{P}$. Then:*

$$\textsc{Instab}_{\mathbb{P}}(\{\langle r_i; \vartheta_i\rangle\}) \geq \textsc{Instab}_{\mathbb{P}}(\{\langle \overline{r_i}; \overline{\vartheta_i}\rangle\})$$

Computing $\textsc{Instab}_{\mathbb{P}}(\{\langle r_i; \vartheta_i\rangle\})$

The aim of this section is to show that computing the value of $\textsc{Instab}_{\mathbb{P}}(\{\langle r_i; \vartheta_i\rangle\})$ is equivalent to computing the set of stable models of a specific logic program. To facilitate the presentation, let us assume that $[0, 1]$ is the set of truth values and the set of rules $\{\langle r_i; \vartheta_i\rangle\}$ is a singleton.

To compute the measure of instability $\textsc{Instab}_{\mathbb{P}}$ we have to obtain what values $\lambda \in [0, 1]$ satisfy that $O_{\mathbb{P}}(\langle r_i; \vartheta_i\rangle, \lambda)$ is stable; we recall that $O_{\mathbb{P}}(\langle r_i; \vartheta_i\rangle, \lambda)$ coincides with $\mathbb{P}$ except in the rule $\langle r_i; \vartheta_i\rangle$, which is changed by $\langle r_i; t(\vartheta_i, \lambda)\rangle$. How can we introduce the parameter λ in $\mathbb{P}$ through propositional symbols? Let α and β be two propositional symbols not occurring in $\mathbb{P}$. Consider the following set of rules:

$$\langle \alpha \leftarrow \neg\beta \quad ; 1 \rangle \tag{1}$$

$$\langle \beta \leftarrow \neg\alpha \quad ; 1 \rangle \tag{2}$$

where the negation is the standard one $n(x) = 1 - x$. The set of stable models of this pair of rules is the set $\{M_\lambda \equiv (\alpha, \lambda); (\beta, 1-\lambda)\}_{\lambda \in [0,1]}$. Notice that for any $\lambda \in [0,1]$ there exists a stable model such that $M_\lambda(\alpha) = \lambda$. We consider now a new residuated logic program $\mathbb{P}^\star$ by modifying r_i as follows[4]

$$r_i^\star \colon \langle p_i \leftarrow \mathcal{B} * t(\vartheta, \alpha) \quad ; 1 \rangle$$

and including the rules (1) and (2). Then the following proposition holds:

Proposition 3. *Let $\mathbb{P}$ be a residuated logic program. M is a stable model of $\mathbb{P}^\star$ if and only if $M|_{\Pi_\mathbb{P}}$ is a stable model of $O_\mathbb{P}(\langle r_i; \vartheta_i \rangle, M(\alpha))$.*

The above proposition shows that there is a univocal correspondence among the stable model of $\mathbb{P}^\star$ and the parameters λ_i such that $O_\mathbb{P}(\langle r_i; \vartheta_i \rangle, \lambda)$ is stable. Therefore we can compute $\textsc{Instab}_\mathbb{P}(\{\langle r_i; \vartheta_i \rangle\})$ by using the stable models of $\mathbb{P}^\star$:

Corollary 1. *Let $\mathbb{P}$ be a residuated logic program. Then:*

$$\textsc{Instab}_\mathbb{P}(\langle r_i; \vartheta_i \rangle) = \inf\{m(\top) - m(M(\alpha)) \colon M \text{ is a stable model of } \mathbb{P}^\star\}$$

5 Conclusions

We have continued our study of fuzzy answer set semantics for residuated logic programs by focusing on the measure of instability of normal residuated programs. Some measures have been provided and initial results have been obtained, in terms of the amount of information that has to be discarded in order to recover stability.

As future work, we will study the dual situation in which stability can be recovered by adding information (as in the framework of consistency restoring rules). In addition, we will extend this methodology to provide explanations for inconsistencies in the data by determining minimal representations of conflicts. In practice, this can be used to identify unreliable data or to indicate missing reactions.

References

1. Balduccini, M., Gelfond, M.: Logic programs with consistency-restoring rules. In: Intl. Symp. on Logical Formalization of Commonsense Reasoning, AAAI 2003, pp. 9–18 (2003)

[4] Technically, the resulting rule is not residuated, but the semantics can be easily adapted to cope with this change.

2. Gebser, M., Schaub, T., Thiele, S., Usadel, B., Veber, P.: Detecting inconsistencies in large biological networks with answer set programming. In: Garcia de la Banda, M., Pontelli, E. (eds.) ICLP 2008. LNCS, vol. 5366, pp. 130–144. Springer, Heidelberg (2008)
3. Gelfond, M., Lifschitz, V.: The stable model semantics for logic programming. In: Proc. of ICLP 1988, pp. 1070–1080 (1988)
4. Grant, J., Hunter, A.: Measuring inconsistency in knowledgebases. J. Intell. Inf. Syst. 27(2), 159–184 (2006)
5. Hunter, A., Konieczny, S.: Approaches to measuring inconsistent information. In: Bertossi, L., Hunter, A., Schaub, T. (eds.) Inconsistency Tolerance. LNCS, vol. 3300, pp. 191–236. Springer, Heidelberg (2005)
6. Hunter, A., Konieczny, S.: Measuring inconsistency through minimal inconsistent sets. In: Proc. of Principles of Knowledge Representation and Reasoning (KR 2008), pp. 358–366. AAAI Press, Menlo Park (2008)
7. Janssen, J., De Cock, M., Vermeir, D.: Fuzzy argumentation frameworks. In: Proc. of IPMU 2008, pp. 513–520 (2008)
8. Lozinskii, E.: Information and evidence in logic systems. Journal of Experimental and Theoretical Artificial Intelligence 6, 163–193 (1994)
9. Madrid, N., Ojeda-Aciego, M.: Towards a fuzzy answer set semantics for residuated logic programs. In: Proc. of WI-IAT 2008. Workshop on Fuzzy Logic in the Web, pp. 260–264 (2008)
10. Madrid, N., Ojeda-Aciego, M.: On coherence and consistence in fuzzy answer set semantics for residuated logic programs. In: Di Gesù, V., Pal, S.K., Petrosino, A. (eds.) Fuzzy Logic and Applications. LNCS, vol. 5571, pp. 60–67. Springer, Heidelberg (2009)
11. Madrid, N., Ojeda-Aciego, M.: On the measure of incoherence in extended residuated logic programs. In: IEEE Intl Conf on Fuzzy Systems (FUZZ-IEEE 2009), pp. 598–603 (2009)
12. Nieves, J.C., Cortés, U., Osorio, M.: Possibilistic-based argumentation: An answer set programming approach. In: Mexican International Conference on Computer Science, pp. 249–260 (2008)
13. Son, T.C., Sakama, C.: Negotiation using logic programming with consistency restoring rules. In: IJCAI 2009: Proc 21st Intl Joint Conf on Artificial intelligence, pp. 930–935. Morgan Kaufmann Publishers Inc., San Francisco (2009)

Implementing Prioritized Merging with ASP

Julien Hue[1], Odile Papini[2], and Eric Würbel[1]

[1] LSIS-CNRS 6168, Université du Sud Toulon -Var. BP 132,
83957 La Garde Cedex France
{hue,wurbel}@univ-tln.fr
[2] LSIS-CNRS 6168, ESIL, Université de la Méditerranée, 163 Av de Luminy,
Case 925 13288 Marseille Cedex 9 France
papini@esil.univmed.fr

Abstract. The paper addresses the extension of the removed sets framework to prioritized removed sets fusion (PRSF). It discusses the links between PRSF and iterated removed sets revision and shows that PRSF satisfy most of the postulates proposed for prioritized merging. An implementation of this new syntactic prioritized fusion operator is proposed thanks to answer sets programming.

Keywords: fusion, iterated revision, ASP reasoning under inconsistency.

1 Introduction

In the last decade, multiple sources belief bases merging has been widely discussed [11,29,3]. Postulates characterizing the rational behavior of merging operations have been proposed [22]. Several merging operations have been proposed that can be divided into two families. The semantic (or model-based) ones which select interpretations that are the "closest" to the original belief bases [22,23,20,12,30,21,10] and the syntactic (or formula-based) ones which select some formulas from the initial bases [25,33,17,13,5].

In some situations, the belief bases are not flat and the beliefs are stratified or equipped with priority levels, in other cases the belief bases are flat but the sources are not equally reliable and there exists a preference relation between sources. In such cases, prioritized merging consists of combining belief bases taking into account the stratification of the belief bases or the preference relation [6]. Prioritized merging has been studied within the framework of propositional logic [12,33,18] as well as within the possibilistic logic one [7,5,8]. The links between iterated revision and prioritized merging has been discussed and rational postulates for prioritized merging have been proposed in [12]. Different iterated revision operations have been studied and it has been shown that the DMA approach [9] then the Qi's approach [27] can be characterized by a lexicographic strategy.

This paper addresses the extension of the Removed Sets framework to prioritized merging where preferences are expressed between belief bases. A new syntactic prioritized fusion operation is proposed and can be expressed in terms

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 138–147, 2010.

of iterated removed sets revision. This operation satisfies most of the postulates for prioritized merging. An implementation stemming from the Answer Sets Programming is described.

The rest of the paper is organized as follows. Section 1 gives a refresher on prioritized merging, removed sets revision, removed sets fusion and logic programming with answer sets semantics. Section 2 presents the prioritized removed sets fusion (PRSF). Section 3 discusses the links between PRSF and iterated removed sets revision and shows that PRSF satisfies the proposed postulates for prioritized merging. Section 4 presents an implementation of PRSF based on logic programming with answer sets semantics before concluding.

2 Background

2.1 Notations

Throughout the paper we consider a propositional language $\mathcal{L}$ over a finite alphabet $\mathcal{P}$ of atoms. A literal is an atom or the negation of an atom. The usual propositional connectives are denoted by $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$ and Cn denotes the logical consequence. A *belief base* K is a finite set of propositional formulas over $\mathcal{L}$. Let $E = \{K_1, \ldots, K_n\}$ be a multi-set of n consistent belief bases to be merged, E is called a *belief profile*. The n belief bases $K_1, \ldots, K_n$ are not necessarily different and the union of belief bases, taking repetitions into account, is denoted by $\sqcup$ and their conjunction (resp. disjunction) are denoted by $\bigwedge$ (resp. $\bigvee$).

2.2 Removed Sets Revision

We briefly recall the Removed Sets Revision (RSR) approach. RSR [32] deals with the revision of a set of propositional formulas by a set of propositional formulas[1]. Let K and A be finite sets of clauses. Removed Sets Revision (RSR) focuses on the minimal subsets of clauses to remove from K, called *removed sets*, in order to restore the consistency of $K \cup A$. More formally: let K and A be two consistent sets of clauses such that $K \cup A$ is inconsistent. Let R be a subset of clauses of K, R is a removed set of $K \cup A$ iff (i) $(K \backslash R) \cup A$ is consistent; (ii) $\forall R' \subseteq K$, if $(K \backslash R') \cup A$ is consistent then $\mid R \mid \leq \mid R' \mid$[2]. Let denote by $\mathcal{R}(K \cup A)$ the collection of removed sets of $K \cup A$, RSR is defined as follows: let K and A be two consistent sets of clauses, $K \circ_{RSR} A =_{def} \bigvee_{R \in \mathcal{R}(K \cup A)} (K \backslash R) \cup A$.

2.3 Removed Sets Fusion

The Removed Sets Revision approach has been extended to the Removed Sets Fusion (RSF) for merging belief bases consisting of well-formed formulas. The key idea of the approach is to remove subsets of well-formed formulas from

[1] We consider propositional formulas in their equivalent conjunctive normal form (CNF).

[2] $\mid R \mid$ denotes the number of clauses of R.

the union of the belief bases, according to some strategy P, in order to restore consistency. Let $E = \{K_1, \ldots, K_n\}$ be a belief profile such that $K_1 \sqcup \cdots \sqcup K_n$ is inconsistent and let IC be integrity constraints. Removed Sets Fusion (RSF) provides, as a result of merging, subsets of formulas of $K_1 \sqcup \cdots \sqcup K_n$ which are consistent with IC.

Definition 1. *Let $E = \{K_1, \ldots, K_n\}$ be a belief profile and IC the integrity constraints such that $K_1 \sqcup \cdots \sqcup K_n \sqcup IC$ is inconsistent, $X \subseteq K_1 \sqcup \cdots \sqcup K_n$ is a* potential Removed Set *of E iff $((K_1 \sqcup \cdots \sqcup K_n) \backslash X) \sqcup IC$ is consistent.*

In order to select the most relevant potential Removed Sets according to a strategy P, a total preorder $\leq_P$ over the potential Removed Sets is defined ($X \leq_P Y$ means that X is preferred to Y according to the strategy P). The associated strict preorder is denoted by $<_P$.

Definition 2. *Let $E = \{K_1, \ldots, K_n\}$ be a belief profile and IC the integrity constraints such that $K_1 \sqcup \cdots \sqcup K_n \sqcup IC$ is inconsistent, $X \subseteq K_1 \sqcup \cdots \sqcup K_n$ is a* Removed Set *of E according to P iff* (i) *X is a potential removed set of E;* (ii) *there is no $Y \subseteq K_1 \sqcup \cdots \sqcup K_n$ such that $Y <_P X$.*

We denote by $\mathcal{F}_{P,IC}\mathcal{R}(E)$ the collection of removed sets of E constrainted by IC according to P. The definition of Removed Sets Fusion is:

Definition 3. *Let $E = \{K_1, \ldots, K_n\}$ be a belief profile and IC the integrity constraints such that $K_1 \sqcup \cdots \sqcup K_n \sqcup IC$ is inconsistent. The fusion operation $\Delta^P_{IC}(E)$ is defined by: $\Delta^P_{IC}(E) = \bigvee_{X \in \mathcal{F}_P\mathcal{R}(E)} \{((K_1 \sqcup \cdots \sqcup K_n) \backslash X) \sqcup IC\}$.*

The usual merging strategies ($Card, \Sigma, Max, GMax$) are captured within our framework by encoding the preference relation between potential removed sets as given in the following table. For the $GMax$ strategy, let X be a potential removed set and K_i be a belief base, we define $p^i(X) = |X \cap K_i|$. Let L^E_X be the sequence composed with every $(p^i(X))_{1 \leq i \leq n}$ in decreasing order. We denote by $\leq_{lex}$ the lexicographic ordering.

P	$X \leq_P Y$
$Card$	$\|X\| \leq \|Y\|$
Σ	$\Sigma_{1 \leq i \leq n} \mid X \cap K_i \mid \leq \Sigma_{1 \leq i \leq n} \mid Y \cap K_i \mid$
Max	$max_{1 \leq i \leq n} \mid X \cap K_i \mid \leq max_{1 \leq i \leq n} \mid Y \cap K_i \mid$
$GMax$	$L^E_X \leq_{lex} L^E_Y$

2.4 Answer Set Programming

In the following, $c, a_i (1 \leq i \leq n), b_j (1 \leq j \leq m)$ are propositional atoms and the symbol *not* stands for *negation as failure*. A *normal logic program* is a set of rules of the form $c \leftarrow a_1, \ldots, a_n, not\ b_1, \ldots, not\ b_m$ where Let r be a rule, we introduce $head(r) = c$ and $body(r) = \{a_1, \cdots, a_n, b_1, \cdots, b_m\}$. Furthermore, let $body^+(r) = \{a_1, \cdots, a_n\}$ denotes the set of positive body atoms and $body^-(r) = \{b_1, \cdots, b_m\}$ the set of negative body atoms, it follows $body(r) = body^+(r) \cup$

$body^-(r)$. Moreover, r^+ denotes the rule $head(r) \leftarrow body^+(r)$, obtained from r by deleting all negative atoms in the body of r.

A set of atoms X is *closed under* a basic program Π iff for any rule $r \in \Pi$, $head(r) \in X$ whenever $body(r) \subseteq X$. The smallest set of atoms which is closed under a basic program Π is denoted by $CN(\Pi)$. The *reduct* or Gelfond-Lifschitz transformation [15], Π^X of a program Π relatively to a set X of atoms is defined by $\Pi^X = \{r^+ \mid r \in \Pi \text{ and } body^-(r) \cap X = \emptyset\}$. A set of atoms X is an *answer set* of Π iff $CN(\Pi^X) = X$.

2.5 ASP Solvers

In the last decade, answer set programming has been considered as a convenient tool to handle non-monotonic reasoning. Moreover, several efficient systems, called ASP solvers, have been developed for computing answer sets, Smodels [26], XSB [28], DLV [14], NoMore [1], ASSAT [24], CMODELS [16], CLASP [2].

In order to extend the expressivity and the efficiency of ASP solvers, logic programs have been extended with new statements [31]:

- *domain definitions* allow for compactly encoding the possible values in a given domain, e.g. the declarations $\#domain\ possible(X)$, $possible(1..n)$. ensure that every occurrence of the variable X will take a value from 1 to n.
- *domain restrictions* can be added in some rules. For example, $short(X) \leftarrow size(Y), X < Y$ is only instantiated for X and Y such that $X < Y$.
- *cardinality optimization* make possible to express that at most, respectively at least, some atoms should appear in the answer sets. For example the rule $h \leftarrow k\ \{a_1, \ldots, a_n\}\ l$ expresses that at least k atoms and at most l atoms among $\{a_1, \ldots, a_n\}$ should appear in the answer sets.
- *optimization statements* allow for selecting among the possible answer sets, the ones that satisfy statements like $minimize\{.\}$ or $maximize\{.\}$. For example, the statement $minimize\{a_1, \cdots, a_n\}$ allows for selecting the answer sets with as few of the atoms in $\{a_1, \cdots, a_n\}$ as possible.

3 Prioritized Removed Sets Fusion

We now present a merging operation which respects the preferences expressed over belief bases. In the context of Removed Sets Fusion, the preferences can be interpreted as a strategy which removes as few formulas as possible in high-ranked belief bases in order to restore consistency.

Let $B = \{B_1, \ldots, B_m\}$ be a belief profile where $B_i,\ 1 \leq i\ \leq m$ is a belief base. B is equipped with a reliability or preference relation (a total pre-order) denoted by $\leq_B$ such that B_i is preferred to B_j iff $B_i \leq_B B_j$. Since several belief bases may be equally reliable, we rearrange the profile B in order to regroup the equally reliable belief bases according to a ranking function, denoted by r. This ranking is such that if $B_i <_B B_j$ then $r(B_i) < r(B_j)$ and if $B_i =_B B_j$ then $r(B_i) = r(B_j)$. In the following, we consider the belief profile $E = \{K_1, \ldots, K_n\}$,

where n, $n \leq m$ is the number of ranks, and $\forall i,\ 1 \leq i \leq n$, $K_i = \cup_j B_j$ such that $r(B_j) = i$. When dealing with integrity constraint, $r(IC) < r(K_1)$.

In order to define a merging strategy which takes into account the ranking of E, we define the following total preorder to compare the potential Removed Sets defined in 1.

Definition 4. *Let $(p_X^1, \ldots, p_X^n)$ be the sequence where $p_X^i = |X \cap K_i|$ is the number of formulas removed from K_i by a potential Removed Set X of E. Let X and X' be two potential Removed Sets of E. The $\leq_{lexipref}$ pre-order is defined by: $X \leq_{lexipref} X'$ iff $(p_X^1, \ldots, p_X^n) \leq_{lex} (p_{X'}^1, \ldots, p_{X'}^n)$.*

We can now define the Removed Sets of E according to the *lexipref* strategy:

Definition 5. *Let X and X' be potential Removed Sets of E, X is a Removed Set of E according to lexipref iff **(i)** X is a potential Removed Set of E; **(ii)** There does not exist X' such that $X' \subset X$; **(iii)** There does not exist X' such that $X' <_{lexipref} X$.*

We denote by $\mathcal{F}_{lexipref,IC}\mathcal{R}(E)$ the set of Removed Sets of E according to *lexipref* and the new merging operation, denoted by $\Delta^{RSF}_{lexipref,IC}(E)$, is the following.

Definition 6. *The merging operation of E constrainted by IC according to the lexipref strategy is such that:*
$\Delta^{RSF}_{lexipref,IC}(E) = \bigvee_{X \in \mathcal{F}_{lexipref,IC}\mathcal{R}(E)} \{((K_1 \sqcup \cdots \sqcup K_n) \backslash X) \sqcup IC\}$.

Example 1. We consider the belief profile $E = \{K_1, K_2, K_3\}$ s.t. $K_1 < K_2 < K_3$ and $IC = \top$ with $K_1 = \{a\}$, $K_2 = \{\neg a \vee b, \neg a \vee c\}$ and $K_3 = \{\neg b, \neg c\}$. Table 1 presents some potential Removed Sets of E, among which the minimal ones according to inclusion, as well as the corresponding $(p_X^1, \ldots, p_X^n)$ sequence.

The potential Removed Set which is minimal according to $\leq_{lexipref}$ is $\{\neg b, \neg c\}$. It is even preferred to $\{a\}$ which removes less formulas and would be preferred according to the Σ strategy. So $\Delta^{RSF}_{lexipref,IC}(E) = \{a, \neg a \vee b, \neg a \vee c\}$.

As shown in [12], merging when preferences are expressed can be dealt with as an iterated revision problem.

Table 1. Potential removed sets of example 1

potential Removed Sets X	p_X^1	p_X^2	p_X^3	potential Removed Sets X	p_X^1	p_X^2	p_X^3
$\{a\}$	1	0	0	$\{\neg a \vee b, \neg a \vee c\}$	0	2	0
$\{a, \neg c\}$	1	0	1	$\{\neg a \vee b, \neg c\}$	0	1	1
$\{a, \neg b\}$	1	0	1	$\{\neg a \vee c, \neg b\}$	0	1	1
$\{a, \neg b, \neg c\}$	1	0	2	$\{\neg b, \neg c\}$	0	0	2

4 Prioritized Merging as an Iterated Revision Operation

We now propose two iterated revision operations based on Removed Sets Revision (RSR) in order to deal with prioritized merging according to two directions. The first version, denoted by $\Delta^{PRSF}_{\alpha,IC}$, performs the revision starting from the least preferred bases to the most preferred ones. The second version, denoted by $\Delta^{PRSF}_{\beta,IC}$, performs in the opposite way, revising the belief bases from the most preferred to the least preferred.

Definition 7. *Let $E = \{K_1, \ldots, K_n\}$ be a belief profile and IC the integrity constraints.*

$\Delta^{PRSF}_{\alpha,IC}(E) = (((K_n \circ_{RSR} K_{n-1}) \circ_{RSR} \cdots \circ_{RSR} K_1) \circ_{RSR} IC)$;
$\Delta^{PRSF}_{\beta,IC}(E) = (K_n \circ_{RSR} (K_{n-1} \circ_{RSR} \cdots \circ_{RSR} (K_1 \circ_{RSR} IC)))$.

The behaviour of the $\Delta^{PRSF}_{\alpha,IC}(E)$ operation is not satisfactory as illustrated by the example 2. On the contrary, the same example shows that the behaviour of the $\Delta^{PRSF}_{\beta,IC}$ operation is closer to our excpectations.

Example 2. We come back to the example 1. In this case, we have $\Delta^{PRSF}_{\alpha,IC}(E) = (K_3 \circ K_2) \circ K_1$. Since $K_3 \circ K_2 = K_3 \sqcup K_2$, the removed sets are $R_1 = \{\neg a \vee b, \neg a \vee c\}$, $R_2 = \{\neg a \vee b, \neg c\}$, $R_3 = \{\neg a \vee c, \neg b\}$, $R_4 = \{\neg b, \neg c\}$, and $(K_3 \circ K_2) \circ K_1 = \{a, \neg a \vee c, \neg b\} \vee \{a, \neg a \vee b, \neg c\} \vee \{a, \neg a \vee b, \neg a \vee c\} \vee \{a, \neg b, \neg c\}$. The result is not satisfactory because the origin of the formulas as well as the preferences attached to the bases are lost during the iteration of the revision process. This operation does not correctly reflect the preferences for prioritized merging.

On the other side, $\Delta^{PRSF}_{\beta,IC}(E) = K_3 \circ (K_2 \circ K_1)$. Since $K_2 \circ K_1 = K_2 \sqcup K_1$, the only removed set is $R = \{\neg b, \neg c\}$, and $K_3 \circ (K_2 \circ K_1) = \{a, \neg a \vee b, \neg a \vee c\}$. This operation gives the same result than the one provided by $\Delta^{RSF}_{lexipref,IC}(E)$.

More generally, the $\Delta^{PRSF}_{\beta,IC}$ and the $\Delta^{RSF}_{lexipref,IC}$ operations lead to the same result and the prioritized merging can be expressed in terms of iterated revision.

Proposition 1. $\Delta^{PRSF}_{\beta,IC}(E) = \Delta^{RSF}_{lexipref,IC}(E)$.

Remark 1. In the context of fusion, considering a belief profile $E = \{K_1, \ldots, K_n\}$, the belief bases K_i are supposed to be consistent. However when grouping the equally reliable belief bases, some belief bases could be inconsistent. This does not affect the $\Delta^{RSF}_{lexipref,IC}$ operation since for each removed set X_i, $|X_i \cap K_k| \neq 0$ where K_k is an inconsistent belief base. This does not affect the $\Delta^{PRSF}_{\beta,IC}$ operation since the iteration of the revision process starts by the revision by IC. On contrast for the $\Delta^{PRSF}_{\alpha,IC}$ operation if K_k is an inconsistent belief base then $K_{k+1} \circ_{RSR} K_k =_{def} K_{k+1} \cup K_k$.

We rephrase within our framework the rational postulates for prioritized merging [12].

$(PMon)$ for $i < n$, $\Delta^{PRSF}_{\beta,IC}(K_1, \cdots K_{i+1}) \vdash \Delta^{PRSF}_{\beta,IC}(K_1, \cdots K_i)$.
$(Succ)$ $\Delta^{PRSF}_{\beta,IC}(K_1, \cdots K_n) \vdash \Delta^{PRSF}_{\beta,IC}(K_1)$.
$(Cons)$ $\Delta^{PRSF}_{\beta,IC}(K_1, \cdots K_n)$ is consistent.
$(Taut)$ $\Delta^{PRSF}_{\beta,IC}(K_1, \cdots K_n, \top) \equiv \Delta^{PRSF}_{\beta,IC}(K_1, \cdots K_n)$.
(Opt) if $\bigwedge E$ is consistent then $\Delta^{PRSF}_{\beta,IC}(K_1, \cdots K_n) = \bigwedge E$.
(RA) $\Delta^{PRSF}_{\beta,IC}(K_1, \cdots K_i) = \Delta^{PRSF}_{\beta,IC}(\Delta^{PRSF}_{\beta,IC}(K_1, \cdots K_{i-1}), K_i)$.

Proposition 2. *$\Delta^{RSF}_{lexipref,IC}$ satisfies the (PMon), (Succ), (Cons), (Taut), (Opt), (RA) postulates* [3].
$\Delta^{PRSF}_{\alpha,IC}$ satisfies the (Succ), (Cons), (Taut), (Opt) postulates.

5 Implementation of Prioritized Removed Sets Fusion

We now propose an implementation of the $\Delta^{RSF}_{lexipref,IC}$ operation stemming from the translation of the merging problem into a logic program with stable model semantics.

This implementation builds a logic program $\Pi_{E,IC}$ which consists of two parts: the first one computing the potential Removed Sets, the second one selecting among them the potential Removed Sets according to the *lexipref* strategy.

The first part was presented in [19]. The generation of Potential Removed Set is based on the generation of the interpretations over the atoms of E. It introduces new atoms called rule atoms. For a formula f, the rule atom r_f is deduced if the formula is not satisfied by the interpretation. The logic program generating all the interpretations and the corresponding sets of rule atoms is denoted $\Pi_{E,IC}$.

1. For every atom $a \in E$, the first step introduces the rules: $a \leftarrow not\ a'$ and $a' \leftarrow not\ a$. These rules build a correspondence between interpretations over the atoms of E and answer sets of the logic program $\Pi_{E,IC}$.
2. The second step introduces the rule atoms. For every formula $f \in K_i$, the following rules are introduced according to the syntax of f : **(i)** If $f =_{def} a$, the corresponding rule is $r^i_f \leftarrow not\ a$; **(ii)** If $f =_{def} \neg f^1$, the corresponding rule is $r^i_f \leftarrow not\ \rho_{f^1}$; **(iii)** If $f =_{def} f^1 \vee \cdots \vee f^j$, the corresponding rule is $r^i_f \leftarrow \rho_{f^1}, \ldots, \rho_{f^j}$; **(iv)** If $f =_{def} f^1 \wedge \cdots \wedge f^j$, the corresponding rules are $r^i_f \leftarrow \rho_{f^1}$, $r^i_f \leftarrow \rho_{f^2}$ to $r^i_f \leftarrow \rho_{f^j}$.

It has been shown in the article cited *supra* that there is a one-to-one correspondence between stable models of $\Pi_{E,IC}$ and the potential Removed Sets of E constrainted by IC. Based on this result, we can translate the notion of preference between potential Removed Sets into a preference between stable models.

Definition 8. *Let X and X' be two stable models of $\Pi_{E,IC}$. The $\leq_{lexipref}$ total preorder between stable models is defined as follows: $X \leq_{lexipref} X'$ iff $(p^1_{(X\cap R^+)}, \ldots, p^n_{(X\cap R^+)}) \leq_{lex} (p^1_{(X'\cap R^+)}, \ldots, p^n_{(X'\cap R^+)})$.*

The potential Removed Sets are compared according to the number of formulas removed in each belief base. The stable models can be compared according to the number of rule atoms representing those formulas. This is the usefulness of rule atoms. It leads to the definition of preferred stable models of $\Pi_{E,IC}$ according to the *lexipref* strategy.

[3] Obviously $\Delta^{RSF}_{lexipref,IC}$ does not satisfy the (IS) postulate since it is a syntactic prioritized merging operation.

Definition 9. *Let X be a set of atoms of E and X' be a stable model of $\Pi_{E,IC}$, X is a preferred stable model of $\Pi_{E,IC}$ according to the lexipref strategy iff the following conditions hold:* ***(i)*** *X is a stable model of $\Pi_{E,IC}$;* ***(ii)*** *there does not exist X' such that $X' \subset X$;* ***(iii)*** *there does not exist X'such that $X' <_{lexipref} X$.*

The problem consisting in determining among the stable models those which are the preferred ones is solved through a set of logic programming statement. The predicate $size(I, J)$ [4] represents the fact that J formulas are coming from K_I in the potential Removed Set. $size(I, J)$ is computed by the following rule which is introduced for every base K_I and every possible U from 1 to m which is the maximum cardinality of a belief base in the profile E.

$$\Pi_E^{lexipref,size} = \left\{ \gamma_1 \ : \ size(V,U) \leftarrow U \ \{f_1^V, \ldots, f_m^V\} \ U. \right\}$$

Therefore the complete program computing the result of $\Delta^{RSF}_{lexipref,IC}(E)$ is the following: $\Pi^{lexipref}_{E,IC} = \Pi_{E,IC} \cup \Pi_E^{lexipref,size} \cup minimize[\ size(1,1) = 1 \times (m+1)^{n-1}, size(1,2) = 1 \times (m+1)^{n-1}, \ldots, size(i,1) = 1 \times (m+1)^{n-i}, size(i,2) = 2 \times (m+1)^{n-i}, \ldots, size(n,m) = m\ ]$. The stable models of $\Pi^{lexipref}_{E,IC}$ are the preferred stable models $\Pi_{E,IC}$ according to the *lexipref* strategy. Moreover, it computes exactly the expected Removed Sets.

Proposition 3. *The set of Removed Sets of $\Pi^{lexipref}_{E,IC}$ is the set of preferred stable models of $\Pi_{E,IC}$ according to the lexipref strategy.*

Example 3. We now present the implementation of example 1.

$$\Pi_{E,IC} = \left\{ \begin{array}{ccc} a \leftarrow not\ a'. & a' \leftarrow not\ a. & b \leftarrow not\ b'. \\ c \leftarrow not\ c'. & c' \leftarrow not\ c. & b' \leftarrow not\ b. \\ r^1_a \leftarrow a'. & r^2_{\neg a \vee b} \leftarrow a, b'. & r^2_{\neg a \vee c} \leftarrow a, c'. \\ r^3_{\neg b} \leftarrow b. & r^3_{\neg c} \leftarrow c. & \end{array} \right\}$$

$$\Pi_E^{lexipref,size} = \left\{ \begin{array}{cc} size(1,1) \leftarrow 1\{r^1_a\}1. & size(2,1) \leftarrow 1\{r^2_{\neg a \vee b}, r^2_{\neg a \vee c}\}1. \\ size(2,1) \leftarrow 2\{r^2_{\neg a \vee b}, r^2_{\neg a \vee c}\}2. & size(3,1) \leftarrow 1\{r^3_{\neg b}, r^3_{\neg c}\}1. \\ size(3,2) \leftarrow 2\{r^3_{\neg b}, r^3_{\neg c}\}2. & \end{array} \right\}$$

$$minimize[size(1,1) = 9, size(2,2) = 6, size(2,1) = 3, size(3,2) = 2, size(3,1) = 1].$$

which has the following stable models (with their associated weight):

$$\begin{array}{cc} \{a', b', c', r^1_a, size(1,1)\} & (9) \\ \{a', b', c, r^1_a, r^3_{\neg c}, size(1,1), size(3,1)\} & (10) \\ \{a', b, c', r^1_a, r^3_{\neg b}, size(1,1), size(3,1)\} & (10) \\ \{a', b, c, r^1_a, r^3_{\neg b}, r^3_{\neg c}, size(1,1), size(3,2)\} & (11) \\ \{a, b', c', r^2_{\neg a \vee b}, r^2_{\neg a \vee c}, size(2,2)\} & (15) \\ \{a, b', c, r^2_{\neg a \vee b}, r^3_{\neg c}, size(2,1), size(3,1)\} & (4) \\ \{a, b, c', r^2_{\neg a \vee c}, r^3_{\neg b}, size(2,1), size(3,1)\} & (4) \\ \{a, b, c, r^3_{\neg b}, r^3_{\neg c}, size(3,2)\} & (2) \end{array}$$

The stable model with the minimal weight is $\{a, b, c, r^3_{\neg b}, r^3_{\neg c}, size(3,2)\}$ which corresponds to the Removed Set $\{\neg b, \neg c\}$ of E according to the *lexipref* strategy.

[4] Variables are represented by words starting by an upper-case letter.

6 Conclusion

We provide a framework, based on Removed Sets Fusion, for merging belief bases where preferences are expressed among belief bases. We propose a new prioritized merging operation, $\Delta^{RSF}_{lexipref,IC}$, which is equivalent to the iterated removed sets revision operation $\Delta^{PRSF}_{\beta,IC}$ and we show that this operation satisfies most of the postulates proposed for prioritized merging in [12]. We provide an implemementation of prioritized merging thanks to answer sets programming. We have now to compare PRSF with the already proposed prioritized merging operations stemming from lexicographic preorder [4], [18]. Moreover an experimental study has to be conducted in order to evaluate the behaviour of the $\Delta^{RSF}_{lexipref,IC}$.

Acknowledgements

Work partially supported by the European Community under project VENUS (Contract IST-034924) of the "Information Society Technology (IST) program of the 6th FP of RTD". The authors are solely responsible for the contents of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data therein.

References

1. Anger, C., Konczak, K., Linke, T.: Nomore: Non-monotonic reasoning with logic programs. In: Flesca, S., Greco, S., Leone, N., Ianni, G. (eds.) JELIA 2002. LNCS (LNAI), vol. 2424, pp. 521–524. Springer, Heidelberg (2002)
2. Baral, C., Brewka, G., Schlipf, J. (eds.): LPNMR 2007. LNCS (LNAI), vol. 4483. Springer, Heidelberg (2007)
3. Baral, C., Kraus, S., Minker, J., Subrahmanian, V.S.: Combining knowledge bases consisting of first order theories. In: Raś, Z.W., Zemankova, M. (eds.) ISMIS 1991. LNCS, vol. 542, pp. 92–101. Springer, Heidelberg (1991)
4. Benferhat, S., Cayrol, C., Dubois, D., Lang, J., Prade, H.: Inconsistency management and prioritized syntax-based entailment. In: Proc. of IJCAI 1993, pp. 640–647 (1993)
5. Benferhat, S., Dubois, D., Kaci, S., Prade, H.: Possibilistic Merging and Distance-based Fusion of Propositional Information. AMAI 34(1-3), 217–252 (2002)
6. Benferhat, S., Dubois, D., Prade, H.: Reasoning in inconsistent stratified knowledge bases. In: Proc. of ISMVL 1996, pp. 184–189 (1996)
7. Benferhat, S., Dubois, D., Prade, H., Williams, M.-A.: A practical approach to revising prioritized knowledge bases. In: KES, pp. 170–173 (1999)
8. Benferhat, S., Kaci, S.: Logical representation and fusion of prioritized information based on guaranteed possibility measures: Application to the distance-based merging of classical bases. Artif. Intell. 148(1-2), 291–333 (2003)
9. Benferhat, S., Kaci, S., Le Berre, D., Williams, M.-A.: Weakening conflicting information for iterated revision and knowledge integration. Artif. Intell. 153(1-2), 339–371 (2004)
10. Bloch, I., Lang, J.: Towards mathematical morpho-logics. In: Technologies for constructing intelligent systems: tools, pp. 367–380. Springer, Heidelberg (2002)

11. Cholvy, L.: Reasoning about merging information. In: Handbook of DRUMS, vol. 3, pp. 233–263 (1998)
12. Delgrande, J., Dubois, D., Lang, J.: Iterated revision as prioritized merging. In: Proc. of KR 2006, pp. 210–220 (2006)
13. Dubois, D., Lang, J., Prade, H.: Possibilistic Logic. In: Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 3, pp. 439–513 (1994)
14. Eiter, T., Leone, N., Mateis, C., Pfeifer, G., Scarcello, F.: The kr system dlv: progress report, comparison and benchmarks. In: Proc. of KR 1998, pp. 406–417 (1998)
15. Gelfond, M., Lifschitz, V.: The stable model semantics for logic programming. In: Proc. of the 5th Int. Conf. on Log. Prog., pp. 1070–1080 (1988)
16. Giunchiglia, E., Lierler, Y., Maratea, M.: Sat-based answer set programming. In: Proc. of AAAI 2004, pp. 25–29 (2004)
17. Hue, J., Papini, O., Würbel, E.: Syntactic propositional belief bases fusion with removed sets. In: Mellouli, K. (ed.) ECSQARU 2007. LNCS (LNAI), vol. 4724, pp. 66–77. Springer, Heidelberg (2007)
18. Hunter, A., Liu, W.: Knowledge base stratification and merging based on degree of support. In: Sossai, C., Chemello, G. (eds.) ECSQARU 2009. LNCS, vol. 5590, pp. 383–395. Springer, Heidelberg (2009)
19. Hue, J., Papini, O., Würbel, E.: Removed sets fusion: Performing off the shelf. In: Proc. of ECAI 2008. Frontiers in Artificial Intelligence and Applications, pp. 94–98. IOS Press, Amsterdam (2008)
20. Konieczny, S.: On the difference between merging knowledge bases and combining them. In: Proc of KR 2000, pp. 135–144. Morgan Kaufmann, San Francisco (2000)
21. Konieczny, S., Lang, J., Marquis, P.: Distance-based merging: A general framework and some complexity results. In: Proc. of KR 2002, pp. 97–108 (2002)
22. Konieczny, S., Pino Pérez, R.: On the logic of merging. In: Proc. of KR 1998, pp. 488–498 (1998)
23. Lafage, C., Lang, J.: Logical representation of preferences for group decision making. In: Proc. of KR 2000, pp. 457–468. Morgan Kaufmann, San Francisco (2000)
24. Lin, F., Zhao, Y.: Assat: computing answer sets of a logic program by sat solvers. Artif. Intell. 157(1-2), 115–137 (2004)
25. Meyer, T., Ghose, A., Chopra, S.: Syntactic representations of semantic merging operations. In: Proc. of Workshop IDK. IJCAI 2001, pp. 36–42 (2001)
26. Niemelä, I., Simons, P.: An implementation of stable model and well-founded semantics for normal logic programs. In: Fuhrbach, U., Dix, J., Nerode, A. (eds.) LPNMR 1997. LNCS, vol. 1265, pp. 420–429. Springer, Heidelberg (1997)
27. Qi, G.: A semantic approach for iterated revision in possibilistic logic. In: Proc. of AAAI 2008, pp. 523–528 (2008)
28. Rao, P., Sagonas, K., Swift, T., Warren, D.S., Friere, J.: Xsb: A system for efficiently computing well-founded semantics. In: Fuhrbach, U., Dix, J., Nerode, A. (eds.) LPNMR 1997. LNCS, vol. 1265, pp. 430–440. Springer, Heidelberg (1997)
29. Revesz, P.Z.: On the semantics of arbitration. Journ. of Alg. and Comp. 7(2), 133–160 (1997)
30. Fagin, R., Kuper, G.M., Ullman, J.D., Vardi, M.Y.: Updating logical databases. In: Advances in computing research, pp. 1–18 (1986)
31. Simons, P.: Extending and implementing the stable model semantics (2000)
32. Würbel, E., Jeansoulin, R., Papini, O.: Revision: An application in the framework of gis. In: Proc. of KR 2000, pp. 505–516 (2000)
33. Yue, A., Liu, W., Hunter, A.: Approaches to constructing a stratified merged knowledge base. In: Mellouli, K. (ed.) ECSQARU 2007. LNCS (LNAI), vol. 4724, pp. 54–65. Springer, Heidelberg (2007)

An Interactive Algorithm to Deal with Inconsistencies in the Representation of Cardinal Information

Brice Mayag[1], Michel Grabisch[1], and Christophe Labreuche[2]

[1] University of Paris 1, 106-112 Boulevard de l'Hôpital, 75013 Paris, France
bmayag@yahoo.fr, michel.grabisch@univ-paris1.fr
[2] T.R.T France, 1 avenue Augustin Fresnel, 91767 Palaiseau Cedex France
christophe.labreuche@thalesgroup.com

Abstract. We present a new interactive algorithm allowing to solve the inconsistencies problem, when the preferences of a decision maker cannot be representable by a numerical function. This algorithm is based on technics of linear programming and the type of preferences we use are cardinal information.

Keywords: Decision making, Preference modeling, Cardinal information, Inconsistencies.

1 Introduction

Decision making aims at helping a decision maker (DM) to select one or more alternatives among several alternatives. During this process, and in many situations, it is important for the DM to construct a preference relation over the set of all alternatives X. Many models have been developed to construct this preference. Some of them, like utility theory, look for a numerical function with good properties (arithmetic mean, Choquet integral, belief functions, ...) which is able to represent faithfully the preferences of the DM on X. This representation requires sometimes to ask to the DM an initial preference on X or when X is very large, a preferential information on a reference subset $X' \subseteq X$.

In this paper, we ask the DM to give, using pairwise comparisons, a cardinal information (a preferential information given with preference intensity) on X and then we test if this preferential information is consistent with a numerical function. If the test leads to inconsistencies, how to help the DM to modify his preferences in order to represent his cardinal information? To answer this question it is desirable to have recommendations understandable by any DM. This is not true with the different theorems [10,11,12,7] on the representation of cardinal information by a numerical function. Indeed, these characterizations are based on the notion of cyclones [5,13] (specific cycles) which is very complex to understand and to detect. Therefore, an alternative to these theorems is to use methods of dealing with inconsistencies based on technics of linear programming [2,8,9].

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 148–157, 2010.

We propose a new interactive algorithm for inconsistency management with cardinal information. Our approach is not in the spirit of the determination of an irreducible inconsistent system (ISS) [3], but to use simple and intuitive methods of constraints relaxation when a linear program is infeasible. The recommendations we suggest to the DM are based on the concepts of augmentation and reduction of a preference able to causing an inconsistency.

The paper is organized as follows: the next section introduces the basic notions we need, then we present in Section 3 our algorithm and we end by an illustrative example.

2 Representation of a Cardinal Information

Let X be a finite set of alternatives (or actions, options). We assume that, given two alternatives x and y the DM is able to judge the difference of attractiveness between x and y when he strictly prefers x to y. Like in the MACBETH [1,4] and GRIP [6] methodologies in Muticriteria Decision Analysis, the difference of attractiveness will be provided under the form of semantic categories d_s, $s = 1, \ldots, q$ defined so that, if $s < t$, any difference of attractiveness in the class d_s is smaller than any difference of attractiveness in the class d_t. MACBETH approach uses the following six semantic categories: d_1 = very weak, d_2 = weak, d_3 = moderate, d_4 = strong, d_5 = very strong, d_6 = extreme. If there is no ambiguity, a category d_s will be simply designated by s.

Under these hypotheses, the preferences given by the DM is expressed by the following relations:

- $P = \{(x, y) \in X \times X$: the DM strictly prefers x to $y\}$, P is an asymmetric relation;
- $I = \{(x, y) \in X \times X$: the DM is indifferent between x and $y\}$, I is an reflexive and symmetric relation;
- For the semantic categories "d_s", "d_t", $s, t \in \{1, ..., q\}$, $s \leq t$, $P_{st} = \{(x, y) \in P$ such that the DM judges the difference of attractiveness between x and y as belonging from the class "d_s" to the class "d_t" $\}$. When $s < t$, P_{st} expresses some hesitation.

Remark 1. In this paper, the relation $P \cup I$ is not necessarily complete.

Definition 1. *The cardinal information on X is the structure $\{P, I, \{P_{st}\}_{s \leq t}\}$.*

We will suppose P to be nonempty for any cardinal information $\{P, I, \{P_{st}\}_{s \leq t}\}$ ("non triviality axiom") and $P = \bigcup_{s,t} P_{st}$. Remark that if the DM wants to say that x is strictly preferred to y, but he hesitates completely on the category, then he will write $xP_{1q}y$.

A cardinal information$\{P, I, \{P_{st}\}_{s \leq t}\}$ is said to be *representable by a numerical function* $f : X \to \mathbb{R}_+$ if the following conditions are satisfied: $\forall x, y, z, w \in X$,

$\forall s,t,u,v \in \{1,\ldots,q\}$ such that $u \leq v < s \leq t$,

$$x\ I\ y \Rightarrow f(x) = f(y), \tag{1}$$

$$x\ P\ y \Rightarrow f(x) > f(y), \tag{2}$$

$$\left.\begin{array}{c}(x,y) \in P_{st}\\(z,w) \in P_{uv}\end{array}\right\} \Rightarrow f(x) - f(y) > f(z) - f(w) \tag{3}$$

De Corte proved in [1] that the previous conditions are equivalent to the existence of q thresholds $\sigma_1,\ldots,\sigma_q$ such that:

$$\forall (x,y) \in I : f(x) = f(y), \tag{4}$$

$$\forall s,t \in \{1,\ldots,q\},\ s \leq t,\ \forall (x,y) \in P_{st} : \sigma_s < f(x) - f(y), \tag{5}$$

$$\forall s,t \in \{1,\ldots,q-1\},\ s \leq t,\ \forall (x,y) \in P_{st} : f(x) - f(y) < \sigma_{t+1}, \tag{6}$$

$$0 < \sigma_1 < \sigma_2 < \cdots < \sigma_q \tag{7}$$

Note that in this representation, the relation (2) disappears so that relation P is no more used explicitly. To know if a cardinal information $\{P, I, \{P_{st}\}_{s\leq t}\}$ on X is representable by a function f, we use the following linear program PL_1:

$$\mathrm{PL}_1 \left\{\begin{array}{lll}
\text{Min} & f(x_0) & \\
s.t. & f(x) = f(y), & \forall (x,y) \in I \quad (c1)\\
& \sigma_i + d_{\min} \leq f(x) - f(y), & \forall (x,y) \in P_{ij}, \forall i,j \in \{1,\ldots,q\},\ i \leq j \quad (c2)\\
& f(x) - f(y) \leq \sigma_{j+1} - d_{\min}, & \forall (x,y) \in P_{ij}, \forall i,j \in \{1,\ldots,q-1\},\ i \leq j \quad (c3)\\
& d_{\min} \leq \sigma_1 & (c4)\\
& \sigma_{i-1} + d_{\min} \leq \sigma_i, & \forall i \in \{2,\ldots,q\} \quad (c5)
\end{array}\right.$$

where x_0 is an alternative of X arbitrarily chosen, and $d_{\min}$ an arbitrary strictly positive constant.

Now, when the cardinal information is inconsistent, i.e. the program PL_1 is infeasible, how to elaborate recommendations for the DM in order to have the consistent judgements? A natural solution is to provide these recommendations by using characterization theorems of the representation of a cardinal information studied in [7,11,12]. But, all these theorems are based on the more complex and specific cycle called "cyclone" [5], which would be difficult to grasp for a DM. Our aim is to propose a new interactive method able to generate recommendations for the DM when PL_1 is infeasible.

3 Our Algorithm

3.1 Step 1: Find the Minimal Number of Constraints to Be Relaxed

To make PL_1 feasible, we choose to relax some of its constraints which can cause an inconsistency. To do this, we associate to each constraint l of PL_1, a binary variable β_l allowing to know whether if the constraint l has to be relaxed or not. The options x and y in constraint l are denoted by x_l and x'_l. Then we find the minimal number of constraints which we will relax by solving the following linear program PL_2:

$$\mathrm{PL}_2 \begin{cases} \text{Min} \sum\limits_{l \in \mathbb{N}_{1,c}} \beta_l & & \\ s.t. \quad f(x_l) - f(x'_l) + M\beta_l \geq 0, & \forall (x_l, x'_l) \in I,\ l \in \mathbb{N}_{1,r^+} & (c1')_1 \\ \quad f(x_l) - f(x'_l) - M\beta_l \leq 0, & \forall (x_l, x'_l) \in I,\ l \in \mathbb{N}_{(r^+ +1),r} & (c1')_2 \\ \quad \sigma_i + d_{\min} \leq f(x_l) - f(x'_l) + M\beta_l, & \forall (x_l, x'_l) \in P_{ij}, \forall i, j \in \mathbb{N}_{1,q}, l \in \mathbb{N}_{(r+1),(r+p_1)} & (c2') \\ \quad f(x_l) - f(x'_l) - M\beta_l \leq \sigma_{j+1} - d_{\min}, & \forall (x_l, x'_l) \in P_{ij}, \forall i, j \in \mathbb{N}_{1,q-1}, l \in \mathbb{N}_{r+p_1+1,c} & (c3') \\ \quad d_{\min} \leq \sigma_1 & & (c4) \\ \quad \sigma_{i-1} + d_{\min} \leq \sigma_i, & \forall i \in \mathbb{N}_{2,q} & (c5) \\ \quad \beta_l \in \{0, 1\}, & \forall l \in \mathbb{N}_{1,c} & (c6) \end{cases}$$

where

- each constraint $f(x) - f(y) = 0$ of PL_1 is replaced in PL_2 by the following two constraints:
  (i) $f(x_l) - f(x'_l) + M\beta_l \geq 0,\ 1 \leq l \leq r^+$ (type $(c1')_1$);
  (ii) $f(x_{l'}) - f(x'_{l'}) - M\beta_{l'} \leq 0,\ r^+ \leq l' \leq r$ (type $(c1')_2$);
  such that $x_l = x_{l'}$ and $x'_l = x'_{l'}$. It is obvious that these two inequalities are always satisfied when $\beta_l = \beta_{l'} = 1$.
- M is a positive large number.
- $r = r^+ + r^-$ with respectively r^+ and r^- the number of constraints of $(c1')_1$ and $(c1')_2$. $r^+ = r^-$ is the number of constraints of type $(c1)$.
- p_1: the number of constraints of type $(c2)'$ corresponding to the number of constraints of $(c2)$ in PL_1.
- p_2: the number of constraints of $(c3)'$ corresponding to the number of constraints of $(c3)$ in PL_1.
- $c = r + p_1 + p_2$;
- $\forall s, t \in \mathbb{N},\ s \leq t,\ \mathbb{N}_{s,t} = \{s, s+1, \ldots, t\}$.

3.2 Step 2: Relaxation by Augmentation or Reduction by p Categories

In this section, we show how to relax each constraint which has its binary variable β_l equals to 1. We suggest two types of relaxation: an increase or decrease of categories and we justify this by the following:

1. Suppose that a preference $(x, y) \in P_{ij}$ causes an inconsistency in PL_1. If the modification of this judgement can restore the consistency, it seems natural to ask the DM to adopt one of these two recommendations:
   - If $(x, y) \in P_{ij}$ belongs to the set of constraints of $(c3')$, increase the category j by replacing this preference by $(x, y) \in P_{ij'}$ with $j < j'$;
   - If $(x, y) \in P_{ij}$ belongs to the set of constraints of $(c2')$, decrease the category i by replacing this preference by $(x, y) \in P_{i'j}$ with $i' < i$.
2. If an indifférence (x_l, x'_l) causes an inconsistency, then PL_2 satisfies either $f(x_l) - f(x'_l) < 0$ or $f(x_l) - f(x'_l) > 0$ (corresponding to $\beta_l = 1$). Therefore if the inequality $f(x_l) - f(x'_l) > 0$ is satisfied in PL_2, we recommend to the DM to change $(x_l, x'_l) \in I$ by $(x_l, x'_l) \in P_{1p}$ where p will be a category to be determined.

We need the following notation in the formal Definition 2 of relaxation by augmentation or reduction by p categories:

- the judgement "$(x, y) \in P_{ij}$" will be represented by the element (x, y, i, j) of $X \times X \times \mathbb{N}_{1,q} \times \mathbb{N}_{1,q}$;
- the judgement "$(x, y) \in I$" will be represented by the element $(x, y, 0, 0)$ of $X \times X \times \mathbb{N} \times \mathbb{N}$.

Definition 2.

1. *"*A reduction of the judgement (x, y, i, j) with p categories*" is the substitution of this judgement by:*
   (a) the judgement $(x, y, i - p, j)$ if $(1 \leq p < i)$;
   (b) the judgement $(y, x, 1, p)$ if $i = j = 0$.
2. *"*An augmentation of the judgement (x, y, i, j) of p categories*" $(1 \leq p \leq q - j)$ is the substitution of this judgement by:*
   (a) the judgement $(x, y, 1, p)$ if $i = j = 0$;
   (b) the judgement $(x, y, i, j + p)$ otherwise.

Using the previous notions, we distinguish two cases:

(i) The judgement is $(x_l, x'_l) \in I$:
- If the binary variable $\beta_l = 1$ of PL_2 is associated to the constraint $f(x_l) - f(x'_l) + M\beta_l \geq 0$ derived from the preference $(x_l, x'_l) \in I$, then the corresponding relaxation is a reduction of the judgement by p categories. We denote by C_{1-1} the set of all "l" satisfying these conditions.
- On the other side, if $\beta_l = 1$ is associated to the constraint $f(x_l) - f(x'_l) - M\beta_l \leq 0$ with $(x_l, x'_l) \in I$, then the proposition of relaxation will be an augmentation of the judgement by p categories. We denote by C_{1-2} the set of all "l" satisfying this type of conditions.

(ii) The judgement is $(x_l, x'_l) \in P_{ij}$:
- If the binary variable $\beta_l = 1$ of PL_2 is associated to the constraint $\sigma_i + d_{\min} \leq f(x_l) - f(x'_l) + M\beta_l$, then we apply the reduction of the judgement. Let us denote by C_2 the set of all "l" satisfying this type of conditions.
- If $\beta_l = 1$ corresponds to $f(x_l) - f(x'_l) - M\beta_l \leq \sigma_{j+1} - d_{\min}$, then we recommend an augmentation of judgement. Let C_3 the set of all "l" satisfying these type of conditions.

3.3 Step 3: Determination of the Number of Categories p Used in the Relaxation

In this section, we suppose the set M of m constraints, which can cause an inconsistency, have been determined through the linear program PL_2. To know for each element l of M, the number of categories necessary for its relaxation by augmentation or reduction, we introduce the binary variables ε_k^l as follows:

1. If the modification of the preference $(x_l, x'_l) \in I$ requires an augmentation of categories, we replace in PL_1 the constraint $f(x_l) - f(x'_l) = 0$ associated to this judgement by the following constraints:

$$\begin{cases} f(x_l) - f(x'_l) \geq 0 \\ f(x_l) - f(x'_l) \leq \sigma_k - d_{\min} + M\ \varepsilon_k^l, \quad \forall k \in \mathbb{N}_{2,q} \end{cases} \tag{8}$$

   Let $h = \sum_{k \in \mathbb{N}_{2,q}} \varepsilon_k^l$.
   - If $h < q-1$, then we recommend to the DM an augmentation of $(x_l, x'_l) \in I$ with $(h+1)$ categories;
   - Otherwise, we suggest him to remove the judgement $(x_l, x'_l) \in I$ in the cardinal information.
2. If the judgement $(x_l, x'_l) \in P_{ij}$ requires an augmentation of categories, we replace in PL_1 the constraint $f(x_l) - f(x'_l) \leq \sigma_{j+1} - d_{\min}$ associated to this preference by the constraints

$$f(x_l) - f(x'_l) \leq \sigma_{j+k} - d_{\min} + M\ \varepsilon_k^l, \quad \forall k \in \mathbb{N}_{2,q-j} \tag{9}$$

   Let $h = \sum_{k \in \mathbb{N}_{2,q-j}} \varepsilon_k^l$.
   - If $h < q-j-1$, we recommend to the DM an augmentation of $(x_l, x'_l) \in P_{ij}$ by $(h+1)$ categories;
   - Otherwise, we propose him to remove the judgement $(x_l, x'_l) \in P_{ij}$.
3. If the preference $(x_l, x'_l) \in I$ requires a reduction of categories, we replace in PL_1 the corresponding constraint $f(x_l) - f(x'_l) = 0$ by the following constraint

$$\begin{cases} f(x'_l) - f(x_l) \geq 0 \\ f(x'_l) - f(x_l) \leq \sigma_k - d_{\min} + M\ \varepsilon_k^l, \quad \forall k \in \mathbb{N}_{2,q} \end{cases} \tag{10}$$

   Let $h = \sum_{k \in \mathbb{N}_{2,q}} \varepsilon_k^l$.
   - If $h < q-1$, we suggest to the DM a reduction of $(x_l, x'_l) \in I$ by $(h+1)$ categories;
   - Otherwise, we suggest to remove the preference $(x_l, x'_l) \in I$.
4. If the judgement $(x_l, x'_l) \in P_{ij}$ requires a reduction of categories, we replace in PL_1 the corresponding constraint $\sigma_i + d_{\min} \leq f(x_l) - f(x'_l)$ by the following:

$$\sigma_{i-k} + d_{\min} \leq f(x_l) - f(x'_l) + M\ \varepsilon_k^l, \quad \forall k \in \mathbb{N}_{1,i-1} \tag{11}$$

   Let $h = \sum_{k \in \mathbb{N}_{1,i-1}} \varepsilon_k^l$.
   - If $h < i-1$, we recommend to the DM a reduction of $(x_l, x'_l) \in P_{ij}$ with $(h+1)$ categories;
   - Otherwise, suggest him to remove $(x_l, x'_l) \in P_{ij}$.

The binary variables ε_k^l introduced in equations (8) to (11) are determined by the following linear program PL_3:

$$\mathrm{PL}_3 \begin{cases} \text{Min} \displaystyle\sum_{k\in\mathbb{N}_{2,q}, l\in C_{1-1}} \varepsilon_k^l + \sum_{k\in\mathbb{N}_{1,q-j}, l\in C_2} \varepsilon_k^l + \sum_{k\in\mathbb{N}_{2,q}, l\in C_{1-2}} \varepsilon_k^l + \sum_{k\in\mathbb{N}_{1,i-1}, l\in C_3} \varepsilon_k^l \\ \text{s.t. } \mathrm{S}_{\text{Relaxed constraints}} \\ \mathrm{S}_{\mathrm{PL}_1} \setminus \mathrm{S}_{\mathrm{PL}_1(m)} \\ \varepsilon_k^l \in \{0,1\}, \forall k \in \mathbb{N}_{2,q}, l \in C_{1-1} \\ \varepsilon_k^l \in \{0,1\}, \forall k \in \mathbb{N}_{1,q-j}, l \in C_2 \\ \varepsilon_k^l \in \{0,1\}, \forall k \in \mathbb{N}_{2,q}, l \in C_{1-2} \\ \varepsilon_k^l \in \{0,1\}, \forall k \in \mathbb{N}_{1,i-1}, l \in C_3 \end{cases}$$

where

- $\mathrm{S}_{\mathrm{PL}_1}$ represents all the constraints of PL_1.
- $\mathrm{S}_{\mathrm{PL}_1(m)}$ represents a subset of $\mathrm{S}_{\mathrm{PL}_1}$ formed by the constraints associated to the constraints of M build by PL_2 that cause an inconsistency.
- $\mathrm{S}_{\text{Relaxed constraints}}$ represents the system formed by all the constraints introduced in (8) to (11).

3.4 Step 4: The Interaction with the DM

We have seen in the previous sections that, if the cardinal information given by the DM is inconsistent, then the linear program PL_3 is solved and its solution is presented to the DM as recommendations to repair the inconsistencies. Therefore, for each judgement $(x_l, x'_l) \in (P \cup I)$ causing an inconsistency, we suggest him an augmentation or reduction of categories to make consistent judgements representable by a numerical function f. Let R be the set of recommendations (judgements with augmentation or reduction) proposed to the DM. In our algorithm, the DM can adopt one of these two positions:

1. *the DM does not agree with the recommendations proposed.*
   He builds a subset $R' \subseteq R$ of judgements for which he decides to conserve his initial judgement. For each element (x_l, x'_l) of R', the DM has no intention to relax the constraint corresponding to (x_l, x'_l) in PL_1. Therefore, the linear program PL_2 will be launched again by considering these constraints as satisfied constraints (by removing their binary variables β_l). There are two possibilities:
   (a) PL_2 with these constraints has a solution. Then we compute the new recommendations. We are thus either in situation 1 or 2.
   (b) PL_2 with these constraints has no solution. This means that the DM cannot conserve R' since they are inconsistent. He needs thus to change R'.
2. *the DM agrees with the recommendation proposed.*
   The linear program PL_1 is launched again by taking into account the new consistent cardinal information given by the DM.

The algorithm is represented by the Figure 1.

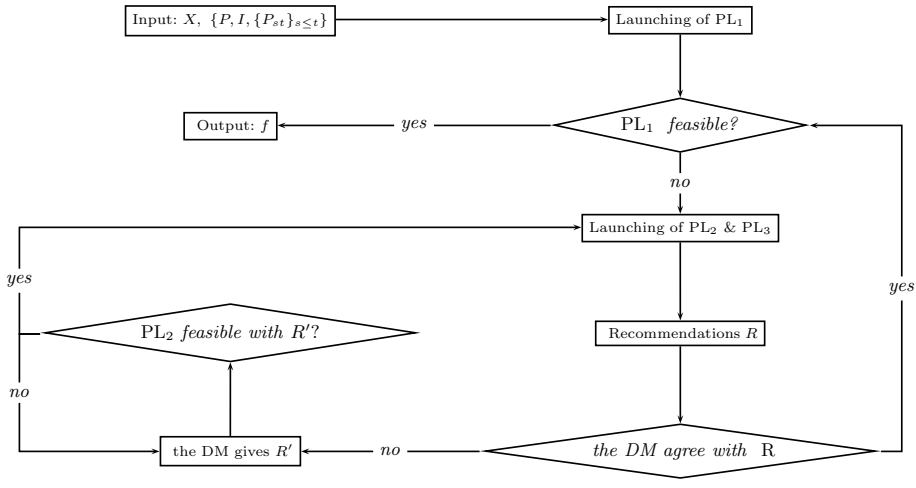


Fig. 1. Interactive algorithm of dealing with inconsistencies

4 An Illustrative Example

$X = \{x_1;\ x_2;\ x_3;\ x_4;\ x_5;\ x_6\}$; $q = 6$. Suppose that the DM gives the following preferences: $I = \{(x_2, x_3); (x_1, x_6)\}$; $P_3 = \{(x_5, x_6)\}$; $P_{12} = \{(x_1, x_3)\}$; $P_{24} = \{(x_1, x_5)\}$; $P_{46} = \{(x_3, x_5)\}$. The consistency of the cardinal information $\{I,\ P_3,\ P_{12},\ P_{24},\ P_{46}\}$ is tested through the linear program PL_1:

$$
PL_1 \begin{cases}
\text{Min } f(x_1) \\
\text{s.t.} \\
f(x_2) - f(x_3) = 0 \\
f(x_1) - f(x_6) = 0 \\
\sigma_3 + d_{\min} \leq f(x_5) - f(x_6) \\
\sigma_1 + d_{\min} \leq f(x_1) - f(x_3) \\
\sigma_2 + d_{\min} \leq f(x_1) - f(x_5) \\
\sigma_4 + d_{\min} \leq f(x_3) - f(x_5) \\
f(x_5) - f(x_6) \leq \sigma_4 - d_{\min} \\
f(x_1) - f(x_3) \leq \sigma_3 - d_{\min} \\
f(x_1) - f(x_5) \leq \sigma_5 - d_{\min} \\
d_{\min} \leq \sigma_1, \quad \sigma_i + d_{\min} \leq \sigma_{i+1}, i = 1, \ldots, 5
\end{cases}
$$

For this test, we set $d_{\min} = 0.001$ and we get PL_1 infeasible. Therefore the algorithm launches the linear program PL_2 in order to find the minimal number of constraints to be relaxed:

$$\mathrm{PL}_2 \begin{cases} \text{Min } \sum_{l=1}^{11} \beta_l \\ \text{s.t.} \\ f(x_2) - f(x_3) + M\beta_1 \geq 0 \\ f(x_1) - f(x_6) + M\beta_2 \geq 0 \\ f(x_2) - f(x_3) - M\beta_3 \leq 0 \\ f(x_1) - f(x_6) - M\beta_4 \leq 0 \\ \sigma_3 + d_{\min} \leq f(x_5) - f(x_6) + M\beta_5 \\ \sigma_1 + d_{\min} \leq f(x_1) - f(x_3) + M\beta_6 \\ \sigma_2 + d_{\min} \leq f(x_1) - f(x_5) + M\beta_7 \\ \sigma_4 + d_{\min} \leq f(x_3) - f(x_5) + M\beta_8 \\ f(x_5) - f(x_6) - M\beta_9 \leq \sigma_4 - d_{\min} \\ f(x_1) - f(x_3) - M\beta_{10} \leq \sigma_3 - d_{\min} \\ f(x_1) - f(x_5) - M\beta_{11} \leq \sigma_5 - d_{\min} \\ d_{\min} \leq \sigma_1, \quad \sigma_i + d_{\min} \leq \sigma_{i+1}, i = 1, \ldots, 5 \\ \beta_l \in \{0, 1\}, \forall l \in \{1, \ldots, 11\} \end{cases}$$

The solution gives $\beta_4 = 1$ and $\beta_l = 0$ for $l \neq 4$. So the only constraint which need to be relaxed is $f(x_1) - f(x_6) = 0$ and its relation corresponds to an augmentation of p categories of the judgement $(x_1, x_6) \in I$. The number p is given by PL_3:

$$\mathrm{PL}_3 \begin{cases} \text{Min } \sum_{l=2}^{6} \varepsilon_l \\ \text{s.t.} \\ f(x_2) - f(x_3) = 0 \\ f(x_1) - f(x_6) \geq 0 \\ f(x_1) - f(x_6) \leq \sigma_2 - d_{\min} - M\varepsilon_2 \\ f(x_1) - f(x_6) \leq \sigma_3 - d_{\min} - M\varepsilon_3 \\ f(x_1) - f(x_6) \leq \sigma_4 - d_{\min} - M\varepsilon_4 \\ f(x_1) - f(x_6) \leq \sigma_5 - d_{\min} - M\varepsilon_5 \\ f(x_1) - f(x_6) \leq \sigma_6 - d_{\min} - M\varepsilon_6 \\ \sigma_3 + d_{\min} \leq f(x_5) - f(x_6) \\ \sigma_1 + d_{\min} \leq f(x_1) - f(x_3) \\ \sigma_2 + d_{\min} \leq f(x_1) - f(x_5) \\ \sigma_4 + d_{\min} \leq f(x_3) - f(x_5) \\ f(x_5) - f(x_6) \leq \sigma_4 - d_{\min} \\ f(x_1) - f(x_3) \leq \sigma_3 - d_{\min} \\ f(x_1) - f(x_5) \leq \sigma_5 - d_{\min} \\ d_{\min} \leq \sigma_1, \quad \sigma_i + d_{\min} \leq \sigma_{i+1}, i = 1, \ldots, 5 \\ \varepsilon_l \in \{0, 1\}, \forall l \in \{2, \ldots, 6\} \end{cases}$$

A solution of PL_3 gives $\varepsilon_2 = \varepsilon_3 = \varepsilon_4 = 1$ and $\varepsilon_5 = \varepsilon_6 = 0$. Therefore we suggest to the DM an augmentation of $(x_1, x_6) \in I$ by 4 categories, i.e. replace $(x_1, x_6) \in I$ by $(x_1, x_6) \in P_{14}$. The DM accepts this recommendation and the

new cardinal information $I = \{(x_2, x_3)\}$; $P_3 = \{(x_5, x_6)\}$; $P_{12} = \{(x_1, x_3)\}$; $P_{14} = \{(x_1, x_6)\}$; $P_{24} = \{(x_1, x_5)\}$; $P_{46} = \{(x_3, x_5)\}$ becomes consistent.

References

1. Bana e Costa, C.A., De Corte, J.-M., Vansnick, J.-C.: On the mathematical foundations of MACBETH. In: Figueira, J., Greco, S., Ehrgott, M. (eds.) Multiple Criteria Decision Analysis: State of the Art Surveys, pp. 409–437. Springer, Heidelberg (2005)
2. Chinneck, J.W.: Computer codes for the analysis of infeasible linear programs. Journal of the Operational Research Society 47 (1996)
3. Chinneck, J.W.: An effective polynomial-time heuristic for the minimum cardinality IIS set-covering problem. Annals of Mathematics and Artificial Intelligence 17 (1996)
4. De Corte, J.M.: Un logiciel d'Exploitation d'Information Préférentielles pour l'Aide à la Décision. Bases Mathématiques et Algorithmiques. PhD thesis, University of Mons-Hainaut, Mons (2002)
5. Doignon, J.P.: Thresholds representation of multiple semiorders. SIAM Journal on Algebraic and Discrete Methods 8, 77–84 (1987)
6. Figueira, J.R., Greco, S., Slowinski, R.: Building a set of additive value functions representing a reference preorder and intensities of preference: Grip method. European Journal of Operational Research 195(2), 460–486 (2009)
7. Mayag, B., Grabisch, M., Labreuche, C.: A characterization of the 2-additive Choquet integral through cardinal information. In: Proceedings of EUROFUSE2009 Workshop Preference Modelling and Decision Analysis (CD), Pamplona, Spain, September 16-18 (2009)
8. Mousseau, V., Dias, L., Figueira, J.R.: Dealing with inconsistent judgments in multiple criteria sorting models. 4OR, 4 (2006)
9. Mousseau, V., Figueira, J.R., Dias, L., Da Silva, C.G., Climaco, J.: Resolving inconsistencies among constraints on the parameters of an MCDA model. European Journal of Operational Research 147 (2003)
10. Mousset, C.: Familles de structures de préférence non complètes. PhD thesis, University of Mons-Hainaut, Belgium (2004)
11. Mousset, C.: Families of relations modelling preferences under incomplete information. European journal of operational research 192(2), 538–548 (2009)
12. Mousset, C., Vansnick, J.-C.: About the representation of a precardinal information. Documents d'Economie et de Gestion de l'Université de Mons-Hainaut (2005)
13. Pirlot, M., Vincke, P.: Semiorders — Properties, Representations, Applications. Kluwer Academic Publishers, Dordrecht (1997)

Characterization of Complete Fuzzy Preorders Defined by Archimedean t-Norms

Ignacio Montes, Davide Martinetti, Susana Díaz, and Susana Montes

Dept. of Statistics and O.R.
University of Oviedo, Spain
imontes@spi.uniovi.es, {diazsusana,martinettidavide.uo,montes}@uniovi.es
http://eio.epv.uniovi.es/diedra/

Abstract. Classical complete preorders can be characterized in several ways. However, when we work with complete fuzzy preorders this equivalences do not hold in general. In previous works we have proven some connections among them when using the minimum and the Łukasiewicz t-norms. In this contribution we generalize the study and we work with two important families (nilpotent and strict t-norms) when defining the fuzzy counterparts of the characterizations of a crisp complete preorder.

1 Introduction

In preference modeling the comparison of a set of alternatives is usually carried out by pairs. The opinion of a decision maker over the set of alternatives can be formalized by a binary relation usually denoted by R and called *weak preference relation*. It is understood as follows: the alternative a is considered at least as good as b if aRb.

Starting from every weak preference relation we can define a preference structure. It is a set of three binary relations that also cover all the possible answers of a decision maker: a symmetric relation identified with the indifference, an asymmetric relation that expresses the strict preference and the dual symmetric relation that holds when the decision maker cannot compare the alternatives.

This last relation is called incomparability relation and in some cases it is not allowed, the decision maker is forced to compare every pair of alternatives. In these cases the weak preference relation is said to be complete, since it connects every pair of alternatives.

Different properties are defined in order to state that the answers of a decision maker are coherent. Maybe transitivity is the most important one. A complete preorder is a complete reflexive relation that is transitive. In this work we focus on these relations and their characterizations. It is well known that the transitivity of a complete weak preference relation is characterized by the transitivity of the associated symmetric and asymmetric relations. This is the best known characterization of a complete preorder, but there are others that we consider in our research.

In [12] six different characterizations of a crisp complete preorder are shown. In previous works (see [5] and [7]) we translated those properties to the fuzzy

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 158–167, 2010.

sets context and we studied whether those new properties characterize a fuzzy complete preorder. We also studied how those relations are connected among them in case they are not related to a fuzzy complete preorder. Those works focus on the minimum and the Łukasiewicz t-norm. The definitions and properties considered for fuzzy relations are based on those two important operators. In this contribution we largely extend the study to other operators. We consider two important families of t-norms: nilpotent and strict t-norms and we compare the six characterizations of a fuzzy complete preorder for those operators.

This communication is organized as follows. In Section 2, we recall the basic ingredients from crisp interval orders and we translate them to fuzzy interval orders. Connections among the different possible definitions for that concept are presented in Section 3. There we present general results for any t-norm, as well as the particular achievements for idempotent, nilpotent and strict t-norms. In the concluding section, we focus our attention in the interpretation and consequences of the obtained results.

2 Basic Concepts

2.1 Crisp Preorders

Every reflexive binary relation R defined on the set A can be seen as a *weak preference relation* if we state that *aRb if and only if a is preferred or indifferent to b*. Given a binary relation R, we will denote by R^t, R^c and R^d the transpose, the complement and the dual relation, respectively. Combining them we can obtain, from any crisp reflexive relation R, three disjoint binary relations: the strict preference relation $P = R \cap R^d$, the indifference relation $I = R \cap R^t$ and the incomparability relation $J = R^c \cap R^d$. If we consider the triplet (P, I, J), that forms a *preference structure*, we can get back the relation R, because $R = P \cup I$. A relation Q on A is said to be *complete* if every pair of elements is connected at least by Q or its transpose, $Q \cup Q^t = A^2$. It is well known that the reflexive relation R is complete if and only if any pair of elements is comparable by means of the relation R, which is equivalent to say that $J = \emptyset$.

Given two relations Q_1 and Q_2 defined on A, their composition, denoted $Q_1 \circ Q_2$, is defined as follows $a(Q_1 \circ Q_2)c \Leftrightarrow aQ_1b \wedge bQ_2c$ for some $b \in A$. Then, the transitivity of a relation Q is equivalent to $Q \circ Q \subseteq Q$.

A binary relation Q is said to be *negatively transitive* if there exists $b \in A$ such that $aQc \Rightarrow (aQb \vee bQc)$. We can represent every binary relation Q defined on a set A as a graph (A, Q), where A is the set of nodes and Q is the set of arcs. So there is an arc from the node a to the node b if and only if aQb and it is represented by (a, b). A *path* of length n in such a graph is a set of n arcs $(a_0, a_1), ..., (a_{n-1}, a_n)$ in (A, Q). A *circuit*, sometimes also called a *cycle*, is a path for which $a_0 = a_n$.

It is well known that the transitivity of a complete reflexive relations R can be characterized in several ways. In particular, the following statements are equivalent [12]:

(i) R is transitive,
(ii) P and I are transitive,
(iii) P is transitive and $P \circ I \subseteq P$,
(iv) P is transitive and $I \circ P \subseteq P$,
(v) P is negatively transitive,
(vi) there is no P in circuits of length ≤ 3 in (A, R).

Note that for any reflexive relation R, the relation P is irreflexive and asymmetric, so there is no P in circuits of length 1 or 2 in (A, R). Then, Property (vi) can be written as *"there is no P in circuits of length 3 in (A, R)"*.

A complete reflexive relation R is a *preorder* if it satisfies the transitive property. Thus, given a complete reflexive relation, the six conditions above provide six different ways of expressing a preorder.

2.2 Fuzzy Binary Relations

Human beings do not express their preferences in a rigid way. When comparing two alternatives, expressions like "a little bit preferred", "quite preferred" are common. Fuzzy relations appeared as a way to catch these intermediate answers. They are $A^2 \to [0,1]$ applications and they generalize crisp relations. The image expresses the strength of the connection between the alternatives. The greatest the value, the strongest the connection. We will keep the same notation as for crisp relations for the transpose, complementary and dual of a fuzzy relation R:

$$R^t(a,b) = R(b,a) \quad R^c(a,b) = 1 - R(a,b) \quad R^d(a,b) = 1 - R(b,a).$$

The intersection and union of fuzzy relations are usually defined by t-norms and t-conorms respectively. We recall that a *t-norm* is a $[0,1]^2 \to]0,1]$, commutative, associative and non decreasing mapping with neutral element 1. Three important examples of t-norms are minimum ($T_{\mathbf{M}}(x,y) = \min(x,y)$), product ($T_{\mathbf{P}}(x,y) = x \cdot y$) and Łukasiewicz ($T_{\mathbf{L}}(x,t) = \max(x+y-1,0)$). Given two fuzzy relations, Q_1 and Q_2, their intersection is defined as $Q_1 \cap_T Q_2(a,b) = T(Q_1(a,b), Q_2(a,b))$.

Analogously, the union of fuzzy relations is defined in general by t-conorms. A *t-conorm* S is a commutative, associative, non decreasing, with neutral element 0, mapping from $[0,1]^2$ into $[0,1]$. If we define the dual of a binary operator U by $U^d(x,y) = 1 - (U(1-x, 1-y))$, then t-norms and t-conorms are dual operators. It is well known that for any t-norm T, the dual operator T^d is a t-conorm. It is called *dual t-conorm* of T.

Automorphisms allow to obtain new t-norms from a given one (and new t-conorms from a given t-conorm). A $[0,1]-$automorphism (automorphism, for short), denoted by φ, is a $[0,1] \to [0,1]$ mapping such that $\varphi(0) = 0, \quad \varphi(1) = 1$ and $\quad \varphi(x) < \varphi(y), \ \forall x < y$. The inverse of the automorphism φ is denoted by φ^{-1} and it is again an automorphism. Given a t-norm T and an automorphism φ, the binary operator T_φ is called φ-transform of T and defined by $T_\varphi(x,y) = \varphi^{-1}(T(\varphi(x), \varphi(y)))$ and it is a t-norm. The minimum is the only t-norm invariant by any φ-transformation, i. e. for any automorphism φ, $T_{\mathbf{M}_\varphi} = T_{\mathbf{M}}$.

A t-norm is continuous if it is continuous in each component. This property is preserved by φ-transformation. What is more, every continuous t-norm can be

expressed as the right combination of the three most important t-norms $T_{\mathbf{M}}$, $T_{\mathbf{P}}$ and $T_{\mathbf{L}}$ and their φ-transforms. Next concepts are involved in the classification of continuous t-norms.

We say that a value $x \in (0,1)$ is a *zero-divisor* of the t-norm T if there exists another value $y \in (0,1)$ such that $T(x,y) = 0$. The value $x \in [0,1]$ is an *idempotent element* of T if $T(x,x) = x$. The values 0 and 1 are idempotent elements of every t-norm T. Based on these elements we can define classes of t-norms. Thus, a continuous t-norm T is *Archimedean* if it satisfies that $T(x,x) < x$ for all $x \in (0,1)$. A t-norm T is *nilpotent* if it is continuous and for any $x \in (0,1)$, x is a zero-divisor of T. A t-norm T is *strict* if it is continuous and it is strictly increasing in each component, that is, if $T(x,\cdot)$ is strictly increasing, for all $x \in (0,1]$. A t-norm T is *idempotent* if for all $x \in (0,1)$, x is an idempotent element.

The only idempotent t-norm is the minimum t-norm. On the other hand, continuous Archimedean t-norms can be classified into nilpotent and strict. The most important example of nilpotent t-norm is the Łukasiewicz t-norm, while the most important example of strict t-norm is the product. In fact, every nilpotent and every strict t-norm can be expressed as a φ-transformation of the Łukasiewicz and the product t-norms, respectively (see e. g. [9]). Thus every continuous Archimedean t-norm is either a transformation of the Łukasiewicz t-norm or a transformation of the product t-norm.

A wide study about t-norms and t-conorms can be found in [11]; all the previously known results we will comment in this subsection are collected in this book.

2.3 Fuzzy Preference Structures

In fuzzy set theory, a reflexive fuzzy relation R on A can also be decomposed into the so-called (additive) fuzzy preference structure, by means of a generator i. The concept of generator was introduced by De Baets and Fodor in [1] as a symmetric (commutative) $[0,1]^2 \to [0,1]$ mapping bounded by the Łukasiewicz t-norm, $T_{\mathbf{L}}$, and the minimum operator, $T_{\mathbf{M}}$, i.e. $T_{\mathbf{L}} \leq i \leq T_{\mathbf{M}}$.

Given a reflexive fuzzy relation R and a generator i, the three components of an additive fuzzy preference structure ([13]), in brief AFPS, are defined as follows: $P(a,b) = R(a,b) - i(R(a,b),R(b,a))$, $I(a,b) = i(R(a,b),R(b,a))$ and $J(a,b) = I(a,b) - (R(a,b) + R(b,a) - 1)$. They satisfy the additive property: $P(a,b)+I(a,b)+P^t(a,b)+J(a,b) = 1$ for all $a,b \in A$. The initial weak preference relation R can be rebuilt as follows: $R(a,b) = P(a,b) + I(a,b)$.

The concept of completeness for fuzzy relations is usually defined by a t-conorm. The fuzzy relation Q is S-complete if $S(Q(a,b),Q(b,a)) \geq 1$ for all $a,b \in A$. The most usual completeness conditions considered are the *strong completeness*, defined by the maximum t-conorm and the *weak completeness* defined by the Łukasiewicz t-conorm. In the fuzzy context, the absence of the associated incomparability relation is not equivalent to any completeness condition in general as it can be seen from the following result.

Proposition 1. [8] *Let R be a reflexive fuzzy relation and let J be the incomparability relation associated to R by means of any generator i. Then $J = \emptyset$ if, and only if, R is weakly complete and $i = T_{\mathbf{L}}$.*

Strong completeness is a more restrictive condition than the absence of incomparability relation and this is a stronger property than weak completeness. As we explained at the beginning of this work we focus on additive fuzzy preference structures without incomparability. Therefore, we handle weakly complete reflexive relations R such that the associated additive fuzzy preference structure (P, I) is defined by the Łukasiewicz generator: $(P, I) = (R^d, R \cap_{T_{\mathbf{L}}} R^t)$.

2.4 Fuzzy Preorders

Any t-norm leads to a different definition of the compositions of fuzzy relations, so there is no unique way of defining it. More precisely, for any t-norm T, the T-composition of two fuzzy relations Q_1 and Q_2 on A is defined by $Q_1 \circ_T Q_2(a, c) = \sup_b T\left(Q(a, b), Q(b, c)\right)$. T-norms allow to generalize the definition of transitivity. Thus, Q is called *T-transitive* if $T\left(Q(a, b), Q(b, c)\right) \leq Q(a, c), \forall a, b, c \in A$. As for crisp relations, the $T-$transitivivy of a fuzzy relation is equivalent to $Q \circ_T Q \subseteq Q$.

By using this concept, we are going to recall or establish the fuzzy generalization of the six conditions considered in Subsection 2.1. Thus, given a t-conorm S, the fuzzy relation Q is negatively S-transitive if $Q(a, c) \leq S(Q(a, b), Q(b, c)$ $\forall a, b, c$ (see for example [9]).

Given a, b and c three alternatives, the absence of strict preference in cycles of length 3 in a crisp context states that $aRb \wedge bRc \wedge cRa \quad \Rightarrow \quad a \not P b \wedge b \not P c \wedge c \not P a$.

Since the intersection of two fuzzy relations depends on the chosen t-norm, it is said that a reflexive fuzzy relation R on A has no P in cycles of length 3 if it holds that $T\left(1 - P(a, b), 1 - P(b, c), 1 - P(c, a)\right) \geq T\left(R(a, b), R(b, c), R(c, a)\right)$. Considering the dual t-conorm S, the last expression can also be written as $1 - S\left(P(a, b), P(b, c), P(c, a)\right) \geq T\left(R(a, b), R(b, c), R(c, a)\right)$. In summary, given a t-norm T and its dual t-conorm S, the six characterizations presented in Subsection 2.1 can be written for fuzzy relations as follows:

Number	Description	Formulation $(\forall a, b, c \in A)$
(I)	R is T-transitive	$T(R(a, b), R(b, c)) \leq R(a, c)$
(II)	P is T-transitive I is T-transitive	$T(P(a, b), P(b, c)) \leq P(a, c)$ $T(I(a, b), I(b, c)) \leq I(a, c)$
(III)	P is T-transitive, $P \circ_T I \subseteq P$;	$T(P(a, b), P(b, c)) \leq P(a, c)$ $T(P(a, b), I(b, c)) \leq P(a, c)$
(IV)	P is T-transitive, $I \circ_T P \subseteq P$;	$T(P(a, b), P(b, c)) \leq P(a, c)$ $T(I(a, b), P(b, c)) \leq P(a, c)$
(V)	P is negatively T-transitive	$P(a, c) \leq S(P(a, b), P(b, c))$
(VI)	every circuit of length 3 in (A, R) contains no P	$1 - S(P(a, b), P(b, c), P(c, a))$ $\geq T(R(a, b), R(b, c), R(c, a))$

3 Fuzzy Preorders for Archimedean t-Norms

3.1 General Results

We would like to study the connections among the previous conditions (I)-(VI). Let us start with a general result, which is fulfilled for any t-norm T (see, e.g., [9,14]).

Proposition 2. *Let R be a fuzzy relation and let T be a t-norm. It holds that*

$$R \text{ is } T\text{-transitive} \Leftrightarrow R^d \text{ is negatively } T\text{-transitive.}$$

Furthermore, if we take the Łukasiewicz t-norm as generator and the relation R is weakly complete, then $P = R^d$. Thus, it holds that in absence of incomparability, for every t-norm properties (I) and (V) are equivalent (see [5]) .

3.2 The Idempotent t-Norm

In previous works ([5,7]) we have studied in detail the case of the minimum t-norm. The surviving implications are presented in Figure 1. Any missing implication in that figure does not hold, as it was proven by appropriate counterexamples.

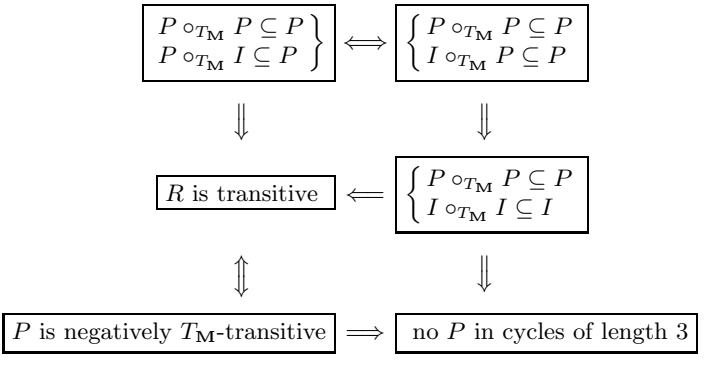


Fig. 1. Relationship among Properties (I)-(VI) for the minimum t-norm

3.3 Nilpotent t-Norms

As we commented in Subsection 2.2, any nilpotent t-norm is a φ-transformation of the Łukasiewicz t-norm. Thus, we will start our study in particular for this t-norm and later we will to extend it to any nilpotent t-norm.

We have also studied (see [7]) some of the connections between the different statements (I)-(VI) using the Łukasiewicz t-norm. Now we present the results that complete this study.

Proposition 3. *Let (P, I) be an additive fuzzy preference structure without incomparability and R its associated weak preference relation. It holds that the*

$T_{\mathbf{L}}$-transitivity of R is not a sufficient condition for the absence of P in cycles of length 3.

It is obvious that the reciprocal implications neither holds, because in other case, as the $T_{\mathbf{L}}$-transitivity of R implies that P and I are $T_{\mathbf{L}}$-transitivity, then the absence of P in cycles of length 3 implies the $T_{\mathbf{L}}$ transitivity of P and I. However, it is known that this is not truth (see [7]).

In the remain cases, the answer is again that the implication cannot be assure in general, as it proves the following result.

Proposition 4. *There exists at least an AFPS without incomparability (P, I) in each of the following situations:*

- *P and I are $T_{\mathbf{L}}$-transitive, but neither $I \circ_{T_{\mathbf{L}}} P \subseteq P$.*
- *P and I are $T_{\mathbf{L}}$-transitive, but neither $P \circ_{T_{\mathbf{L}}} I \subseteq P$.*
- *P is $T_{\mathbf{L}}$-transitive and $P \circ_{T_{\mathbf{L}}} I \subseteq P$, but $I \circ_{T_{\mathbf{L}}} P \subseteq P$.*
- *P is $T_{\mathbf{L}}$-transitive and $I \circ_{T_{\mathbf{L}}} P \subseteq P$, but $P \circ_{T_{\mathbf{L}}} I \subseteq P$.*
- *P is $T_{\mathbf{L}}$-transitive, $P \circ_{T_{\mathbf{L}}} I \subseteq P$ and $I \circ_{T_{\mathbf{L}}} P \subseteq P$, but I is not $T_{\mathbf{L}}$-transitive.*

We summarize all the results in Figure 2.

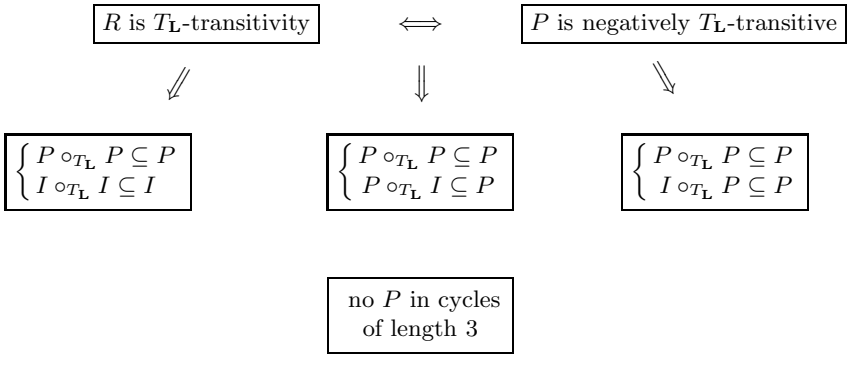


Fig. 2. Relationship among Properties (I)-(VI) for the Łukasiewicz t-norm. The missing implications do not hold.

Now, we will try to extend the known results for the Łukasiewicz t-norm to any nilpotent t-norm. The equivalence between (I) and (V) holds trivially. In light of the previous results, we will focus on the connection between (I) and (II), (III) and (IV). No other implication can hold for any nilpotent t-norm, since it does not hold for the Łukasiewicz t-norm. However, condition (I) does not guarantee any of the three conditions (II), (III) or (IV) for a whole special class of nilpotent t-norms.

Remark 1. Let us consider a nilpotent t-norm $T_{\mathbf{L}}^{\varphi}$ such that $x + y < 1$ and $\varphi(x) + \varphi(y) > 1$. The reflexive relation R defined on $A = \{a, b, c\}$ as follows

is $T_{\mathbf{L}}^{\varphi}$-transitive. However, its associated (by the Łukasiewicz generator) strict preference relation P is not, since $P(a,a) = 0 \not\geq T_{\mathbf{L}}^{\varphi}(x,y) = T_{\mathbf{L}}^{\varphi}(P(a,b), P(b,a))$.

R	a	b	c
a	1	$1-y$	$1-y$
b	$1-x$	1	$1-y$
c	$1-x$	$1-x$	1

P	a	b	c
a	0	x	x
b	y	0	x
c	y	y	0

Let us notice that in the complementary case ($\nexists (x,y) \in [0,1]^2$ such that $x+y<1$ and $\varphi(x)+\varphi(y)>1$), we also could obtain a counterexample for this implication.

3.4 Strict t-Norms

We have completed the study of the connections between (I)-(VI) using $T_{\mathbf{M}}$, $T_{\mathbf{L}}$ and in general for a nilpotent t-norm. Now we are going to complete this study with any strict t-norm. Finally, we will consider the particular case of the product t-norm ($T_{\mathbf{P}}$). Once this is done, the connections between the statements for any continuous Archimedean t-norm will be totally characterized. It is known, by Proposition 2, that statements (I) and (V) are equivalent for every t-norm, so in particular for a strict t-norm. Let us complete the others connections. We begin studying if the T-transitivity of R or the T-transitivity of P and I implies some of the others statements for any strict t-norm T.

Proposition 5. *Let T be a strict t-norm. Then there exists at least an AFPS without incomparability (P,I) in each of the following situations:*

- *R is T-transitive, but P is not T-transitive.*
- *P and I are T-transitive, but R is not T-transitive.*
- *P and I are T-transitive, but $P \circ_T I \not\subseteq P$.*
- *P and I are T-transitive, but $I \circ_T P \not\subseteq P$.*

We have obtained that the T-transitivity of R does not imply neither (II), nor (III), nor (IV), and furthermore, the T-transitivity of P and I does not imply neither (I), nor (III), nor (IV).

Let us study the possible connections that the absence of P in cycles of length 3 has with the other statements.

Proposition 6. *Let T be a strict t-norm. Then, there exists at least a weakly complete reflexive relation R in each of the following situations, where (P,I) denotes its associated AFPS obtained by means of the generator $i = T_{\mathbf{L}}$:*

- *P and I are T-transitive, but P is in a cycle of length 3 of R.*
- *R is T-transitive, but P is in a cycle of length 3 of R.*
- *There is no P in cycles of length 3 of R, but R is not T-transitive.*
- *There is no P in cycles of length 3 of R, but P is not T-transitive.*

It only remains to see if conditions (III) or (IV) imply some of the others. For this purpose, we need to see firstly the next lemma.

Lemma 1. *Let (P, I) be an additive fuzzy preference structure without incomparability. If P is T-transitive, with T a strict t-norm, it holds that*

- $\forall a, b \in A$, $P(a, b) = 0$ *or* $P(b, a) = 0$.
- $\forall a, b, c \in A$ *it holds that* $min\{P(a, b), P(b, c), P(c, a)\} = 0$.

With this lemma we can prove the following theorem.

Proposition 7. *Let (P, I) be an additive fuzzy preference structure without incomparability and let R be its associated weak preference structure. If P is T-transitivity, with T a strict t-norm, and $P \circ_T I \subseteq P$, then R is a crisp relation.*

As consequence of this result we have that statement (III) implies everyone else. Furthermore, if we change the condition $P \circ_T I \subseteq P$ for the condition $I \circ_T P \subseteq P$, the theorem still holds. Then, statements (III) and (IV) are equivalent and they implies everyone else.

In particular, the previous results hold for the product t-norm. Moreover, no better results can be obtained in this particular case. In fact, the counterexamples used to prove the statements included in Propositions 5 and 6 are done with $T_{\mathbf{P}}$. Thus, the connections that hold in general for $T_{\mathbf{P}}$ are exactly the same that hold for a strict t-norm. In Figure 3 is presented the summary of the obtained results.

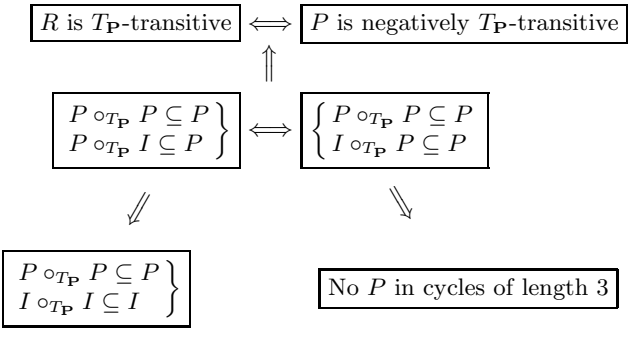


Fig. 3. Connections among Properties (I)-(VI) for the product t-norm. The missing implications do not hold.

4 Conclusions

The main contribution of this work is that we have been able to complete to any idempotent and Archimedean continuous t-norm the previous studies about the connections among different possible definitions of the concept of fuzzy preorder. In the cases the implications are not fulfilled, we have found appropriate counterexamples. With theses studies we could choose the best definition of fuzzy preorder in each situation. As a future work, we will like to be able to study other operators: general continuous t-norms, conjunctors, etc. and also extend our studies when other ways of expressing the negation are considered.

Acknowledgements

The research reported on in this paper has been partially supported by project FEDER-MEC-MTM2007-61193.

References

1. De Baets, B., Fodor, J.: Generator triplets of additive fuzzy preference structures. In: Proc. of 6th. Internat. Workshop on Relational Methods in Computer Sicence, Tilburg, The Netherlands (2001)
2. De Baets, B., Van De Walle, B., Kerre, E.: Fuzzy preference structures without incomparability. Fuzzy Sets and Systems 76, 333–348 (1995)
3. Dasgupta, M., Deb, R.: Factoring fuzzy transitivity. Fuzzy Sets and Systems 118, 489–502 (2001)
4. Díaz, S., De Baets, B., Montes, S.: On the transitivity of indifference in the framework of additive fuzzy preference structures. In: De Baets, B., Kaynak, O., Bilgiç, T. (eds.) IFSA 2003. LNCS, vol. 2715, pp. 87–94. Springer, Heidelberg (2003)
5. Díaz, S., De Baets, B., Montes, S.: On some characterizations of complete fuzzy preorders. In: Proceedings of the fourth conference of the EUSFLAT, Barcelona, Spain, pp. 1039–1044 (2005)
6. Díaz, S., De Baets, B., Montes, S.: General results on the decomposition of transitive fuzzy relations. Fuzzy Decision and Decision Making 9, 1–29 (2010)
7. Díaz, S., Martinetti, D., Montes, I., Montes, S.: Connection among some characterizations of complete fuzzy preorders. In: Proceedings of the ninth conference of the ISDA, Pisa, Italy, pp. 839–844 (2009)
8. Díaz, S., Montes, S., De Baets, B.: Transitivity bounds in additive fuzzy preference structures. IEEE Transactions on Fuzzy Systems 15(2), 275–286 (2007)
9. Fodor, J., Roubens, M.: Fuzzy Preference Modelling and Multicriteria Decision Support. Kluwer Academic Publishers, Dordrecht (1994)
10. García-Lapresta, J.L., Rodríguez-Palmero, C.: Some algebraic characterizations of preference structures. Journal of Interdisciplinary Mathematics 7, 233–254 (2004)
11. Klement, E.P., Mesiar, R., Pap, E.: Triangular norms. Kluwer Academic Publishers, Dordrecht (2000)
12. Pirlot, M., Vincke, P.: Semiorders: Properties, Representations, Applications. Kluwer Academic Publishers, Dordrecht (1997)
13. Van de Walle, B., De Baets, B., Kerre, E.: Characterizable fuzzy preference structures. Ann. Oper. Res. 80, 105–136 (1998)
14. Wang, X., Xue, Y.: Notes on transitivity, negative transitivity, semitransitivity and Ferrers property. Journal of Fuzzy Mathematics 12, 323–330 (2004)

Rectification of Preferences in a Fuzzy Environment

Camilo A. Franco de los Ríos, Javier Montero, and J. Tinguaro Rodríguez

Faculty of Mathematics, Complutense University, Madrid 28040, Spain
francodelosrios@gmail.com

Abstract. In this paper we consider a complete fuzzy preference structure, defined by means of a constructive approach associated to the necessary learning process in a decision making problem. Preference relations are successively assigned over a given set of alternatives with the possibility of revision and rectification, subject to a certain set of beliefs which sustain the construction of different viewpoints. Measures of information are then considered in this process for a formal identification of uncertainty due to ignorance, a concept that is closely related to ambiguity as defined by Fishburn (1993). Finally, a coherence criterion is proposed in order to study the importance of preferences, in such a way that an indirect and comparative analysis between them is needed so outcomes may be coherently organized.

Keywords: fuzzy preferences, rectification, beliefs, viewpoints, uncertainty due to ignorance, ambiguity, coherence measures.

1 Introduction

Different notions of preference can be established between distinct classes of entities, actions, circumstances or propositions, in order to understand how decision makers view the world and its inherent uncertainty. Within this vision, motivated by a subjective vision over classical logic [6], the set of possible worlds is viewed as a subset of alternatives, in such a way that preferences are stated on the basis of individual or collective beliefs towards a certain subset of alternatives, taking into account positive and negative arguments for each comparison. In particular, the *base of beliefs*, a set of formulas where each formula represents an individual belief, delimits the set of possible worlds in such a way that any added belief will impossible worlds be rejected. It is out of question the importance of opening the possibility of revision and rectification of these beliefs, since it is a primary learning characteristic of any intelligent system [18].

In this article the concept of fuzzy environment is supported from a constructive approach providing the setting for fuzzy logic and its applications. In this case, the field of application is preference theory, considering the construction of a preference order, according to a characteristic viewpoint, under uncertainty due to ignorance or ambiguity. The distinction that we follow here is the one that almost a century ago (1921) Knight [15] and Keynes [14] presented between measurable uncertainty (probabilities) and unmeasurable uncertainty (weight of evidence), arguing that ambiguity,

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 168–178, 2010.

understood as lack of knowledge [7], [8], has an important role in rational learning and decision processes.

This paper is organized as follows: section 2 discusses the key aspects of uncertainty due to ignorance and reviews ambiguity and coherence as information measures. In section 3, a complete preference structure is defined under characteristic viewpoints. Finally, section 4 examines the possibility of rectification in a decision process with fuzzy preference relations, where information measures have the main function of preserving coherence when knowledge is ambiguous.

2 Uncertainty and Ignorance

The search for alternative qualitative and symbolic models is extremely relevant for a further development of preference theory, in order to complement, explain and generalize the classical body of knowledge founded in the theory of (expected) utility. Diverse studies in social sciences (see [7], [10]) where the problem of decision under ignorance is examined according to the approach formulated in [14] and [15], offer some experimental evidence supporting that preferences do not depend only on the degree of uncertainty over the quality of information, but also over its source. This distinction motivates a careful study of uncertainty, stressing the relationship of the concept of ignorance with ambiguity [8] and fuzziness [24], different to probabilistic uncertainty.

To illustrate the idea stated above we recall the classical example of the "two colors" (see [7], [14], [15]). Consider two boxes containing red and black balls. The first box contains 50 red balls and 50 black balls, while the second box has 100 black and red balls without known proportion. Although most decision makers will assign the same probability of reaching for a red ball in any of the boxes (0.5), the weight of the argument in favour of this conclusion is greater in the first case than in the second one. The fact that in general people prefer to bet for the box with known proportion, suggests that subjective probabilities of reaching for a red or black ball are greater in the box where its proportion is known, revealing a behavioural pattern not fully consistent with standard subjective probability and expected utility theory.

As shown in [7], there is a pronounced difference between the type of uncertainty that can be measured by probabilistic functions, which can be referred to as *risk*, and the type that can not be directly measured by such functions. So, in the presence of uncertainty we have the probability of occurrence of an event, and on the other hand, a degree of confidence over the different attributes of information. Decisions rest upon beliefs about the state of nature in situations where non probabilistic uncertainty is present. In common human decision making, for example, it is clear that probabilistic reasoning gets less weight, since quite often we process mixtures of different kinds of uncertainties (this is the case, for example, of most statistical polls, as soon as they make use of linguistic terms that should bring a specialized non probabilistic analysis).

Restricting ourselves to non probabilistic uncertainty, in particular to a state of knowledge of the decision maker that can be referred to as uncertainty due to ignorance, we realize that such *ambiguity* affects decisions in a way that contradicts additivity of subjective probability. To clarify this idea we remind below the definition of

ambiguity [8] and study its relation with fuzziness [24], in order to represent decision situations where the boundary between the set of possible and impossible worlds, constructed upon beliefs that support a finite set of viewpoints, is not always crisp.

Ambiguity, as a primitive concept, has been proposed as a comparative relation on a set of elements X where $P(X)$ is the set of all subsets of X. The set measure $\alpha:P(X)\rightarrow[0,1]$ is called an ambiguity function if it satisfies the following three axioms [8]:

A1. $\alpha(\varnothing)=0$

A2. $\alpha(A)=\alpha(A^c)$

A3. $\alpha(A\cup B) + \alpha(A\cap B) \leq \alpha(A) + \alpha(B)$

Minimum ambiguity is assigned to the null set, such that $\alpha(\varnothing)=\alpha(X)=0$. The complement of A is denoted by A^c, so the central idea of this characterization is expressed by (A2), where ambiguity is a measure of a certain attribute of information shared by any set A and its complementation. Let us remind that for probabilities, if p stands for a probability function, $p(A)+p(A^c)=1$, while for ambiguity, $\alpha(A)=\alpha(A^c)$, reflecting the original intuition [8] that whatever underlies the ambiguity of a set also underlies the ambiguity of its complement. Finally, the third axiom expresses submodularity: the idea that the union of two sets A, B may reduce or cancel ambiguities associated to each one (considered separately), in such a way that if A and B are disjoint, then $\alpha(A\cup B)\leq\alpha(A)+\alpha(B)$ and if $\alpha(A\cup B)=X$ then $\alpha(A\cap B)\leq\alpha(A)+\alpha(B)$.

The above idea of ambiguity measures can be translated into fuzzy set theory [25]. Fuzzy logic focuses on situations where the boundary between the set of possible and impossible worlds is not always crisp, and a fuzzy subset R may be naturally described by the membership and non-membership degree of each element x in X, denoted by $R^+(x)$ and $R^-(x)$ respectively. As suggested in [24], fuzziness may rise from the lack of an absolute distinction between R^+ and R^-. Note that R^- does not necessarily correspond to the classical conception of complementation, since (see for example [2], [16], [19]) non-membership intensities can be better understood as some kind of *orthogonal* but positively measurable degrees over the elements of X. A more detailed study on fuzziness and its different interpretations for fuzzy sets ([25], [26]) and intuitionistic fuzzy sets [2] can be found in [22].

Following [24], where R^- is taken as the complement in such a way that $R^-(x)=R^c(x)=1-R(x)$, if we assume that R is a fuzzy subset defined over the finite set X with cardinality n, the intersection of $R^+(x)$ and $R^-(x)$ is defined by the set $I(x)$, $I=R^+\cap R^-$ (see [4] in order to stress the relevance of the related concept of overlapping when dealing with fuzzy information). As a result, if the set $I(x)$ is empty, then R^+ and R^- are clearly distinct and no fuzziness occurs (membership intensities are either 0 or 1) but when $I(x)$ is positive, the larger it becomes, the greater the fuzziness associated with the set R.

Let us remind the formal characterization of measures of fuzziness, Fuzz(R), for any fuzzy subset R (see [24]):

F1. Fuzz(R)=0 if and only if R is a crisp subset.

F2. Fuzz(R) has its maximal defined as $R(x)=0.5$ for all x.

F3. For any fuzzy subset R^* such that $R^*(x) \geq R(x)$ if $R(x) \geq 0.5$ and $R^*(x) \leq R(x)$ if $R(x) \leq 0.5$ then $\text{Fuzz}(R) \geq \text{Fuzz}(R^*)$: R^* is called a sharpened version of R.

As shown in [24], a measure of fuzziness can be also an ambiguity measure in the sense of Fishburn [8], where $R(x) \cup S(x) = \max(R(x), S(x))$, $R(x) \cap S(x) = \min(R(x), S(x))$ and $R^c(x) = 1 - R(x)$ for any fuzzy sets R and S. As a consequence, ambiguity measures are also valid for measuring fuzziness.

Once the concept of uncertainty due to ignorance has been associated to the concept of ambiguity and fuzziness, we can analyze its role in a preference ordering construction process. Usually preference evaluation *vía* binary relations considers only the two alternatives at stake and their principal attributes, paying less attention to second order characteristics, which receive greater relevance in situations where ambiguity is present. For this reason we need to analyze the different states of knowledge (from absolute certainty to total ignorance) of a decision maker.

A comparative analysis should allow us to distinguish between different states of knowledge, judging the set of constructed preferences based on expert knowledge or maximal certainty. In consequence, diverse attributes may receive positive and higher values of relevance in a comparative evaluation than in an isolated one. In order to study the importance of preference relations and their outcomes, it is necessary to remind some basic definitions and introduce a preference structure within a fuzzy environment.

3 Complete Preference Structure

A decision making problem is here understood as the construction of an ordering process over a finite set of alternatives A, where fuzzy binary relations are defined between alternatives a and b by a degree of truth for the predicate R, "a is at least as good as b", assigned according to a complete and partially ordered valuation set L. Following standard approaches (see [9], [17]), each fuzzy preference relation R can be understood as a composition of four different relations and its corresponding intensities. These components are "strict preference of a over b": P, "strict preference of b over a": P^{-1}, "indifference between a and b": I, and "incomparability between a and b": J.

A fuzzy preference binary relation for a set of alternatives A is characterized by a complete valuation space L and a function R such that $R : A \times A \rightarrow L$. For simplicity, we consider the case in which $L=[0,1]$. Here we consider the three axioms proposed in [9] and [17]. The "Independence of Irrelevant Alternatives" axiom states that for every pair of alternatives a, b, the values of $P(a,b)$, $I(a,b)$ and $J(a,b)$ depend only on the values of $x=R(a,b)$ and $y=R(b,a)$. The existence of the continuous functions,

$$p, i, j : [0,1]^2 \rightarrow [0,1]$$

can be stated, in such a way that $P(a,b)=p(x,y)$, $I(a,b)=i(x,y)$ and $J(a,b)=j(x,y)$. The "Positive Association" axiom says that functions $p(x,n(y))$, $i(x,y)$ and $j(n(x),n(y))$ are non-decreasing over both arguments. The "Symmetry" axiom states the symmetry of the functions $i(x,y)$ and $j(x,y)$. The complete preference structure is therefore described as:

$$R(x,y)=\langle p(x,y),p(y,x),i(x,y),j(x,y)\rangle \tag{1}$$

where $R(x,y)\geq 1$. In this way, the aggregated intensity is being decomposed into four intensities that should cover the total possible intensity.

In order to build up the relation R over the available alternatives, a rational process is needed to identify viewpoint $h\in H$, under which preferences can be valued. The principle of rationality that we follow here is the one of *minimum action*, where an individual searches for the viewpoint with lowest dimension (minimal number of criteria) so a consistent preference ordering can be constructed for a finite subset of alternatives. For an exploration of different approaches to dimension for some classes of orders see for example [12].

Each viewpoint h is constructed over the attributes of the available information on the elements of the set A, interpreted according to the base of beliefs B, over which the set of criteria C can be defined. Here h represents an independent state of mind for undertaking the preference analysis. The following definition is based on [13]:

Definition. A viewpoint $h\in H$ for a subset of alternatives $S_h\subseteq A$ is characterized by an outcome space Ω_h, a set of criteria C_h where each criterion $c_h:S_h\rightarrow \Omega_h$ maps alternatives to their outcomes, and a partial order given by the set of fuzzy preference relations R over Ω_h.

The set of criteria C_h is a finite set whose elements can be combined for the construction of any viewpoint $h\in H$, where each criterion determines a new dimension for the outcome space Ω_h.

Differences between viewpoints $h,v\in H$ over the same set of alternatives S, case where $S_h=S_v$, may refer to differences in preferences, in outcome spaces or in criteria [3], [13]. If a set of alternatives S is organized by a unique viewpoint h, we say it is the *characterizing viewpoint* of S, denoted by S_h. If more viewpoints are candidates for ordering the same set S, then we say that a process of *rectification* is possible. Rectification allows the identification of a new order on the alternatives in S, revealing a change in the preferences of the decision maker. When an intelligent agent rectifies, a new viewpoint is being searched. This is a task where the decomposition or combination of viewpoints may be necessary. Therefore, two viewpoints can be combined or decomposed for constructing a new enlarged or reduced viewpoint, where the set of alternatives, preferences, outcome spaces or criteria may change. Still, maximum coherence needs to be preserved.

The complete structure for this preference constructive approach can now be defined as,

$$Z=\langle A,H,R,L\rangle$$

where H is the set of all possible viewpoints under which the alternatives in A can be completely ordered according to the structure given by R and its valuation set L.

4 Information Measures and Preference Rectification

The process of decision making implies a comprehensive identification of the most relevant attributes of information, learning how to order the available alternatives and classifying them (see [1], [16]) with a maximum degree of coherence. Stressing the importance of a comparative analysis, a decision making process is here called a learning process if and only if rectification of preferences is possible. Rectification is then understood as the reconstruction of a preferential order for a finite set of alternatives, where it may be necessary an expansion (composition), or on the contrary, a contraction (decomposition) of viewpoints.

When any pair of alternatives is compared and a preference relation is constructed, an order is assigned over $S \subseteq A$ according to certain viewpoint *h*. Following the idea that a viewpoint may have multiple extensions so a preferred outcome may at last be found [13], the decision maker has to be able to rectify along the sequential learning process. Such learning process is here characterized as the elicitation of knowledge from single attributes of information, where these attributes are identified step by step. In this way, the most relevant attributes can change and viewpoints may need modification (combined, contracted or enlarged).

The sequential learning process described above needs to take into account its natural ambiguity or degree of ignorance. Let us take for example two viewpoints *h* and *v*. If we compare two alternatives *a*, *b* under both viewpoints, and the intensity of the predicate "*a* is at least as good as *b* under viewpoint *h*", $R_h(a,b)$, is stronger than the intensity of $R_v(a,b)$, then we are less certain about the value of $R(a,b)$ when the viewpoint *v* is used than when *h* is used. This means that *v* is more ambiguous than *h.*

Following [26], [27], we can measure if there is a possibility that $R_v(a,b)$ is true given that we know $R_h(a,b)$. This possibility measure (see [6], [27]),

$$\rho_{R_h(a,b)}\left(R_v\left(a,b\right)\right)=\max\left(R_w\left(a,b\right)\right)$$

expresses the degree of intersection between *h* and *v*, where $w=h\wedge v$. Analogously, we can measure if there is a possibility that $R_v^-\left(a,b\right)$, where $R^-(x)=R^c(x)=1-R(x)$, is true given that we know $R_h(a,b)$. The measure of certainty [27] (also called measure of necessity [6])

$$\varsigma_{R_h(a,b)}\left(R_v\left(a,b\right)\right)=1-\rho_{R_h(a,b)}\left(R_v^-\left(a,b\right)\right)$$

expresses the degree of inclusion of *h* in *v*. In addition,

$$\upsilon_{R_h(a,b)}\left(R_v\left(a,b\right)\right)=\rho_{R_h(a,b)}\left(R_v\left(a,b\right)\right)-\varsigma_{R_h(a,b)}\left(R_v\left(a,b\right)\right)$$

is an ambiguity measure as defined in A1-A3 (see [24]), where the greater that υ is, the more ambiguity or uncertainty due to ignorance exists. In this way, when ambiguity on viewpoint *h* is large, a greater possibility exists for rectifying over the current order given by $R_h(a,b)$. This measure helps to identify when some preference order rests upon a viewpoint with a low degree of confidence. A new viewpoint that takes

into consideration other attributes, second order characteristics, can be necessary for a robust decision process to take place.

Now, if we want to assign a greater value of confidence to the viewpoint with a greater degree of truth, we need a measure of confidence, dual of uncertainty (due to ignorance), such that a greater value can be assigned to the outcome that uses information with a higher intensity of certainty. The idea is to find a measure that explicitly assigns, by a comparative analysis, a greater value if a high level of confidence exists. Such a measure, closely related to ambiguity as it has been examined above, will directly distinguish the agent's states of knowledge, by a confidence scale that goes from total ignorance to total certainty, based on expert knowledge or maximal certainty.

We then propose that the confidence measure we are looking for can be analyzed as a coherence measure. Recall that ambiguity [8] and coherence measures [20] are explicitly related by extension theorems proven for any strong negation (such that a k-bijection verifying n(x)=k-1(k(1)-k(x)) exists, see [5], [20], [21]). These coherence measures evaluate how fuzzy and similar are any pair of fuzzy sets, where the degree of fuzziness is characterized by F1-F3. For our purposes, the basic idea can be phrased as the nearer a fuzzy preference relation is to the sets $\varnothing$ or Ω, the higher coherence measure is expected.

Given the referential set *X*, with $R_j, R_i \in X$, representing the set of all fuzzy preference relations $Q^f(X)$ and given a strong negation *n* on $Q^f(X)$, coherence measures can be then defined over the set of fuzzy preferences: the function $\zeta : Q^f(X) \times Q^f(X) \to [0,1]$ is a coherence measure, if and only if the following three axioms are satisfied (see [20], [21]):

C1. $\zeta\left(R_i, R_j\right) = \zeta\left(R_j, R_i\right)$

C2. $\zeta\left(R_i, n\left(R_j\right)\right) = n\left(\zeta\left(R_j, R_i\right)\right)$

C3. $\zeta(\varnothing, X) = 0$

The first condition (C1) states the symmetry of this measure and the third one (C3) guarantees minimum coherence. About (C2), analyzing the relation between ambiguity and coherence measures, we can see a complementary approach between A2 and C2 in the sense that the coherence between R_i and R_j^- is the opposite of the existing one between R_i and R_j. Remind that if we want to assign a greater value of confidence to the viewpoint with a greater degree of truth, we need some measure of confidence, something like the dual of ambiguity. Therefore, coherence is here understood as a comparative confidence measure for any pair of fuzzy sets, in our case preference relations, such that a common link between fuzzy preference relations can be identified for any viewpoint.

Let us recall the above ambiguity measure $\upsilon_x\left(y\right)$, taking $x=R_h(a,b)$ and $y=R_v(a,b)$. If ambiguity is represented by the difference between possibility and

necessity measures, then confidence should take into account the aggregated weight of these two measures. And examining confidence as here presented, we can pursue a coherence measure close to

$$\frac{1-\upsilon_x(y)}{2}. \tag{2}$$

Therefore, if $\upsilon_x(y)=\rho_x(y)-\varsigma_x(y)$ and $\rho_x(y)=\max(x,y)$ and $\varsigma_x(y)=1-\max(x,y^c)$, then $\upsilon_x(y)=\max(x,y)+\max(x,y^c)-1$. Condition C1 requires that the measure is symmetric, but it can be seen that for the necessity measure the order of the arguments does matter. In other words, the degree of inclusion of *x* in *y* is not the same as the degree of inclusion of *y* in *x*: $1-\max(x,y^c)\neq 1-\max(y,x^c)$. For this reason, because coherence is conceived as a relation defined over pairs of fuzzy sets, the max function can be replaced by some kind of distance representing how far is each set from any other one.

If we take for example the Euclidian distance $d(x,y)=\frac{1}{m}\cdot\sum_{i=1}^{m}|x_i-y_i|$, the measure defined by (2), where $d(x,y)$ and $d(x,y^c)$ represent a separation degree between *x* and *y* and between *x* and everything not being *y*, respectively, becomes

$$\beta(x,y)=\frac{1+d(x,y^c)-d(x,y)}{2}. \tag{3}$$

Coherence of fuzzy preferences, as defined by C1-C3 and (3), has been initially explored in [11] (the origin of (3) as a coherence measure is founded in [20]), establishing a possible criterion where all outcomes can be evaluated. This proposal is useful for any viewpoint, offering relevant information about the importance of its preferences, according to the confidence on their truth values. In order to maintain coherence along the learning process, this *coherence criterion* allows the decision agent to add an extra dimension to the outcome space of a given viewpoint (see example 1 below), and it can also be used to examine the degree of coherence between different viewpoints so the possibility of combining, contracting or enlarging them can be studied (see example 2 below).

Example 1. Under the state of knowledge of uncertainty because of ignorance, the coherence criterion enables the decision agent to assign an ordinal value over the set of alternatives, judging the set of constructed preferences against one certain relation *R**, which stands as an organizing predicate [23] and denotes maximal certainty or expert knowledge. Given the existence of such a predicate, there exists an associated coherence criterion for any viewpoint, establishing an independent dimension where the outcome space can be coherently organized.

Consider the set of preference relations R_h under viewpoint h such that a decision agent is interested in identifying weak preference with a greater degree of confidence (case where it is difficult to differentiate strict preference from indifference). Therefore, for any preference relation where the components p and i have simultaneously high membership intensities, a higher degree of coherence, as defined by C1-C3, will be assigned to it. In order to apply the coherence criterion we use an organizing predicate R^* such as $R^*=\{1,0,1,0\}$. As it can be seen, R^* represents the most certain knowledge for the values defined by (1).

Now, consider the following fuzzy preference relations:

$$R_h(x,y)=\{0.5,0.5,0.5,0.0\}$$
$$R_h(w,z)=\{0.9,0.8,0.8,0.1\}$$
$$R_h(u,z)=\{0.2,0.2,0.9,0.7\}$$

This set of preferences reveals some uncertainty due to ignorance and notice that there is not a complete order between them. Comparison of preference is missing between alternatives (x,z). Through a third element it would be possible now to examine how these alternatives relate with each other. According to the coherence criterion (3), the corresponding confidence-ordinal values over these set of fuzzy preferences relations under viewpoint h are:

$$\beta(R_h(x,y),R^*)=0.625$$
$$\beta(R_h(w,z),R^*)=0.70$$
$$\beta(R_h(u,z),R^*)=0.55$$

In this case, the optimal outcome is $R_h(w,z)$, because it has a greater coherence measure than all the others. The coherence criterion, which is here presented as a dimension where preferences may be organized according to the degree of confidence on their truth values, identifies the fuzzy preference relation with the strongest intensity and certainty.

Example 2. Let's consider three viewpoints $h,v,w\in H$ over the same set of alternatives $a,b\in A$, such that both a, b are members of S_h, S_v and S_w. Remember that we say that a process of rectification is possible when two or more viewpoints are candidates for ordering the same set of alternatives. In this case $S_h = S_v = S_w$, so if an intelligent agent rectifies then his viewpoint changes. For this reason some decision aid is necessary for the agent to identify the optimal viewpoint to be used, based in his beliefs but also on the confidence (over his state of knowledge) that the set of preferences reveal. The viewpoint may change, but maximum coherence needs to be preserved.

Assume the agent believes viewpoint h is the most important, but when evaluating preference between a and b under h, confidence in its outcome suffers of some ambiguity. This can be caused by many reasons, for example, if h is composed by price-quantity criteria, the agent may suffer from lack of perfect knowledge over price stability. So, if the agent is comparing different goods, he can also reveal his preferences under viewpoint v, composed by fair-trade attributes, and viewpoint w, which values presentation

and marketing. The agent will then have to decide and the viewpoint he uses will determine the final outcome, which has to be an optimal answer to his necessities.

Consider the following preferences,

$$R_h(a,b)=\{0.7,0.3,0.5,0.3\}$$

$$R_v(a,b)=\{0.2,0.8,0.2,0.4\}$$

$$R_w(a,b)=\{0.9,0.2,0.9,0.1\}$$

If the agent wants to choose the most certain viewpoint, such that alternative a is at least as good as alternative b with maximum certainty, he can use the same organizing predicate R^* defined in example 1. Otherwise, a confidence order can be constructed between viewpoints h, v and w so the possibility of merging viewpoints, due to their relative coherence, can be examined. Using the coherence measure defined in (3):

$$\beta(R_h,R^*)=0.65 \qquad \beta(h,v)=0.425$$

$$\beta(R_v,R^*)=0.30 \qquad \beta(h,w)=0.65$$

$$\beta(R_w,R^*)=0.875 \qquad \beta(v,w)=0.30$$

As a result, the optimal outcome according to R^* is given by viewpoint w and the agent will find reasons to rectify. In the other case and complementary to the situation just described, the decision maker will be also motivated to combine viewpoints h and w, in order to find a new enlarged viewpoint where the decision maker can be more certain about his/her knowledge and preferences.

5 Final Comments

A complete fuzzy preference structure has been considered for decision making, characterized as a learning process, where preference relations need to be successively assigned over a set of alternatives, always allowing revision and rectification of beliefs. An analytical framework has been set for studying preference construction and the identification of an optimal outcome under characterizing viewpoints.

References

1. Amo, A., Montero, J., Biging, G., Cutello, V.: Fuzzy classification systems. European Journal of Operational Research 156, 459–507 (2004)
2. Atanassov, K.: Intuitionistic Fuzzy Sets. Theory and Applications. Physica-Verlag, New York (1999)
3. Bouyssou, D., Marchant, T., Pirlot, M., Tsoukiàs, A., Vincke, P.: Evaluation and Decision Models with Multiple Criteria, vol. 86. Springer, Heidelberg (2006)
4. Bustince, H., Fernández, J., Mesiar, R., Montero, J.: Overlap function. Non Linear Analysis 72, 1488–1499 (2010)
5. Cubillo, S., Bellido, S.: A first approximation to Ng-coherence measures. In: Proceedings EUSFLAT 2003, pp. 709–713 (2003)

6. Dubois, D., Prade, H.: Formal representations of uncertainty. In: Bouyssou, D., Dubois, D., Pirlot, M., Prade, H. (eds.) Decision-Making Process. ISTE, London (2009)
7. Ellsberg, D.: Risk ambiguity and the savage axioms. Quarterly Journal of Economics 75, 643–669 (1961)
8. Fishburn, P.: The axioms and algebra of ambiguity. Theory and Decision 34, 119–137 (1993)
9. Fodor, J., Roubens, M.: Fuzzy Preference Modelling and Multicriteria Decision Support. Kluwer Academic Publishers, Dordrecht (1994)
10. Fox, C., Tversky, A.: Ambiguity aversion and comparative ignorance. The Quarterly Journal of Economics 110, 585–603 (1995)
11. Franco de los Ríos, C., Montero, J.: On the use of coherence measures for fuzzy preference relations. In: Eurofuse Workshop 2009, pp. 15–19 (2009)
12. González-Pachón, J., Gómez, D., Montero, J., Yáñez, J.: Soft dimension theory. Fuzzy Sets and Systems 137, 137–149 (2003)
13. Junker, U.: Preferences in an open world. In: First International Conference, ADT 2009, pp. 215–224 (2009)
14. Keynes, J.: A Treatise on Probability. MacMillan, London (1963)
15. Knight, F.: Risk, Uncertainty, and Profit. University of Chicago Press, Chicago (1971)
16. Montero, J., Gómez, D., Bustince, H.: On the relevance of some families of fuzzy sets. Fuzzy Sets and Systems 158, 2429–2442 (2007)
17. Montero, J., Tejada, J., Cutello, C.: A general model for deriving preference structures from data. European Journal of Operational Research 98, 98–110 (1997)
18. Nebel, B.: A knowledge level analysis of belief revision. In: Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning, pp. 301–311 (1989)
19. Öztürk, M., Tsoukiàs, A.: Modelling uncertain positive and negative reasons in decision aiding. Decision Support Systems 43, 1512–1526 (2007)
20. Sancho-Royo, A., Verdegay, J.: Coherence measures on finite fuzzy sets. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems 8, 641–663 (2000)
21. Sancho-Royo, A., Verdegay, J.: Fuzzy coherence measures. International Journal of Intelligent Systems 20, 1–11 (2005)
22. Szmidt, E., Kacprzyk, J.: Entropy for intuitionistic fuzzy sets. Fuzzy Sets and Systems 118, 467–477 (2001)
23. Trillas, E.: On the use of words and fuzzy sets. Information Sciences 176, 1463–1487 (2006)
24. Yager, R.: On a measure of ambiguity. International Journal of Intelligent Systems 10, 1001–1019 (1995)
25. Zadeh, L.: Fuzzy sets. Information and Control 8, 338–353 (1965)
26. Zadeh, L.: The concept of a linguistic variable and its application to aproximate reasoning. Information Sciences 8, 199–249 (1975)
27. Zadeh, L.: Fuzzy sets and information granularity. In: Adavances in Fuzzy Set Theory and Applications, pp. 3–18. North-Holland, Amsterdam (1979)

Identification of Speakers by Name Using Belief Functions

Simon Petitrenaud, Vincent Jousse, Sylvain Meignier, and Yannick Estève

Laboratoire d'Informatique de l'Université du Maine
Avenue Laennec, 72085 Le Mans Cedex, France
{simon.petit-renaud,vincent.jousse,
sylvain.meignier,yannick.esteve}@univ-lemans.fr

Abstract. In this paper, we consider the extraction of speaker identity (first name and last name) from audio records of broadcast news. Using an automatic speech recognition system, we present improvements for a method which allows to extract speaker identities from automatic transcripts and to assign them to speaker turns. The detected full names are chosen as potential candidates for these assignments. All this information, which is often contradictory, is described and combined in the Belief Functions formalism, which makes the knowledge representation of the problem coherent. The Belief Function theory has proven to be very suitable and adapted for the management of uncertainties concerning the speaker identity. Experiments are carried out on French broadcast news records from a French evaluation campaign of automatic speech recognition.

Keywords: Speaker identification, speaker recognition, information fusion, belief functions.

1 Introduction

In order to allow later retrieval of recorded information, large collections of audio documents have to be indexed. The system presented in this paper focuses on speaker identification by their full name in audio documents. The speaker identity detection is composed of several steps and is in most cases subject to uncertainty and confusion. The first step to automatically get audio documents indexing consists in detecting speakers turns and regrouping those uttered by the same speaker. It is generally based on a first stage of segmentation that consists in partitioning the regions of speech into homogeneous audio segments which contains ideally the voice of only one speaker, followed by a clustering stage that consists in giving the same label to segments uttered by the same speaker. A speaker turn starts when a speaker is starting to speak and ends when another speaker is starting to speak, or a song or advertising is starting. Speaker turns are regrouped by class of same but *anonymous* speakers.

The next step is to automatically transcribe the content of speaker turns into words and is complemented by an annotation for some words as "named entities"

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 179–188, 2010.

or categories. Some words are particularly identified as "PERSON". The more promising way to identify speakers by their *real* full name consists in extracting them from the automatic speech recognition system (ASR) [1,8,4,5]. The general principle is to determine if a detected named entity as a "PERSON" refers to a speaker of the document or to a person who does not speak in the document. This principle assumes that the names are often pronounced, as in broadcast news.

Our article takes place in this framework. The system we developed in [4,5] uses uttered full names to assign them to anonymous speakers from identified speaker turns. The principle is to assign one of these four labels: *"current turn"*, *"previous turn"*, *"following turn"* or *"another person"* to each detected full name. But this approach ignored the potential conflict information on the speakers within a same speaker turn. In this paper, we propose to improve the consistency of the system and to better combine the various information on the potential speakers. The formalism of Belief Functions seemed to be particularly suited to managing these conflicts and combining this information.

First, we briefly present the automatic transcription system used, before describing the reference system for speaker named identification. Then, we discuss the shortcomings of this model and the improvements of our model using belief functions. Finally, we propose metrics for evaluating such systems, and we comment on the results obtained on the ESTER I evaluation campaign [2].

2 Speaker Identification Based on a Transcription System

2.1 Transcription System

The main hypothesis initially proposed in [1] assumes that a detected full name in a speaker turn allows to identify the current turn or a directly contiguous turn (previous or following turn). However, some full names identify farther speaker turns or persons that are not involved in the document. The used identification method is based on previously transcribed and enriched documents. This transcript needs to cut the document into segments which are then classified in anonymous speakers. These segments, grouped into speaker turns are transcribed and the named entities are annotated.

2.2 Semantic Classification Trees

The speaker identification method uses a binary decision tree based on the principle of semantic classification trees (SCT) [3]. A SCT automatically learns lexical rules from full names detected in the training corpus, with the left and right surrounding words. A SCT is used for each occurrence of full names detected in the transcripts. This tree allows to associate to each occurrence of a full name the probability to correspond to one of the four envisaged hypotheses: *"current turn"*, *"previous turn"*, *"following turn"* or *"another person"*. These probabilities are determined during the learning of the tree and reflect the observed cases in the training corpus.

2.3 Reference Combination Method

The final goal of the system is to assign a full name to each anonymous speaker. Let us recall here the combination method of information provided by the trees we proposed in [4]. Let $\mathcal{E} = \{e_1, \ldots, e_I\}$ denotes the *closed* set of full names hypotheses to assign to a speaker. These candidates come from an exhaustive list of possible speakers known by the system. The set $\mathcal{O} = \{o_1, \ldots, o_J\}$ corresponds to the successive occurrences of full names detected in the transcripts, $\mathcal{T} = \{t_1, \ldots, t_K\}$ is the set of the speaker turns in chronological order, and $\mathcal{C} = \{c_1, \ldots, c_L\}$ is the set of anonymous speakers to be labeled. Thus, the main goal is to assign a full name of $\mathcal{E}$ to a speaker of $\mathcal{C}$. Each speaker c_l may involve one or several times in a broadcast, that corresponds to several speaker turns: $c_l = \{t \in \mathcal{T} \mid c_l \text{ is the speaker of turn } t\}$. In a same turn, several occurrences of full names may be detected. For each occurrence of a full name o_j (for $j = 1, \ldots, J$) detected in a given speaker turn t_k, let us define by $P(o_j, t_k)$ the probability that o_j is current speaker. Thus, $P(o_j, t_{k-1})$ and $P(o_j, t_{k+1})$ represents the probability that o_j is respectively the speaker of the previous and the following turn. By hypothesis, the probability that o_j is another speaker is: $1 - \sum_{r=-1}^{1} P(o_j, t_{k+r})$. At this stage, a filter is made by the comparison of genders: if the gender of the full name e_i and the speaker c_l are different, the corresponding occurrence is ignored. Let $g(e_i)$ and $g(c_l)$ be the respective gender (female, male or unknown) of e_i and c_l. The speaker gender is detected by the acoustic segmentation and classification system with high reliability and the full name gender is determined by the first name (generally without ambiguity) from a linguistic base of first names.

In [4], to assign a full name e_i to a speaker c_l, we have computed a "score" for each full name e_i, denoted as $s_l(e_i)$. This score is no more a probability, but is simply a sum of probabilities concerning the speaker turns of c_l and taking gender constraints into account:

$$s_l(e_i) = \sum_{\{(o_j,t) | o_j = e_i,\ t \in c_l, g(e_i) = g(c_l)\}} P(o_j, t) \tag{1}$$

2.4 Decision Process

The goal is now to assign a full name e_i to each speaker c_l. Let $f : \mathcal{C} \rightarrow \mathcal{E}$ be the assignment function of full names to speakers. The principle of our solution, proposed in [4], is actually to find a coherent matching between full names and speakers. Let $\mathcal{D} = \{c_l \in \mathcal{C} \mid \forall e_i \in \mathcal{E}, s_l(e_i) = 0\}$, be the set of speakers with no potential candidate. Several strategies may be used to sort the competing speakers c_l for a given full name e_i. The more natural way seems to choose the full name which has the maximum score for a given speaker e_i (if there is at least one non-null score). Let us define the rule $\mathbf{R_1}$ by:

$$\begin{gathered} \forall c_l \in \mathcal{C} \setminus \mathcal{D}, e_i^* = \arg\max_{e_i \in \mathcal{E}} s_l(e_i) \Rightarrow f(c_l) = e_i^* \\ \forall c_l \in \mathcal{D}, f(c_l) = \text{Anonymous}. \end{gathered} \tag{2}$$

An issue is that the same full name e_i^* may be assigned to several speakers. We proposed to reorganize the sharing of full names among speakers. Let the coefficient β_{il} define the relative score of e_i among all the possible candidates for assignment to c_l: $\beta_{il} = \frac{s_l(e_i)}{\sum_{q=1}^{I} s_l(e_q)}$ if $c_l \notin \mathcal{D}$ and $\beta_{il} = 0$ if $c_l \in \mathcal{D}$. A concrete example is given in Table 1. The full name "Jacques Derrida" has been assigned to three different speakers from the decision rule in Equation 2. In this example, c_{13} has the best score, and "Jacques Derrida" should be assigned to c_{13} ; but the score represents only 35% of the total scores among all the possible candidates for c_{13}, whereas the score for c_{15} represents 80% of total scores. Then, for the final decision, we have proposed to use the product of score $s_l(e_i)$, by coefficient β_{il} (rule $\mathbf{R_2}$):

$$\mathrm{SC}_l(e_i) = s_l(e_i)\beta_{il}. \tag{3}$$

Finally, in the same example, "Jacques Derrida" is assigned to c_{15} and the speakers c_{13} and c_{14} will be labeled with other full names. The algorithm is iterative: all the full names are taken *a priori* into account and sorted according to their score $\mathrm{SC}_l(e_i)$. First, the full name with the maximum score (denoted e_i^*) is chosen, and if several speakers are associated to the same e_i^*, then e_i^* is assigned to the speaker whose score $SC_l(e_i^*)$ is maximum. Then, all chosen full names are deleted from the list of speakers that are not yet assigned in this first iteration. In a second iteration, remaining full names are examined in the same way for the remaining speakers and so on, until all speakers are assigned, or their list is empty. Table 2 shows the result of this algorithm for the preceding example.

Table 1. Example of an initial multiple assignment

Speaker	Full name e_i^*	$s_l(e_i^*)$	β_{il}	$\mathrm{SC}_l(e_i^*)$
c_{13}	Jacques Derrida	**8.58**	35%	3.00
c_{14}	Jacques Derrida	1.67	56%	0.94
c_{15}	Jacques Derrida	4.94	**80%**	**3.95**

Table 2. Example of the decision process with two iterations (decision in bold type, scores in parenthesis)

Speaker	e_i^* (1st iteration)	2nd iteration
c_{13}	J. Derrida (3.00)	**N. Demorand** (0.25)
c_{14}	J. Derrida (0.94)	**A. Adler** (0.56)
c_{15}	**J. Derrida** (3.95)	-
c_{16}	**O. Duhamel** (1.15)	-

2.5 Drawbacks of the Combination Method

The combination method described above has several serious drawbacks, even though it has yielded good results [4]. First, the concept of score is difficult to

interpret, the quantities obtained in Equations 2 and 3 do not represent a degree of confidence, or a probability that a full name is a given speaker. They lead to a lack of clarity of the decision. Equation 3 represents a compromise that is difficult to justify.

But the main drawback concerns the lack of uncertainty management in the combination method: particularly, conflict information in a given speaker turn is not taken into account. The available information is not correctly combined as a whole. No link is made between the different information provided by the classification tree, particularly when a same speaker pronounces several different full names and can therefore lead to erroneous results. Table 3 presents an example of a speaker turn t_k where 8 occurrences are detected. The probabilities correspond to the next speaker turn t_{k+1}, who is a male. A female full name is therefore eliminated and two full names are rejected because they do not belong to the list. Some occurrences are redundant, because they correspond to a repeated full name and only one occurrence has a relative high probability. Two full names are still competing and they represent a significant incompatibility. These full names have high scores: Jean-Claude Pajak (1.25) and Jacques Chirac (0.87). These scores are close to those obtained if we had some information without ambiguity, for example a turn with only one occurrence with a high probability. This example highlights the fact that this method does not take into account the contradictory information provided by some speaker turns. A probabilistic formalism based on conditional probabilities could be considered for this kind of situation, but the lack of *a priori* information makes this type of modeling difficult. Even though the classification tree outputs are probabilistic, belief theory seemed more appropriate and less restrictive, particularly in the flexibility of its use.

Table 3. Score contribution in a speaker turn (t_{k+1} is a male)

Occurrence o_j	gender	belongs to the list	$P(o_j, t_{k+1})$	score
Oscar Temaru	*M*	*No*	*0.29*	–
Hamid Karzaï	*M*	*No*	*0.29*	–
Jacques Chirac	M	Yes	0.29	0.87
Jacques Chirac	M		0.29	
Jacques Chirac	M		0.29	
Jean-Claude Pajak	M	Yes	0.29	
Jean-Claude Pajak	M		**0.96**	1.25
Véronique Rebeyrotte	*F*	*Yes*	*0.29*	–

3 Belief Functions for Speaker Recognition

The contribution of this article lies in the combination process of different information, especially from the classification tree.

3.1 Belief Function Theory

In this section, we briefly recall some notions of the belief function theory [6,7]. In this article, we adopt the point of view proposed by Smets: the Transferable Belief Model (TBM) [7]. The aim of this model is to determine the belief concerning different propositions, from some available information. Let Ω be a finite set, called frame of discernment of the experience. The representation of the uncertainty is made by the means of the concept of belief function, defined as a function m from 2^{Ω} to $[0,1]$ such as $\sum_{A\subseteq\Omega} m(A) = 1$. The quantity $m(A)$ represents the belief exactly allowed to proposition A. The subsets A of Ω such as $m(A) > 0$ are called the *focal elements* of m. One of most important operations in the TBM is the procedure for aggregating operator to combine several belief functions defined in a same frame of discernment [7]. In particular, the combination of two belief functions m_1 and m_2 "independently" defined on Ω using the conjunctive binary operator $\cap$, denoted as $m' = m_1 \cap m_2$, is defined as [7]:

$$\forall A \subseteq \Omega,\ m'(A) = \sum_{B\cap C=A} m_1(B)m_2(C). \tag{4}$$

Repeatedly, we may define the combination of n functions $m_1, \ldots, m_n$ on Ω by: $m = m_1 \cap \ldots \cap m_n$. Once a belief function m is defined, it is possible to transform it into a probability distribution particularly for decision aspects. One of these, called *pignistic* probability and denoted by P_m, is defined for all $\omega \in \Omega$ as [7], if $m(\emptyset) \neq 1$:

$$P_m(\{\omega\}) = \sum_{A\subset\Omega} \frac{m(A)}{|A|(1-m(\emptyset))}\delta_A(\omega), \tag{5}$$

where $|A|$ denotes the cardinality of A, $\delta_A(\omega) = 1$ if $\omega \in A$ and $\delta_A(\omega) = 0$ if $\omega \notin A$.

3.2 Definition of Belief Masses

In this article, we propose to improve the system described in [4] by taking into account the *coherence* of the whole information provided by *contiguous* speaker turns. As we have seen before, in a *same* turn, several occurrences corresponding to different full names may be detected.

First, we focus on a turn t_k with n_k occurrences and owing to speaker c_l. Let n_{k+r} be the number of occurrences for the previous turn ($r = -1$) and the following one ($r = 1$). Let $\{o^k_{j,r}\}$, with $r = -1, 0, 1$ and $j = 1, \ldots, n_{k+r}$, be the occurrences of the detected full names in these three turns. Each occurrence $o^k_{j,r}$, corresponding to a label e_i, represents some knowledge concerning the speaker of the turn t_k that can be described by a simple support belief function $m^{jr}_{t_k}$ on $\mathcal{E}$, focused on e_i and $\mathcal{E}$:

$$\begin{cases} m^{jr}_{t_k}(\{e_i\}) = \alpha_{ij} P(o^k_{j,r}, t_{k-r}) \text{ si } o^k_{j,r} = e_i \\ m^{jr}_{t_k}(\mathcal{E}) = 1 - \alpha_{ij} P(o^k_{j,r}, t_{k-r}), \end{cases} \tag{6}$$

Table 4. Mass distribution of the belief function in a speaker turn

Focal elements	$m_{t_{k+1}}(\{e_i\})$
Jacques Chirac	0.018
Jean-Claude Pajak	**0.348**
$\emptyset$	**0.624**
$\mathcal{E}$	0.010

where $\alpha_{il} \in [0,1]$ is a confidence measure of gender compatibility between e_i and c_l. If the genders are known with certainty, $\alpha_{il} = 0$ if $g(e_i) \neq g(c_l)$ and $\alpha_{il} = 1$ if $g(e_i) = g(c_l)$. If the first names are ambiguous (like Dominique in French) or unspecified, or if the speaker gender is uncertain, $\alpha_{il} \in]0,1[$ is estimated from a database of first names and the training corpus. Table 4 presents the belief function concerning the speaker of turn t_{k+1} in the example seen in section 2.5. The belief mass of "Jean-Claude Pajak" is still high while the one of the other candidate is very low, and the degree of conflict is important since the mass of the empty set is high.

3.3 Combination by Speaker

The first combination step consists in aggregating the whole information in a given speaker turn. The combination of the $n_{k-1} + n_k + n_{k+1}$ belief functions focused on the t_k and obtained by Equation 6 is made with conjunctive *non normalized* Dempster rule (Equation 4), in order to ensure associativity and commutativity of the combination: we obtain a belief function m_{t_k} that represents a more synthetic knowledge of speaker identity provided in turn t_k, defined by:

$$m_{t_k} = \bigcap_{r=-1}^{1} \bigcap_{j=1}^{n_{k+r}} m_{t_k}^{jr}. \tag{7}$$

The second combination step consists in aggregating the results obtained by each speaker turn for the whole broadcast news. The more relevant and natural consists in keeping on combining all the belief functions focused on the same speaker c_l with the same conjunctive Dempster rule and therefore combining all the belief functions corresponding to the speaker turns t_k of this speaker. Thus, we obtain a global belief function M_l which represents the state of belief concerning speaker c_l for the whole broadcast news, and defined by:

$$M_l = \bigcap_{t_k \in c_l} m_{t_k} \tag{8}$$

3.4 Decision Rule

We use a similar procedure presented in section 2.4, but the decision process is simplified and unified thanks to the use of belief functions. We transform the

belief function M_l into the pignistic probability P_{M_l} (Equation 5) and we obtain the following rule **R**:

$$\begin{gathered} \forall c_l \in \mathcal{C} \setminus \mathcal{D}, e_i^* = \arg\max_{e_i \in \mathcal{E}} \mathrm{P}_{M_l}(e_i) \Rightarrow f(c_l) = e_i^* \\ \forall c_l \in \mathcal{D}, f(c_l) = \text{Anonymous}. \end{gathered} \tag{9}$$

Then, since some full names may initially be assigned to several speakers, we apply the same decision process developed in 2.4, replacing scores SC_l by pignistic probabilities P_{M_l}. If we come back to the proposed example in 2.4, the full name "Jacques Derrida" is again initially assigned to three speakers c_{13}, c_{14} and c_{15} (see Table 5). Finally, "Jacques Derrida" is also assigned to c_{15}, because this speaker has the most important pignistic probability.

Table 5. Decision with two iterations (decision in bold, belief masses in parentheses)

Speaker	e_i^* (1st iteration)	2nd iteration
c_{13}	J. Derrida (0.89)	**N. Demorand** (0.11)
c_{14}	J. Derrida (0.71)	**A. Adler** (0.25)
c_{15}	**J. Derrida** (0.99)	-
c_{16}	**O. Duhamel** (0.88)	-

4 Evaluation of the Proposed System

4.1 Data Description

The system evaluation are realized on French broadcast news records from the French ESTER 1 phase II evaluation campaign [2]. The data were recorded from 5 French radios and *Radio Télévision Marocaine* and last from 10 to 60 minutes. They are divided in 3 corpora used for the SCT training, the system development and the evaluation: the training corpus of 76h (7416 speaker turns, 11292 detected full names), the development corpus of 30h (2931 speaker turns, 4533 full names) and the test corpus of 10h (1082 speaker turns, 1541 full names). This corpus contains two radios which are not present in the training and the development corpora. It was also recorded 15 months after the previous data.

4.2 Metrics

The results are evaluated comparing the generated hypothesis and the reference. This comparison highlights five cases:

- Identity is correct (C_1): the identity hypotheses corresponds to the correct one in the reference.
- Substitution error (S): the identity hypotheses differs from the one found in the reference.

- Deletion error (D) : no identity is proposed although the speaker is identified in the reference.
- Insertion error (I): an identity is proposed although the speaker is not identified in the reference.
- No identity (C_2): no identity is proposed, and there is no identity for this speaker in the reference.

Among the measures defined in [8,4], the one that seems to best summarize the results is the global error rate Err computed from these 5 quantities:

$$Err = \frac{S + I + D}{S + I + D + C_2 + C_1}. \tag{10}$$

The errors may be computed in terms of duration or in terms of number of speakers.

4.3 System Evaluation

During experiments, the system is supposed to know all the full names that may be the speakers. This list is composed 1008 full names (the set $\mathcal{E}$). Comparison between the reference system (c.f. [4] and section 2.3), and the proposed system is made on manual transcripts and segmentations. However, the named entities detection is automatic and may have some errors. The reference system is described in section 2.3 with two rules using scores $s_l(e_i)$ and $SC_l(e_i)$ (c.f. Equations 2 and 3) and our model is based on belief functions (Equation 9).

As Table 6 shows, in the new model, the error rate in terms of duration ($ErrDur$) is 3 points less than reference system with rule $\mathbf{R}_2$ and 7 points less with rule $\mathbf{R}_1$. Not only the use of belief functions is more easily interpretable, but also it allows to eliminate much errors. Concerning the number of identified speakers, the result is even more obvious: the new system correctly labels much more speakers than the base system, and also improves the reference system. In conclusion, taking account global information on speakers within a speaker turn, and before the decision, allows to significantly improve results both in terms of duration and in terms of numbers of speakers.

Table 6. Comparison between the proposed system and the reference system according to the decision rule on the test corpus of ESTER 1 phase II campaign (1541 detected full names); **ErrDur**: error rate in duration; **ErrSpk**: error rate in number of speakers

System	ErrDur	ErrSpk
Reference (rule $\mathbf{R_1}$)	20.6%	20.2%
Reference (rule $\mathbf{R_2}$)	16.6%	19.5%
Proposed (rule $\mathbf{R}$)	**13.7%**	**14.9%**

5 Conclusion

The speaker identification method proposed in this article allows to extract speaker identities from transcriptions. The identification is realized thanks to a semantic classification tree which helps to give the full names found in the transcription to speakers in a recording. In this article, we propose a new system that consistently combines different information about the potential speakers in the form of belief functions. Particularly, the system manages possible conflict of information on the speakers within a speaker turn and takes into account the uncertainty concerning the gender. Experiments have been realized on a French broadcast news and the system have very good performances. Future work will focus on developing solutions to deal with automatic outputs containing errors. Different kind of uncertainty, dues to segmentation error, classification in speakers or to the bad transcription of full names will be taken into account. We will also study the realistic case of open systems when the list of possible speakers is unknown.

References

1. Canseco-Rodriguez, L., Lamel, L., Gauvain, J.-L.: A comparative study using manual and automatic transcriptions for diarization. In: Automatic Speech Recognition and Understanding, San Juan, pp. 415–419 (2005)
2. Galliano, S., Geffroy, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G.: The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In: European Conference on Speech Communication and Technology (2005)
3. Kuhn, R., De Mori, R.: The application of semantic classification trees to natural language understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence 17(5), 449–460 (1995)
4. Jousse, V., Petitrenaud, S., Meignier, S., Estève, Y., Jacquin, C.: Automatic named identification of speakers using diarization and ASR systems. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Taipei, pp. 4557–4560 (2009)
5. Mauclair, J., Meignier, S., Estève, Y.: Speaker diarization: about whom the speaker is talking? In: IEEE Odyssey (2006)
6. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
7. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66, 191–234 (1994)
8. Tranter, S.E.: Who really spoke when? Finding speaker turns and identities in broadcast news audio. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1013–1016 (2006)

Constructing Multiple Frames of Discernment for Multiple Subproblems*

Johan Schubert

Division of Information Systems,
Swedish Defence Research Agency,
SE-164 90 Stockholm, Sweden
schubert@foi.se
http://www.foi.se/fusion/

Abstract. In this paper we extend a methodology for constructing a frame of discernment from belief functions for one problem, into a methodology for constructing multiple frames of discernment for several different subproblems. The most appropriate frames of discernment are those that let our evidence interact in an interesting way without exhibit too much internal conflict. A function measuring overall frame appropriateness is mapped onto a Potts spin neural network in order to find the partition of all belief functions that yields the most appropriate frames.

Keywords: Dempster-Shafer theory, belief function, representation, frame of discernment, clustering, Potts spin, conflict, simulated annealing.

1 Introduction

In this paper we extend a methodology for constructing a frame of discernment for *one* problem [1] into a methodology for constructing multiple frames of discernment from a set of belief functions [2, 3] for *several* different subproblems. These belief functions are assumed to concern different subproblems that should be handled separately. Previously, we have developed methods for clustering belief functions that are mixed-up [4–8] based on their pairwise conflicts. These methods were developed to manage simple support and consonant belief functions. The case with nonconsonant belief functions can be handled by decomposition into simple support functions followed by clustering of the decomposed parts [9]. If the number of clusters K (in Fig. 4.) is unknown, it can be estimated by observing the change in the logarithm of a meta frame appropriateness function (*MFA*) for different number of clusters [10, p. 90], or inferred using specification [5] and *a priori* information [11], or managed by particle filtering methods [12].

The methodology for constructing a frame of discernment is extended by adopting a measure of frame appropriateness for a single problem into handling multiple subproblems. This new function is mapped onto a Potts spin neural network. We reuse a previously developed methodology for clustering large amounts of belief function in such a manner as to find the best frames of discernment for these subproblems. When

* This work was supported by the FOI research project "Real-Time Simulation Supporting Effects-Based Planning", which is funded by the R&D programme of the Swedish Armed Forces.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 189–198, 2010.

the mixed-up belief functions are partitioned into subsets corresponding to the different subproblems, a frame of discernment is constructed within each subproblem using the methodology developed for a single problem [1].

In Sec. 2 we study the problem of construction alternative frames of discernment. In Sec. 3 we extend this methodology to multiple subproblems. In Sec. 4 we review Potts spin theory. We then put everything together by mapping the multiple frame construction problem onto Potts spin (in Sec. 5). In Sec. 6 we present an algorithm for constructing multiple frames. Finally, in Sec. 7 conclusions are drawn.

2 Constructing Alternative Frames of Discernment

Let us assume we have a set of evidence $\chi = \{m_i\}$ that originates from *one* problem with yet undetermined representation. The focal elements of each belief function m_i contain pieces of that representation. Our task is to find the most appropriate frame of discernment that lets our evidence "interact in an interesting way" without "exhibit too much internal conflict" in the words of Glenn Shafer [3, p. 280].

This will usually not be the union of all cores of m_i as different cores may hold non-exclusive elements. For example, one belief function may assign support to a focal element "Red" in relation to the color of a car. Another belief function may assign support to a focal element "Fast" in relation to speed of that car. Obviously, "Red" and "Fast" are not both elements of the frame of discernment as they are not exclusive. However, the "(Red, Fast)" pair might be an element of the frame.

Our task of finding the most appropriate frame of discernment becomes finding the most appropriate cross product of different unions of cores. Let us begin by introducing the representation needed to construct a frame of discernment from input data.

For an example of the material in Sec. 2 see [1].

2.1 Representation

Assume we have a set of evidence χ. We observe the core C_i of each available belief function m_i. We assume that the core of each belief function is a subset of exclusive but not exhaustive elements of a so far unconstructed frame of discernment.

Let $C = \{C_i\}$ be the set of all cores of χ, where C_i is the core of m_i, the ith piece of evidence. We have

$$C_i = \bigcup_j \{A_j | m_i(A_j) > 0\} \tag{1}$$

where A_j is a focal element of m_i.

Let $\Omega = \{\Omega_k\}$ be the set of all possible set partitions of C (the set of all cores), where Ω_k is the kth possible partition of C. We have

$$\Omega_k = \{\omega_l\} \tag{2}$$

where the ω_l's are disjoint subset of C, i.e., $\forall l.\ \omega_l \subseteq C$ such that

$$\bigcup_l \omega_l = \{C_i\} \equiv C \tag{3}$$

and $\omega_m \cap \omega_n = \varnothing$ whenever $m \neq n$.

Let $\Theta = \{\Theta_k\}$ be the set of all possible cross products relating to Ω, such that Θ_k is the cross product of all unions of elements of the partition Ω_k, (2). We have

$$\Theta_k = \times \{\theta_l\} \tag{4}$$

where θ_l is the union of the elements in ω_l, $\omega_l \in \Omega_k$, and θ_l must be an exclusive set of elements. We have

$$\forall l.\ \theta_l = \cup \omega_l = \bigcup_i \{C_i | C_i \in \omega_l\} \tag{5}$$

such that

$$\bigcup_l \theta_l = \bigcup_l \{\cup \omega_l\} = \bigcup_i \{C_i\} = \cup C \tag{6}$$

where all θ_l's observe two different crucial type conditions:

Type Condition 1. No element of any θ_p may belong to any other cross product elements θ_q, i.e.,

$$\theta_p \cap \theta_q = \varnothing, \tag{7}$$

whenever $p \neq q$.

This will eliminate any frame that obviously distributes elements of the same type over different cross product elements.

Type Condition 2. Every cross product element θ_l must be an exclusive set, i.e.,

$$e_m \cap e_n = \varnothing, \tag{8}$$

whenever $\forall m,n \exists l.\ e_m, e_n \in \theta_l$.

2.2 Abridgment

For all possible frames of discernment $\{\Theta_k\}$, where $|\Theta_k| > 1$, we may include further assumptions that make the frames tighter. This may lead to more interesting interaction between the belief functions and lead to firmer conclusions provided that the conflict does not increase in any significant way. Every frame is based on assumptions. The frame we begin with is based on the assumption that the elements of that frame are all disjunct possible alternatives, and that no other possibilities exists. Whether a tighter or looser frame is to be preferred is a matter of appropriateness. Most often this will be a point of balance where meaningful interaction is weighted against too much conflict.

Let us study one particular frame of discernment Θ_i from the remaining set of possible frames Θ that observe both type condition 1 and 2, (7) and (8), respectively. We have $\Theta_i = \times \{\theta_l\}$. At least one cross product element θ_l must be abridged to construct a new abridged frame of Θ_i. We have a set of all possible abridgments of Θ_i,

$$\Theta_i' = \{\Theta_{ij}'\}_j = \{\times \{\theta_{lj}'\}\}_j \tag{9}$$

where $\theta_{lj}' \in 2^{\theta_l}$ and 2^{θ_l} is the power set of θ_l, $\theta_{lj}' \neq \varnothing$, and $\exists j.\ \theta_{lj}' \neq \theta_l$.

2.3 Enlargement

We may make enlargements to any frame of discernment $\{\Theta_k\}$. The only enlargement we can perform is to enlarge a particular cross product element θ_l with an element of unstated meaning. Let us denote these elements Λ_l, one for each θ_l.

Let us again look at $\Theta_i = \times \{\theta_l\}$. For each cross product element θ_l there is one possible enlargement: enlarging θ_l by Λ_l. At least one cross product element θ_l must be enlarged to construct a new enlarged frame of Θ_i. The set of all possible enlargements

of Θ_i becomes $\Theta_i'' = \{\Theta_{ij}''\}_j = \{\times \{\theta_{lj}''\}\}_j$ where $\theta_{lj}'' \in \{\theta_l, \theta_l + \{\Lambda_l\}\}$ and $\exists j.\ \theta_{lj}'' \neq \theta_l$.

2.4 Appropriate Representation

We evaluate the alternative frames of discernment on the grounds of being appropriate for yielding interesting interactions among the available belief functions without exhibiting too much internal conflict. A measure of frame appropriateness was defined in [1]. This measure gives an equal weight to both conditions were both must be appropriate simultaneously (see [1]).

Definition 1. Let Θ_k be a frame of discernment and let $\{m_j\}$ be a set of all available belief functions defined on Θ_k. We define a measure of frame appropriateness of Θ_k, denoted as $FA(\Theta_k)$, by

$$FA(\Theta_k | \{m_j\}) = \left[1 - Con(\oplus\{m_j|\Theta_k\})\right]\left[1 - \frac{AU(\oplus\{m_j|\Theta_k\})}{\log_2|\Theta_k|}\right], \tag{10}$$

where *Con* is the conflict in Dempster's rule and *AU* is the functional called the aggregated uncertainty. We have $Con \in [0, 1]$, $AU \in [0, \log_2|\Theta_k|]$ and $FA \in [0, 1]$.

The aggregated uncertainty functional *AU* [13–15] is defined as

$$AU(\text{Bel}) = \max_{\{p_x\}_{x \in \Theta}} \left\{ -\sum_{x \in \Theta} p(x)\log_2 p(x) \right\}, \tag{11}$$

where $\{p_x\}_{x \in \Theta}$ are all probability distributions such that $p_x \in [0, 1]$ for all $x \in \Theta$,

$$\sum_{x \in \Theta} p(x) = 1 \tag{12}$$

and

$$\text{Bel}(A) \leq \sum_{x \in A} p(x) \tag{13}$$

for all $A \subseteq \Theta$.

2.5 An Algorithm for Computing *AU*

An algorithm for computing *AU* was found by Meyerowitz *et al.* [16]. For the sake of completeness we cite the algorithm here, in the way it is described by Harmanec *et al.* [17], Fig. 1. The computational time complexity of *AU* is $O(2^{|\Theta|})$.

Input: a frame of discernment X, a belief function *Bel* on X.
Output: $AU(Bel)$, $\{p_x\}_{x \in X}$ such that $AU(Bel) = -\sum_{x \in X} p_x \log_2 p_x$, $p_i \geq 0$, $\sum_{x \in X} p_x = 1$, and $Bel(A) \leq \sum_{x \in X} p_x$ for all $\emptyset \neq A \subseteq X$.
Step 1. Find a non-empty set $A \subseteq X$, such that $Bel(A) / |A|$ is maximal. If there are more than such sets A than one, take the one with maximal cardinality.
Step 2. For $x \in A$, put $p_x = Bel(A) / |A|$.
Step 3. For each $B \subseteq X - A$, put $Bel(B) = Bel(B \cup A) - Bel(A)$.
Step 4. Put $X = X - A$.
Step 5. If $X \neq \emptyset$ and $Bel(X) > 0$, then go to Step 1.
Step 6. If $Bel(X) = 0$ and $X \neq \emptyset$, then put $p_x = 0$ for all $x \in X$.
Step 7. Calculate $AU(Bel) = -\sum_{x \in X} p_x \log_2 p_x$.

Fig. 1. An algorithm for computing $AU(Bel)$

2.6 An Algorithm for Constructing a Frame of Discernment

Using the results of the preceding sections we develop an algorithm for constructing and evaluating all possible frames of discernment. This algorithm will first generate the possible frames using different partitions of the set of all cores. From these possible frames we generate abridgments and enlargements. The frames are evaluated using *FA*, (10), in Fig. 2. The most appropriate frame that maximizes *FA* is then selected.

Input: a set of belief functions χ.
Output: Possible frames of discernment $\{\Theta_i\}$, $\{\Theta'_{ij}\}$, $\{\Theta''_{ij}\}$. Frame appropriateness $\forall ij$. $FA(\Theta_i|\chi)$, $FA(\Theta'_{ij}|\chi)$, $FA(\Theta''_{ij}|\chi)$.
Step 1. $\forall i$. generate C_i using (1). Set $C = \{C_i\}$.
Step 2. $\forall k$. generate Ω_k using (2). Set $\Omega = \{\Omega_k\}$.
Step 3. $\forall k$. generate Θ_k using (4). Set $\Theta = \{\Theta_k\}$.
Step 4. $\forall ij$. generate $\{\Theta'_{ij} \mid \forall kl. Con(\oplus\{m_j|\Theta'_{kl}\}) < 1\ ,\ \Theta'_{kl} \supset \Theta'_{ij}\}$ using (9).
Step 5. $\forall k$. If $Con(\oplus\{m_j|\Theta_k\}) > 0$ then $\forall j$. generate Θ''_{ij}. Set $\Theta''_i = \{\Theta''_{ij}\}_j$.
Step 6. Compute frame appropriateness $\forall ij$. $FA(\Theta_i|\chi)$, $FA(\Theta'_{ij}|\chi)$, $FA(\Theta''_{ij}|\chi)$ using (10).

Fig. 2. An algorithm for generating and evaluating appropriate frames of discernment

Brute force implementation of *FA* has a computational time complexity of $O(|\chi|^{|\chi|}2^{|\Theta|})$. Implementing step 2–4 in an iterative way may reduce the term $|\chi|^{|\chi|}$ of the time complexity.

3 Constructing Multiple Frames of Discernment

In this section we extend the methodology from Sec. 2 into a new methodology for constructing several multiple frames of discernment for different subproblems. This is done by extending *FA* (10) to several subsets. Let us define such a function of overall frame appropriateness.

Definition 2. Let the meta frame appropriateness function,

$$MFA(\{\chi_a\}_a) \triangleq \prod_a FA_a\,, \tag{14}$$

over several subproblems χ_a be the product of the frame appropriateness functions FA_a for these subproblems χ_a.

In order to find the best frames of discernment for these subproblems we maximize $MFA(\{\chi_a\}_a)$,

$$\max MFA(\{\chi_a\}_a) = \max \prod_a FA_a\,. \tag{15}$$

For computational reasons the actual maximization of *MFA* is done in several steps.

First, let us map *MFA* onto a Potts spin neural network that will cluster all belief functions into subsets using an approximation of *MFA* as a distance measure in such a manner that *MFA* is maximized. This will partition the belief functions into subsets that should be handle separately is such a way that it gives us the best overall frames of discernment for the subproblems.

Secondly, for each subproblem separately, a frame of discernment is constructed using the algorithm for constructing an appropriate frame of discernment, Fig. 2. With

these frames of discernment each subproblem can be solved separately by combining all belief functions in the subset.

4 Potts Spin Theory

The Potts spin problem [18] consists of minimizing an energy function

$$E = \frac{1}{2} \sum_{i,\, j\, =\, 1}^{N} \sum_{a\, =\, 1}^{q} J_{ij} S_{ia} S_{ja} \tag{16}$$

by changing the states of the spins S_{ia}'s, where $S_{ia} \in \{0, 1\}$ and $S_{ia} = 1$ means that belief function i is in cluster a. This model serves as a clustering method if J_{ij} is used as a penalty factor when report i and j are in the same cluster.

The minimization is carried out by simulated annealing. In simulated annealing temperature is an important parameter. The process starts at a high and continues by gradually lowering the temperature. As the temperature is lowered the spins gradually become more influenced by the interactions J_{ij}'s so that a minimum of the energy function (16) is reached. This gives us the best partition of all belief functions into the clusters with minimal energy function.

For computational reasons we use a mean field model, where spins are deterministic with $V_{ia} = \langle S_{ia} \rangle$, $V_{ia} \in [0, 1]$. The Potts mean field equations are formulated [19] as

$$V_{ia} = \frac{e^{-H_{ia}[V]/T}}{\sum_{b\, =\, 1}^{K} e^{-H_{ib}[V]/T}} \tag{17}$$

where

$$H_{ia}[V] = \sum_{j\, =\, 1}^{N} J_{ij} V_{ja} - \gamma V_{ia} \,. \tag{18}$$

In order to minimize the energy function, (17) and (18) are iterated until a stationary equilibrium state has been reached for each temperature. Then, the temperature is lowered step–by–step by a constant factor until $\forall i, a \,.\, V_{ia} \approx 0, 1$ in the stationary equilibrium state.

The time complexity of Potts spin clustering was shown to be $O(N^2 K^2)$ in terms of the number of belief functions N (= $|\chi|$) and the number of clusters K [7].

5 Mapping a Multiple Frame Construction Problem onto Potts Spin

In order to map the meta appropriateness function *MFA* onto a Potts spin network we need to rewrite *MFA* as a sum of terms similar to the energy function being minimized in (16). To find the best set of frames of discernment we maximize $MFA(\{\chi_a\}_a)$. This can be rewritten as a sum of terms over the subsets χ_a

$$\max MFA(\{\chi_a\}_a) = \max \prod_a FA_a, \tag{19}$$

$\Leftrightarrow$

$$\begin{aligned}\max \log \prod_a FA_a &= \max \log \prod_a \left[1 - Con(\oplus\{m_j|\Theta_a\})\right]\left[1 - \frac{AU(\oplus\{m_j|\Theta_a\})}{\log_2|\Theta_a|}\right] \\ &= \max \sum_a \log\left[1 - Con(\oplus\{m_j|\Theta_a\})\right] + \log\left[1 - \frac{AU(\oplus\{m_j|\Theta_a\})}{\log_2|\Theta_a|}\right] \\ &= \min \sum_a -\log\left[1 - Con(\oplus\{m_j|\Theta_a\})\right] - \log\left[1 - \frac{AU(\oplus\{m_j|\Theta_a\})}{\log_2|\Theta_a|}\right].\end{aligned} \tag{20}$$

Furthermore, we must also rewrite *MFA* as a sum of terms over pairwise simple support functions as all interactions in Potts spin are pairwise.

It was shown in [20] that minimizing a sum of $-\log(1 - s_k s_l)$ terms is an approximation correct to leading order, i.e., all second order terms in $\{s_i\}$ are unchanged in this approximation. The first term in the last line of (20) can be rewritten as

$$\sum_a -\log\left[1 - Con(\oplus\{m_j|\Theta_a\})\right] = \sum_a -\log\left[1 - \left(\sum_{\substack{k,l \\ S_k, S_l \in \chi_a}} s_k s_l - X\right)\right], \tag{21}$$

while the actual function being minimized in the neural network is

$$\begin{aligned}\sum_a \sum_{\substack{k,l \\ S_k, S_l \in \chi_a}} -\log(1 - s_k s_l) &= \sum_a -\log\left[\prod_{\substack{k,l \\ S_k, S_l \in \chi_a}} (1 - s_k s_l)\right] \\ &= \sum_a -\log\left[1 - \left(\sum_{\substack{k,l \\ S_k, S_l \in \chi_a}} s_k s_l - Y\right)\right],\end{aligned} \tag{22}$$

where X and Y are higher order terms.

The second term in the last line of (20) can be rewritten as

$$\sum_a -\log\left[1 - \frac{AU(\oplus\{m_j|\Theta_a\})}{\log_2|\Theta_a|}\right] \approx \sum_a -\log\left[1 - \left(\sum_{\substack{k,l \\ m_k, m_l \in \chi_a}} \frac{AU(\oplus\{m_k, m_l|\Theta_a\})}{(|\chi_a| - 1)\cdot \log_2(|\Theta_a|)} - W\right)\right] \tag{23}$$

when $|\chi_a| \geq 2$, and where W are higher order terms in AU (i.e., $(AU)^p$, $p \geq 2$).

When calculating $AU(\oplus\{m_j|\Theta_a\})$ and $AU(\oplus\{m_k, m_l|\Theta_a\})$ in (23) each leading term comes in twice in AU. However, since $AU(\oplus\{m_k, m_l|\Theta_a\})$ is summed up for all pairs m_k, m_l in χ_a the leading terms come in $2\cdot(|\chi_a| - 1)$ times in the equation. To compensate for this multiple counting we must include a factor $1/(|\chi_a| - 1)$. This approximation is correct in its first order terms.

Thus, the function being minimized is

$$\sum_{a} \sum_{\substack{k,l \\ m_k, m_l \in \chi_a}} -\log\left(1 - \frac{AU(\oplus\{m_k, m_l|\Theta_a\})}{(|\chi_a|-1)\cdot \log_2(|\Theta_a|)}\right) = \sum_{a} -\log\left[\prod_{\substack{k,l \\ m_k, m_l \in \chi_a}} \left(1 - \frac{AU(\oplus\{m_k, m_l|\Theta_a\})}{(|\chi_a|-1)\cdot \log_2(|\Theta_a|)}\right)\right]$$

$$= \sum_{a} -\log\left[1 - \left(\sum_{\substack{k,l \\ m_k, m_l \in \chi_a}} \frac{AU(\oplus\{m_k, m_l|\Theta_a\})}{(|\chi_a|-1)\cdot \log_2(|\Theta_a|)} - Z\right)\right] \tag{24}$$

which is identical in its first order terms to (23).

Thus, maximizing the meta frame appropriateness function $MFA(\{\chi_a\}_a)$ (15), is equivalent in its first order terms to minimizing

$$\min \sum_{a} \sum_{\substack{k,l \\ S_k, S_l \in \chi_a}} \left[-\log(1 - s_k s_l) - \log\left(1 - \frac{AU(\oplus\{m_k, m_l|\Theta_a\})}{(|\chi_a|-1)\cdot \log_2(|\Theta_a|)}\right)\right]. \tag{25}$$

An algorithm for minimizing (25) is shown in Fig. 4. This is an adoption from [7]. Here the interactions J_{kl} are identical to (25). All parameters of Fig. 4. are immediate except for the number of clusters K. Its determination is domain dependent and can be found in several different ways as discussed in Sec. 1, e.g., using the method of [10, p. 90].

6 An algorithm for constructing multiple frames

Let us describe an algorithm for constructing the best frames of discernment for several multiple subproblems, χ_a, Fig. 3.

Input: A set χ of simple support functions or consonant belief functions.
Output: Frames of discernment $\{\Theta_a\}_a$ for all subproblems χ_a.
Step 1. Instantiate all interactions J_{ij} (in Fig. 4.) between all pairs in χ, using (10) in Fig. 1.
Step 2. Partition χ by minimizing *MFA* (14) using the Potts spin clustering algorithm, Fig. 4.
Step 3. For each subproblem χ_a use the algorithm to construct the most appropriate frame of discernment Θ_a, Fig. 2. Return $\{\Theta_a\}_a$.

Fig. 3. An algorithm for constructing multiple frames of discernment for multiple subproblems

Using this algorithm will construct a set of frames of discernment $\{\Theta_a\}_a$ for several subproblems χ_a that should be handled separately. This set of frames is best in terms of minimizing the overall frame appropriateness *MFA* over all subproblems.

7 Conclusions

We have extended a methodology for constructing a frame of discernment from incoming belief functions for *one* problem, into a methodology for constructing multiple frames of discernment for *several* different subproblems. This lets our evidence interact in an interesting way within each subproblem without exhibit too much internal conflict. This dual task is achieved simultaneously for all subproblems by maximizing a function of overall frame appropriateness over all subproblems.

INITIALIZE

K (number of clusters); N ($= |\chi|$) (number of simple support functions);

$J_{ij} = \left[-\log(1 - s_i s_j)\delta_{|A_i \cap A_j|} - \log\left(1 - \frac{\mathrm{AU}(\oplus\{m_i, m_j | \Theta_a\})}{(|\chi_a| - 1) \cdot \log_2(|\Theta_a|)}\right)\right] \quad \forall i, j$, where

$|\chi_a| = \frac{N}{K}$ and $|\Theta_a| = |C_i \cup C_j|$; $s = 0$; $t = 0$; $\varepsilon = 0.001$; $\tau = 0.9$; $\gamma = 0.5$;

$T^0 = T_c$ (a critical temperature) $= \frac{1}{K} \cdot max(-\lambda_{min}, \lambda_{max})$, where λ_{min} and λ_{max} are the extreme eigenvalues of M, where $M_{ij} = J_{ij} - \gamma\delta_{ij}$;

$V_{ia}^0 = \frac{1}{K} + \varepsilon \cdot rand[0,1] \quad \forall i, a$;

REPEAT–1

- REPEAT–2

  $\forall i$ Do:

  - $|\chi_a| = \max\left\{2, \sum_{i=1}^{N} V_{ia}^s\right\}$;
  - $H_{ia}^s = \sum_{j=1}^{N} J_{ij} V_{ja}^{\begin{cases} s+1, & j < i \\ s, & j \geq i \end{cases}} - \gamma V_{ia}^s \quad \forall a$;
  - $F_i^s = \sum_{a=1}^{K} e^{-H_{ia}^s / T^t}$;
  - $V_{ia}^{s+1} = \frac{e^{-H_{ia}^s / T^t}}{F_i^s} + \varepsilon \cdot rand[0,1] \quad \forall a$;
  - $s = s + 1$;

  UNTIL–2

  $\frac{1}{N}\sum_{i,a}\left|V_{ia}^s - V_{ia}^{s-1}\right| \leq 0.01$;

- $T^{t+1} = \tau \cdot T^t$;
- $t = t + 1$;

UNTIL–1

$\frac{1}{N}\sum_{i,a}(V_{ia}^s)^2 \geq 0.99$;

RETURN

$\left\{\chi_a \middle| \forall S_i \in \chi_a.\ \forall b \neq a\ V_{ia}^s > V_{ib}^s\right\}$;

Fig. 4. Clustering algorithm

References

[1] Schubert, J.: Constructing and Reasoning about Alternative Frames of Discernment. In: Proceedings of the Workshop on the Theory of Belief Functions, paper 24, pp. 1–6 (2010), http://www.ensieta.fr/belief2010/papers/p24.pdf

[2] Dempster, A.P.: A Generalization of Bayesian Inference. Journal of the Royal Statistical Society B 30, 205–247 (1968)

[3] Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)

[4] Schubert, J.: On Nonspecific Evidence. International Journal of Intelligent Systems 8, 711–725 (1993)

[5] Schubert, J.: Specifying Nonspecific Evidence. International Journal of Intelligent Systems 11, 525–563 (1996)

[6] Schubert, J.: Managing Inconsistent Intelligence. In: Proceedings of the Third International Conference on Information Fusion, pp. TuB4/10–16. International Society of Information Fusion, Sunnyvale (2000)

[7] Bengtsson, M., Schubert, J.: Dempster-Shafer Clustering using Potts Spin Mean Field Theory. Soft Computing 5, 215–228 (2001)

[8] Schubert, J.: Clustering Belief Functions Based on Attracting and Conflicting Metalevel Evidence Using Potts Spin Mean Field Theory. Information Fusion 5, 309–318 (2004)

[9] Schubert, J.: Clustering Decomposed Belief Functions Using Generalized Weights of Conflict. International Journal of Approximate Reasoning 48, 466–480 (2008)

[10] Ahlberg, S., Hörling, P., Johansson, K., Jöred, K., Kjellström, H., Mårtenson, C., Neider, G., Schubert, J., Svenson, P., Svensson, P., Walter, J.: An Information Fusion Demonstrator for Tactical Intelligence Processing in Network-Based Defense. Information Fusion 8, 84–107 (2007)

[11] Schubert, J.: Finding a Posterior Domain Probability Distribution by Specifying Nonspecific Evidence. International Journal of Uncertainty, Fuzziness and Knowledge- Based Systems 3, 163–185 (1995)

[12] Schubert, J., Sidenbladh, H.: Sequential Clustering with Particle Filters - Estimating the Number of Clusters from Data. In: Proceedings of the Eighth International Conference on Information Fusion, paper A4-3, pp. 1–8. IEEE, Piscataway (2005)

[13] Chau, C.W.R., Lingras, P., Wong, S.K.M.: Upper and Lower Entropies of Belief Functions Using Compatible Probability Functions. In: Komorowski, J., Raś, Z.W. (eds.) ISMIS 1993. LNCS (LNAI), vol. 689, pp. 306–315. Springer, Heidelberg (1993)

[14] Maeda, Y., Ichihashi, H.: An Uncertainty Measure with Monotonicity under the Random Set Inclusion. International Journal of General Systems 21, 379–392 (1993)

[15] Harmanec, D., Klir, G.J.: Measuring Total Uncertainty in Dempster-Shafer Theory: A novel approach. International Journal of General Systems 22, 405–419 (1994)

[16] Meyerowitz, A., Richman, F., Walker, E.: Calculating Maximum-Entropy Probability for Belief Functions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2, 377–389 (1994)

[17] Harmanec, D., Resconi, G., Klir, G.J., Pan, Y.: On the Computation of Uncertainty Measure in Dempster-Shafer Theory. International Journal of General Systems 25, 153–163 (1996)

[18] Wu, F.Y.: The Potts model. Reviews of Modern Physics 54, 235–268 (1982)

[19] Peterson, C., Söderberg, B.: A New Method for Mapping Optimization Problems onto Neural Networks. International Journal of Neural Systems 1, 3–22 (1989)

[20] Schubert, J.: Fast Dempster-Shafer Clustering Using a Neural Network Structure. In: Bouchon-Meunier, B., Yager, R.R., Zadeh, L.A. (eds.) Information, Uncertainty and Fusion. SECS, vol. 516, pp. 419–430. Kluwer Academic Publishers, Boston (1999)

Conflict Interpretation in a Belief Interval Based Framework

Clément Solau[1], Anne-Marie Jolly[1], Laurent Delahoche[2], Bruno Marhic[2], and David Menga[3]

[1] Polytech'Orléans, 45072, Orléans, France
clementsolau@hotmail.com
[2] LTI, Picardie Jules Verne University, 80025 Amiens, France
laurent.delahoche@u-picardie.fr
[3] EDF R&D, 92141 Clamart, France
david.menga@edf.fr

Abstract. This paper combines belief functions in an interval based framework with the purpose of comparing two manners of transferring the conflict. We made a focus on the classical problem of conflict interpretation in this particular context. In connection with this, the normalization introduced by the Dempster's rule is compared with an alternative method. Since conflict is often considered as a quantity carrying a signification, it is a relevant reason for analyzing it in a belief interval based framework and explaining why the normalization is always used. Results are provided based on a voluntarily simple example.

Keywords: Belief functions, interval analysis, conflict signification, state estimation and constraint satisfaction problem.

1 Introduction

We introduce the problem of how to construct a solution space to efficiently regulate the energy in the context of smart home. This problem can be reduced to a state estimation problem integrating constraints and numerous measures with the property of being uncertain. It is equivalent to a CSP (standing for Constraint Satisfaction Problem) where two supports will be combined, the first is the interval framework and the second is the belief function framework. To sum up, it is equivalent to solving a CSP within a belief interval framework first introduced in [5] where the transferable belief model is generalized and extended to the case where focal sets are in the form of intervals. The aim of this aggregation is to assign non-uniform quantity of knowledge on subintervals. A fundamental link between the interval theory and belief function can be the conjunctive rule of combination (if the sources are reliable), which have the meaning of respectively intersection for the first theory and mass aggregation phase for the second theory. These concepts have already been introduced in [6] with the notion of CSP and in [10] where focal sets are IVFS. In [6], the authors introduced a combination rule that permits to merge mass functions with intervals in a common framework. In our case, to combine our data, a very similar manner has been used with few adaptations. The intervals containing the solution set to be approximated are contracted using the forward-backward

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 199–208, 2010.

propagation. We have made the choice of an homogeneous masses repartition for the combination stage. Finally, we compare two manners of transferring the mass allocated to the empty set (i.e. conflict) based on a voluntary simple CSP to interpret the conflict. Last, we provide experimental results before concluding.

2 Brief Outline of Interval Analysis

A real interval written $[x]$is a subset of R with the attribute that any number that lies between two numbers included in the set is also contained in the set. Mathematically, it corresponds to $[x]$which is defined so that: $[x]=[\underline{x},\overline{x}]=\{x\in R|\underline{x}\le x\le\overline{x}\}$ .Since any real or non-empty interval is a set, the set theoretic operations can be extended to intervals. On the one hand, it is thus possible to define the intersection between two non-empty intervals $[x]$ and $[y]$so that $[x]\cap[y]=\{z\in R|\ z\in[x]\text{and}\,z\in[y]\}$. On the other hand, it is similarly possible to define the union of two non-empty intervals $[x]$ and $[y]$defined by $[x]\cup[y]=\{z\in R\mid z\in[x]\text{or}\,z\in[y]\}$.

2.1 Interval Arithmetic and Computation

Similarly to the description mentioned before, the any classical arithmetic operators such that $\Diamond\in\{+,-,*,/\}$can also be extended into an interval framework verifying $[x]\Diamond[y]=[\{x\Diamond y\in R|x\in[x]\text{and}\ y\in[y]\}]$. A function is a combination of variables linked with operators (+,- ,*……). Hence, it is now possible to claim that functions can also be widened to interval.

2.2 Constraint Satisfaction Problems (CSP)

A Constraint Satisfaction Problem H is a problem involving a vector of variables X whose value is drawn from the initial domain D. Mathematically speaking, it consists in computing X such that $f(X)=0$ still holds. Hence, it is thus possible to define the solution S set so that $S=\{x\in[X]|f(X)=0\}$ which is not systematically box-shaped. Intervals (or boxes if many dimensions) are almost systematically used for approximating the solution set S and they will strongly contribute to finding a relevant result. Contractors aim at approximating a solution set S by computing in a minimal time the smallest box $[X']\subseteq[X]$such that $S\subseteq[X']$.

2.3 Constraint Propagation Algorithm

The constraint propagation technique guarantees an enclosure of S with a complexity that can be polynomial in time. The domain is reduced by considering the constraints in isolation and in turn. Since inclusion functions have already been defined in part 2.1, it is possible to define an interval counterpart of the previously described decomposition: this is the propagation phase. The algorithm is stopped when domain cannot be reduced any further. Examples can be found in [2].

3 Belief Function Theory

The Dempster-Shafer Theory (sometimes abbreviated to DST) framework is a mathematical theory of evidence introduced in [8] and allowing to both model and combine evidence. Let Θ be the frame of discernment which contains hypotheses that denote the exclusive and exhaustive solutions to the problem. The frame of discernment Θ can be described as follows:

- $\Theta = \{ H_i, i = 1,...,n \}$ where $H_i \cap H_j = \varnothing \ \ \forall i, \forall j$ with $i \neq j$ (i.e. exclusivity) (1.1)
- Θ must also satisfy: $\bigcup_{i=1,...n} H_i = \Theta$ (i.e. sufficiency) (1.2)

Therefore, it is possible to deduce from Θ the power set 2^Θ which contains all the subsets resulting from Θ:

$$2^\Theta = \{A \mid A \subseteq \Theta\} = \{\varnothing, H_1, H_2, ..., H_n, H_1 \cup H_2, ..., \Theta\} \tag{2}$$

The previously defined power set 2^Θ represents the frame of definition on which the 'basic belief masses' (bbm) are defined. Besides, a collection of bbm is called 'a basic belief assignment' (bba) or 'a mass function'. A belief function is formally written and defined as follows:

$$m : 2^\Theta \to [0,1] \text{such that} \sum_{A \in 2^\Theta} m(A) = 1 \tag{3}$$

where the bbm $m(A)$ represents the part of belief that is exactly allocated to the proposition A. A focal element of m is an element $A \in 2^\Theta$ that also satisfies $m(A) > 0$. The belief allocated to A is independent from the belief assigned to any sub-hypotheses composing A (which models the doubt) in the case where A is a disjunction. However, the case where several sources of information examine data, several sources of evidence appear. Before making any decision, the opinions of the different sources must be fused by using operators on all the sources of evidence**.** The combination operators are generally used depending on the framework of the application [4]. In this paper we will propose an alternative (see paragraph 4.3) to the Dempster's rule of combination given that intervals will be combined. Specifically the combination is computed from the two sets of masses m_1 and m_2 in the following manner:

$$m_{12}(A) = (m_1 \oplus m_2)(A) = \frac{1}{1-K} \sum_{B \cap C = A \neq \varnothing} m_1(B) m_2(C) \ \text{ where } K = \sum_{B \cap C = \varnothing} m_1(B) m_2(C) \tag{4}$$

K denotes the conflict between the two mass sets (i.e. the mass assigned to the empty set). In Dempster's rules of combination, the conflict is normalised. This means that each mass that is allocated to a non-empty set is divided by the normalization factor (i.e. 1-K). It has the effect of completely ignoring any mass associated with conflict to the empty set. However, the computational complexity can be greater when considering uncertainty (i.e. greater number of focal sets). A method that permits to decrease the number of focal sets exist [7] : this is the summarization. Let *m* be a mass function

on Θ with n focal sets $H_1, H_2, ..., H_n$ ordered such that $m(H_1) \geq m(H_2) \geq ..... \geq m(H_n)$. Let us define $1 \leq k \leq n$ as being the number of focal sets we want to keep into account in our focal set reduction problem. Hence a reduced mass function m' with k focal sets is built as it is described below :

$$m'(H_i) = m(H_i) \text{ with } i = 1,.....,k-1 \text{ and } m'(H_0) = \sum_{i=k}^{n} m(H_i) \tag{5}$$

Where H_0 is defined as $\bigcup_{i=k}^{n} H_i$ with H_i an interval.

4 Belief Functions on Real Numbers

4.1 Basic Principles

This part deals with belief function theory applied to real numbers and firstly introduced in [5]. Let $[\alpha, \beta]$ be a non-empty interval verifying $\alpha < \beta$ and $[\alpha, \beta] \subseteq R$. Similarly, let us consider $I_{[\alpha,\beta]}$ as the set of closed intervals in $[\alpha, \beta] \subseteq R$ and it is defined as follows:

$$I_{[\alpha,\beta]} = \left\{ [x, y] \middle| \, x \geq \alpha, x \leq y \leq \beta \right\} \tag{6}$$

To make interval analysis compliant with the mass function theory, all the intervals will be half-closed and thus disjointed with each others. Similarly to (6), let $\forall H_1, H_2, ..., H_n \in J_{[a,z[}$ with $J_{[a,z[} = \left\{ [x, y[\middle| \, x \geq a, x \leq y < z \right\}$ be the set such that:

$$H_i \cap H_j = \varnothing \text{ and } i, j \in \{1,2,.....,n\}, \text{ with } i \neq j \tag{7}$$

This consideration is illustrated in the following picture:

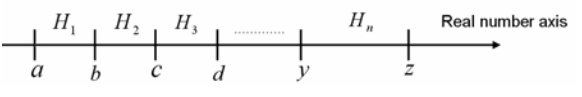


Fig. 1. Definition of intervals composing Θ

with $H_1 = [a, b[, H_2 = [b, c[, H_3 = [c, d[$ and $H_n = [y, z[$ and thus verifying the property that each interval is disjoint with each other (1.1). In the sequel, $[a, z[$ will be broken down into numerous smaller intervals and the collection of these smaller intervals will be denoted by Θ and defined as follows: $\Theta = \{H_i : H_i \in J_{[a,z[}, i = 1,2,.....,n\}$. Nevertheless, each $H_i, H_j \in \Theta$ satisfies (1.1) and $\bigcup_{i=1}^{n} H_i = \Theta$ (1.2). Thus, a bba can be built on Θ with Θ being the frame of discernment. All the elements of Θ will be in the form of intervals or at least of unions of intervals. 2^{Θ} will be made up of unions of elements belonging to Θ. Thus it is possible to build a mass function having intervals as focal elements. Let $m^{\Theta} : 2^{\Theta} \to [0,1]$ be a basic belief assignment that satisfies (3) with $X_i \in 2^{\Theta}$. Hence, all

the X_i associated with $m^{\Theta}(X_i)>0$ are the focal elements with X_i being an interval. Indeed each $X_i \in 2^{\Theta}$ can be decomposed as being a union of intervals belonging to Θ (in other words : $X_i = \bigcup_j H_j$). The granularity on Θ will strongly affect 2^{Θ}. Indeed in an interval such as [0, 1] there is an infinity of existing Θ.

4.2 Rule of Combination

Let m_1^{Θ} and m_2^{Θ} be two belief functions defined on Θ provided by the distinct sources of information 1 and 2 and expressing a part of knowledge on closed intervals. Hence, Dempster's rule of combination (4) can be extended to the mass function having intervals as focal sets:

$$m_{12}^{\Theta}([z]) = \frac{1}{1-K} \sum_{\{i,j|[x_i]\cap[y_j]=[z]\neq\varnothing\}} m_1^{\Theta}([x_i])m_2^{\Theta}([y_j]) \quad \text{where} \quad K = \sum_{\{i,j|[x_i]\cap[y_j]=\varnothing\}} m_1^{\Theta}([x_i])m_2^{\Theta}([y_j]) \quad (8)$$

K is defined as in (4). In this paper, the Dempster's rule is used and will be compared with an alternative manner (see paragraph 4.3) for transferring the conflict (i.e. the K factor).

4.3 Alternative Method

This part will explain our alternative which aims at taking into account the conflict and assessing its effect on the overall result. For the sake of simplicity this alternative will be presented through an example. Let m_1^{Θ} and m_2^{Θ} be the two belief functions to be combined with intervals as focal elements. These mass functions will be combined using the following alternative method:

$$m_{12}^{\Theta}([z]) = \sum_{\{i,j|[x_i]\cap[y_j]=[z]\}} m_1^{\Theta}([x_i])m_2^{\Theta}([y_j]) \text{ where } m_{12}^{\Theta}([h]) = \sum_{\{i,j|[x_i]\cap[y_j]=\varnothing\}} m_1^{\Theta}([x_i])m_2^{\Theta}([y_j]) \text{ with } [h] = \bigcup_{[z_i]\neq\varnothing}^{i} [z_i] \quad (9)$$

Graphically, it can be interpreted as shown in the following picture:

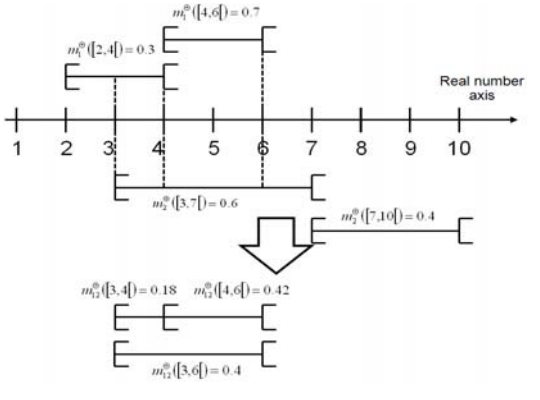


Fig. 2. Alternative method

As can be seen, the mass originally committed to the empty set is now transferred to the union of all the focal sets associated with a non-empty mass with respect to (9). We desired to introduce this notion in order to take into account the conflict and to assess the effect of this alternative by providing results in the sequel.

4.4 Belief Counterpart of Operators and Functions

As mentioned in the part 1, the classical computations involving different operators (+,-,…) has been generalised to intervals. In connection with this consideration, [9] introduced the extension of the classical common operators to mass function in order to get a mass counterpart resulting from the calculation. This is expressed as below:

$$m^{\Theta}([z]) = \sum_{\{i,j | [x_i] \Diamond [y_j] = [z]\}} m_i^{\Theta}([x_i]) m_j^{\Theta}([y_j]) \tag{10}$$

where $[z],[x_i]\text{and}[y_j]$ represent closed intervals and $\Diamond$ an operator such as $\Diamond \in \{+,-,*,/,\cup,\cap,.....\}$. The interval framework already defined the interval counterpart of function and introduced the notion of inclusion functions. Similarly, it is also possible to deduce from (10) the quantity of mass that will be transferred to the focal elements (i.e. interval) resulting from a function f. This consideration has been introduced by [6] and is expressed as follows:

$$m^{z}([z]) = \sum_{\{i_1,....,i_n | [z]=[f]([x_{i_1}],....,[x_{i_n}])\}} m_{i_1}^{x_1}([x_{i_1}]).....m_{i_n}^{x_n}([x_{i_n}]) \tag{11}$$

Since $[x_{i_j}]$ are intervals, $[z]$ is also an interval.

5 Global State Estimation Problem

5.1 Basic Principles

This CSP aims at estimating the position of a person moving forward. The movement of the pedestrian is assumed to be a one dimension displacement. Along his way, the only information that the pedestrian has, is the coordinates of two known points placed on the same axis as the one on which the person is walking. These two points also measure the distance and send it to the pedestrian which can deduce an approximate of its position. Let us consider the one dimensional time discrete system of equations:

$$\begin{cases} x_t = f(x_{t-1}, \Delta x_{t-1}) \\ z_t = g(x_t, w_t) \end{cases} \Leftrightarrow \begin{cases} x_t = x_{t-1} + \Delta x_{t-1} \\ \begin{pmatrix} {}^1z_t \\ {}^2z_t \end{pmatrix} = \begin{pmatrix} x_t \\ x_t \end{pmatrix} - \begin{pmatrix} x_{B1} \\ x_{B2} \end{pmatrix} + \begin{pmatrix} {}^1w_t \\ {}^2w_t \end{pmatrix} \end{cases} \tag{12}$$

where the first equation f relates the current state x_t to the previous one x_{t-1} linked by a bounded displacement (i.e. including an error factor). The function f models this consideration and is sometimes referred to as being the state transition (or prediction) function. The ${}^1z_t$ and ${}^2z_t$ variables indicate the observation of the pedestrian

deduced from a distance (i.e. $x_t - x_{B1}$) separating the pedestrian and fixed the points whose coordinates are known by the pedestrian (i.e. x_{B1} and x_{B2}). These observations are used to correct the estimation. The function $g(x_t, w_t)$ is referred to as the observation (or correction) function with w_t an observation noise.

5.2 Our Approach

Let us assume that the initial state is globally bounded as follows: $x_0 \rightarrow [x_0[$. However, the repartition of belief is always uniform. This is why we decided to mix it with belief functions that can provide relevant information in order to ripen our estimation. Indeed, in the interval analysis framework, the probability distribution is uniform, meaning that there is no preconception of any part of $[x_0[$ containing the set of solution S (which is a point) to be estimated. Therefore, for the sake of precision of the state estimation, we choose to sub-divide the interval representing and guaranteeing the initial state into smaller intervals verifying (1.1) and (1.2). It is therefore possible to associate to each interval $[x_0^i[$ with $i = 1,....,n$ a basic belief assignment. Thus each part of the global interval $[x_0[$ is coupled with a basic belief assignment. However, the manner used to construct mass functions is not explained in this paper even if it is a topic of a great importance that impacts on the algorithms results. Moreover, the distance measured by the exteroceptive telemetric sensors is also assumed to yield bounded measurements $[{}^i z_t[$ with $i \in \{1,2\}$. The intervals defined by $g^{-1}([{}^i z_t[) = [{}^i z_t[+ x_{Bi} - {}^i w_t$ with $i \in \{1,2\}$ enclose an estimation of the current position of the pedestrian and are expressed in the form of intervals (i.e. bounded error). A similar process to the one adopted for $[x_t[$ is undertaken for the subdivision of $[{}^i z_t[$ and mass are allocated to subparts verifying (1.1) and (1.2). The mass function will be denoted m^{1_z} and m^{2_z} assumed to be reliable [11]. The distribution is derived from a Gaussian probability distribution. Hence at each sample time, this system can be established:

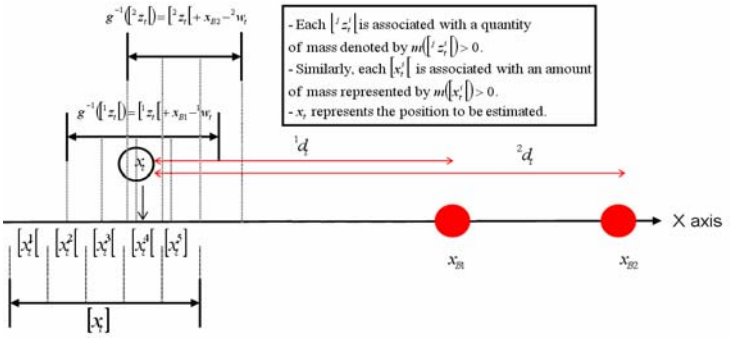


Fig. 3. Overall system

Therefore, it is possible to deduce an interval counterpart associated with masses of the previously defined system of equation (12). This interval counterpart will allow us to provide at each iteration *t* a guaranteed estimate of the state in the form of an interval denoted by $[x_t[$ with respect to the previous state $[x_{t-1}[$. Given that intervals are linked with masses, it is also possible to deduce a belief counterpart used for estimating the state. The computations on both the intervals and mass functions are achieved concurrently but the link relies on the fact that focal sets are intervals. Hence, the CSP can be solved by a contractor and associated with a belief counterpart. The entire CSP will be solved by treating parts of intervals (and thus mass functions) in turn:

Input: $t_{\max}, [x_{t-1}[, m([x_{t-1}[), [\Delta x_{t-1}], x_{B_1}, x_{B_2}$ *and* $nb\,\text{int} = 0$

Output : $[x_t[, m([x_t[)$

1 : Build the sub-intervals $[^1z_t^j[, [^2z_t^{j'}[$ *and* $m([^1z_t^j[), m([^2z_t^{j'}[)$;

2 : for i=1 to n=nb($[x_{t-1}[$*) do*

3 : for j=1 to m=nb($[^1z_t[$*) do*

4 : for k=1 to m'=nb($[^2z_t[$*) do*

5 : nbint=nbint+1;

6 : $[x_t^{nb\,\text{int}}[= \text{contractor}([x_{t-1}^i[, [^1z_t^m[, [^2z_t^{m'}[, [\Delta x_{t-1}])$;

7 : $m([x_t^{nb\,\text{int}}[) = m([x_{t-1}^i[) * m([^1z_t^m[) * m([^2z_t^{m'}[)$;

8 : *endfor; endfor; endfor;*

9: $m([x_t[) = Normalisation(m([x_t[)$ *or* $m([x_t[) = Alternative(m([x_t[)$

10: $m([x_t[) = Summarization(m([x_t[))$

11: $x_t^i = \sum_{i=1}^{nb\,\text{int}} C_t^i * m([x_t^i[)$; // With C_t^i the center of the interval $[x_t^i[$

12 : *do until* $t = t_{\max}$;

Fig. 4. Algorithm

The computational complexity of the algorithm is cubic which is why summarization has been added as in [7].

6 Experimental Results

An evolution of the conflict versus the pedestrian position is given :

This example is based on 300 iterations. Graphically, it is easy to notice that the amount of conflict is globally lower when using the alternative method. This is because the conflict generated at the previous iteration has been transferred to a disjunction covering all the elements associated with a non-empty mass (10). A comparison between the evolutions of errors generated by the two methods is provided:

The results obtained with the normalization are far better than those provided by the alternative method. The average error of the alternative method is more than two times greater than the error provided by the normalization. The percentage of the highest error (71%) is also more than two times greater than the one obtained when

normalizing (29%). Since the results obtained when allocating the conflict to a disjunction are far weaker, it is obvious to claim conflict should not have been kept into account in our process (because sources are reliable). It means that conflict should have been normalized because it does not carry any information when mass functions are coupled with intervals. In connection with this, the interval analysis framework clearly claims that if a solution exits, it compulsory belongs to the interval resulting from the contractor algorithm output (see also part 2.3.). It insinuates that $S \subseteq [x_t[$ with S the state to be estimated and $[x_t[$ the enclosing interval resulting from the constraint propagation algorithm. To sum up, it is possible to state that we were wrong to put a certain quantity of belief on specific parts (i.e. those generating conflict) of intervals. According to the figure (3), it was meaningless to allocate mass to the interval [7,10[from which the conflict appears. It still raises the question about how to construct relevant mass distributions when applied to intervals.

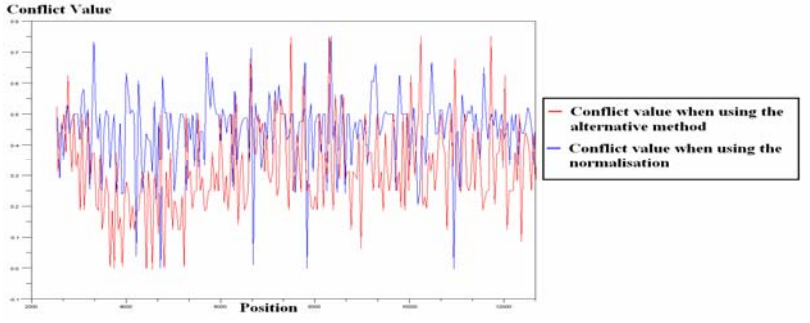


Fig. 5. Conflict evolution

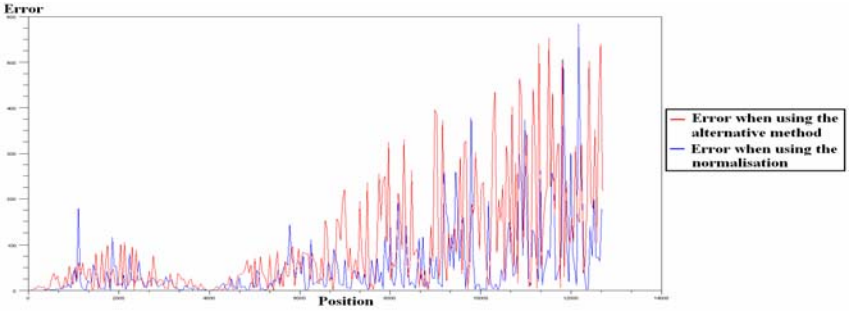


Fig. 6. Evolution of the error on the estimate versus position

7 Conclusion

This paper aimed at comparing two methods with the purpose of interpreting the conflict which is a classical problem in the belief function community. The comparison is based on a simple CSP in order to demonstrate that the normalisation technique is far

better than any other method. The results provided by this method clearly confirm and verify our intuition and we tried also to explain theoretically why normalisation is always used when working with an interval framework. However, the conflict signification in an interval based framework has been given and differs with respect to the application [4].

References

1. Moore, R.E.: Interval Analysis. Prentice-Hall, Englewood Cliffs (1966)
2. Jaulin, L., et al.: Applied Interval Analysis. Springer, London (2001)
3. Benhamou, F., et al.: Revising Hull and Box Consistency. In: ICLP 1999 (1999)
4. Sadiq, R., et al.: Investigating evidential reasoning for the interpretation of microbial water quality in a distribution network. Springer, Heidelberg (2006)
5. Smets, P.: Belief Functions on Real Numbers. Journal of Approximate Reasoning (2004)
6. Nassreddine, G., et al.: State estimation using belief function theory. In: LFA 2007, Nîmes, France, pp. 115–124 (2007)
7. Nassredine, G., et al.: State estimation using interval analysis and belief function theory: Application to dynamic vehicle localization. Man and Cybernetics B (2009)
8. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton (1976)
9. Yager, R.: Arithmetic and other operations on Dempster-Shafer structures. International Journal of Man-Machine Studies, 357–366 (1986)
10. Rocha, L.M.: Interval Based Evidence Sets. In: ISUMA-NAFIPS 1995, pp. 624–629. IEEE Press, Los Alamitos (1995)
11. Klein, J., et al.: Hierarchical and conditional combination of belief functions induced by visual tracking. International Journal of Approximate Reasoning (2009)

Evidential Data Association Filter

Ahmed Dallil[1,2], Mourad Oussalah[2], and Abdelaziz Ouldali[1]

[1] Polytechnic Military School, BP 17 Bordj El Bahri Algiers, Algeria
[2] University of Birmingham, Electronics, Electrical and Computer Engineering Edgbaston, Birmingham B15 2TT, UK

Abstract. The problem of multi-target tracking in clutter environment has been shown to be very challenging for both the measurement to track association and the targets' state estimation. Several approaches have been put forward to deal with such issue. Especially, the family of joint probabilistic data association and its modified versions has been very popular in the field. This paper advocates the use of the theory of belief function to tackle the measurement-to-track association as well as the estimation problems. The proposal generates the basic belief mass assignment using a Bayesian approach, while the decision making process is based on the extension of the frame of hypotheses. Our method has been tested for a nearly constant velocity target and compared to both the nearest neighbor filter and the joint probabilistic data associations filter in a highly ambiguous cases. The results demonstrate the feasibility of the proposal and show improved performances compared to the aforementioned alternative commonly used methods.

Keywords: Belief functions, data association, target tracking.

1 Introduction

Considerable research has been undertaken in the field of estimation theory in relation to the multi-target tracking (MTT) problem. This is of interest in both military and civilian applications. The data association is a central problem in MTT since validated measurements should be correctly assigned to theirs associated track (s) in order to obtain correct measurements for state estimation. More specifically, the data association involves the issues of data validation, assigning the correct validated measurement to the corresponding target being observed, target initiation, target maintenance and confirmation, among others. Especially, the importance of the data association step is highlighted by its straightforward incidence to the state estimation problem in the sense that any wrongly associated measurement would yield target loss or miss-estimation. Nevertheless, the two issues estimation and data association are ultimately linked to each other. Indeed, without good estimation, the data association is at risk, and, in turn, a wrong data association influences negatively the targets' state estimation. In the literature, several methods have been proposed to deal with this problem [1], [2] and [3]. Most of these methods are developed in the frame work of classical probability. Among these methods, one shall mention the well known nearest neighbor filter (NNF) and the conventional sub optimal joint probabilistic data association filter (JPDAF).

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 209–217, 2010.

The former is based on the idea that we associate the closest measurement to the underlying track, while the latter avoids the direct problem of a single measurement-track association by incorporating, through a weighted linear combination, all the latest measurements into the estimate. The theory of evidence, also called belief functions theory or Dempster-Shafer theory has been introduced by Glen Shafer [4], based on the seminal work of Dempster [5], as a new approach for representing uncertainty. Nowadays this formalism is considered as one of the most interesting and promising approach for handling uncertainty [6] and [7]. In recent publication [8] and [9], a novel approach based on the use of belief theory for data association in MTT was presented. It was reported to be suitable for this problem since it allows us to reason with uncertainty and suggests a way of combining imperfect data and expressing ignorance. Furthermore, the conflict between the sources to be combined can be expressed in this theory. In our paper, we introduce a new general algorithm for data association in MTT in the frame work of belief functions theory. We called this method the evidential data association filter (EDAF). The solution adopted for the representation of the information in our method is influenced by the algorithm used in [10] where for each target, a set of local frame of discernment, constituted of the two propositions –"the target is associated to the given measurement" and "the target is not associated to the given measurement"-, is defined for each measurement. While a new approach for mass generation based on the use of cheap JPDAF [11] and probabilistic reasoning is put forward in this paper. The proposal contrasts with distance based approaches investigated in previous evidential-based data association methods [12] and [13]. The feasibility and performances of the proposal have been investigated using nearly constant velocity targets in Monte Carlo simulations. Comparisons with NNF and JPDAF have been carried out throughout this study. The results demonstrate the feasibility of the proposal and provide a comprehensive delimitation of the advantage and limitations of the proposal. Section 2 summarizes the main concepts in evidence theory that will be throughout this paper. Section 3 of this paper outlines the main features of the suggested evidential data association filter. Section 4 describes the decision-making process where the best target matching a given measurement is identified. Section 5 copes with the mass function generation used in the EDAF. Simulation results and comparison with NNF and JPDAF are carried out in Section 6.

2 Basics of Belief Theory

The theory of belief as formalized by Shafer [4] allows the distribution of elementary mass known as basic belief mass function or assignment or distribution (bbm), which quantifies the amount of support given to some groups of hypothesis of the frame of discernment Ω. Assigning a mass $m(A)$ to a subset A of Ω gives some support to exactly the subset A, and not to a more specialized subset. Typically, any mass m is such that:

$$\sum_{A\subset\Omega} m(A) = 1 \tag{1}$$

This contrasts with discrete probability model where the sum is only over singletons of Ω. Also belief theory provides a method for combining the measure of evidence from different sources using the conjunctive rule (Dempster's rule) of combination.

This rule combines two independents distributions of evidence defined in the same frame of discernment Ω to a new distribution of evidence. More specifically, let m_1 and m_2 be two bbm defined in the same frame of discernment Ω. The new distribution of evidence noted m_{12} is given by:

$$m_{12}(A) = m_1 \otimes m_2(A) = \sum_{\substack{X \in \Omega \\ Y \in \Omega}} {}_{X \cap Y = A} \, m_1(X) \times m_2(Y) \tag{2}$$

In the case where X and Y are disjoint, it indicates the occurrence of a conflict among the information supplied by the sources modelled by the mass m_1 and m_2 because one is strictly supporting a set of hypothesis, and the other one supports a completely disjoint set of hypothesis. If we are in an open world (truth can be elsewhere), it sounds good to give support to external hypothesis, not in Ω so the mass $m_1(X) \times m_2(Y)$ would go to ϕ as a measure of conflict between the combined beliefs. In a closed world, to verify (1), it is necessary to normalize the mass function with the lost mass corresponding to the amount of the conflict. So, the normalized masse m^* becomes:

$$m^*{}_{12}(A) = \frac{m_{12}(A)}{1 - m_{12}(\emptyset)} \tag{3}$$

3 Formulation of the Proposed Method

Let $\Omega = \{P_1, P_2 \ldots P_I\}$ be the set of predicted targets and $\Theta = \{A_1, A_2 \ldots A_J\}$ the set of validated measurements, at time k. To study the association relation between these two sets, a basic idea is to define, for each target P_i, i=1… I, and each observation A_j, j=1… J, a basic belief mass distribution on the set $\Theta_i = \Theta \times \{P_i\}$. This distribution gives an initial belief about the association of the observation A_j to the target P_i by assigning a mass $m(.)$ to the sets $\{A_j\} \times \{P_i\}$ and $\overline{\{A_j\}} \times \{P_i\}$ as follows:

- $m(\{A_j\} \times \{P_i\}) \triangleq m_{ij}(A_j)$*:* The proposition: *a target P_i generates observation A_j (association).*
- $m\left(\overline{\{A_j\}} \times \{P_i\}\right) \triangleq m_{ij}(\bar{A}_j)$*:* The proposition: *a target P_i does not generate A_j (no association).*

Since we have J measurements, we obtain J initial beliefs for each target P_i, which are defined over the same frame of discernment Θ_i. These initial beliefs will be combined to make a decision about the return of target P_i. This is done by using the conjunctive rule of combination which produces a single bbm distribution over the power set of Θ_i. The results of mass combination for all measurements A_j and all targets P_i are summarized in Table 1, which represents the belief matrix. As may be seen, for each target P_i we obtain a new distribution defined only on singletons. Each element $M_{i,j}$ represents the mass value for the proposition "*Measurement A_j originated from target P_i*.". One may notice from the last row that a mass is being assigned to the empty set ϕ. This reflects partial or total ambiguity in the correlation between considered target and measurements. It can be interpreted in part as missed target hypothesis.

Table 1. Belief matrix: results of mass combination

	P_1	P_2	…	P_I
A_1	$M_{1,1}$	$M_{2,1}$	…	$M_{I,1}$
A_2	$M_{1,2}$	$M_{2,2}$	…	$M_{I,2}$
….	…	…	…	…
A_J	$M_{1,J}$	$M_{2,J}$	…	$M_{I,J}$
ϕ	$M_{1,\phi}$	$M_{2,\phi}$	…	$M_{I,\phi}$

An example of generation of the belief matrix for a target P_L and two measurements A_1 and A_2 is presented in Table 2.

Table 2. Belief matrix for one target P_L and two measurements A_1 and A_2

	P_L
A_1	$M_{L,1}=m_{L1}(A_1)\times m_{L2}(\overline{A_2})$
A_2	$M_{L,2}=m_{L1}(\overline{A_1})\times m_{L2}(A_2)$
ϕ	$M_{L,\phi}=m_{L1}(A_1)\times m_{L2}(A_2) + m_{L1}(\overline{A_1})\times m_{L2}(\overline{A_2})$

By recurrence, we obtain the following generalization for computing masses for I predicted targets and J validated measurements:

$$M_{i,j} = m_{ij}(A_j) \times \prod_{\substack{k=1 \\ k\neq j}}^{J} m_{ik}(\bar{A}_k) \tag{4}$$

$$M_{i,\emptyset} = 1 - \sum_{j=1}^{J} m_{ij}(A_j) \times \prod_{\substack{k=1 \\ k\neq j}}^{J} m_{ik}(\bar{A}_k) \tag{5}$$

4 Decision Making

Using the results of Table 1, we can make decisions about the association of observation A_j to target P_i by associating in each column the observation with the highest mass value to the related target. But this method leads to a suboptimal solution because it takes into consideration only local information regardless of other targets. Our idea is to take a global decision using all the elements of the belief matrix. Each column of the belief matrix is considered as a source of information, which define a bbm distribution on the singleton in the frame of discernment Θ_i. We want to combine these I bbm to obtain a single distribution. A direct combination is not possible because the I bbm are defined on different frames of discernment. For this reason, we extend each virtual distribution by applying a cylindrical extension of its focal

elements to obtain a single frame of discernment $\theta = \Theta_1 \times \ldots \times \Theta_{i-1} \times \Theta_i \times \Theta_{i+1} \times \ldots \times \Theta_I$ common to all I distributions. Also, each mass assigned to a subset $\{(A_j, P_i)\}$ of Θ_i becomes assigned to the subset $\Theta_1 \times \ldots \times \Theta_{i-1} \times \{(A_j, P_i)\} \times \Theta_{i+1} \times \ldots \times \Theta_I$ of θ. These I distributions, which are defined on the same frame of discernment, must be combined to a new distribution $m_T(.)$ on θ to make a global opinion about the association between the I targets and the J observations. This is done using the normalized conjunctive rule of combination. We obtain:

$$m_T(\{(A_{k1}, P_1), \ldots, (A_{ki}, P_i), \ldots, (A_{kI}, P_I)\}) = \prod_{i=1}^{I} \frac{M_{i,k_i}}{1 - M_{i,\emptyset}} \qquad \forall\, ki \in \{1, \ldots, J\} \tag{6}$$

where $\{(A_{k1}, P_1), \ldots, (A_{ki}, P_i), \ldots, (A_{kI}, P_I)\}$ Denotes the proposition: "the targets $P_1 \ldots P_i \ldots P_I$ generate the observations $A_{k1} \ldots A_{ki} \ldots A_{kI}$ respectively." The element of θ with the highest mass value is chosen as the solution to the association problem with the additional assumption that no two tracks are associated with the same observation.

5 Mass Function Generation

The bbm distribution generation is a crucial step in establishing the belief model. We start with:

$$m_{ij}(A_j) = \beta_{ij} \tag{7}$$

$$m_{ij}(\bar{A}_j) = 1 - \beta_{ij} \tag{8}$$

Where β_{ij} represent the probability for assigning measurement A_j to target P_i. We suggest using an approximation of the value pointed out in the development of suboptimal JPDAF [14]:

$$\beta_{ij} = G_{ij} / \sum_{j=1}^{J} G_{ij} \tag{9}$$

Where G_{ij} is the likelihood function associated to the assignment of observation j to track i. If we assume a Gaussian distribution for the innovation vector y_{ij} the likelihood function G_{ij} is:

$$G_{ij} = exp - \frac{y_{ij}^T \Gamma_i^{-1} y_{ij}}{2} \Big/ (2\pi)^{\frac{M}{2}} \sqrt{|\Gamma_i|} \tag{10}$$

Γ_i is the innovation covariance matrix for track I, and M is the dimension of the observation vector.

6 Simulation Results

We want to evaluate the performance of our method trough a comparison with both the NNF and the JPDAF in terms of both correct association percent and RMSE (precision). The simulations are done for three kinematic targets moving in a two dimensional space with a nearly constant velocity. For each target, this amount to:

$$X(k+1) = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix} X(k) + \begin{bmatrix} \frac{T^2}{2} & 0 \\ T & 0 \\ 0 & \frac{T^2}{2} \\ 0 & T \end{bmatrix} W(k) \tag{11}$$

The state vector $X(k)$ consists of the position and the velocity components in Cartesian coordinates at time k:

$$X(k) = \begin{bmatrix} X_1(k) \\ X_2(k) \\ X_3(k) \\ X_4(k) \end{bmatrix} = \begin{bmatrix} x(k) \\ \dot{x}(k) \\ y(k) \\ \dot{y}(k) \end{bmatrix} \tag{12}$$

T denotes the sampling period, $W(k)$ is the state noise assumed to be Gaussian and centered with covariance matrix $Q = 0.03 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Each target is characterized by its noisy position (range and bearing) $Z(k)$ given by:

$$Z(k) = \begin{bmatrix} \sqrt{X_1^2(k) + X_3^2(k)} \\ \tan^{-1} \frac{X_3(k)}{X_1(k)} \end{bmatrix} + V(k) \tag{13}$$

$V(k)$ is an additive noise, independent from $W(k)$, and with covariance matrix $R = \begin{bmatrix} 100 & 0 \\ 0 & 0.003 \end{bmatrix}$.

To analyse the performance of the EDAF in ambiguous situation, we consider three closest targets (Fig1):

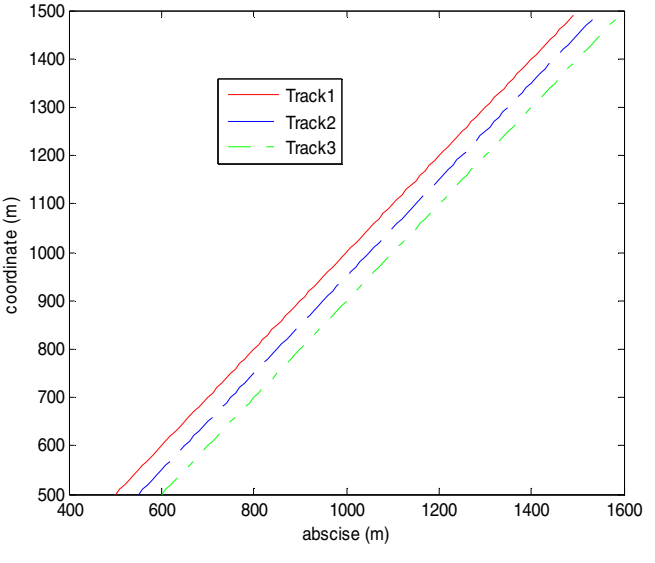


Fig. 1. Clean trajectories of the three targets

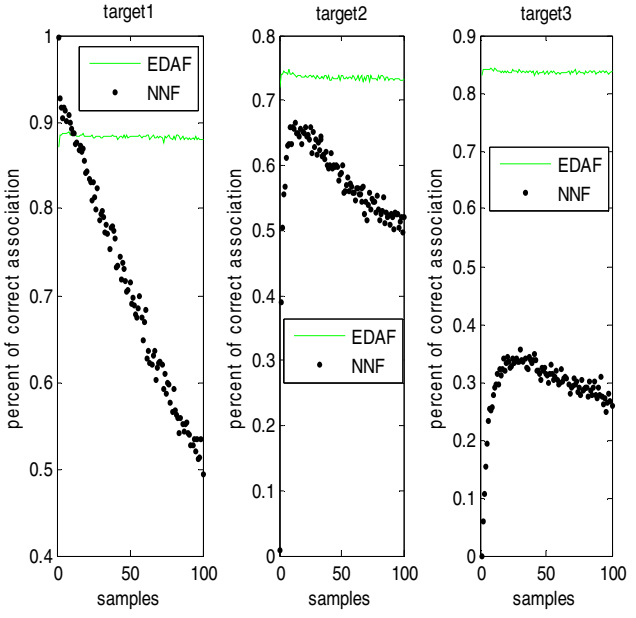


Fig. 2. Percent of correct association with the EDAF and the NNF

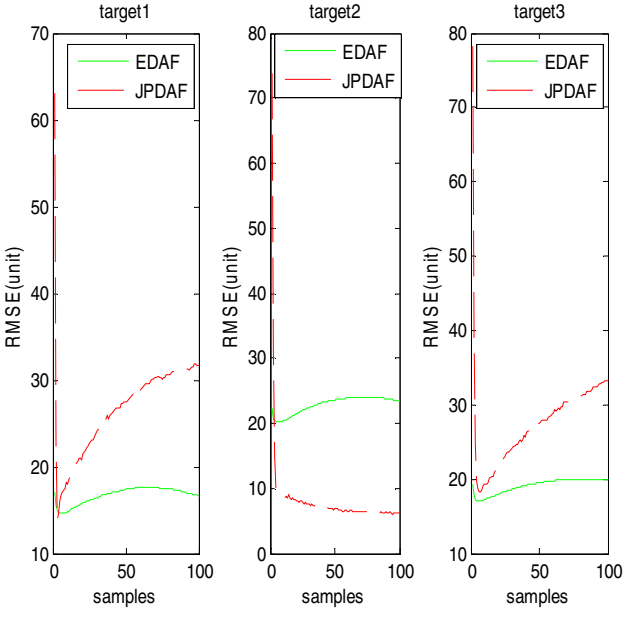


Fig. 3. RMSE position with the EDAF and the JPDAF

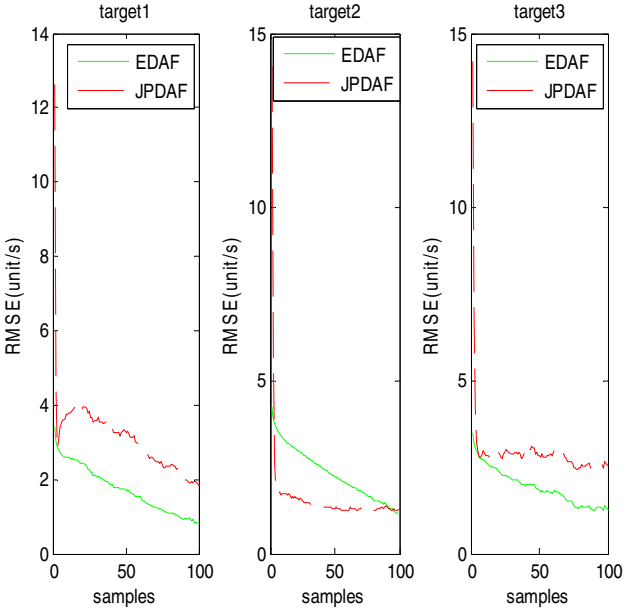


Fig. 4. RMSE velocity with the EDAF and the JPDAF

- with initial positions: track1 (500,500), track2 (550,500) and track3 (600,500)
- and with initial velocities: track1 (100,100), track2 (100,100) and track3 (100,100)

In the present work, an extended Kalman filter is used to generate independently for each target, state prediction and estimation. When new measurements are available, the predictions are associated with the incoming observations by using our method. The results of 500 Monte Carlo simulations are depicted in Fig 2, Fig 3 and Fig 4 As may be seen the proposed method is better than NNF in terms of percent of correct association (Fig 2). In addition the RMSE of the proposed method is better than the one of JPDAF in the global case; this is true for both position (Fig 3) and velocity (Fig 4).

7 Conclusion

In the present letter, we propose a new and simple method, for data association in multi target tracking, based on the theory of evidence. In our method the representation of the information is made using belief functions. The bbm is computed using a practical Bayesian approach and decision making is based on the extension of the frame of discernment. Simulations have been provided to evaluate the performance of our method in ambiguous association situation. It is shown that the tracking performance is better for the proposed algorithm than for JPDAF and NNF.

References

1. Kirubarajan, T., Bar-Shalom, Y.: Probabilistic Data Association Techniques for Target Tracking in Clutter. IEEE Proceeding 92(3), 536–557 (2004)
2. Bar-Shalom, Y. (ed.): Multitarget-Multisensor Tracking: Advanced Applications, vol. I. Artech House, Dedham (1990); Reprinted by YBS Publishing (1998)
3. Pulford, G.W.: Taxonomy of multiple target tracking methods. IEE Proc. Radar Sonar Navig. 152(5), 291–304 (2005)
4. Shafer, G.: Mathematical theory of evidence. Princeton University Press, Princeton (1976)
5. Dempster, A.: A generalization of bayesian inference. Journal of Royal Statistical Society series B 30, 205–247 (1968)
6. Smithson, M.: Ignorance and Uncertainty. Springer, Heidelberg (1989)
7. Dempster, A.P.: The Dempster–Shafer calculus for statisticians. Internat. J. Approx. Reason. 48, 365–377 (2008)
8. Blasch, E., Westerkamp, J., Hong, L., Layne, J., Garber, J., Shaw, A.: Identifying moving HRR signatures with an ATR belief data association filter. In: Proc. SPIE, Orlando FL, April 2000, vol. 4053, pp. 479–488 (2000)
9. Ayoun, A., Smets, P.: Data association in multi-target detection using the transferable belief model. Inter. J. Intell. Systems 16(1), 1167–1182 (2001)
10. Megherbi, N., Ambellouis, S., Colot, O., Cabestaing, F.: Multimodal data association based on the use of belief functions for multiple target tracking. In: 8th International conference on Information Fusion. Philadelphia (2005)
11. Fitzgerald, R.J.: Development of Practical PDA Logic for Multitarget Tracking by Microprocessors. In: Bar-Shalom, Y. (ed.) Multitarget-Multisensor Tracking: Advanced Applications, ch. 1. Artech House, Norwood (1990)
12. Gruyer, D., Royere, C., Cherfaoui, C.: Heterogeneous multi-criteria combination with partial or full information. In: 6th international conference on information fusion Cairns, Australia (2003)
13. Mourllion, B., Gruyere, D., Royere, C., Theroude, S.: Multi-Hypotheses Tracking Algorithm Based on the Belief Theory. In: 8th international conference on information fusion Philadelphia, NJ, USA (2005)
14. Aoki, E.H., Kienitz, K.H.: Suboptimal JPDA for Tracking in the Presence of Clutter and Missed Detections. In: 12th International Conference on Information Fusion Seattle, WA, USA (2009)

Maintaining Evidential Frequent Itemsets in Case of Data Deletion

Mohamed Anis Bach Tobji[1] and Boutheina Ben Yaghlane[2]

[1] LARODEC, ISG, University of Tunis
[2] LARODEC, IHEC, University 7 November in Carthage

Abstract. Incremental Maintenance of Frequent Itemsets (IMFI) consists in maintaining a set of extracted patterns when mined data are updated. This field knew considerable improvement in the last decade. However, it is not sufficiently tackled when mined data are imperfect, especially where imperfection is modelled by the evidence theory. In this work, we maintain incrementally the set of initially extracted itemsets both in cases of insertion and deletion of evidential data. Experimentations led on our method show satisfying results.

1 Introduction

Recently, several Frequent Itemset Mining (FIM) techniques that process imperfect data emerged. Mined data are *probabilistic* [10], *possibilistic* [8], *fuzzy* [4] and *evidential* [9,2].

When a FIM operation is performed, the obtained patterns describe the data only at the moment of mining. But data are dynamic due to continuous users' updates. Therefore, the mined patterns become invalid at some point. These updates could invalidate some already frequent itemsets, and turn some invalid itemsets to frequent ones. Hence, we are obliged to maintain the set of frequent itemsets initially mined. There are two approaches to handle this problem; the classic approach and the incremental one.

The classic approach consists in performing a new FIM operation over the updated database. The drawback of this method is it ignores the initial set of frequent itemsets. On the other hand, the incremental approach consists in computing the new set of frequent itemsets starting from the initial ones, hence its incremental feature. This method is less costly in term of execution time.

The field of Incremental Maintenance of Frequent Itemsets (IMFI) has attracted attention of several researches. The literature is abundant in this way [5,13]. However, IMFI in uncertain databases is not sufficiently tackled in spite of its importance and applicability to interesting fields where data are frequently updated. In this paper we extend the work of [3] where authors introduced a new method for maintaining incrementally frequent itemsets when data updates consists only of data insertion. Our new solution takes into account, besides

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 218–227, 2010.

the data insertion, the data deletion. Consequently, the new method handles any type of data update whether it is data insertion, data deletion and even data modification since a modification could be performed via deletion (of the data to be modified) and then insertion (of the changed data). We recall that our method handles imperfect data that is modelled via the Dempster-Shafer theory [6].

The rest of the paper is organized as follows: in section 2, we briefly introduce the basic concepts of the FIM from evidential data. In section 3, we introduce the problem of IMFI in evidential data and the solution we propose. The section 4 contains the experimentations led on our algorithm and finally section 5 is the conclusion of the work.

2 Mining Frequent Itemsets in Evidential Databases

2.1 Evidential Databases

In the literature, several papers have introduced what we call *belief databases*, *Dempster-Shafer databases*, or simply *evidential databases* [7,9]. This type of database contains uncertain and/or imprecise data that is represented via the evidence theory [6].

An evidential database contains n attributes and d lines. Each attribute i ($1 \leq i \leq n$) has a domain D_i of discrete values. Each attribute k among the n ones could store *evidential* values. An instance of the attribute k in a line j is an *evidential value* V_{kj} which is a *basic belief assignment* (bba) defined as follows:

$$m_{kj} : 2^{D_k} \rightarrow [0,1] \text{ with: } m_{kj}(\emptyset) = 0 \text{ and } \sum_{x \subseteq D_k} m_{kj}(x) = 1$$

Table 1 is an example of evidential database. Note that in such database, we can store imperfect information that could be modelled via the probability theory (such in the cell $A1$) or the possibility theory (such in the cell $C1$). Evidential databases derive their robustness from the evidence theory that is able to represent a large variety of data imperfection types.

Table 1. Evidential database example

id	A	B	C
1	$A_1(0.6)$ $A_2(0.4)$	$B_1(0.4)$ $\{B_5, B_6, B_7\}(0.6)$	$C_1(0.5)$ $\{C_1, C_2\}(0.5)$
2	$A_1(0.2)$ $A_3(0.3)$ $\{A_2, A_3\}(0.5)$	$B_1(0.4)$ $\{B_2, B_3\}(0.6)$	C_2

2.2 Basic Concepts

The basic concepts of FIM in evidential databases were established in [9] and then refined in [2]. New definitions of the concepts of *item*, *itemset* and *support* were adapted to the imperfect nature of the data. The preliminaries of this FIM model are briefly presented here, we recommend readers to consult the paper [2] for more explanation:

An *evidential item* denoted iv_k is one focal element in a body of evidence V_{kj} which corresponds to the evidential attribute k. Thus, it is defined as a subset of D_k ($iv_k \in 2^{D_k}$). For example, in table 1, C_1 is an evidential item, $\{C_1, C_2\}$ too.

An *evidential itemset* is a set of evidential items that correspond to different attributes. For example, $A_1B_1\{C_1, C_2\}$ is an evidential itemset. Formally, an evidential itemset X is defined as: $X \in \prod_{1 \leq i \leq n} 2^{D_i}$

The *inclusion operator* for evidential itemsets is defined as follows: let X and Y be two evidential itemsets. The i^{th} items of X and Y are respectively denoted by i_X and i_Y.

$$X \subseteq Y \text{ if and only if: } \forall i_X \in X, i_X \subseteq i_Y$$

For example, the itemset $A_1B_1\{C_1, C_2\}$ includes the itemset $A_1B_1C_1$.

The body of evidence of one evidential database EDB is defined on the frame of discernment $\Theta = \prod_{1 \leq i \leq n} 2^{D_i}$. Its set of focal elements is composed of all possible evidential itemsets existing in the database, and the mass function m_{EDB} is defined as follows: Let X be an evidential itemset and d be the size of EDB:

$$m_{DB} : \Theta \rightarrow [0, 1] \text{ with } m_{DB}(X) = \frac{1}{d} \sum_{j=1}^{d} \textcircled{\cap}_{i \leq n} m_{ij}(X) = \frac{1}{d} \sum_{j=1}^{d} \prod_{iv_i \in X} m_{ij}(iv_i)$$

Note that $\textcircled{\cap}$ is the operator of the conjunctive rule of combination introduced in [12]. The belief function is naturally defined as follows: $Bel_{DB}(X) = \sum_{Y \subseteq X} m_{DB}(Y)$

Example 1. In the database example (table 1) the mass of the evidential itemset $A_1B_1\{C_1, C_2\}$ is the sum of its line masses in the database divided by $d = 2$ so $m_{DB}(A_1B_1\{C_1, C_2\}) = 0.06$. Its belief in the database is the sum of all database masses of evidential itemsets that are included in it, which are $A_1B_1C_1(0.06)$, $A_1B_1C_2(0.04)$ and $A_1B_1\{C_1, C_2\}(0.06)$ so $Bel_{DB}(A_1B_1\{C_1, C_2\}) = 0.16$.

According to [9,2], the support of an itemset X in the evidential database, is its belief measurement in the database *BoE*.

2.3 The FIM Problem

Let EDB be an evidential database, X be an evidential itemset and Θ be the cross product of all attributes' domains. The FIM problem consists in extracting the set F that contains the frequent evidential itemsets in EDB whose supports exceed the user-defined support threshold denoted by min_{supp}. Formally, the

set F is defined as follows: $F = \{X \subseteq \Theta / X.support_{EDB} \geq min_{supp}\}$ where $X.support_{EDB}$ is the support of the itemset X in the database EDB.

3 Maintaining Frequent Itemsets in Evidential Databases

3.1 Problem Definition

The problem of IMFI in perfect databases was introduced in [5]. We present here the same problem in the context of evidential databases. It is formally defined as follows:

Let EDB be an evidential database and D be its size. Let F be the set of frequent itemsets in EDB and min_{supp} the support threshold under which F was mined. After some updates of EDB -consisting in inserting the increment edb^+ of size d^+, and deleting the decrement edb^- of size d^--, we obtain $EDB\prime = EDB \cup edb^+ \setminus edb^-$. The size of $EDB\prime$ is denoted by $D\prime$. The problem of IMFI consists in computing $F\prime$: the set of frequent evidential itemsets in $EDB\prime$ under the initial support threshold min_{supp}.

3.2 Classification of Itemsets

In this section, we study the different kinds of itemsets that exist in the updated database $EDB\prime$. The classification done below helps us to constitute the set $F\prime$ of new frequent itemsets. it is based on the status (frequent or not) of the itemsets in EDB and then their status in $EDB\prime$.

Indeed, when EDB is updated by inserting edb^+ and deleting edb^-, some itemsets that were infrequent in EDB will emerge to be frequent in $EDB\prime$, and vice versa; some itemsets that were frequent in EDB will become infrequent under the threshold min_{supp}. Below is the exact classification of the itemsets in $EDB\prime$:

- **Winner Itemsets** are itemsets X that were infrequent in EDB ($X.support_{EDB} < min_{supp} \times D$) and become frequent in $EDB\prime$ thanks to the updates.
- **Loser Itemsets** are itemsets X that were frequent in EDB ($X.support_{EDB} \geq min_{supp} \times D$) and become infrequent in $EDB\prime$ because of the updates.
- **Persistent Itemsets** are itemsets X that were frequent in EDB ($X.support_{EDB} \geq min_{supp} \times D$) and remain frequent in $EDB\prime$ in spite of the updates.
- **Invalid Itemsets** are itemsets X that were infrequent in EDB ($X.support_{EDB} < min_{supp} \times D$) and remain also infrequent in $EDB\prime$.
- **Hidden Itemsets** are itemsets X that are composed of non-singleton items (such as $A_4\{B_5, B_6\}$) and occur in the data increment but not in the initial database. These itemsets could be frequent in EDB without being present in F because they didn't occur in EDB at least once.

An itemset belongs necessarily to one of these classes. Now, the goal of the IMFI is to compute the set $F\prime$ that is composed of the sets W and P of respectively *winner* and *persistent* itemsets, but also of some *hidden* itemsets. These "winner" hidden itemsets are composed (vs. singleton) itemsets that occurred only in the data increment and whose supports in $EDB\prime$ exceed the support threshold.

3.3 Description of Our Incremental Method

Unlike the IMFI method presented in [3], our new solution takes into account deletion updates. It produces the itemsets in $F\prime$ level-by-level in the itemset lattice, i.e., it mines the frequent itemsets of size k before extracting the itemsets of size $k+1$ and so on. It proceeds exactly as follows:

First, we generate the set of candidate itemsets of size k, denoted by C_k, from the set $F\prime_{k-1}$ (frequent itemsets of size $k-1$ in $EDB\prime$)[1] via the *Apriori_Gen* function [1]. C_k contains three types of itemsets; (1) itemsets of F_k whose supports in EDB are known, these itemsets compose the set PP_k of potentially persistent itemsets, (2) itemsets composed of singleton items (like $A_1B_2C_1$ or $A_2B_2C_1$ but not $A_2\{B_2, B_3\}$) that were not frequent in EDB and so future winner or invalid itemsets, they compose the set PW_k of potentially winner itemsets, and (3) itemsets composed of non singleton items, that didn't occur in EDB and thus we have no information about their frequency (or not) in the initial database. These itemsets are handled in the set SS, the frequent ones among them will be stored in the set FS.

Thus, the set C_k is split into three complementary sets PP_k, PW_k and SS_k; the set $PP_k = C_k \cap F_k$ including candidate itemsets that are in F_k (potentially persistent itemsets) and the set $PW_k \cup SS_k = C_k \setminus F_k$ including the remainder candidate itemsets. Then the set PW_k will contain the itemsets that are composed of singleton items, and SS_k will contain the rest. After preparing our three sets of candidate itemsets (PW_k, PP_k and SS_k), we scan the decrement edb^- and we update supports of itemsets in PP_k (so we get their supports in $EDB \setminus edb^-$). We also obtain the supports of the itemsets of PW_k and SS_k in edb^-. At this step, we can prune the set PP_k thanks to the proposition 1. Indeed, if the support of an itemset in $EDB \setminus edb^-$ is less than the *Persistents Pruning Threshold* (see proposition 1) denoted by $ppt = min_{supp} \times D\prime - d^+$, then it could not be frequent in $EDB\prime$. This kind of itemsets is deleted from PP_k. After, we scan the increment edb^+ to update the supports of the itemsets of PP_k, PW_k and SS_k. After this scan, we can already distinguish between persistent itemsets ($X.support_{EDB\prime} \geq min_{supp} \times D\prime$) and loser ones ($X.support_{EDB\prime} < min_{supp} \times D\prime$). The set P_k of persistent itemset is already computed after the light scans of the data decrement edb^- and increment edb^+. Once the persistent itemsets are found, we will try to compute the winner itemsets and the frequent hidden itemsets to complete the set $F\prime$. Thanks to the proposition 2, we prune the set PW_k by eliminating all itemsets whose supports in $edb^+ \setminus edb^-$ do not exceed the *Winners Pruning Threshold* denoted by wpt

[1] Assume that the set C_1 in the first level contains all possible items in $EDB\prime$.

(see proposition 2). This optimization is very important because it makes the return to the initial database less heavy. Indeed, from the itemsets of PW_k, only those whose supports exceed the *wpt* will be updated when scanning the initial database $EDB \setminus edb^-$, to obtain their supports in the whole of $EDB\prime$. After this scan, we can filter the winner itemsets from the invalid ones by comparing their supports in $EDB\prime$ to the minimum support threshold $min_{supp} \times D\prime$. Finally, we are obliged to compute the supports of the itemsets of SS_k in the whole of $EDB \setminus edb^-$, to get the set FS of frequent ones in the updated database. The computation of FS is a costly operation that we cannot avoid.

Proposition 1 (The Persistents Pruning Threshold). *Let X be a frequent itemset in EDB. X could not persist if:* $X.support_{EDB} - X.support_{edb^-} < min_{supp} \times D\prime - d^+$

Proof. X is frequent in $EDB\prime$ if and only if: $X.support_{EDB\prime} \geq min_{supp} \times D\prime \Leftrightarrow X.support_{EDB} + X.support_{edb^+} - X.support_{edb^-} \geq min_{supp} \times D\prime \Leftrightarrow X.support_{EDB} - X.support_{edb^-} \geq min_{supp} \times D\prime - X.support_{edb^+}$ **(1)** We know that the support of X in edb^+ is at the most d^+, and so $X.support_{edb^+} \leq d^+$ **(2)**

(1) and **(2)** $\Rightarrow X.support_{EDB} - X.support_{edb^-} \geq min_{supp} \times D\prime - d^+$

Proposition 2 (The Winners Pruning Threshold). *Let be X an itemset that was infrequent in EDB. X could not win if:* $X.support_{edb^+} - X.support_{edb^-} < min_{supp} \times (d^+ - d^-)$

Proof. X is frequent in $EDB\prime$ if and only if: $X.support_{EDB\prime} \geq min_{supp} \times D\prime \Leftrightarrow X.support_{EDB} + X.support_{edb^+} - X.support_{edb^-} \geq min_{supp} \times D\prime \Leftrightarrow X.support_{edb^+} - X.support_{edb^-} \geq min_{supp} \times D\prime - X.support_{EDB}$ **(1)** Now, we know that X is infrequent in $EDB \Leftrightarrow X.support_{EDB} < min_{supp} \times D$ **(2)**

(1) and **(2)** $\Rightarrow X.support_{edb^+} - X.support_{edb^-} \geq min_{supp} \times D\prime - min_{supp} \times D \Leftrightarrow X.support_{edb^+} - X.support_{edb^-} \geq min_{supp} \times (d^+ - d^-)$

In the next section, we present the data structure we use to accelerate the support computation. The data structure is adapted to the evidential character of the data, it allows also the optimization of the early pruning of loser and invalid itemsets.

3.4 Used Data Structure

To compute the supports of the candidate itemsets, we use the *RidLists* data structure introduced in [2]. It consists in storing for each evidential item the list of the couples (1) *record identifier* of the line that contains the item and (2) *the belief* of the item in the corresponding record. Table 2 presents the *RidLists* of only the evidential items $\{A_2, A_3\}$ and $\{C_1, C_2\}$. Once we have the *RidLists* representation of the evidential database, we can compute the support of any itemset via the intersection of the lists of its items. Its support is the sum of the product of the believes of the shared records identifiers of its items [2].

Table 2. The *RidLists* of the items $\{A_2, A_3\}$ and $\{C_1, C_2\}$

item	***rid* list**
$\{A_2, A_3\}$	$(1, 0.4)(2, 0.8)$
$\{C_1, C_2\}$	$(1, 1)(2, 1)$

3.5 The Incremental Maintenance Algorithm

Our algorithm proceeds level-by-level as described above. In the first iteration, we compute the frequent evidential items. It is a particular iteration compared to the other ones, i.e., when $k \geq 2$. Indeed, when $k \geq 2$ the set C_k -from which starts our method- is computed from the set $F\prime_{k-1}$ via the *Apriori_Gen* function. Thus, this latter function could not generate the set C_1, hence the particularity of the first iteration.

Algorithm 1 presents the procedure that computes the frequent evidential items. It starts not from a candidate set, but from both the items of the data increment and the set of initial frequent items F_1. This method allows to compute the sets PP_1 and PW_1, but also the set SS of *super items*, i.e., non-singleton items that are not in F_1. This latter set is very special because it includes the only items whose we do not know any information about their frequency in EDB. In other words, we do not know if these items are frequent in EDB or not. We present here the procedure *ComputeFrequentItems* followed by a detailed example that explains more explicitly our method.

Algorithm 1. ComputeFrequentItems

Require: RL_{EDB} as RIDLIST, RL_{edb^+} as RIDLIST, RL_{edb^-} as RIDLIST, F_1 as Set of Items, min_{supp} as Real
Ensure: $F_1\prime$ as Set of Items

1 $PP_1 \leftarrow F_1$
2 **for all** item i in PP_1 **do**
3 compute $i.support_{edb^-}$
4 **if** $i.support_{EDB} - i.support_{edb^-} < min_{supp} \times D\prime - d^+$ **then**
5 Delete i from PP_1
6 **else**
7 Compute $i.support_{edb^+}$
8 **if** $i.support_{EDB\prime} \geq min_{supp} \times D\prime$ **then**
9 Add i to P_1
10 **else**
11 Delete i from PP_1
12 **end if**
13 **end if**
14 **end for**
15 **for all** item i in RL_{edb^+} and not in F_1 **do**
16 **if** i is a superset **then**
17 Add i to SS
18 **else**
19 Compute $i.support_{edb^+}$
20 Compute $i.support_{edb^-}$
21 **if** $i.support_{edb^+} - i.support_{edb^-} \geq min_{supp} \times (d^+ - d^-)$ **then**
22 Add i to PW_1
23 **end if**
24 **end if**
25 **end for**
26 **for all** item i in PW_1 **do**
27 Compute $i.support_{EDB \setminus edb^-}$
28 **if** $i.support_{EDB\prime} \geq min_{supp} \times D\prime$ **then**
29 Add i to W_1
30 **end if**
31 **end for**
32 **for all** item i in SS **do**
33 Compute $i.support_{EDB\prime}$
34 **if** $i.support_{EDB\prime} \geq min_{supp} \times D\prime$ **then**
35 Add i to FS
36 **end if**
37 **end for**
38 $F_1\prime \leftarrow P_1 \cup W_1 \cup FS$

Algorithm 2 is the process of our incremental maintenance solution that generates the set $F\prime$.

Algorithm 2. ComputeFrequentItemsets

Require: RL_{EDB} as RIDLIST, RL_{edb+} as RIDLIST, RL_{edb-} as RIDLIST, F as Set of Itemsets, min_{supp} as Real
Ensure: $F\prime$ as Set of Itemsets
1 ComputeFrequentItems($RL_{EDB}, RL_{edb+}, F_1, min_{supp}, F_1\prime$)
2 $k \leftarrow 1$
3 **while** $F_k\prime \neq \emptyset$ **do**
4 $k \leftarrow k+1$; $C_k \leftarrow Apriori\_Gen(F_{k-1}\prime)$
5 $PP_k \leftarrow F_k \cap C_k$;$PW_k \leftarrow C_k \setminus F_k$
6 **for all** itemset $X \in PP_k$ **do**
7 Compute $X.support_{edb-}$
8 **if** $X.support_{EDB} - X.support_{edb-} < min_{supp} \times D - d^+$ **then**
9 Delete X from PP_k
10 **else**
11 Compute $X.support_{edb+}$
12 **if** $X.support_{EDB\prime} < min_{supp} \times D\prime$ **then**
13 Delete X from PP_k
14 **else**
15 Add X to P_k
16 **end if**
17 **end if**
18 **end for**
19 **for all** itemset $X \in PW_k$ **do**
20 Compute $i.support_{edb+}$
21 Compute $i.support_{edb-}$
22 **if** X is composed of singleton-items **then**
23 **if** $X.support_{edb+} - X.support_{edb-} < min_{supp} \times (d^+ - d^-)$ **then**
24 Delete X from PW_k
25 **else**
26 Compute $X.support_{EDB \setminus edb-}$
27 **if** $X.support_{EDB\prime} \geq min_{supp} \times D\prime$ **then**
28 Add X to W_k
29 **end if**
30 **end if**
31 **else**
32 Add X to SS
33 **end if**
34 **end for**
35 **for all** itemset $X \in SS$ **do**
36 Compute $X.support_{EDB \setminus edb-}$
37 **if** $X.support_{EDB\prime} \geq min_{supp} \times D\prime$ **then**
38 Add X to FS
39 **end if**
40 **end for**
41 $F_k\prime \leftarrow P_k \cup W_k \cup FS$
42 **end while**

4 Performance Analysis

In order to evaluate the performance of our solution, we implemented the methods of [2] -denoted by *FIMED*- and of [9] -denoted by *HPSS*- as well as our solution. These three algorithms were tested over several synthetic databases [2]. For short, we content with the tests led on the database *D4000I500C14%U10*. The size of this database is 4000 lines, it contains 500 items distributed on 14 columns. The percent of the records that include evidential values is 10%.

In the first experimentation, we perform a simple FIM operation over the first 3500 records under a fixed support threshold, and we stored the set F of frequent itemsets. Then, we insert the last 500 records (those we eliminated in the initial FIM operation), and we delete the first 500 records. After that, we maintain the set F after the updates (insertion and deletion) we did, to produce the set $F\prime$ of frequent itemsets in the updated database. We used the three implemented methods, that are FIMED and HPSS (classic approach) and our solution (incremental approach). Indeed, when we perform maintenance via our solution, we use the initial data set (that includes 3500 records), the increment data (500 records), the decrement data (500 records) and the set F of the initial frequent itemsets in the initial database. These operations are performed for several support thresholds.

The figure 1 shows how the incremental solution is the less costly in term of execution time. That is logical since this solution takes a considerable advantage

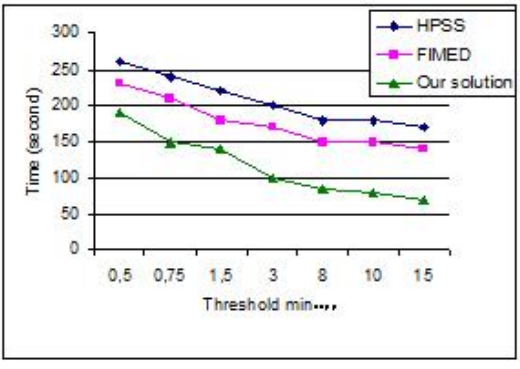


Fig. 1. Comparison of the performances of FIMED, HPSS and our solution

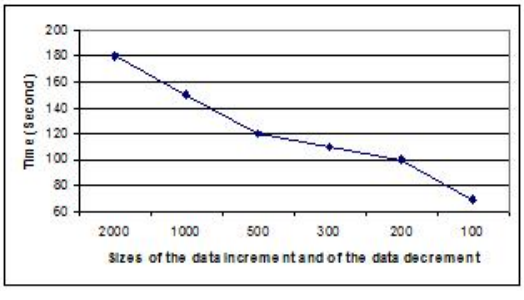


Fig. 2. Effect of the data increment and decrement sizes on the solution's performance

from the result of the initial mining operation, i.e., the set F of initially frequent itemsets. Indeed, the difference between the incremental solution and the classic ones is essentially in the candidate itemsets generation phase where our method lightens the set of candidates thanks to the pruning techniques presented in section 3.3. The second experimentation is about the effect of the data increment and data decrements sizes on our solution's performance. In figure 2, we note that the more the sizes of the data increment and decrement are important, the less the incremental solution is efficient. Indeed, our solution takes a considerable advantage from the persistent itemsets computation. When the set P_k is large, its complement in C_k, namely $PW_k \cap SS_k$ is small. In this case, the return to the initial database is light which accelerates the computation of $F\prime$. Now, the more the decrement and the increment are important, the more $PW_k \cap SS_k$ are also important. More simply, when the changes (insertion and deletion) are important, the difference between the initial frequent itemsets (namely the set F) and the new ones (namely the set $F\prime$) is also important and the incremental solution takes less advantage from the results of the previous FIM operation.

5 Conclusion

This work aims to solve the problem of incremental maintenance of frequent itemsets when mined data are updated. It is an alternative to the classic solution that consists in mining again the whole of the updated data without taking into consideration the results of the previous mining operations. The experimentations led on our method showed satisfying results.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499 (1994)
2. Bach Tobji, M.-A., Ben Yaghlane, B., Mellouli, K.: A New Algorithm for Mining Frequent Itemsets from Evidential Databases. In: Proceedings of the 12th International Conference on IPMU, Malaga, Spain, pp. 1535–1542 (2008)
3. Bach Tobji, M.-A., Ben Yaghlane, B., Mellouli, K.: Incremental Maintenance of Frequent Itemsets in Evidential Databases. In: Sossai, C., Chemello, G. (eds.) ECSQARU 2009. LNCS, vol. 5590, pp. 457–468. Springer, Heidelberg (2009)
4. Chen, G., Wei, Q.: Fuzzy association rules and the extended mining algorithms. Information Sciences-Informatics and Computer Science 147(1-4), 201–228 (2002)
5. Cheung, D.-W., Han, H., Ng, N.-T.: Maintenance of Discovered Association Rules in Large Databases: An Incremental Update technique. In: Proceedings of The Twelfth International Conference on Data Engineering, pp. 106–114 (1996)
6. Dempster, A.P.: Upper and Lower Probabilities Induced by a Multivalue Mapping. Annals of Mathematical Statistics 38(2), 325–339 (1967)
7. Dai, B., Bell, D.A., Hughes, J.G.:Query processing in temporal evidential databases. In: Proceeding of the 24th Seminar on Current Trends in Theory and Practice of Informatics Milovy, Czech Republic, pp. 383–390 (1997)
8. Djouadi, Y., Redaoui, S., Amroun, K.: Mining Association Rules under Imprecision and Vagueness: towards a Possibilistic Approach. In: Intl. Fuzzy Systems Conference, vol. (23-26), pp. 1–6 (2007)
9. Hewawasam, K.K.R.G.K., Premaratne, K., Subasingha, S.P., Shyu, M.-L.: Rule Mining and Classification in Imperfect Databases. In: Proceedings of the Seventh International Conference on Information Fusion, pp. 661–668 (2005)
10. Leung, C.K.-S., Carmichael, C.L., Hao, B.: Efficient Mining of Frequent Patterns from Uncertain Data. In: Proceedings of the Seventh IEEE International Conference on Data Mining, pp. 489–494 (2007)
11. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
12. Smets, P.: The Application of the Transferable Belief Model to Diagnostic Problems. Int. J. Intelligent Systems 13, 127–158 (1998)
13. Teng, W.-G., Chen, M.-S.: Incremental Mining on Association Rules. Studies in Fuzziness and Soft Computing, 125–162 (2005)

TS-Models from Evidential Clustering

Rui Jorge Almeida[1] and Uzay Kaymak[1,2]

[1] Erasmus University Rotterdam, Erasmus School of Economics, Rotterdam, The Netherlands
[2] Eindhoven University of Technology, School of Industrial Engineering, Eindhoven, The Netherlands
rjalmeida@ese.eur.nl, u.kaymak@ieee.org

Abstract. We study how to derive a fuzzy rule-based classification model using the theoretical framework of belief functions. For this purpose we use the recently proposed Evidential c-means (ECM) to derive Takagi-Sugeno (TS) models solely from data. ECM allocates, for each object, a mass of belief to any subsets of possible clusters, which allows to gain a deeper insight in the data while being robust with respect to outliers. Some classification examples are discussed, which show the advantages and disadvantages of the proposed algorithm.

1 Introduction

Classification problem is an important subject for a variety of fields, including pattern recognition, artificial intelligence, credit risk and direct marketing. In a classification problem the aim is to assign class labels to a set of data instances described by multiple features. A possible method to solve classification problems is to use a fuzzy rule based model, built from data [1,2,3]. Fuzzy models use if-then rules and logical connectives to establish relations between the variables defined for the model of the system. The fuzzy sets in the rules serve as an interface amongst qualitative conceptualization in the model, and the numerical input and output variables. The if-then rules provide a transparent description of the system, that may reflect a possible nonlinearity of the system. The rule-based nature of the model allows for a linguistic description of the knowledge.

One way of obtaining Takagi-Sugeno fuzzy models is product-space fuzzy clustering. A clustering algorithm finds a partition matrix which best explains and represents the unknown structure of the data with respect to the model that defines it [4]. Different clustering algorithms can be used, which will yield different information and insights about the underlying structure of the data.

Uncertainty in the data is a challenge for classification [5]. Several approaches have been proposed to deal with this problem, using the framework of belief functions. In [6,7,8], it was proposed to represent the partial knowledge regarding the class membership of an object using a basic belief assignment. A classification method based on the decision tree approach that takes into consideration the uncertainty characterized by the classes of the training examples, as well as the uncertainty of their attribute values was proposed in [9].

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 228–237, 2010.

In this paper we study the use of Evidential C-Means (ECM) for system identification. For this it is necessary to relate information represented in the framework of the theory of beliefs, as understood in the transferable belief model [10], and fuzzy sets. Previous studies have shown that consonant beliefs and fuzzy sets are related [11]. In our approach, we consider that the obtained credal partition captures relevant information for correct interpretation of data substructure, and we discuss the possibility to map the obtained credal partition to a fuzzy set, providing linguistic interpretation and labels to the obtained structure. Using the credal partition it is possible to highlight the points that unambiguously belong to one cluster, and the points that lie at the boundary of two or more clusters. We try to convene this added information into the rule based classification system.

The paper is organized as follows. Section 2 reviews briefly the main concepts underlying the theory of belief functions. Section 3 presents its use for deriving a credal partition from object data. Section 4 presents the method used for classification and system identification in this work. The experimental setup and the results are presented in Section 5 and Section 6, respectively. A brief discussion of the results is in Section 7. Finally the conclusions are given in Section 8.

2 Belief Functions

Dempster-Shafer theory of evidence, is a theoretical framework for reasoning with partial and unreliable information. In the following, we briefly recall some of the basics of the belief function theory. More details can be found in [12,10,13].

Let Ω be a finite set of elementary values ω called the frame of discernment. The basic belief assignment (bba) [12] is defined as a function m from 2^{Ω} to $[0,1]$, satisfying:

$$\sum_{A \subseteq \Omega} m(A) = 1 \ , \tag{1}$$

which represents the partial knowledge regarding the actual value taken by ω. The subsets A of Ω such that $m(A) > 0$ are the focal sets of m. Each focal set A is a set of possible values for ω, and the value $m(A)$ can be interpreted as the part of belief supporting exactly that the actual event belongs to A.

A bba m such that $m(\emptyset) = 0$ is said to be normal [12]. This condition may be relaxed by assuming that ω might take its value outside Ω, which means that Ω might be incomplete [14]. The quantity $m(\emptyset)$ is then interpreted as a mass of belief given to the hypothesis that ω might not lie in Ω. A bba m can be equivalently represented by a plausibility function $pl : 2^{\omega} \mapsto [0,1]$, defined as

$$\mathrm{pl}(A) \triangleq \sum_{B \cap A \neq \emptyset} m(B) \quad \forall A, B \subseteq \Omega \ . \tag{2}$$

The plausibility $\mathrm{pl}(A)$ represents the potential amount of support given to A.

The decision making problem regarding the selection of one single hypothesis in Ω, is solved in the transferable belief model framework, by using a pignistic probability, BetP, defined, for a normal bba, by [13]:

$$\text{BetP}(\omega) \triangleq \sum_{\omega \in A} \frac{m(A)}{|A|} \quad \forall \omega \in \Omega \ , \tag{3}$$

where $|A|$ denotes the cardinality of $A \subseteq \Omega$. It is shown, that this is the only transformation between belief function and a probability function satisfying elementary rationality requirements, in which each mass of belief $m(A)$ is equally distributed among the elements of A [15].

3 Evidential c-Means

In [6], the Evidential c-Means (ECM) algorithm was proposed to derive a credal partition from object data. In this algorithm the partial knowledge regarding the class membership of an object i is represented by a bba m_i on the set Ω. This representation makes it possible to model all situations ranging from complete ignorance to full certainty concerning the class label of the object. This idea was also applied to relational data in [8] and proximity data [7].

Determining a credal partition $M = (m_1, m_2, \ldots, m_n)$ from object data, using ECM, implies determining, for each object i, the quantities $m_{ij} = m_i(A_j)(A_j \neq \emptyset, A_j \subseteq \Omega)$ in such a way that m_{ij} is low (high) when the distance d_{ij} between i and the focal set A_j is high (low). The distance between an object and any non empty subset of Ω is defined by associating to each subset A_j of Ω the barycenter $\bar{\mathbf{v}}$ of the centers associated to the classes composing A_j. It is assumed that each class ω_k is represented by a center $\mathbf{v}_k \in \mathbb{R}^p$. Specifically,

$$s_{kj} = \begin{cases} 1, \text{ if } \omega_k \in A_j \\ 0 \ \text{ otherwise} \end{cases} . \tag{4}$$

The barycenter $\bar{\mathbf{v}}_j$ associated to A_j is:

$$\bar{v}_j = \frac{1}{\tau_j} \sum_{k=1}^{c} s_{kj} \mathbf{v}_k \ , \tag{5}$$

where $\tau_j = |A_j|$ denotes the cardinality of A_j. The distance d_{ij} is then defined as $d_{ij}^2 \triangleq ||\mathbf{x}_i - \bar{\mathbf{v}}_j||$. The proposed objective function for ECM, used to derive the credal partition M and the matrix V containing the cluster centers, is given by:

$$J_{ECM}(M, V, A) = \sum_{i=1}^{n} \sum_{\{j/A_j \subseteq \Omega, A_j \neq \emptyset\}} \tau_j^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta} \ , \tag{6}$$

subject to

$$\sum_{\{j/A_j \subseteq \Omega, A_j \neq \emptyset\}} m_{ij} + m_{i\emptyset} = 1 \ \forall i = 1, n \ , \tag{7}$$

where $\beta > 1$ is a weighting exponent that controls the fuzziness of the partition, δ controls the amount of data considered as outliers, and $m_{i\emptyset}$ denotes $m_i(\emptyset)$,the amount of evidence that the class of object i does not lie in Ω. The weighting coefficient τ_j^α aims at penalizing the subsets in Ω of high cardinality and the exponent α allows to control the degree of penalization. The second term of (7) is used to give a separate treatment term for the empty set. This focal element is in fact associated to a noise cluster, which allows to detect atypical data. The minimization of (7), can be done using the Lagrangian method.

The credal partition provides different structures, that can give different types of information about the data. A possibilistic partition could be obtained by computing from each bba m_i the plausibilities $\mathrm{pl}_i(\{w_k\})$ of the different clusters, using (2). The value $\mathrm{pl}_i(\{w_k\})$ represents the plausibility that object i belongs to cluster k. In the same way, a probabilistic fuzzy partition may be obtained by calculating the pignistic probability $\mathrm{BetP}_i(\{w_k\})$, using (3) induced by each bba m_i.

Furthermore, a hard credal partition can be obtained, by assigning each object to the set of clusters with the highest mass. This allows to divide the partition space into a maximum of 2^c groups. Formally, the $X(A_j)$ for $j = 1, \ldots, 2^c$ defines a hard credal partition of the n objects [6]:

$$X(A_j) = \{i / m_i(A_j) = \max_k m_i(A_k)\} \; . \tag{8}$$

Finally, it is possible to characterize each cluster ω_k by a set of objects which can be classified in ω_k without any ambiguity and the set of objects which could possibly be assigned to ω_k. These two sets ω_k^L and ω_k^U, are defined as the lower and upper approximations of ω_k respectively [6], and they are defined as:

$$\omega_k^L = X\{\omega_k\}, \qquad \text{and} \qquad \omega_k^U = \underset{j/\omega_k \in A_j}{U} X(A_j) \; . \tag{9}$$

The information obtained from the credal partition and its approximations can be considered intuitive and simple to interpret. In this work, we try to incorporate the added degrees of freedom and information obtained from the credal partition, in the rule based classification systems.

4 Rule Based Classification

4.1 Model Structure

A fuzzy classification system consists of a set of fuzzy IF-THEN rules combined with a fuzzy inference mechanism. This type of rules can be viewed as an extension of the Takagi-Sugeno fuzzy model [16], and it can be described by N rules of the following type [17]:

$$R_k^q : \text{ If } x_1 \text{ is } F_{k1} \text{ and } \ldots \text{ and } x_n \text{ is } F_{kn} \text{then } d_q(\mathbf{x}) = g_k^q(\mathbf{x}) \; , \tag{10}$$

where $g_k, k = 1, 2, \ldots, N$ is the consequent function for rule R_k, and $d_q(\mathbf{x})$ is a discriminant function associated with each class $\omega_q, q = 1, \ldots, Q$. Note that the

index q indicates that the rule is associated with the output of class q. The output of the classifier assigns the class label corresponding to the maximum value of the discriminant functions. The antecedent parts of the rules are the same for different discriminants, but the consequents may be different. The output of each discriminant function $d_q(\mathbf{x})$ can be interpreted as a score for the associated class q given the input feature vector $\mathbf{x}$. The degree of fulfillment $\beta_i(x)$ of the ith rule, is computed using the intersection operator in the cartesian product space of the antecedent variables as $\beta_i(x) = \mu_{F_{i1}}(x_1) \wedge \mu_{F_{i2}}(x_2) \wedge \ldots \wedge \mu_{F_{ip}}(x_p)$. Other t-norms, such as the product, can be used instead of the minimum operator.

4.2 Model Parameters

To form the fuzzy system model from the data set with n data samples, given by $X = [x_1, x_2, \ldots, x_n]^T$, $Y = [y_1, y_2, \ldots, y_n]^T$ where each data sample has a dimension of p ($n \gg p$), the structure is first determined and afterwards the parameters of the structure are identified. The number of rules characterizes the structure of a fuzzy system and in our case corresponds to the number of partitions obtained from the clustering algorithm.

In this work, we use ECM to partition the space using the framework of belief function and map the obtained credal partition as a fuzzy set. When clustering the object data with ECM, several clustering structures can be obtained, as explained in Section 3. In this work we focus on the partitioning structure obtained from the credal partition, which we expect to better describe the data and its underlying structure.

Using c clusters, the credal partition obtained from ECM partitions the space in at most 2^c intervals, with a center associated with each interval. In contrast, using the well-known FCM [18], the space is partition in, at most, c intervals. The added information from the credal partition allows to reveal objects that unambiguously belong to a given cluster and the set of objects that lie at the boundaries of each cluster. Since the values of the credal partition m are in $[0,1]^p$, this value can be perceived as an assignment to each subset of the partition. Thus we can obtain the following mapping $\varphi : m \in [0,1]^p \mapsto A_{ij}$, using one the following functions:

$$\varphi_1 : \mu_{F_{ij}}(x_{jk}) = \mathrm{proj}_j(m_{ik}) \tag{11a}$$

$$\varphi_2 : \mu_{F_{ij}}(x_{jk}) = X(Aj) \tag{11b}$$

$$\varphi_3 : \mathrm{supp}(F) = \omega_k^L,\ \mathrm{core}(F) = \omega_k^U \tag{11c}$$

where $\mathrm{supp}(F) = \{x \in X | \mu_F(x) > 0\}$, $\mathrm{core}(F) = \{x \in X | \mu_F(x) = 1\}$ and proj_j is a pointwise projection of the partition matrix M onto the axes of the antecedent variables x_j.

The obtained point-wise fuzzy sets F_{ij} can now be approximated by appropriate parametric functions, such as Gaussian functions, resulting in the antecedent membership functions. Although, the obtained antecedent membership functions from each one of the mappings presented in (11) will be different from one another, there are situations where they can be very similar. For example, the

mappings (11b) and (11c) will be equivalent for cases where the obtained credal partition does not indicate that some points lie in the boundaries of two or more clusters. The number of obtained rules using the proposed method is, at most 2^c-1, as we exclude, $m(\emptyset)$. The consequent parameters for each rule are obtained by means of linear least square estimation, which concludes the identification of the classification system.

5 Experimental Setup

We used several data sets to test the proposed modelling approach. The Wisconsin Breast Cancer (WBC) database [19] is composed of 699 objects, 10 features and has missing values and an uneven distribution of classes [20]. Also, two databases related with bankruptcy prediction were used. The Altman data set [21] has 66 objects and 5 features, and the CR data set [22] which contains extreme values, missing values and a very skewed class distribution. The CR data set is composed of 1817 companies and has 51 features.

In our experiments we used minimum amount of improvement $\varepsilon = 0.0001$, maximum number of iterations 100, $\beta = 3$ for both ECM and FCM, $\alpha = 1$ and $\delta = 2000$. The initialization for the cluster prototypes centers in ECM were obtained with FCM, as suggested in [6]. Since all cases are binary classification, two clusters were used for all models. In [6] an index for choosing the number of clusters to be used with ECM is discussed. All trials terminated with the convergence criteria after a few iterations.

In all tests we used a simple holdout method for validation. 50 trials were made and only the results obtained with the testing set are reported below. The overall performance of the models is measured by the classification accuracy.

As already stated, the Wisconsin Breast Cancer and the CR database have missing values. It was considered that these missing values are missing completely at random (MCAR) [23], and thus imputation of values is the usual course of action. The missing values were inferred using the expectation-maximization (EM) algorithm, as explained in [22].

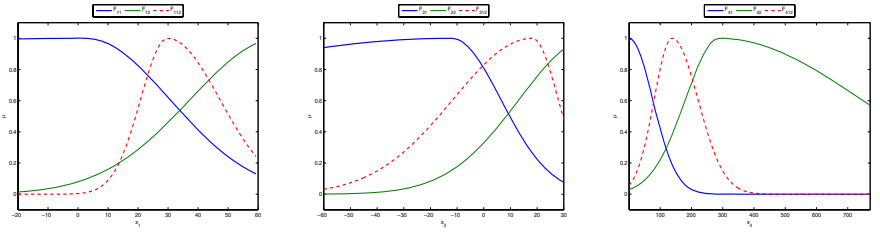
6 Examples

Table 1 exhibits the obtained mean accuracy and respective variance, for each class and the global accuracy of different fuzzy models, for all the data sets under study. These models were obtained using the three methods explained in (11) for mapping the antecedent fuzzy sets. Also, for comparison purposes, the results obtained for a fuzzy rule-based model that uses the FCM to derive the antecedent fuzzy sets is shown.

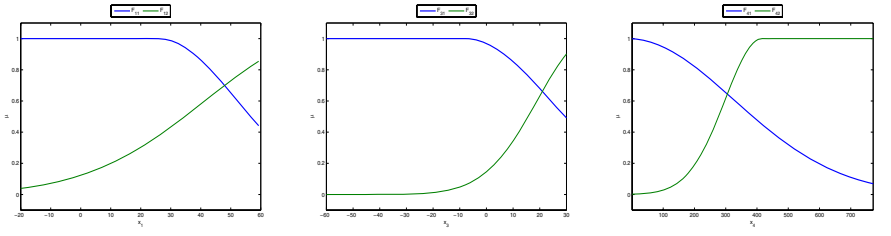
The obtained global accuracy of the model obtained using ECM, is comparable to the results obtained using FCM algorithm. Depending on the type of mapping used, different accuracies for each class are obtained. For the case of the CR database, the mapping φ_1 and φ_2 give better results in terms of the class Bankrupt, which is arguably, the more relevant class in study. In general,

Table 1. Classification accuracy for the databases in study- mean (variance)

Database	Method	φ_1	φ_2	φ_3	FCM
WBC	Class 0	0.978 (0.016)	0.971 (0.014)	0.976 (0.014)	0.972 (0.018)
	Class 1	0.936 (0.032)	0.961 (0.023)	0.946 (0.026)	0.956 (0.027)
	Global	0.964 (0.016)	0.968 (0.013)	0.966 (0.013)	0.966 (0.013)
Altman	Bankrupt	0.960 (0.067)	0.990 (0.032)	1.000 (0.000)	0.980 (0.042)
	Not Bankrupt	1.000 (0.189)	0.730 (0.048)	0.620 (0.079)	0.600 (0.094)
	Global	0.980 (0.111)	0.860 (0.021)	0.810 (0.039)	0.790 (0.046)
CR	Bankrupt	0.465 (0.068)	0.478 (0.105)	0.326 (0.132)	0.361 (0.082)
	Not Bankrupt	0.988 (0.008)	0.985 (0.003)	0.989 (0.005)	0.989 (0.005)
	Global	0.966 (0.009)	0.964 (0.006)	0.961 (0.006)	0.962 (0.005)



(a) ECM-feature 1 *(b) ECM-feature 3* *(c) ECM-feature 4*



(d) FCM-feature 1 *(e) FCM-feature 3* *(f) FCM-feature 4*

Fig. 1. Antecedent Membership Functions for ECM with mapping φ_2 and FCM

the mapping φ_1 and φ_2 give better results. Also note that using the mapping φ_3, only two rules are obtained, while for the other mappings, three rules are obtained.

Figure 1 shows the obtained antecedent membership functions using ECM and mapping φ_2, as well as using FCM, for three features of the Altman database. It is interesting to note that although only two clusters are used, three membership functions are obtained for ECM, corresponding to the focal sets $\{\omega_1, \omega_2, \Omega\}$. It is our opinion that this conveys more information than using, for instance, the FCM algorithm. Linguistic terms can easily be assigned on the cases of for

the membership functions obtained with ECM, e.g. *High* for F_{k1}, *Low* for F_{k2} and *Medium* for F_{k12}. This proposed method seems to provide more information regarding the underlying structure of the model of the model, using the same number of clusters. Specifically for the case of bankruptcy prediction it is very appealing to be able to have a model that identifies the cases which are in between classes, as in general, these cases are the ones that require further assessment [22].

7 Discussion

During the experiments performed in all data sets, it was noticed that if the parameters of ECM were not chosen carefully, no object would be assigned as belonging to the boundaries of two or more clusters. This would result in results that are not as good as the ones obtained with FCM. Furthermore, it was noticed that ECM works very good in databases that contain many points in between clusters, i.e. noisy data. This is due to the fact that the credal partition gives relevant information about the points that unambiguously belong to one cluster, and the points that lie at the boundary of two or more clusters. Since these points can be seen as noise, FCM may have problems clustering them.

The proposed method seems to provide more information regarding the underlying structure of the model, using the same number of clusters. In the case of the (noisy) CR database, better results were obtained with the use of ECM, specially for the bankruptcy class while the overall accuracy remains comparable. In the general case of bankruptcy prediction it is very appealing to derive a model that identifies the cases which are in between classes, as in general, these cases are the ones that require further assessment [22]. More research is needed into transforming the information about more difficult cases (i.e. boundary cases) in specific rules.

The proposed models, obtained by extracting rules from the credal partition obtained with ECM, are computationally only slightly more complex compared to models using FCM, because a small number of clusters is used. This increase is due to the fact that in the case of FCM c rules are derived, while in the case of ECM, at most 2^c rules are derived, for the cases of the mapping φ_1 and φ_2. Note that the added rules refer to the points that lie at the boundary of two or more clusters. The assignment of linguistic terms to the obtained credal partitions, can be seen as an quantification of the degree of belief, and can be easily done by inspecting the obtained membership functions. It was noted that in some cases, due to the overlap of membership functions, the linguistic interpretation may be difficult, and further optimization of the models is required [24].

8 Conclusions

This paper discusses the use of the credal partition obtained from the Evidential C-Means based on the theoretical framework of belief functions, in deriving rule based classification models. We compare the performance of the proposed

methodology with the models obtained by using the fuzzy partition matrix derived from the well know Fuzzy C-Means algorithm. The results with the proposed methodology are similar to the results obtained using FCM, but have the advantage that the use of ECM seems to provide more information about the system, which is successfully translated into rules. However, more future research is needed to assess all the characteristics of the proposed method. We will also concentrate on comparing the partitions obtained with ECM with the partitions obtained from FCM using 2^c clusters.

Acknowledgments. The authors thank the anonymous reviewers for their valuable comments. This work was (partially) supported by the European Science Foundation through COST Action IC0702.

References

1. Wang, L., Mendel, J.: Generating fuzzy rules by learning from examples. IEEE Transactions on Systems, Man and Cybernetics 22(6), 1414–1427 (1992)
2. Wu, T., Chen, S.: A new method for constructing membership functions and fuzzy rules from training examples. IEEE Transactions on Systems, Man and Cybernetics, Part B 29(1), 1083–4419 (1999)
3. Rojas, I., Pomares, H., Ortega, J., Prieto, A.: Self-organized fuzzy system generation from training examples. IEEE Transactions on Fuzzy Systems 8(1), 23–36 (2000)
4. Hartigan, J.A.: Clustering Algorithms. Wiley-Interscience, New York (1974)
5. Elouedi, Z., Mellouli, K., Smets, P.: Classification with belief decision trees. In: AIMSA 2000: Proceedings of the 9th International Conference on Artificial Intelligence, London, UK, pp. 80–90. Springer, Heidelberg (2000)
6. Masson, M.H., Denoeux, T.: ECM: An evidential version of the fuzzy c-means algorithm. Pattern Recognition 41(4), 1384–1397 (2008)
7. Masson, M.H., Denux, T.: RECM: Relational evidential c-means algorithm. Pattern Recogn. Lett. 30(11), 1015–1026 (2009)
8. Denoeux, T., Masson, M.H.: EVCLUS: evidential clustering of proximity data. IEEE Transactions on Systems, Man, and Cybernetics, Part B 34(1), 95–109 (2004)
9. Elouedi, Z., Mellouli, K., Smets, P.: Belief decision trees: theoretical foundations. Int. J. Approximate Reasoning 28, 91–124 (2001)
10. Smets, P., Kennes, R.: The transferable belief model. Artif. Intell. 66(2), 191–234 (1994)
11. Smets, P.: The transferable belief model and possibility theory. In: NAFIPS 1990, pp. 215–218 (1990)
12. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton (1976)
13. Smets, P.: The transferable belief model for quantified belief representation. In: Gabbay, D., Smets, P. (eds.) Handbook of Defeasible Reasoning and Uncertainty Management Systems. Kluwer, Dordrecht (1998)
14. Smets, P.: The combination of evidence in the transferable belief model. IEEE Trans. Pattern Anal. Mach. Intell. 12(5), 447–458 (1990)
15. Smets, P.: Decision making in the TBM: the necessity of the pignistic transformation. Int. J. Approx. Reasoning 38(2), 133–147 (2005)

16. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its aplication to modeling and control. IEEE Transactions on Systems, Man and Cybernetics 15(1), 116–132 (1985)
17. van den Berg, J., Kaymak, U., van den Bergh, W.M.: Fuzzy classification using probability-based rule weighting. In: Proceedings of 2002 IEEE International Conference on Fuzzy Systems, Honolulu, Hawaii, May 2002, pp. 991–996 (2002)
18. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
19. Mangasarian, O.L., Wolberg, W.H.: Cancer diagnosis via linear programming. SIAM News 23(5), 1–18 (1990)
20. Murphy, P.M.: UCI repository of machine learning databases and domain theories (1985), http://www.ics.uci.edu/-mlearn/MLRepository.html http://www.ftp.ics.uci.edu:pub/machine-learning-databases
21. Altman, E.I.: Financial ratios, discriminate analysis and the prediction of corporate bankruptcy. Journal of Finance 23(4), 589–609 (1968)
22. Almeida, R.J., Vieira, S., Milea, V., Kaymak, U., Sousa, J.M.C.: Knowledge discovery in the prediction of bankruptcy. In: IFSA World Congress / EUSFLAT Conference (IFSA-EUSFLAT 2009), Lisbon, Portugal, July 2009, pp. 1785–1790 (2009)
23. Timm, H., Döring, C., Kruse, R.: Differentiated treatment of missing values in fuzzy clustering. In: De Baets, B., Kaynak, O., Bilgiç, T. (eds.) IFSA 2003. LNCS, vol. 2715, pp. 354–361. Springer, Heidelberg (2003)
24. Setnes, M., Roubos, H.: GA fuzzy modeling and classification: Complexity and performance. IEEE Transactions on Fuzzy Systems 8(5), 509–522 (2000)

Measuring Impact of Diversity of Classifiers on the Accuracy of Evidential Ensemble Classifiers

Yaxin Bi and Shengli Wu

School of Computing and Mathematics, University of Ulster
Newtownabbey, Co. Antrim, BT37 0QB, UK
{y.bi,s.wu1}@ulster.ac.uk

Abstract. Diversity being inherent in classifiers is widely acknowledged as an important issue in constructing successful classifier ensembles. Although many statistics have been employed to measure diversity among classifiers to determine whether it correlates with ensemble performance in the literature, most these measures are incorporated and explained in the non-evidential context. In this paper, we first introduce a modelling for formulating classifier outputs as triplet mass functions and an unform notation for defining diversity measures, we then present our studies on the relationship between diversity obtained by four pairwise and non-pairwise diversity measures and accuracy of classifiers combined in different orders in the framework of belief functions. Our experimental results demonstrate that the negative correlation between the diversity and accuracy is stronger than the positive one, which is not in favor of the claim that increasing diversity could lead to reduction of generalization error of classifier ensembles.

Keywords: diversity, ensemble learning, belief functions and classifier combination.

1 Introduction

The combination of multiple classifiers/ensemble approach is rather powerful decision making and classification techniques that have been used successfully for modelling many practical problems. In the modelling of classifiers combination, many researchers believe the diversity being inherent in the classifiers plays an important role in constructing successful ensemble classifiers. Unfortunately to date there exists no general accepted theoretical framework underpinning the development of methods for capturing diversity among ensemble classifiers. Although many statistics have been employed to measure diversity with the intension to determine whether it correlates with ensemble performance in the literature, results are varied. Most commonly these measures are incorporated and explained in the context of majority voting, linear sum and non-evidential framework [1]. Presently there is little effort concerning how diversity measured by statistics imparts ensemble performance in the framework of the Dempster-Shafer (DS) theory of evidence [2], where classifier outputs are modeled as pieces

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 238–247, 2010.

of evidence that are thereby combined by Dempster's rule of combination. In this paper, we present our studies on measuring diversity among classifiers and then experimentally examine the relationship between diversity obtained by four pairwise and non-pairwise diversity measures and accuracy of combined classifiers by Dempster's rule of combination.

Early studies on the relationship between diversity and ensemble performance have stimulated considerable interest and they can be categorized into the contexts of regression and classification. In the context of classification, Kuncheva et al. carried out an experimental study on the relationship between diversity and accuracy [1]. Their results show although there are proven connections between diversity and accuracy in some special cases, there is no strong linear and non-linear correlation between diversity and accuracy. In [3], Tang, et al. conducted a follow-up comprehensive study. They investigate the correlation among the six statistical measures used in [1] and relate these measures to the concept of margin proposed in [4], which is explained a key factor to the success of Boosting algorithms. The experimental results indicate that large diversity may not consistently correspond to better ensemble performance and the information available from varying diversity cannot provide a consistent guidance for making an ensemble of classifiers to achieve good generalization performance.

In the previous study[5], we have developed new evidence structures called a *triplet* and *quartet* and a formalism for modelling classifier outputs as triplet and quartet mass functions, and we also established a range of formulae for combining these mass functions in order to arrive at a consensus decision. However in that study we did not address the issues of how diversity impacts the performance of classifiers combined by Dempster's rule of combination. In this work, we carry out an analysis of effects of diversity on the quality of ensemble classifiers that are independently generated by 13 machine learning algorithms and are combined using a purely evidential combination function in decreasing and mixed orders. We use the triplet as an underlying evidence structure for representing classifier outputs and study the correlation between diversity and accuracy by the Spearmans rank correlation method over 9 benchmark data sets. The experimental results demonstrate that the positive correlation between the diversity and accuracy in decreasing order is weaker than that in mixed order, which could not confirm the role of diversity in reducing general errors of classifier ensembles. Our results conjectures that the order of combining classifiers could be regarded as a useful factor in constructing successful classifier ensembles.

2 Representation of Classifier Outputs

In ensemble approaches, a learning algorithm is provided with a training data set made up of $D \times C = \{\langle d_1, c_1\rangle, \cdots, \langle d_{|D|}, c_q\rangle\}$ $(1 \leq q \leq |C|)$ for deriving some unknown function f such that $f(d) = c$. D is a set of instances in the form of $(d_{i_1}, \cdots, d_{i_n})$ where d_{i_j} is either a nominal or ordinal value. C is set of class labels c_i. Given a set of training data $D \times C$, a learning algorithm is aimed at learning a function φ in terms of classifier from the training data set, where classifier φ is an approximation to an unknown function f.

Given a new instance d, a classification task is to make decision for d using φ about whether d belongs to class c_i. Instead of single-class assignment, we regard such a classifying process as a mapping:

$$\varphi : D \rightarrow C \times [0,\ 1] \tag{1}$$

where $C \times [0,\ 1] = \{(c_i,\ s_i) \mid c_i \in C,\ 0 \leq s_i \leq 1\}$, s_i is a numeric value. The greater the value of class s_i, the greater the amount of belief given to the proposition of instance d belonging to c_i. Simply we denote a classifier output by $\varphi(d) = \{s_1, \cdots, s_{|C|}\}$. Given an ensemble of classifiers, $\varphi_1, \varphi_2, \cdots, \varphi_M$, all the classifier outputs on instance d can be organized into a matrix as illustrated in formula (2).

$$\begin{pmatrix} \varphi_1(d) \\ \varphi_2(d) \\ \vdots \\ \varphi_M(d) \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1|C|} \\ s_{21} & s_{22} & \cdots & s_{2|C|} \\ \vdots & \vdots & \cdots & \vdots \\ s_{M1} & s_{M2} & \cdots & s_{M|C|} \end{pmatrix} \tag{2}$$

3 Basics of the Dempster-Shafer (DS) Theory of Evidence

The DS theory of evidence remedies the limitations of the traditional Bayesian belief model to allow the explicit representation of uncertainty and management of conflict information involved in the decision making process [2]. It formulates a proposition (alternative) set as a frame of discernment, denoted by Ω and its power set 2^{Ω} is all the subsets of Ω.

Definition 1. Let Ω be a frame of discernment. Let m be a mass function, which is defined as a assignment function assigning a numeric value in $[0, 1]$ to $X \in 2^{\Omega}$ with two conditions below.

$$1)\ m(\emptyset) = 0, \quad 2)\ \sum_{X \subseteq \Omega} m(X) = 1$$

where X is called a focal element, focus or singleton if X is one element subset with $m(X) > 0$. Given the general representation of classifier outputs in formula (2) and Definition 1, we define an application-specific mass function below.

Definition 2. Let $\Omega = \{c_1, \ldots, c_{|\Omega|}\}$ be a frame of discernment and let $\varphi(d)$ be a list of scores, an application-specific mass function is defined a mapping function, $m : 2^{\Omega} \rightarrow [0, 1]$ as follows:

$$m(\{c_i\}) = \frac{s_i}{\sum_{j=1}^{|\Omega|} s_j} \tag{3}$$

where $1 \leq i \leq |\Omega|$.

This mass function expresses the degrees of belief with respect to determining class labels to which a given instance could be assigned.

Definition 3. Let Ω be a frame of discernment. Let m_1 and m_2 be two mass functions defined for $X, Y \subseteq \Omega$. Dempster's rule of combination (or Dempster's rule) is, denoted by $\oplus$, defined as

$$(m_1 \oplus m_2)(A) = \frac{\sum_{X \cap Y = A} m_1(X) m_2(Y)}{\sum_{X \cap Y \neq \emptyset} m_1(X) m_2(Y)} \tag{4}$$

where operator $\oplus$ is also called the *orthogonal sum*. $N = \sum_{X \cap Y \neq \emptyset} m_1(X) m_2(Y)$ is the *normalization constant*. $E = 1 - N$ is called the *conflict factor*. This rule strongly emphasizes the agreement between multiple independent sources and ignores all the conflicting evidence through a normalization factor.

Given the formulation of classifier outputs in formula (2), by formula (3), we can rewrite $\varphi(d)$ as $\varphi(d) = \{m(\{c_1\}), m(\{c_2\}), \cdots, m(\{c_{|C|}\})\}$. Based on the triplet structure in [5], we can further represent $\varphi(d)$ by a triplet structure as $\varphi(d) = \langle \{u\}, \{v\}, C \rangle$ and then rewrite formula (2) as (5) below,

$$\begin{pmatrix} \varphi_1(d) \\ \varphi_2(d) \\ \vdots \\ \varphi_M(d) \end{pmatrix} = \begin{pmatrix} m_1(\{u_1\}) & m_1(\{v_1\}) & m_1(C) \\ m_2(\{u_2\}) & m_2(\{v_2\}) & m_2(C) \\ \vdots & \vdots & \vdots \\ m_M(\{u_M\}) & m_M(\{v_M\}) & m_M(C) \end{pmatrix} \tag{5}$$

More details about the triplet formulation and calculation can be found in [5].

4 Diversity Measures

This section defines statistical diversity measures in an uniform way. Formally suppose we are given M classifiers denoted by $\varphi_1, \cdots, \varphi_M$, a set of classes $C = \{c_1, \cdots, c_{|C|}\}$ and a test set $T = \{x_1, \cdots, x_{|T|}\}$. For $x \in T$, each classifier produces an output vector $\varphi_i(x)$, which is modeled as a binary output, i.e. $\varphi_i(x) = 1$ if φ_i correctly classifies x, $\varphi_i(x) = 0$ if φ incorrectly classifies x. We also denote $\widehat{\varphi}(x) = \{\varphi(x) | \varphi_i(x) = 1, 1 \leq i \leq M, x \in T\}$. With this notation, we define four statistical diversity measures below.

4.1 Kappa (κ) Statistic

The κ statistic is the most widely used pairwise method to measure the level of agreement between classifiers [7]. Given two classifiers φ_i and φ_j and a test data set T, we can construct a global contingency table where entry $n(c_h, c_k)$ contains the number of instances $x \in T$ for $\varphi_i(x) = c_h$ and $\varphi_j(x) = c_k$. If φ_i and φ_j are identical on the data set, then all non-zero counts will appear along the diagonal of the table, otherwise there will be a number of counts off the diagonal. Now we define

$$\kappa_{i,j} = \frac{\mu_1 - \mu_2}{1 - \mu_2} \tag{6}$$

where μ_1 is the probability that two classifiers agree and μ_2 is a correction term for μ_1, estimating the probability that the two classifiers agree simply by chance.

The pairwise κ statistic over the whole set of classifiers over T is then defined as follows:

$$\kappa = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \kappa_{i,j} \tag{7}$$

$\kappa = 0$ when the agreement of two classifiers equals that expected by chance, $\kappa = 1$ when two classifiers agree on all the testing instances, and negative values of κ mean that an agreement is less than expected by chance.

4.2 Disagreement Measure

The disagreement measure is used to characterize the diversity between a classifier and its complementary classifier [1] [3]. Formally let $n(a, b)$ be the number of which the binary outputs of φ_i and φ_j are a and b on test instances, where a and b take only two values on 1 and 0, respectively. The disagreement between two classifiers is measured by:

$$dis_{i,j} = \frac{n(0,1) + n(1,0)}{n(0,0) + n(0,1) + n(1,0) + n(1,1)} \tag{8}$$

The pairwise disagreement diversity among the whole set of classifiers over T is then defined as an average over all pairs of classifiers below:

$$dis = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} dis_{i,j} \tag{9}$$

The diversity increases as values of the disagreement measure increase.

4.3 Q-Statistic

The Q-statistic (qs) is a well studied measure in statistics [1] [3]. The disagreement between two classifiers over a testing set T is measured by:

$$Q_{i,j} = \frac{n(0,0)n(1,1) - n(1,0)n(0,1)}{n(0,0)n(1,1) + n(1,0)n(0,1)} \tag{10}$$

The definition of $n(a, b)$ is the same as in Section 4.2. For two classifiers φ_i and φ_j, $Q_{i,j}$ is a measurement of diversity between φ_i and φ_j. When $Q_{i,j} = 1$ indicates that all the class labels assigned by φ_i for instances $x \in T$ are exactly the same as ones assigned by φ_j. $Q_{i,j} = -1$ means that all the class labels recognized by φ_i for instances $x \in T$ are entirely different from those that φ_j

recognizes. For a set of classifiers, the averaged $Q_{i,j}$ statistics for all the pairs of classifiers over T is measured by

$$qs = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} Q_{i,j} \tag{11}$$

4.4 Kohavi-Wolpert Variance

In [6], Kohavi and Wolpert proposed a formula for representing the classification errors of classifiers. This formula is built on the basis of the bias-variance decomposition of errors of classifiers. The expression of the variability of a predicted class label $c \in C$ for an instance $x \in T$ is

$$variance_x = \frac{1}{2}(1 - \sum_{i=1}^{|C|} P(c = c_i|x)^2) \tag{12}$$

Averaging *variance* over the whole set of test data T, we have a revised measure, denoted by kw, which can be used to measure the diversity among the whole set of classifiers:

$$kw = \frac{1}{|T|M^2}(\sum_{i=1}^{|T|} |\widehat{\varphi}(x_i)|(M - |\widehat{\varphi}(x_i)|)) \tag{13}$$

The diversity increases as the kw variance increases.

5 Experimental Evaluation

5.1 Experimental Settings

In our experiments, we used nine data sets downloaded from the UCI machine learning repository, including *anneal*, *audiology*, *balance car*, *glass*, *autos*, *segment*, *soybean* and *wine*. All the selected data sets have at least three or more classes as required by the triplet structure.

For base (individual) classifiers, we used thirteen learning algorithms which all are taken from the Waikato Environment for Knowledge Analysis (Weka) version 3.4, including *AOD*, *NaiveBayes*, *SOM*, *IB1*, *IBk*, *KStar*, *DecisionStump*, *J48*, *RandomForest*, *DecisionTable*, *JRip*, *NNge*, and *PART*. These algorithms were simply chosen on the basis of their performance in three randomly picked data sets. Parameters used for each algorithm was at the Weka default settings described in[8].

For the combination of classifiers in decreasing order, we first rank all the 13 classifiers, and then we combine the best with the second best, denoted by 2c, and combine the combined result of the 2 classifiers with the third best, denoted by 3c, and so forth, until combine the combined result of the 12 classifiers with the 13th classifiers, denoted by 13c as delineated in Fig. 1. With respect to the

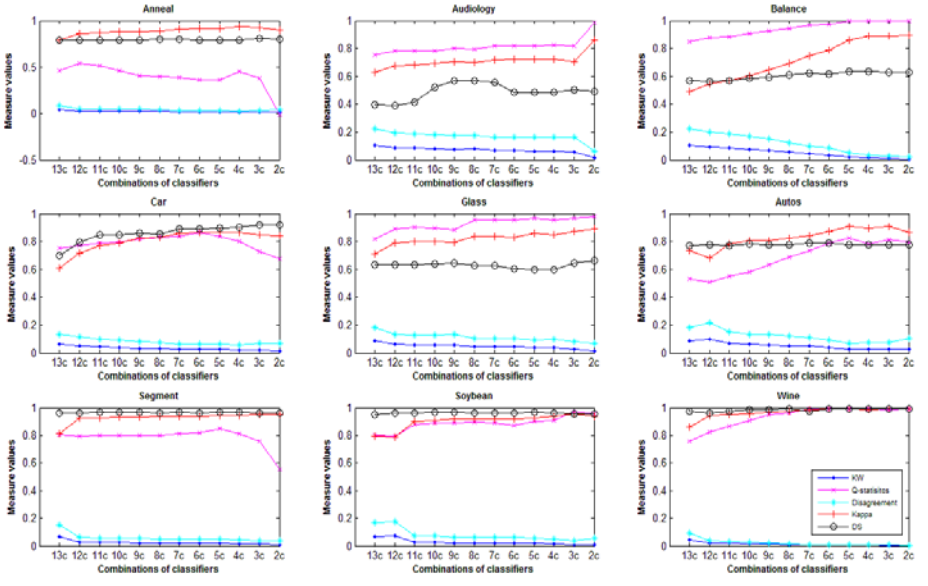


Fig. 1. Diversity and accuracy of the corresponding combinations of 13 classifiers over the 9 data sets in decreasing order

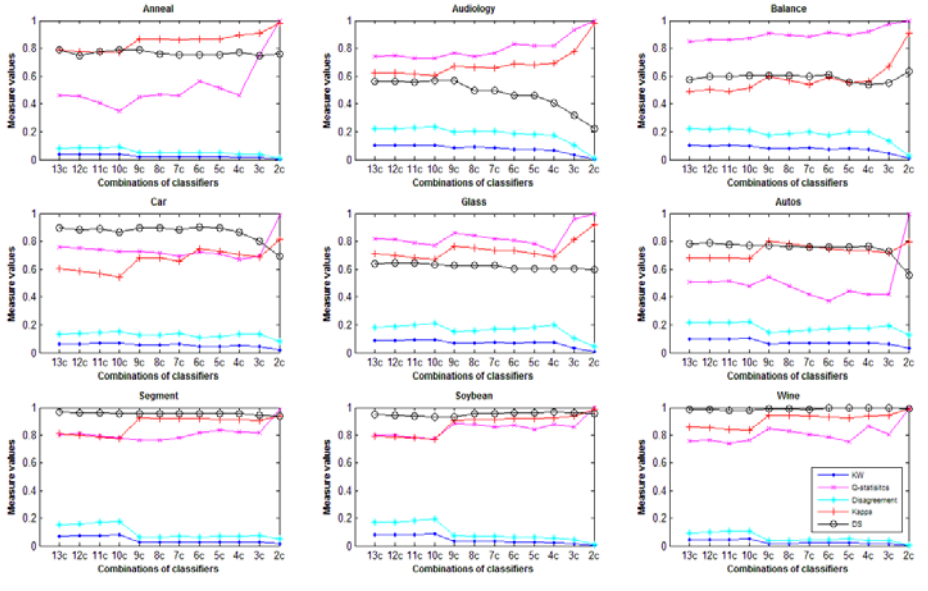


Fig. 2. Diversity and accuracy of the corresponding combinations of 13 classifiers over the 9 data sets in mixed order

Table 1. Correlation between diversity and combined accuracies of individual classifiers using Dempster's rule in decreasing (left) and mixed (right) orders

Dataset	kw	qs	dis	κ	kw	qs	dis	κ
anneal	-0.3893	-0.5307	-0.2593	0.2705	0.4684	-0.4861	0.4442	-0.4444
audiology	-0.3332	0.2215	-0.3321	0.3303	**0.9818**	**-0.9682**	**0.9573**	**-0.9311**
balance	**-0.9731**	**0.9858**	**-0.9720**	**0.9768**	-0.1789	0.0420	-0.3588	0.352
car	**-0.9530**	0.0664	**-0.9330**	**0.9593**	**0.8003**	**-0.7737**	**0.6616**	-0.5170
glass	0.0253	-0.2486	0.0793	-0.0797	**0.6870**	-0.3240	0.5365	-0.5362
autos	-0.3120	0.4156	-0.3657	0.3447	**0.7708**	**-0.8755**	**0.5998**	-0.4814
segment	0.0975	**0.5778**	0.0312	-0.0309	**0.6408**	**-0.7968**	0.5554	-0.5566
soybean	-0.4667	0.2020	-0.4647	0.4360	**-0.6776**	0.4193	**-0.6621**	**0.6491**
wine	**-0.6079**	**0.7458**	**-0.5881**	**0.5899**	**-0.7616**	0.3411	**-0.7278**	**0.7311**
Av	-0.4347	0.2706	-0.4227	0.4219	0.3036	-0.3803	0.2230	-0.1928
Abs(Av)	0.4620	0.4438	0.4473	0.4464	0.6632	0.5586	0.6116	0.5778

combination of classifiers in mixed order, the order of classifiers is random. We first pick up two classifiers to combine, specified by $2c$, and then combine the combined result of the 2 classifiers with the third classifier that is randomly chosen, denoted by $3c$, until combine the previous result with the last classifiers, denoted by $13c$ as depicted in Fig. 2.

To assess how the diversity of 12 groups of classifiers and the corresponding ensemble accuracy is actually correlated, we carried out a correlation analysis on each pair of the groups of the classifier diversity and the combined accuracy over the nine data sets, resulting in 12 pairs of the correlation coefficient $r \in [-1, 1]$ and $p\text{-}value \in [0, 1]$. A positive r indicates a positive correlation between diversity and accuracy, whereas a negative r indicates a negative correlation between them. Here negative correlation indicates that one increases while the other decreases in values. The closer the value of is to 0, the smaller the correlation. *p-value* indicates the degree of that the correlation is statistically significant.

5.2 Experimental Results

In this experimental study, we carried out three groups of experiments, including the combinations of different groups of classifiers using Dempter's rule of combination in decreasing and mixed orders; the diversity being inherent in the different groups of classifiers in decreasing and mixed orders; and the relationship between the diversity and accuracy. For the sake of comparison, we place the results of the diversity and accuracy in the same order into the same figures and present them in Figs. 1 and 2, respectively.

From these figures, we can observe that the curves of the combined accuracy, which are marked by "DS", are slightly different. For decreasing order, the smaller the number of the combinations of classifiers, the better the combined performance of classifiers, and they converge to the combination of the best with the second best. While the combinations of classifiers in mixed order appears to be opposite, the larger the number of the combinations of classifiers, the better the

combined performance of classifiers. However this phenomenon is not apparent when the combined accuracy exceeds 85%.

In Fig.1 according to the behaviors of the curves and the nature of the four diversity measures, the diversity curves of 12 groups of classifiers from $2c$ to $13c$ over the 9 data sets can be characterized into two groups: one is measured by qs and κ, and the other is measured by kw and dis. It can be observed that the fitness between kw and dis is better than that between qs and κ, and the curve margins between qs and κ are larger than those between kw and dis. Roughly speaking, for the former group the curves decrease as more classifiers are added, whereas for the latter group, the curves increase with the addition of more classifiers, i.e. both of them go towards the closer from $2c$ to $13c$. These results suggest that the order of classifiers has an impact on the agreement among the classifiers − the more classifiers are added into the groups of classifiers, the more diversity appears among the groups.

Inspecting Fig.2 the diversity curves of 12 groups of classifiers can be similarly divided into two groups. It is clear to see that when more classifiers are added into the groups of classifiers, the first group of the curves roughly decreases, while the second group increases but both of them have some fluctuations. There is no consistent trend that can visually be identified. For example, there are sharp changes between $9c$ and $10c$ in the cases of *glass*, *autos*, *segment*, *soybean*, *wine*, the later three seems to be correlated to the combined accuracy that is over 85%, but the first two cases could not confirm this. In addition, we notice that when the accuracy of the combined classifiers is less 65%, the curves of the diversity measured by qs are above those obtained by κ in the cases of *audiology*, *balance*, *car*, *glass*, which is similar to that in decreasing order. This result could indicate that qs is not as effective as the other three measures in measuring diversity among the groups of classifiers when the combined accuracy of the groups of classifiers is less than 65%.

Table 1 presents a correlation analysis on the relationship between the ensemble accuracy and the diversity by using the Spearmans rank method. The left group is the results of the decreasing order and the right one is the results of the mixed order. Based on the properties of the diversity measures, when the qs and κ coefficients are positive and kw and dis are negative, they represent a positive correlation between the diversity and accuracy, otherwise they express a negative correlation. From the decreasing order results, we can find that the diversity obtained by the four diversity measures are not very strong since the correlation coefficients only over 3 of the 9 data sets is statistically significant ($p \leq 0.05$), which are shown in *bold*, and the average coefficients and their absolute values are lower than the critical value 0.577 (making $p \leq 0.05$). From the mixed order results, it can be observed that the negative correlation between the diversity and accuracy appear to be very strong because the the kw coefficients over 7 of the 9 data sets is statistically significant ($p \leq 0.05$) and for dis there are 5 of the 9 data sets being statistically significant. The strong correlation reveals the fact that the larger diversity measured by kw and dis corresponds to the poor performance of the combined classifiers, which intuitively complies with the

results depicted in Fig. 2. Therefore both of the results are consistent with the findings reported in [1] [3] that increasing diversity might not consistently lead to reduction of generalization error of classifier ensembles.

6 Summary and Future Work

In this study we report a range of experiments on 9 benchmark data sets with the different groups of classifiers generated by 13 machine learning algorithms and combined by Dempster's rule of combination in different orders. In order to quantify the relationship between the diversity and accuracy, the correlation coefficients with respect to the accuracy and each of the four diversity results over the different classifier combinations were calculated. The experimental results reveal the fact that the strong negative correlation in mixed order does not favorites the claim that increasing diversity would lead to reduction of generalization error of classifier ensembles. This fact is, however, not fully supported by the weaker correlation in decreasing order that warrants a further investigation. Meanwhile our studies could postulate that the decreasing order could be a better way to combine classifiers. To solid these findings, we are carrying out a further experimental study.

References

1. Kuncheva, L., Whitaker, C.J.: Measures of diversity in classifier ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning 51, 181–207 (2003)
2. Shafer, G.: A Mathematical theory of evidence. Princeton Univ. Press, Princeton (1976)
3. Tang, E.K., Suganthan, P.N., Yao, X.: An analysis of diversity measures. Machine Learning 65(1), 247–271 (2006)
4. Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S.: Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics 26, 1651–1686 (1998)
5. Bi, Y., Guan, J., Bell, D.: The Combination of Multiple Classifiers Using an Evidential Approach. Artificial Intelligence 17, 1731–1751 (2008)
6. Kohavi, R., Wolpert, D.: Bias plus variance decomposition for zero-one loss functions. In: Saitta, L. (ed.) Machine Learning: Proc. 13th International Conference, pp. 275–283. Morgan Kaufmann, San Francisco (1996)
7. Fleiss, J.L., Cuzick, J.: The reliability of dichotomous judgments: unequal numbers of judgments per subject. Applied Psychological Measurement 3, 537–542 (1979)
8. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

Multiplication of Multinomial Subjective Opinions

Audun Jøsang[1] and Stephen O'Hara[2]

[1] UNIK Graduate Center
University of Oslo, Norway
josang@unik.no
[2] 21st Century Systems Inc.
USA
sohara@21csi.com

Abstract. Multinomial subjective opinions are a special type of belief functions, where belief mass can be assigned to singletons of the frame as well as to the whole frame, but not to overlapping subsets of the frame. The multiplicative product of two multinomial opinions applies to the Cartesian product of the two corresponding frames. The challenge when multiplying multinomial opinions is that the raw product initially produces belief mass terms on overlapping subsets which does not fit into the opinion requirement of only having belief mass on singletons and on the whole frame. It is therefore necessary to reassign belief mass from overlapping subsets to singletons and to the frame in a way that preserves consistency for multinomial opinions. This paper describes a method for computing multinomial products of opinions according to this principle.

1 Introduction

Arguments in subjective logic are called *"subjective opinions"* or just *"opinions"* for short[1,2], and are traditionally denoted as ω. A binomial opinion applies to a single proposition/state in a frame. A multinomial opinion applies to the whole frame, i.e. to all the propositions/states in the frame. A binomial opinion is represented by the quadruple consisting of belief mass, disbelief mass, uncertainty mass and base rate, denoted as $\omega = (b, d, u, a)$. A multinomial opinion is represented by the composite function consisting of a belief vector, uncertainty mass and a base rate vector, denoted as $\omega = (\vec{b}, u, \vec{a})$. The uncertainty mass is interpreted as *"uncertainty about probabilities"*, i.e. as the second order complement probability of the first order probability expectation values.

It is relatively straightforward to define operators for subjective opinions that generalize classical binary logic and probabilistic operators. The literature describes a variety of practical operators that provide a basis for modeling and analyzing situations where input arguments are incomplete or affected by uncertainty. Binomial and multinomial opinions are equivalent to Beta and Dirichlet probability density functions respectively. Through this equivalence subjective logic provides a calculus for reasoning with probability density functions. In addition to generalizing the set of basic operators traditionally used in binary logic and classical probability calculus, subjective logic also contains some non-traditional operators which are specific to subjective logic.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 248–257, 2010.

In this manuscript we describe multinomial opinion multiplication which previously has not been described in the literature. Multinomial products are useful e.g. when combining opinions about different aspects of the same phenomenon or object.

It is straightforward to compute multinomial products in traditional probability calculus, which simply consists of multiplication of the argument probability vectors. In binary logic, the product of two binary frames with their binary truth values produces a quaternary frame with corresponding truth values, which in turn can be multiplied with other frames. The related analytical framework of the Dempster-Shafer belief theory [5] traditionally does not define multiplication of bbas that apply to separate frames, but it would be straightforward to do. The approach would simply be to multiply the belief mass terms of the argument belief functions and assign the product belief masses to the corresponding subsets of the product frame. Multiplication of binomial opinions has also been described in the literature [3].

The challenge with multinomial multiplication in subjective logic is that it initially produces belief mass terms that do not fit into the multinomial representation of $\omega = (\vec{b}, u, \vec{a})$. What is needed therefore is a transformation of the initial product terms into a product opinion that conforms with the required representation. This approach to computing the multinomial opinion product is described below.

2 The Multinomial Opinion Representation

Uncertainty comes in many flavours, and a good taxonomy is described in [6]. In subjective logic, the uncertainty relates to probability values. For example, let the probability estimate of a future event x be expressed as $P(x) = 0.5$, e.g. for obtaining heads when flipping a coin. In subjective logic, the probability P expressed without uncertainty is interpreted as dogmatic and expresses a crisp value, even though the outcome of the event itself might be totally unpredictable. The probability of an event is thus separated from the certainty/uncertainty of its probability. With this separation subjective logic can be applied in case of an event with very likely outcome but where the probability of the outcome still can be totally uncertain. This is possible by including the base rate of an event in the belief representation. For example the *a priori* likelihood that a given person selected at random is immune against tetanus[1] is close to 1, simply due to the base rate of tetanus immunity in the population. However, before actually testing the person, the immunity is still uncertain. The extreme case of an absolutely likely event that still has an uncertain probability is theoretically possible but is at the same time a singularity in subjective logic.

A general multinomial opinion is a composite function consisting of a belief vector $\vec{b}$, an uncertainty mass u and a base rate vector $\vec{a}$. These components are defined next.

Definition 1. Belief Mass Vector
Let $X = \{x_i | i = 1, \ldots k\}$ be a frame of cardinality k and let $\vec{b}$ be a vector function from the singletons of X to $[0, 1]^k$ satisfying:

$$\vec{b}(\emptyset) = 0 \quad \textit{and} \quad \sum_{x \in X} \vec{b}(x) \leq 1 \, . \tag{1}$$

[1] Assuming a random person from the population of the developed world.

Then $\vec{b}$ is called a belief mass vector, or belief vector for short.

The parameter $\vec{b}(x_i)$ is interpreted as belief mass on x_i, i.e. the amount of positive belief that x_i is true. The belief vector can be interpreted as a sub-additive probability function because the sum can be less than one. Additivity is achieved by including the uncertainty mass defined below.

Definition 2. Uncertainty Mass
Let $X = \{x_i | i = 1, \ldots k\}$ be a frame with a belief vector $\vec{b}$. Let u be a function from X to $[0, 1]$ representing uncertainty over X satisfying:

$$u + \sum_{x \in X} \vec{b}(x) = 1 \, . \tag{2}$$

The parameter u is then called an uncertainty mass.

The uncertainty mass can be interpreted as the lack of committed belief about the truth of any of the propositions of X. In other words, uncertainty mass reflects that the belief owner does not know which of the propositions of X in particular is true, only that one of them must be true.

In case the belief vector is subadditive, i.e. $\sum_{x \in X} \vec{b}(x) < 1$, the base rate vector together with base rates will determine the probability expectation values over X. The base rate vector is defined below.

Definition 3. Base Rate Vector
Let $X = \{x_i | i = 1, \ldots k\}$ be a frame and let $\vec{a}$ be a vector function from the singletons of X to $[0, 1]^k$ representing non-informative a priori probability over X satisfying:

$$\vec{a}(\emptyset) = 0 \quad \textit{and} \quad \sum_{x \in X} \vec{a}(x) = 1 \, . \tag{3}$$

Then $\vec{a}$ is called a base rate vector.

Having defined the belief vector, the uncertainty mass and the base rate vector, the general opinion can be defined.

Definition 4. Subjective Opinion
Let $X = \{x_i | i = 1, \ldots k\}$ be a frame, i.e. a set of k exhaustive and mutually disjoint propositions x_i. Let $\vec{b}$ be a belief vector, let u be the corresponding uncertainty mass, and let $\vec{a}$ be the base rate vector over X, all seen from the viewpoint of a subject entity A. The composite function $\omega_X^A = (\vec{b}, u, \vec{a})$ expresses A's subjective beliefs over X. This represents the traditional belief notation of opinions.

We use the convention that the subscript on the multinomial opinion symbol indicates the frame to which the opinion applies, and that the superscript indicates the subject owner of the opinion so that ω_X^A denotes A's opinion about X. Subscripts can be omitted when it is clear and implicitly assumed to which frame an opinion applies, and superscripts can be omitted when it is irrelevant who the owner is.

Assuming that the frame X has cardinality k, then the belief vector $\vec{b}$ and the base rate vector $\vec{a}$ will have k parameters each. The uncertainty parameter u is a simple scalar. A multinomial opinion over a frame of cardinality k will thus contain $2k + 1$ parameters. However, given the constraints of Eq.(2) and Eq.(3), the multinomial opinion will actually only have $2k - 1$ degrees of freedom. A binomial opinion will for example be 3-dimensional.

The introduction of the base rate vector allows the probabilistic transformation to be independent from the internal structure of the frame. The probability expectation of multinomial opinions is a vector expressed as a function of the belief vector, the uncertainty mass and the base rate vector.

Definition 5. Probability Expectation Vector
Let $X = \{x_i | i = 1, \ldots k\}$ be a frame and let ω_X be an opinion on X with belief vector $\vec{b}$, uncertainty mass u, and base rate vector $\vec{a}$. The function $\vec{\mathrm{E}}_X$ from the singletons of X to $[0, 1]^k$ expressed as:

$$\vec{\mathrm{E}}_X(x_i) = \vec{b}(x_i) + \vec{a}(x_i)u \ . \tag{4}$$

is then called the probability expectation vector over X.

It can be shown that $\vec{\mathrm{E}}_X$ satisfies the additivity principle:

$$\vec{\mathrm{E}}_X(\emptyset) = 0 \quad \text{and} \quad \sum_{x \in X} \vec{\mathrm{E}}_X(x) = 1 \ . \tag{5}$$

The base rate vector of Def.3 expresses non-informative *a priori* probability, whereas the probability expectation function of Eq.(4) expresses informed probability estimates, i.e. that are based on evidence which comes in addition to the base rates.

Given a frame of cardinality k, the default base rate of each element in the frame is $1/k$, but it is possible to define arbitrary base rates for all elements of the frame, as long as the additivity constraint of Eq.(3) is satisfied.

Two different multinomial opinions on the same frame will normally share the same base rate vectors. However, it is obvious that two different observers can assign different base rates to the same frame, in addition to assigning different beliefs to the frame. This naturally reflects different views, analysis and interpretations of the same context and situation seen by different observers.

The largest multinomial opinions that can be easily visualized are trinomial, in which case it can be represented as a point inside an equal-sided tetrahedron (pyramid with triangular base), as shown in Fig.1 below.

In Fig.1 the vertical elevation of the opinion point inside the tetrahedron represents the uncertainty mass. When considering a triangular side plane and the opposite vertex corresponding to a given state x_i, the orthogonal distances from the plane to the opinion point represents the belief mass value on the state x_i. This geometric structure is commonly called a *barycentric* coordinate system, so named by August Ferdinand Möbius (1827). It can be shown that the opinion point is the center of mass when it is

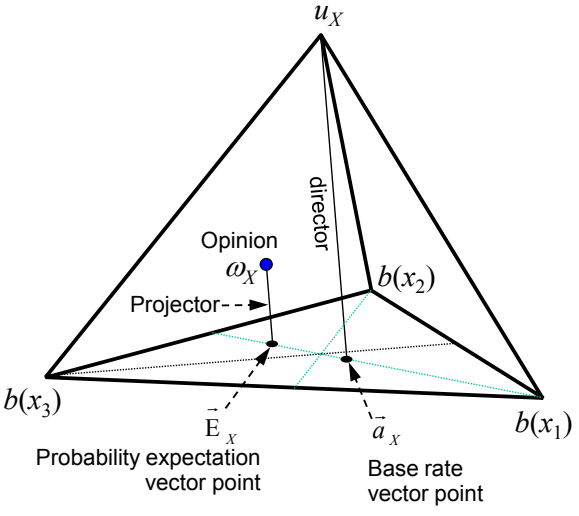


Fig. 1. Opinion tetrahedron with example opinion

assumed that the uncertainty mass and the belief masses are placed on their respective vertices.

The base rate vector $\vec{a}_X$ is indicated as a point on the base plane. The line that joins the tetrahedron apex and the base rate vector point represents the director. The probability expectation vector point is geometrically determined by drawing a projection from the opinion point parallel to the director onto the base plane.

In general, the triangle and tetrahedron belong to the *simplex* family of geometrical shapes. Multinomial opinions on frames of cardinality k can in general be represented as a point in a simplex of dimension $(k + 1)$. For example, binomial opinions can be represented inside a triangle which is a 3D simplex, and trinomial opinions can be represented inside a tetrahedron which is a 4D simplex.

The 2D aspect of paper and computer display units makes it impractical to visualize larger than 4D simplexes, meaning that opinions larger than trinomial do not lend themselves to traditional visualization.

3 Products of Multinomial Opinions

Evaluating the products of two separate multinomial opinions involves the Cartesian product of the respective frames to which the opinions apply. Let ω_X and ω_Y be two independent multinomial opinions that apply to the separate frames

$$X = \{x_1, x_2, \ldots x_k\} \text{ with cardinality } k \tag{6}$$

$$Y = \{y_1, y_2, \ldots y_l\} \text{ with cardinality } l\ . \tag{7}$$

The Cartesian product $X \times Y$ with cardinality kl is expressed as the matrix:

$$X \times Y = \begin{pmatrix} (x_1, y_1), & (x_2, y_1), & \cdots & (x_k, y_1) \\ (x_1, y_2), & (x_2, y_2), & \cdots & (x_k, y_2) \\ . & . & \cdots & . \\ . & . & \cdots & . \\ (x_1, y_l), & (x_2, y_l), & \cdots & (x_k, y_l) \end{pmatrix} \tag{8}$$

We now turn to the product of the multinomial opinions. The raw terms produced by $\omega_X \cdot \omega_Y$ can be separated into different groups.

1. The first group of terms consists of belief masses on singletons of $X \times Y$:

$$b^{\mathrm{I}}_{X\times Y} = \begin{cases} b_X(x_1)b_Y(y_1), & b_X(x_2)b_Y(y_1), & \ldots & b_X(x_k)b_Y(y_1) \\ b_X(x_1)b_Y(y_2), & b_X(x_2)b_Y(y_2), & \ldots & b_X(x_k)b_Y(y_2) \\ . & . & \ldots & . \\ b_X(x_1)b_Y(y_l), & b_X(x_2)b_Y(y_l), & \ldots & b_X(x_k)b_Y(y_l) \end{cases} \tag{9}$$

2. The second group of terms consists of belief masses on rows of $X \times Y$:

$$b^{\mathrm{Rows}}_{X\times Y} = \big(u_X b_Y(y_1), \quad u_X b_Y(y_2), \quad \ldots \quad u_X b_Y(y_l) \big) \tag{10}$$

3. The third group of terms consists of belief masses on columns of $X \times Y$:

$$b^{\mathrm{Columns}}_{X\times Y} = \big(b_X(x_1) u_Y, \quad b_X(x_2) u_Y, \quad \ldots \quad b_X(x_k) u_Y \big) \tag{11}$$

4. The last term is simply the belief mass on the whole product frame:

$$u^{\mathrm{Frame}}_{X\times Y} = u_X u_Y \tag{12}$$

The singleton terms of Eq.(9) and the term on the whole frame are unproblematic because they conform with the opinion representation of having belief mass only on singletons and on the whole frame. In contrast, the terms on rows and columns apply to overlapping subsets which is not compatible with the required opinion format, and therefore need to be reassigned. Some of it can be reassigned to singletons, and some to the whole frame. There are several possible strategies for determining the amount of uncertainty mass to be assigned to singletons and to the frame. Two methods are described below.

3.1 Determining Uncertainty Mass

1. **The Method of Assumed Belief Mass:** The simplest method is to assign the belief mass from the terms of Eq.(10) and Eq.(11) to singletons. Only the uncertainty mass from Eq.(12) is then considered as uncertainty in the product opinion, expressed as:

$$u_{X\times Y} = u_X u_Y \ . \tag{13}$$

   A problem with this approach is that it in general produces less uncertainty than intuition would dictate.

2. **The Method of Assumed Uncertainty Mass:** A method that preserves more uncertainty is to consider the belief mass from Eq.(10) and Eq.(11) as potential uncertainty mass that together with the uncertainty mass from Eq.(12) can be called intermediate uncertainty mass. The intermediate uncertainty mass is thus:

$$u^{\mathrm{I}}_{X\times Y} = u^{\mathrm{Rows}}_{X\times Y} + u^{\mathrm{Columns}}_{X\times Y} + u^{\mathrm{Frame}}_{X\times Y} \tag{14}$$

The probability expectation values of each singleton in the product frame can easily be computed as the product of the expectation values of each pair of states from X and Y, as expressed in Eq.(15).

$$\begin{aligned} \mathrm{E}((x_i, y_j)) &= \mathrm{E}(x_i)\mathrm{E}(y_j) \\ &= (b_X(x_i) + a_X(x_i)u_X)(b_Y(y_j) + a_Y(y_j)u_Y) \end{aligned} \tag{15}$$

We also require that the probability expectation values of the states in the product frame can be computed as a function of the product opinion according to Eq.(16).

$$\mathrm{E}((x_i, y_j)) = b_{X\times Y}((x_i, y_j)) + a_X(x_i)a_Y(y_j)u_{X\times Y} \tag{16}$$

In order to find the correct uncertainty mass for the product opinion, each state $(x_i, y_j) \in X \times Y$ will be investigated in turn to find the smallest uncertainty mass that satisfies both Eq.(16) and Eq.(17).

$$\frac{b^{\mathrm{I}}_{X\times Y}((x_i, y_j))}{u^{\mathrm{I}}_{X\times Y}} = \frac{b_{X\times Y}((x_i, y_j))}{u_{X\times Y}} \tag{17}$$

The uncertainty mass that satisfies both Eq.(16) and Eq.(17) for state (x_i, y_j) can be expressed as:

$$u^{(i,j)}_{X\times Y} = \frac{u^{\mathrm{I}}_{X\times Y}\mathrm{E}((x_i, y_j))}{b^{\mathrm{I}}_{X\times Y}((x_i, y_j)) + a_X(x_i)a_Y(y_j)u^{\mathrm{I}}_{X\times Y}} \tag{18}$$

The product uncertainty can now be determined as the smallest $u^{(i,j)}_{X\times Y}$ among all the states, expressed as:

$$u_{X\times Y} = \min\left\{ u^{(i,j)}_{X\times Y} \text{ where } (x_i, y_j) \in X \times Y \right\} \tag{19}$$

3.2 Determining Belief Mass

Having determined the uncertainty mass, either according to Eq.(13) or according to Eq.(19), the expression for the product expectation of Eq.(15) can be used to compute the belief mass on each element in the product frame, as expressed by Eq.(20).

$$b_{X\times Y}((x_i, y_j)) = \mathrm{E}((x_i, y_j)) - a_X(x_i)a_Y(y_j)u_{X\times Y} \tag{20}$$

It can be shown that the additivity property of Eq.(21) is preserved.

$$u_{X\times Y} + \sum_{(x_i,y_j)\in X\times Y} b_{X\times Y}((x_i, y_j)) = 1 \tag{21}$$

From Eq.(20) it follows directly that the product operator is commutative. It can also be shown that the product operator is associative.

4 Example

We consider the scenario where a GE (Genetic Engineering) process can produce Male (M) or Female (F) eggs, and that in addition, each egg can have genetical mutation S or T independently of its gender. This constitutes two binary frames $X = \{\mathrm{M}, \mathrm{F}\}$ and $Y = \{\mathrm{S}, \mathrm{T}\}$, or alternatively the quaternary product frame $X \times Y = \{\mathrm{MS}, \mathrm{MT}, \mathrm{FS}, \mathrm{FT}\}$. Sensor A observes whether each egg is M or F, and Sensor B observes whether the egg has mutation S or T.

Assume that an opinion regarding the gender of a specific egg is derived from Sensor A data, and that an opinion regarding its mutation is derived from Sensor B data. Sensors A and Sensor B have thus observed different and orthogonal aspects, so their respective opinions can be combined with multiplication. This is illustrated in Fig.2.

The result of the opinion multiplication can be considered as an opinion based on a single observation where both aspects are observed at the same time. Let the observation opinions be:

$$\text{Gender } \omega_X^A : \begin{cases} \vec{b}_X^A = (0.8, 0.1) \\ u_X^A = 0.1 \\ \vec{a}_X^A = (0.5, 0.5) \end{cases} \qquad \text{Mutation } \omega_Y^B : \begin{cases} \vec{b}_Y^B = (0.7, 0.1) \\ u_Y^B = 0.2 \\ \vec{a}_Y^B = (0.2, 0.8) \end{cases} \tag{22}$$

The Cartesian product frame can be expressed as:

$$X \times Y = \begin{pmatrix} \mathrm{MS}, & \mathrm{FS} \\ \mathrm{MT}, & \mathrm{FT} \end{pmatrix} \tag{23}$$

According to Eq.(15) the product expectation values are:

$$\mathrm{E}(X \times Y) = \begin{pmatrix} 0.629, & 0.111 \\ 0.221, & 0.039 \end{pmatrix} \tag{24}$$

Below are described the results of both methods proposed in Sec.3.1.

1. When applying the method of *Assumed Belief Mass* where the uncertainty mass is determined according to Eq.(13), the product opinion is computed as:

$$b_{X\times Y} = \begin{pmatrix} 0.627, & 0.109 \\ 0.213, & 0.031 \end{pmatrix}, \quad u_{X\times Y} = 0.02, \quad a_{X\times Y} = \begin{pmatrix} 0.1, & 0.4 \\ 0.1, & 0.4 \end{pmatrix} \tag{25}$$

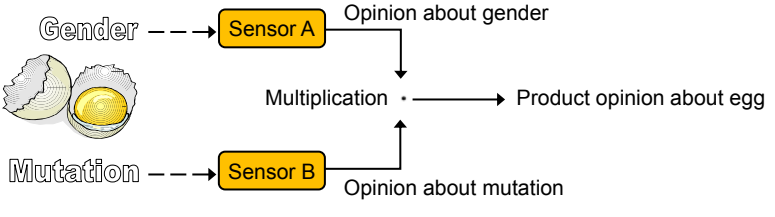


Fig. 2. Multiplication of opinions on orthogonal aspects of GE eggs

2. When applying the method of *Assumed Uncertainty* where the uncertainty mass is determined according to Eq.(18) and Eq.(19), the product opinion is computed as:

$$b_{X\times Y} = \begin{pmatrix} 0.620, & 0.102 \\ 0.185, & 0.003 \end{pmatrix}, \quad u_{X\times Y} = 0.09, \quad a_{X\times Y} = \begin{pmatrix} 0.1, & 0.4 \\ 0.1, & 0.4 \end{pmatrix} \tag{26}$$

The results indicate that there can be a significant difference between the two methods, and that the safest approach is to use the *assumed uncertainty* method because it preserves the most uncertainty in the product opinion.

5 Discussion and Conclusion

Multiplication of multinomial opinions is useful in many situations, such as when combining input from sensors that observe different aspects of a target. Two methods for computing the product of multinomial opinions are presented in this paper, where the method of assumed uncertainty is recommended because it preserves the most uncertainty and thereby better reflects the uncertainty of the input arguments.

Subjective opinions are related to general bbas. One of the differences is that a bba can assign belief mass to any subset of a frame, whereas an opinion can only assign belief mass to singletons and to the whole frame. The other difference is that bbas do not include base rates, whereas opinions do. Consequently opinions represent both a subset of, and an extension of general bbas.

Opinions can be derived from bbas if it can be assumed that base rates can be defined separately [4]. It is thus possible to use general bbas as input to subjective logic models in general and to multiplication of opinions in particular.

The advantage of subjective logic over traditional probability calculus and probabilistic logic is that real world situations can be modeled and analyzed more realistically. The analyst's partial ignorance and lack of information can be taken explicitly into account during the analysis, and explicitly expressed in the conclusion. When used for decision support, subjective logic allows decision makers to be better informed about uncertainties affecting the assessment of specific situations. At the same time subjective logic is compatible with traditional statistical analysis.

While the belief representation of opinions is less flexible than that of general bbas, it has the advantage that traditional statistical analysis can be directly applied and that the set of operators such as conditional deduction and abduction can be used for modeling Bayesian networks and the transitivity operator can be used to model trust networks.

References

1. Jøsang, A.: A Logic for Uncertain Probabilities. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9(3), 279–311 (2001)
2. Jøsang, A.: Conditional Reasoning with Subjective Logic. Journal of Multiple-Valued Logic and Soft Computing 15(1), 5–38 (2008)
3. Jøsang, A., McAnally, D.: Multiplication and Comultiplication of Beliefs. International Journal of Approximate Reasoning 38(1), 19–51 (2004)

4. Jøsang, A., O'Hara, S., O'Grady, K.: Base Rates for Belief Functions. In: Proceedings of the Workshop on the Theory on Belief Functions(WTBF 2010), Brest (April 2010)
5. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
6. Smithson, M.: Ignorance and Uncertainty: Emerging Paradigms. Springer, Heidelberg (1988)

Evaluation of Information Reported: A Model in the Theory of Evidence

Laurence Cholvy

ONERA, 2 avenue Edouard Belin,
31055 Toulouse, France
cholvy@onera.fr

Abstract. This paper deals with the evaluation of a piece of information when successively reported by several agents. It describes a model based on the Theory of Evidence in which the evaluation of a reported piece of information is defined by a plausibility degree and depends on validity degrees of agents which report it.

Keywords: Information reported, evaluation, validity, Theory of Evidence.

1 Introduction

Before making a decision, any rational agent tries to know what is the current state of the world. For doing so, the agent has to acquire information about the current state of the world and this can be done by different ways. First, the agent can itself acquire information if it has got the capacity for doing so. For instance, in order to know if I take my umbrella before going out, I can glance at the sky through the window. The agent can also get information it needs via another agent which can provide it. For instance, in order to know if I take my umbrella before going out, I can look at the web site of Météo-France. Sometimes, the process is more complex and the agent gets information it needs via a long chain of agents. This is the case when, in order to know if I take my umbrella before going out, I read in my newspaper, the forecast provided by Météo-France. Or when I ask my neighbour to read the forecast in his newspaper. Here, between the agent which provides the report (Météo-France) and I who need it, there is a sequence of agents: the newspaper, the neighbour, each one in its turn reporting information to the next one.

Once a piece of information is acquired, the problem of evaluating it, i.e. evaluating its truth, cannot be ignored. This means asking the question: how true is this piece of information ? If it is not easy to answer this question when the report is acquired via another agent, it is even less when it is reported by several successive agents. Indeed, how can I estimate how true is the report "rainy weather today" given by my neighbour after he reads in his newspaper the forecast provided by Météo France ?

The question of computing the evaluation of a piece of information reported by several successive agents is the object of this paper. Our objective is to propose

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 258–267, 2010.

a method for characterizing information evaluation by taking into account some properties of agents. Furthermore, in order to be as general as possible, we consider that information evaluation is not a binary value. Consequently, we use the Theory of Evidence as model.

This paper is organized as follows. Section 2 presents some works we drew inspiration from. Section 3 presents our model. Finally, section 4 is devoted to a discussion, points outs the limitations of this work and its possible extensions.

2 State of the Art

A domain in which the question of information evaluation is very important is military Intelligence. For this reason, NATO has defined a standard in order to guide intelligence officers to associate information with its evaluation [11]. According to this standard, this evaluation is a pair of two values. The first one corresponds to the reliability degree of the source which provides the piece of information and the second value refers to the information credibility. Informal comments define such values. More precisely, the reliability degree of a source depends on how, from its past use, we can trust the source for delivering true information. The credibility degree of a piece of information depends on the fact that it is confirmed or not by several sources and also depends on the fact that it is more or less in conflict with other information.

This way of defining information evaluation has been criticized in [4] where it is shown that the two values are not independent.

Several proposals have been proposed in order to circumvent the main limitations of the NATO guidelines [4], [9]. For instance, [9] suggests to define the overall confidence in a reported information by taking into account not only the source reliability and information credibility but also information plausibility and the source competence. Information plausibility refers to the degree of match between a reported information and our knowledge of the world. The furthest information is from what we know about the world, the less plausible information is. The source competence is generally topic dependent and refers to the skills, experience of the source with regard to the topic tackled by information it reports. However none of the proposals deals with reported information as informally defined in introduction and which is precisely the main point of interest of this article.

As far as we know, the work described in [8] is the only one which studies the case of reported information. This work considers French Press Agency (AFP) dispatches, each press dispatch being represented by a sentence, the most representative one, which can mention several successive sources like, for instance: "President X said that country Y is probably developing nuclear power". Here, AFP reports "country Y is probably developing nuclear power" reported by X. In this work, parameters which are considered as having an influence on the evaluation of such information are: *the quality of the source*, which can be defined from its past use if the source is already known, or be defined a-priori if we know which type the source is. Notice that this notion of quality is close to

the notion of reliability already mentioned; *the opinion of the source on its own report* which is drawn from text analysis and in particular, from the analysis of the subjective modalities mentioned in the report such "I am certain that ..." or "this is highly probable", "this is impossible", "I am fearing that ..."...; *relations which may exist between the agent which makes the report and agents mentioned in this report*; *relations which may exist between successive sources.* Such relations can be neutral, hostile or well-willing. For instance, in case of non neutral relation, the source may make a non sincere report, thus propagating a piece of information it knows to be false.

We have also found interesting contributions in the domain of logic. Among them, [7] and [2] are worth mentioning.

In his work about the notion of trust, [7], R. Demolombe studies the relations which exist between a piece of information, its truth and the mental attitudes of the agent which produces this piece of information. The formalism he uses is a modal logic, [3], some operators of which are: B_i ($B_i p$ means "agent i believes that p"), I_i^j ($I_i^j p$ means "agent i informs agent j that p"). Operator B_i obeys KD system which is quite usual for beliefs and operator I_i^j only obeys rule of equivalence substitutivity. Before focusing on the notion of trust which is his main subject, the author defines several properties agents can have, called epistemic properties, among which the following are interesting for our problem:

- *Sincerity*: Agent i is sincere with regard to j for information p iff, if i informs j that p, then i believes p. I.e. a sincere agent believes what he says. Thus $sincere(i, j, p) \equiv I_i^j p \rightarrow B_i p$.
- *Competence* Agent i is competent about p iff, if i believes p, then p is true. I.e. the beliefs of a competent agent are true. Thus $competent(i, p) \equiv B_i p \rightarrow p$.
- *Validity*: Agent i is valid with regard to j for p iff, if i informs j about p, then p is true. I.e. a valid agent tells the truth. Thus $valid(i, j, p) \equiv I_i^j p \rightarrow p$.

Thus we have: $sincere(i, j, p) \wedge competent(i, p) \rightarrow valid(i, j, p)$.

In [2], Ph. Capet also uses modal logic to characterize the ability of agents for lying. Several definitions of lies are proposed, from the simplest and thus the most questionable one to the richest one. Besides the operators B_i and I_i^j[1], he considers an intention operator Int_i, where $Int_i \phi$ means that agent i intends that ϕ. The three definitions successively given in the thesis are:

- (1) $lies(i, j, p) \equiv I_i^j p \wedge \neg p$. According to this definition, agent i lies when he tells some other agent j something which is false. This is a bit strong and it can be shown that this definition characterizes in fact a notion of "untruth". This is the case of forecasts found on an odd web site and which happen to be false. One cannot consider that this site lies, but we can conclude that this site is not valid (according to Demolombe's terminology) at least for p.
- (2) $lies(i, j, p) \equiv I_i^j p \wedge B_i \neg p$. According to this definition, and agent i lies when he tells some other agent j something he himself does not believe. However this definition does not consider the j's intention. This is why it is refined in:

[1] The author denotes it A_i^j but we prefer to re-use the previous notation.

- (3) $lies(i,j,p) \equiv I_i^j p \wedge B_i \neg p \wedge Int_i B_j p$. According to this definition, agent i lies when he tells some other agent j something he does not believe and when he intends that j believes it. This definition can also be refined by assuming an inference mechanism in j, say:
- (3′) $lies(i,j,p) \equiv I_i^j p \wedge B_i \neg p \wedge Int_i B_j f(p)$ where $f(p)$ is a fact which depends on p such as $p \wedge q$, $\neg p$ etc. This is the case for instance when an expert reports that *the leak in the nuclear plant will have no consequence on water pollution* to make people believe that nuclear plants are safe even if he thinks the contrary.

The two previous works are worth mentioning since they clarify some properties of information sources in regard with their ability of telling true information or telling false information.

In recent works, [6], D. Dubois and T. Denoeux address very close questions by using the Theory of Evidence [10]. In this theory, the concept which corresponds to the notion of evaluation is the concept of plausibility. In their work, the authors propose a mechanism for computing the plausibility of a piece of information which is emitted by an agent i given our uncertain belief about i's reliability. In this work, the reliability of an agent is defined by its relevance and its sincerity. For Dubois and Denoeux, an agent is relevant if it is competent in the topic of the piece of information it provides; an agent is sincere if it does not lie (a non-sincere agent says the opposite of what it believes). Thus Dubois and Denoeux's notion of "reliability" and Demolombe's notion of "validity" are very close, despite being modelled in different formalisms.

Consider that an agent i provides information ϕ. The belief one has about i's reliability is used in Dubois and Denoeux's model as follows. If i is not competent in the topic of ϕ, then ϕ is replaced by the tautology $\phi \vee \neg\phi$; If i is competent in the topic of ϕ, then, if it is sincere then we keep ϕ, else ϕ is replaced by $\neg\phi$.

Competence and sincerity can be considered as two independent notions. Thus, if p is the probability of i's being competent and q is the probability of i's being sincere, then the plausibility of ϕ can be shown to be equal to $p.q + 1 - p$.

In their work, Dubois and Denoeux assume that any piece of information is provided by a single agent. They do not assume that information is reported by several successive agents. However, like Dubois and Denoeux, we think that the Theory of Evidence is an interesting formalism when one has to deal with uncertainty. In what follows, we show how we use the Theory of Evidence to deal with graded validity (we prefer to use this term instead of the term reliability) and how to get a graded plausibility of reported information.

3 Modelisation in the Theory of Evidence

In this section, we suggest to model our problem in the framework of the Theory of Evidence. This choice is justified by the fact that this formalism offers two interesting concepts which are the concept of mass assignment, which allows us to express degrees of beliefs on information and the concept of plausibility function which will allow us to quantify the evaluation of a piece of information.

We assume that the reader is familiar with the Theory of Evidence and also with the propositional logic.

3.1 First Case: One Agent

In this first case, we consider that an agent i reports a piece of information ϕ. This is denoted $R_i\phi$. The question we deal with is: how true is ϕ ? In order to answer this question, we take as a starting point the notion of validity introduced by Demolombe in [7] and we generalize it in the framework of the Theory of Evidence, thus expressing a graded validity.

We consider a propositional language the two letters of which are: ϕ and $R_i\phi$, representing respectively the facts "information ϕ is true" and "agent i reported information ϕ". The four interpretations of this language are $\{w_1, w_2, w_3, w_4\}$. w_1 represents the situation in which i has reported information ϕ and ϕ is true. It is denoted $w_1 = \{R_i\phi, \phi\}$; w_2 represents the situation in which i has reported information ϕ and ϕ is false. It is denoted $w_2 = \{R_i\phi, \neg\phi\}$; w_3 represents the situation in which i did not report information ϕ and ϕ is true. It is denoted $w_3 = \{\neg R_i\phi, \phi\}$.; w_4 represents the situation in which i did not report information ϕ and ϕ is false. It is denoted $w_4 = \{\neg R_i\phi, \neg\phi\}$. We consider as discernment frame, the set $\Theta = \{w_1, w_2, w_3, w_4\}$.

Definition 1. Consider an agent i and a piece of information ϕ. We consider that i is *valid for* ϕ *at the degree* d_i, $d_i \in [0, 1]$, written $valid(i, \phi, d_i)$ if and only if our beliefs can be modelled by the mass assignment $m^{v(i,\phi,d_i)}$ defined by:

$$m^{v(i,\phi,d_i)}(w_1 \vee w_3 \vee w_4) = d_i$$
$$m^{v(i,\phi,d_i)}(w_2 \vee w_3 \vee w_4) = 1 - d_i$$

Let us recall that assigning a mass on a disjunction of w_i is equivalent to assigning this mass on any propositional formula satisfied by all the w_i in the disjunction. The equivalence is proved in [5]). Consequently, the mass assignment defined in the previous definition can be reformulated by:

$$m^{v(i,\phi,d_i)}(R_i\phi \rightarrow \phi) = d_i$$
$$m^{v(i,\phi,d_i)}(R_i\phi \rightarrow \neg\phi) = 1 - d_i$$

Thus, according to definition 1, we consider that i is valid for ϕ at the degree d_i if and only if the degree of our beliefs in the fact that, "if i reports ϕ then ϕ is true" is d_i and the degree of our belief in the fact "if i reports ϕ then ϕ is false" is $1 - d_i$.

In this work, we consider that for any agent i and any information ϕ, degree d_i is unique. This means that we implicitly assume that the current environment in which we evaluate the validity does not influence our evaluation of it. This is obviously simplistic.

The following particular cases are worth detailing:

- $(d_i = 1)$ We say that i is *valid* for ϕ if and only if i is valid for ϕ at the degree 1. Thus we have $m^{v(i,\phi,1)}(R_i\phi \rightarrow \phi) = 1$. I.e. we are certain that if i reports ϕ then ϕ is true.

– $(d_i = 0)$ We say that i is *invalid* for ϕ if and only if i is valid for ϕ at the degree 0. Thus we have $m^{v(i,\phi,0)}(R_i\phi \rightarrow \neg\phi) = 1$. I.e. we are certain that if i reports ϕ then ϕ is false.

Definition 2. We denote $m^{R_i\phi}$ the mass assignment defined by: $m^{R_i\phi}(R_i\phi) = 1$ (or equivalently, $m^{R_i\phi}(w_1 \vee w_2) = 1$)

The mass function defined by this definition represents the fact that, for sure, agent i has reported information ϕ.

Definition 3. Let us consider that agent i is valid for ϕ at the degree d_i. After i reports ϕ, our beliefs are modelled by the mass assignment m obtained by Dempster's combination of $m^{v(i,\phi,d_i)}$ and $m^{R_i\phi}$. I.e.,

$$m = m^{v(i,\phi,d_i)} \oplus m^{R_i\phi}$$

Notice that we have:

$$m(R_i\phi \wedge \phi) = d_i$$
$$m(R_i\phi \wedge \neg\phi) = 1 - d_i$$

Definition 4. Consider that agent i, valid for ϕ at degree d_i, reports ϕ. Thus the evaluation of information ψ is defined by: $pl(\psi)$ where pl is the plausibility function associated with the mass assignment m.

According to this definition, we define the evaluation of any piece of information by its plausibility as given by the plausibility function associated with the mass assignment m. In particular, we have:

Proposition 1. $pl(\phi) = d_i$
Thus, the evaluation of a reported piece of information is the degree of validity of the agent which reported it. The higher this degree the higher this evaluation is.

Proposition 2. $pl(\neg\phi) = 1 - d_i$
Consequently, if $d_i > 0.5$, ϕ is more plausible than $\neg\phi$.

It must be noticed that the mass function m given by definition 3, which models our beliefs after agent i, valid for ϕ at the degree d_i reports ϕ, is identical to the mass function defined obtained by Dubois and Denoeux's model [6] if one considers an agent i reporting information ϕ, where i is supposed to be totally competent and sincere at the degree d_i.

3.2 Second Case: Two Agents

Here, we consider that agent j reports that agent i has reported ϕ. This is denoted: $R_jR_i\phi$. The question is: how true is ϕ ? or saying it differently, what is the influence of validity degrees of i and j on the evaluation of ϕ ?

Now, we consider a propositional language the letters of which are: ϕ, $R_i\phi$, and $R_jR_i\phi$. This language has got 8 interpretations $w_1 ... w_8$. The frame of

discernment is the set $\Theta = \{w_1, ...w_8\}$. We do not detail these w_i because, as before, we will assign mass on formulas and not on disjunctions of w_i.

Definition 5. Consider that $R_j R_i \phi$, $valid(i, \phi, d_i)$ and $valid(j, R_i\phi, d_j)$. Then, our beliefs are defined by the mass assignment denoted m defined by:

$$m = m^{v(i,\phi,d_i)} \oplus m^{v(j,R_j\phi,d_j)} \oplus m^{R_j R_i \phi}$$

Thus, when $R_j R_i \phi$, our beliefs are defined by combining, by Dempster's rule of combination: our beliefs on the fact that i is valid at the degree d_i and our beliefs on the fact that j, known to be valid at the degree d_j has reported information $R_i\phi$.

Proposition 3

$$m(R_j R_i \phi \wedge R_i \phi \wedge \phi) = d_i.d_j$$
$$m(R_j R_i \phi \wedge R_i \phi \wedge \neg\phi) = (1 - d_i)d_j$$
$$m(R_j R_i \phi \wedge \neg R_i \phi) = 1 - d_j$$

Proposition 4

$$pl(\phi) = (d_i - 1).d_j + 1$$
$$pl(\neg\phi) = 1 - d_i d_j$$

The two following cases are worth examining:

- $(d_j = 1)$ In this case, $valid(j, R_i\phi, 1)$. Thus we get $pl(\phi) = d_i$ and $pl(\neg\phi) = 1 - d_i$. We come to the "one agent" case (see before).
- $(d_j = 0)$ In this case, $valid(j, R_i\phi, 0)$, i.e. $R_i\phi$ is false. We have: $pl(\phi) = 1$ and $pl(\neg\phi) = 1$. Thus we cannot decide among ϕ and $\neg\phi$ which is the most plausible.

Proposition 5. $pl(\phi) > pl(\neg\phi) \Longleftrightarrow d_j \neq 0$ and $d_i > 0.5$

I.e. we can conclude that ϕ is strictly more plausible than $\neg\phi$ iff j is not invalid, and i is valid at a degree strictly higher than 0.5.

Example. Let us illustrate this on the example given in introduction. Consider that the reports about the weather are given to me by my neighbour who read them in his newspaper. If I consider that my neighbour is valid (i.e. he really tells me what he reads in his newspaper) and if I consider his newspaper valid (the forecast is always true in this newspaper) then, I can conclude that the forecast my neighbour gives me is true. If I consider that my neighbour is valid (i.e. he really tells me what he reads in his newspaper) and if I consider his newspaper invalid (the forecast is always false in this newspaper), then, I can conclude that the forecast my neighbour gives me is false. But If I consider that my neighbour is invalid (i.e he does not tell me what he reads in his newspaper because for instance the newspaper were not distributed today) I cannot conclude: the forecast he reports may be true but it may be false as well.

3.3 General Case

We consider here the case when agent i_n reports that agent i_{n-1} has reported that agent ..... i_1 has reported ϕ. This is denoted $R_{i_n}...R_{i_1}\phi$. The question is again the influence of the degrees of validity of agents on the evaluation of information.

We consider a propositional language the $n+1$ letters of which are ϕ, $R_{i_1}\phi$, $R_{i_2}R_{i_1}\phi$, ..., $R_{i_n}...R_{i_1}\phi$. This language has got 2^{n+1} interpretations which form the frame of discernment we consider but we do not detail them because, as before, we assign masses to formulas.

Definition 6. Assume $R_{i_n}...R_{i_1}\phi$, $valid(i_1, \phi, d_1)$, $valid(i_2, R_{i_1}\phi, d_2)$,.., $valid(i_n, R_{i_{n-1}}...R_{i_1}\phi, d_n)$. Then, our beliefs are defined by the following mass assignment:

$$m = m^{v(i_1,\phi,d_1)} \oplus ... m^{v(i_{n-1},R_{i_{n-2}}...R_{i_1}\phi,d_{n-1})} \oplus m^{v(i_n,R_{i_{n-1}}...R_{i_1}\phi,d_n)} \oplus m^{R_{i_n}...R_{i_1}\phi}$$

By this definition, when $R_{i_n}...R_{i_2}R_{i_1}\phi$, our beliefs are defined by combining: our beliefs in the fact that i_1 is valid at degree d_1 and ....; our beliefs in the fact that i_{n-1} is valid at degree $d_{i_{n-1}}$ and our beliefs in the fact that i_n, known to be valid at degree d_{i_n} has reported $R_{i_{n-1}}...R_{i_2}R_{i_1}\phi$.

Proposition 6

$$pl(\phi) = (d_1 - 1).d_2...d_n + 1$$
$$pl(\neg\phi) = 1 - d_1...d_n$$

Proposition 7. $pl(\phi) > pl(\neg\phi) \iff \forall i = 2...n\ d_i \neq 0$ and $d_1 > 0.5$
I.e. we can conclude that ϕ is strictly more plausible than $\neg\phi$ iff i_2 ... i_n are not invalid and i_1 is valid at a degree strictly greater than 0.5.

4 Discussion

The main contribution of our paper is a model for characterizing the evaluation of reported information. This model assumes that the evaluation of reported information mainly depends on the ability of the reporting agents to report something true or to report something false. More precisely, it depends on the degree at which the reporting agents are reporting something true (i.e. are valid). The Theory of Evidence is used to express this model and its notion of plausibility allows us to characterize the evaluation of reported information.

This work takes credit for addressing a problem which has received little attention. However, many assumptions we made could be relaxed.

For instance, we could enrich the model by adding a supplementary level of uncertainty for representing our uncertain belief about the reported information itself. Indeed here, we assume that for sure, we face a reported information (see definition 2). The mass assignment $m^{R_i\phi}\phi$ could be extended to express uncertainty as follows: $m^{R_i\phi}\phi(R_i\phi) = x$ and $m^{R_i\phi}(\Theta_i) = 1 - x$, for $x \neq 1$.

We also have to study the parameters which influence the degrees at which agents report true information (i.e. their degree of validity). This paper considers these degrees as given, but in real application, we will have to provide guides to define them. For doing so, it will be necessary to take into account more knowledge about reporting agents.

As for the different choices we make (choices of a combination method for combining assignments defined on the same frame, choice of a method for combining assignments defined on different frames...), each of them can be discussed. Analysing the plausibility function we would get by making other choices defines an interesting research direction.

Another interesting open issue is to extend the type of information reported by the agents. In particular, we are currently investigating means of handling information of the type "agent i reports that he believes that ϕ is highly probable".

Finally, in the immediate future, we are going to study a variant of the present model in which the degree of validity will be splitted in two independent degrees: the degree of validity of a source as well as its degree of invalidity.

Acknowledgements

This work has been granted by ANR (Agence Nationale de Recherche) under project CAHORS.

Many thanks to Frédéric Pichon and Thomas Delavallade for their comments on this paper.

I am also very grateful to the anonymous referees whose comments helped me to improve this paper.

References

1. Appriou, A., Janez, F.: Theory of Evidence and non-exhaustive frames of discernment: Plausibilities correction methods. International Journal of Approximate reasoning (18), 1–19 (1998)
2. Capet, P.: Logique du mensonge, Thèse de doctorat de philosophie, Université de Paris 3 (2005)
3. Chellas, B.F.: Modal logic: An introduction. Cambridge University Press, Cambridge (1980)
4. Cholvy, L.: Information Evaluation in fusion: a case study. In: Proceedings of 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2004), Perugia (July 2004)
5. Cholvy, L.: Using Logic to Understand relations between DSmT and Dempster-Shafer Theory. In: Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU 2009), Verona (July 2009)
6. Dubois, D., Denoeux, T.: Pertinence et Sincérité en Fusion d'Informations. In: Rencontres Françaises sur la Logique Floue et ses Applications, LFA 2009, Annecy (Novembre 2009)
7. Demolombe, R.: Reasoning about trust: a formal logical framework. In: Proc. 2nd International Conference iTrust, Oxford (2004)

8. Jacquelinet, J.: Pertinence et cotation du renseignement, Mémoire de stage de l'ENSIIE, effectué sous la direction de Ph. Capet, Thalès (2007)
9. Revault d'Allonnes, A., Besombes, J.: Critères d'évaluation contextuelle pour le traitement automatique, Atelier Qualité des Données et des Connaissances (QDC). Strasbourg, France, Janvier (2009)
10. Shafer, G.: A mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
11. NATO: Intelligence Reports. STANAG 2511 (January 2003)

Gradual Evaluation of Granules of a Fuzzy Relation: R-related Sets

Slavka Bodjanova[1] and Martin Kalina[2]

[1] Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX 78363, USA
kfsb000@tamuk.edu
[2] Department of Mathematics, Slovak Technical University, Sk-813 68 Bratislava, Slovakia
kalina@math.sk

Abstract. A fuzzy relation R between elements from two finite universes is considered. Granules of R created by pairs of subsets from the two universes are evaluated. A rough extension of R based on a generalized rough set model is proposed. This extension allows us to introduce the notions of R-related sets, strongly R-related sets and R-compatible sets. R-related sets can be further used for evaluations of R-relationship between partitions of elements from two related universes.

Keywords: fuzzy relation, rough sets, aggregation functions, granulation, approximation, decision making.

1 Introduction

Granulation involves decomposition of whole to parts. In the early eighties Pawlak introduced the theory of rough sets [3] which quickly became one of the most important mathematical frameworks in granular computing [9,10,11,12]. The basic structure of rough set theory is an approximation space consisting of the universe of discourse U and an equivalence relation defined on U. The equivalence relation and the induced equivalence classes may be regarded as the available information (knowledge base) for approximation of an arbitrary subset X of U. The lower approximation is the union of all equivalence classes included in X and the upper approximation is the union of all equivalence classes having a non-empty intersection with X. The lower and the upper approximations create a rough set. Many generalizations and modifications of rough sets theory have been proposed [4,5,8].

We refer to the generalization suggested by Yao, Wong and Wang [7]. In their model the equivalence relation on the single universe is replaced by an arbitrary binary relation between two universes. We use this model in order to answer the following question arising in applications: If two finite universes U and W are connected by a fuzzy relation R, to what degree is a concept defined on U related to a concept defined on W? If concepts are precise (e.g., well defined categories of some external attributes observed or measured on U and W) they can be

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 268–277, 2010.

represented by crisp subsets of U and W. We focus on gradual evaluation of the relationship of a subset X from U to a subset Y from W. We evaluate the degree to which X is related (R-related) to Y at a level of importance $\alpha \in (0,1]$ by using the generalized rough set approximation of Y in knowledge base (approximation space) created by elements from W to which elements from X are related at least to the degree α.

In Section 2 we propose a rough extension of a fuzzy relation R between two universes to an interval-valued fuzzy relation between subsets from the two universes considering a level $\alpha \in (0,1]$. Then we introduce the notions of R_α-related sets, strongly R_α-related sets and R_α-compatible sets and we discuss some of their properties. Overall evaluation of the R-relationship between two related sets can be obtained by aggregation of their R_α-relationships over all $\alpha \in (0,1]$ by an appropriate aggregation function [2]. The choice of aggregation function depends on the area of application. In Section 3 we explain how R-related sets can be used in a quick (rough) assessment of the relationship between partitions of elements from two related universes. We conclude our paper with an illustrative example in Section 4.

2 Rough Extension of a Fuzzy Relation

2.1 Related Sets and Strongly Related Sets

Let U and W be finite nonempty universes of discourse. The family of all crisp subsets of U and the family of all fuzzy subsets of U will be denoted by 2^U and $\mathcal{F}(U)$, respectively. A fuzzy set $R \in \mathcal{F}(U \times W)$ is referred to as a fuzzy relation from U to W. The value of $R(x,y)$ gives the strength of the relationship of an element $x \in U$ to an element $y \in W$. A pair of sets $(X,Y) \in 2^U \times 2^W$ induces granule R_{XY} of R, which is the fuzzy relation R restricted to the Cartesian product $X \times Y \subset U \times W$. A numerical evaluation of granule R_{XY} gives information about the strength of the relationship between X and Y. Some evaluations based on aggregation of membership grades of R_{XY} by two aggregation functions were examined in [1].

In this paper we consider approximation of R by its crisp α-level sets R_α, $\alpha \in (0,1]$. Let us recall that $R_\alpha \subset U \times W$ is defined as follows: for all $(x,y) \in U \times W$, $R_\alpha(x,y) = 1$ if $R(x,y) \geq \alpha$ and $R(x,y) = 0$ otherwise. Then, regardless of the area of application, we say that x is related to y at level α (or simply R_α-related) if $R_\alpha(x,y) = 1$. For each $x \in U$ the R_α-neighborhood of x is the set

$$r_\alpha(x) = \{y \in W : R_\alpha(x,y) = 1\}. \tag{1}$$

Then, using the generalized rough set model [7], a set $Y \in 2^W$ can be represented in terms of elements from U which are R_α-related with elements in Y by pair of sets

$$\underline{R}_{\alpha,\,Y} = \{x \in U :\ r_\alpha(x) \subset Y\}, \tag{2}$$

$$\overline{R}_{\alpha,\,Y} = \{x \in U :\ r_\alpha(x) \cap Y \neq \emptyset\}. \tag{3}$$

The sets $\underline{R}_{\alpha,\,Y}$ and $\overline{R}_{\alpha,\,Y}$ are called the lower and the upper approximations of Y, respectively. The pair $(\underline{R}_{\alpha,\,Y}, \overline{R}_{\alpha,\,Y})$ is called a generalized rough set. Several other generalizations of the rough set model over two universes were suggested in [6]. In applications, α may be interpreted as the level of importance. Further in this paper we will use notations $|.|$ for cardinality of a set, N for the set of all positive integers and Y^c for the complement of Y. We propose the following rough extension of R at level $\alpha \in (0,1]$.

Definition 1. *Let $R \in \mathcal{F}(U \times W)$ and $\alpha \in (0,1]$. Then the mapping $\psi_R^\alpha : 2^U \times 2^W \to [0,1]^2$ defined for all $(X,Y) \in 2^U \times 2^W$ by $(\underline{\psi}_R^\alpha(X,Y), \overline{\psi}_R^\alpha(X,Y))$ where*

$$\underline{\psi}_R^\alpha(X,Y) = \frac{|\underline{R}_{\alpha,\,Y} \cap X|}{|X|}, \tag{4}$$

$$\overline{\psi}_R^\alpha(X,Y) = \frac{|\overline{R}_{\alpha,\,Y} \cap X|}{|X|}, \tag{5}$$

will be called the rough extension of R at level α.

Note that $\underline{\psi}_R^\alpha(X,Y)$ is the proportion of elements from X which have all their R_α-related elements from W included in Y. It can be interpreted as the degree to which X is strongly R_α-related to Y (or necessity that X is R-related to Y at level α.) On the other hand $\overline{\psi}_R^\alpha(X,Y)$ is the proportion of elements from X which have at least one of their R_α-related elements from W included in Y. It can be interpreted as the degree to which X is R_α-related to Y (or possibility that X is R-related to Y at level α.) In the case of single element sets $X = \{x\}$ and $Y = \{y\}$ we have that $\overline{\psi}_R^\alpha(x,y) = R_\alpha(x,y)$.

Definition 2. *Let $R \in \mathcal{F}(U \times W)$, $\alpha \in (0,1]$ and $(X,Y) \in 2^U \times 2^W$. We say that X is R_α-related to Y if $\overline{\psi}_R^\alpha(X,Y) = 1$. We say that X is strongly R_α-related to Y if $\underline{\psi}_R^\alpha(X,Y) = 1$.*

Proposition 1. *Let $R \in \mathcal{F}(U \times W)$. For all $(X,Y) \in 2^U \times 2^W$ and all $\alpha \in (0,1]$ the following hold:*
i) $\psi_R^\alpha(X,Y) = (1,1)$ if and only if for all $x \in X$: $R(x,y) \geq \alpha \Rightarrow y \in Y$,
ii) $\psi_R^\alpha(X,Y) = (0,1)$ if and only if for each $x \in X$ there exist $y_1 \in Y$ and $y_2 \in Y^c$ such that $R(x,y_1) \geq \alpha$ and $R(x,y_2) \geq \alpha$,
iii) $\psi_R^\alpha(X,Y) = (0,0)$ if and only if $R(x,y) < \alpha$ for all $(x,y) \in X \times Y$.

Proposition 2. *Let $\alpha \in (0,1]$. Consider fuzzy relations $R, R_1 \in \mathcal{F}(U \times W)$ and a pair of subsets $(X,Y) \in 2^U \times 2^W$. Then*
i) if $R_1(x,y) \leq R(x,y)$ for all $(x,y) \in X \times Y$ we obtain that

$$\overline{\psi}_{R_1}^\alpha(X,Y) \leq \overline{\psi}_R^\alpha(X,Y),$$

ii) if $R_1(x,y) \leq R(x,y)$ for all $(x,y) \in X \times Y$ and $R_1(x,y) \geq R(x,y)$ for all $(x,y) \in X \times Y^c$ we obtain that

$$\underline{\psi}_{R_1}^\alpha(X,Y) \leq \underline{\psi}_R^\alpha(X,Y).$$

Proposition 3. *Let $R \in \mathcal{F}(U \times W)$. For all $(X, Y) \in 2^U \times 2^W$ and all $\alpha \in (0, 1]$ we have that*

$$\underline{\psi}_R^{\alpha}(X, Y) + \overline{\psi}_R^{\alpha}(X, Y^c) = \overline{\psi}_R^{\alpha}(X, W). \tag{6}$$

From equation (6) it follows that if $\overline{\psi}_R^{\alpha}(X, W) = 1$, i.e., X is R_α-related to the whole universe W then

$$\underline{\psi}_R^{\alpha}(X, Y) = 1 - \overline{\psi}_R^{\alpha}(X, Y^c), \tag{7}$$

which is a well known relationship in rough set theory. This property can be used for evaluation of $\underline{\psi}_R^{\alpha}(X, Y)$ when $\overline{\psi}_R^{\alpha}(X, Y)$ is known. For convenient evaluation of $\overline{\psi}_R^{\alpha}(X, Y)$ we can use the formula

$$\overline{\psi}_R^{\alpha}(X, Y) = \frac{\sum_{x \in X} \max_{y \in Y}(R_\alpha(x, y))}{|X|}. \tag{8}$$

Proposition 4. *Consider $R \in \mathcal{F}(U \times W)$, $(X, Y) \in 2^U \times 2^W$ and $\alpha_1, \alpha_2 \in (0, 1]$ such that $\alpha_1 \leq \alpha_2$. Then $\overline{\psi}_R^{\alpha_1}(X, Y) \geq \overline{\psi}_R^{\alpha_2}(X, Y)$, and if $\overline{\psi}_R^{\alpha_1}(X, W) = \overline{\psi}_R^{\alpha_2}(X, W)$ then also $\underline{\psi}_R^{\alpha_1}(X, Y) \leq \underline{\psi}_R^{\alpha_2}(X, Y)$.*

Proposition 5. *Let $R \in \mathcal{F}(U \times W)$. For all $(X, Y) \in 2^U \times 2^W$, all $D \subset Y$ and all $\alpha \in (0, 1]$ we have that $\underline{\psi}_R^{\alpha}(X, D) \leq \underline{\psi}_R^{\alpha}(X, Y)$ and $\overline{\psi}_R^{\alpha}(X, D) \leq \overline{\psi}_R^{\alpha}(X, Y)$.*

From Proposition (5) it follows that if $\overline{\psi}_R^{\alpha}(X, Y) = 1$ then $\overline{\psi}_R^{\alpha}(X, Z) = 1$ for all $Z \supset Y$. Note that for a nonempty $C \subset X$ we may obtain $\overline{\psi}_R^{\alpha}(C, Y) \geq \overline{\psi}_R^{\alpha}(X, Y)$ or $\overline{\psi}_R^{\alpha}(X, C) \leq \overline{\psi}_R^{\alpha}(X, Y)$. However, if $\overline{\psi}_R^{\alpha}(X, Y) = 1$ then $\overline{\psi}_R^{\alpha}(C, Y) = 1$ for all $\emptyset \neq C \subset X$.

Proposition 6. *Let $R \in \mathcal{F}(U \times W)$. For all $(X, Y) \in 2^U \times 2^W$ and $\alpha \in (0, 1]$*

$$\overline{\psi}_R^{\alpha}(X, Y) = \begin{cases} 1 & \text{if } \alpha \leq \min_{x \in X}(\max_y(R(x, y)), \\ 0 & \text{if } \alpha > \max_{(x,y) \in X \times Y}(R(x, y)), \\ \delta \in (0, 1) & \text{otherwise.} \end{cases} \tag{9}$$

There is a variety of ways to obtain an overall evaluation $eval_R(X, Y)$ from gradual evaluations $eval_R^{\alpha}(X, Y)$, $\alpha \in (0, 1]$. For example, if for all $\alpha_1, \alpha_2 \in (0, 1]$ we have that $eval_R^{\alpha_1} \geq eval_R^{\alpha_2}$ when $\alpha_1 \leq \alpha_2$ then

$$eval_R(X, Y) = \max\{\alpha : eval_R^{\alpha}(X, Y) = 1\}. \tag{10}$$

If $eval_R^{\alpha_1} \leq eval_R^{\alpha_2}$ when $\alpha_1 \leq \alpha_2$ then

$$eval_R(X, Y) = \min\{\alpha : eval_R^{\alpha}(X, Y) = 1\}. \tag{11}$$

If we want to take into account different weights of $eval_R^{\alpha}(X, Y)$ at different levels α, we can use formula

$$eval_R(X, Y) = \sum_{t=1}^{k} (\alpha_t - \alpha_{t+1}) eval_R^{\alpha_t}(X, Y), \tag{12}$$

where $1 = \alpha_1 > \alpha_2 > \ldots > \alpha_k > \alpha_{k+1} = 0$, $k \in N$, is a sequence which includes all distinct membership grades of granule R_{XY}.

2.2 Compatible Sets

The inverse relation of $R \in \mathcal{F}(U \times W)$ is relation $R^{-1} \in \mathcal{F}(W \times U)$ such that $R^{-1}(y,x) = R(x,y)$ for all $(y,x) \in W \times U$. We may also evaluate the degree to which $Y \in 2^W$ is R_α^{-1}-related to $X \in 2^U$.

Definition 3. *Let $R \in \mathcal{F}(U \times W)$ and $\alpha \in (0,1]$. We say that $X \in 2^U$ and $Y \in 2^W$ are R_α-compatible if X is R_α-related to Y and Y is R_α^{-1}-related to X. We say that $X \in 2^U$ and $Y \in 2^W$ are strongly R_α-compatible if X is strongly R_α-related to Y and Y is strongly R_α^{-1}-related to X.*

The degrees of R_α-compatibility and strong R_α-compatibility between $X \in 2^U$ and $Y \in 2^W$ can be evaluated by coefficients

$$\overline{\gamma}_R^\alpha(X,Y) = \min\{\overline{\psi}_R^\alpha(X,Y),\ \overline{\psi}_{R^{-1}}^\alpha(Y,X)\}, \tag{13}$$

and

$$\underline{\gamma}_R^\alpha(X,Y) = \min\{\underline{\psi}_R^\alpha(X,Y),\ \underline{\psi}_{R^{-1}}^\alpha(Y,X)\}, \tag{14}$$

respectively.
The degree of R-compatibility and the degree of strong R-compatibility between X and Y can be evaluated by applying formula (12) to formulas (13) and (14), respectively.

In some applications, especially when R represents relationships between resources and goals, we might be interested in some special subsets of the related sets in question. We introduce the notions of the core of a set from one universe with respect to a set from another related universe.

Definition 4. *Let $R \in \mathcal{F}(U \times W)$, $\alpha \in (0,1]$ and $(X,Y) \in 2^U \times 2^W$. Then the subset of Y*

$$Y_{core}^\alpha = \bigcup_{x \in X} r_\alpha(x) \cap Y \tag{15}$$

will be called the R_α-core of Y with respect to X. The subset of X

$$X_{core}^\alpha = \bigcup_{y \in Y} r_\alpha^{-1}(y) \cap X \tag{16}$$

will be called the R_α^{-1}-core of X with respect to Y.

Proposition 7. *Let $R \in \mathcal{F}(U \times W)$, $\alpha \in (0,1]$ and $(X,Y) \in 2^U \times 2^W$. Then the following hold true:*

$$\overline{\psi}_R^\alpha(X,Y) = \overline{\psi}_R^\alpha(X, Y_{core}^\alpha), \tag{17}$$

and

$$\overline{\psi}_{R^{-1}}^\alpha(Y_{core}^\alpha, X) = 1. \tag{18}$$

Analogous properties hold for R_α^{-1}-core of X with respect to Y.

3 Related Partitions

Now we consider a partition P consisting of clusters $\{P_1, \ldots, P_k\}$ of elements from U and a partition Q created by clusters $\{Q_1, \ldots, Q_m\}$ of elements from W, $k, m \in N$. We can use the R-relationships of individual clusters from P to individual clusters from Q for evaluation of R-relationship of P to Q.

In applications we encounter two basic types of partitions. Firstly, we may have partitions where it is desirable to have the relationships of all clusters from P to all clusters from Q as strong as possible. In this case we can evaluate the relationship of P to Q by aggregation of overall evaluations

$$\{eval_R(P_i, Q_j),\ P_i \times Q_j \in P \times Q,\ i = 1, \ldots k,\ j = 1, \ldots, m\} \tag{19}$$

by an aggregation function (e.g., arithmetic mean, minimum or maximum). If $eval_R(P_i, Q_j)$ is the degree of R-compatibility between clusters P_i and Q_j then by aggregation of (19) we obtain the degree of R-compatibility between partitions P and Q.

Secondly, we may have partitions of U and W with the same number of clusters, say k, and we would like to have a good match of each cluster $P_i \in P$ to its corresponding cluster $Q_i \in Q$, $i = 1, \ldots, k$. Therefore we will evaluate the degree of R_α-matching of P to Q by aggregation of overall evaluations

$$\{eval_R(P_i, Q_i),\ P_i \times Q_i \in P \times Q,\ i = 1, \ldots, k\} \tag{20}$$

by a selected aggregation function. If $eval_R(P_i, Q_i)$ is the degree of R-compatibility between clusters P_i and Q_i then by aggregation of (20) we obtain the degree of R-matching between partitions P and Q.

4 Application

Quantitative evaluations of relationship between sets of related elements presented in the previous sections can be used in all applications where the following concept of gradual relationship is justified:

Considering $R \in \mathcal{F}(U \times W)$, a set $X \in 2^U$ is related to a set $Y \in 2^W$ at level $\alpha \in (0, 1]$, if each element $x \in X$ is related to at least one element $y \in Y$ to the degree $R(x, y) \geq \alpha$. Formally,

$$X \text{ is } R_\alpha\text{-related to } Y \Longleftrightarrow \min_{x \in X}(\max_{y \in Y} R(x, y)) \geq \alpha.$$

If in addition $\max(R(x, y), (x, y) \in X \times Y^c) < \alpha$, then X is strongly R_α-related to Y. Our method is based on simple calculations and provides clear interpretation of numerical results. We will illustrate it with the following example.

Example. The manager of a university tutoring center collected information from students attending the center and tutors of mathematics working in the center. Students were asked to describe their preferences for individual tutors on

Table 1. Preference of students for tutors

$x\backslash y$	y_1	y_2	y_3	y_4	y_5	y_6
x_1	0.6	0.5	0.4	0.6	0.4	1
x_2	0.8	0.7	0.5	0.2	0.5	0.6
x_3	0.5	0.7	1	0	0.2	0.3
x_4	0.6	0.7	0.9	0.3	0	0.2
x_5	0.4	0.6	1	0.8	0.7	0
x_6	0.1	0.3	0.6	0.9	0	0.3
x_7	1	0	0.5	0.5	0.5	0.5
x_8	0.7	0.2	0.8	0.9	0.2	1
x_9	0.3	0.6	0.2	0.4	0.3	0.7
x_{10}	0.5	0.4	0.2	0.9	0.6	0.7

the scale $[0, 1]$, where 0 means no preference and 1 indicates the full preference. The results from a small set of students $U = \{x_1, \ldots, x_{10}\}$ and a small set of tutors $W = \{y_1, \ldots, y_6\}$ are described by a fuzzy relation $R \in \mathcal{F}(U \times W)$ in Table 1. The manager also collected information about the major field of study of students and tutors, using categories SC=science and NS=non-science, the most convenient time for students to attend the center (M=morning, A=afternoon) and the most convenient time for tutors to work in the center (M, A) in the second half of the semester. He wants to answer the following questions:
Question 1. Are tutors from both groups (SC, NS) preferred reasonably well by both groups of students (SC, NS)? Namely, is each group of tutors preferred by each group of students at least to the degree 0.5?
Question 2. If two schedules of tutors are available for the second half of semester (M, A working hours) which one is a better match with preferences of students partitioned according to their most convenient time for attendance of the center?

Answer to Question 1: The threshold 0.5 can be interpreted as follows: It is expected that each student will have in each group of tutors at least one tutor preferred to the degree at least 0.5. If X is the set of students majoring in science and Y is the set of tutors with a science major, we need to check whether

$$\min(\max\{\alpha : \overline{\psi}_R^{\alpha}(X_i, Y_j) = 1\}, X_i \in \{X, X^c\}, Y_j \in \{Y, Y^c\} \geq 0.5.$$

In Table 2 we have results of evaluations of R_α-relationship of the set $X = \{x_1, x_2, x_3, x_4\} \subset U$ to the set $Y = \{y_1, y_2, y_3\} \subset W,\ \alpha \in (0, 1]\}$. Then

$$eval_R(X, Y) = \max\{\alpha : \overline{\psi}_R^{\alpha}(X, Y) = 1\} = 0.6.$$

Table 2. R_α-relationship of X to Y

α	$(0, 0.3]$	$(0.3, 0.6]$	$(0.6, 0.8]$	$(0.8, 0.9]$	$(0.9, 1]$
$\overline{\psi}_R^{\alpha}(X, Y)$	1	1	3/4	2/4	1/4

Analogously we can evaluate preferences of X to Y^c (non-science tutors), X^c (non-science students) to Y and finally X^c to Y^c. We have the following results: $eval_R(X, Y^c) = 0.3$, $eval_R(X^c, Y) = 0.5$ and $eval_R(X^c, Y^c) = 0.5$.

Because of the low degree of preference of students with science major for tutors with non-science major (0.3), the manager's expectations are not satisfied. However, evaluations by formula (10) are influenced by extreme values (outliers). The effect of outliers can be reduced by using for evaluation of the overall preference formula (12). Then

$$\overline{\psi}_R(X, Y) = (1 - .9)\frac{1}{4} + (0.9 - 0.8)\frac{2}{4} + (0.8 - .6)\frac{3}{4} + (0.6 - 0)1 = 0.825,$$

which is the weighted average proportion of students from X who have at least one preferred tutor in Y to the degree $\alpha \in (0, 1]$. The values of $\overline{\psi}_R$ for pairs of groups of students and groups of tutors are in Table 3. Then we conclude:

Table 3. Values of $\overline{\psi}_R$

students\tutors	Y	Y^c
X	0.825	0.55
X^c	0.75	0.80

Partition P of students according to their major field of study (science, non-science) is R-related to the partition Q of mathematics tutors (science majors, non-science majors) to the degree

$$\overline{\psi}(P, Q) = \min\{0.825, 0.55, 0.75, 0.916\} = 0.55.$$

This means that each group of students prefers each group of tutors at least to the degree 0.55, which agrees with the manager's expectations, if his threshold 0.5 is interpreted as follows: on average, at least 50 percent of students from each group will have in each group of tutors at least one tutor preferred to the degree $\alpha \in (0, 1]$. We can see that the highest preference is in the group of students with science majors for science mathematics tutors (0.825) followed by the preference of students with non-science majors for non-science mathematics tutors (0.8). Students with non-science majors show reasonable preference for science tutors (0.75), but the preference of students with science majors for non-science tutors is only moderate (0.55).

Answer to Question 2: We evaluate R-matching of partition P of students given by clusters $P_1 = \{x_3, x_4, x_5, x_6\}$ = morning hours and $P_2 = \{x_1, x_2, x_7, x_8, x_9, x_{10}\}$ = afternoon hours to partitions Q and T of tutors according to their working hours. Partition Q is created by clusters $Q_1 = \{y_2, y_3, y_4\}$ = morning and $Q_2 = \{y_1, y_5, y_6\}$ = afternoon hours. Partition T is created by clusters $T_1 = \{y_1, y_3, y_6\}$ = morning and $T_2 = \{y_2, y_4, y_5\}$ = afternoon hours.

Partition P matches partition Q to the degree

$$\overline{\mu}(P, Q) = \min\{\overline{\psi}_R(P_1, Q_1), \overline{\psi}_R(P_2, Q_2)\} = \min\{0.95, 0.90\} = 0.90.$$

The degree of strong matching of P to Q is

$$\underline{\mu}(P,Q) = \min\{\underline{\psi}_R(P_1,Q_1), \underline{\psi}_R(P_2,Q_2)\} = \min\{0.425, 0.2\} = 0.2.$$

On the other hand, P matches T to the degree

$$\overline{\mu}(P,T) = \min\{\overline{\psi}_R(P_1,T_1), \overline{\psi}_R(P_2,T_2)\} = \min\{0.875, 0.70\} = 0.70$$

and the degree of strong matching is

$$\underline{\mu}(P,Q) = \min\{\underline{\psi}_R(P_1,T_1), \underline{\psi}_R(P_2,T_2)\} = \min\{0.175, 0.03\} = 0.03.$$

Because $\overline{\mu}(P,Q) > \overline{\mu}(P,T)$ and also $\underline{\mu}(P,Q) > \underline{\mu}(P,T)$, the manager should choose schedule Q.

5 Conclusion

Evaluation of relationships between crisp sets proposed in this paper can be easily extended to evaluation of relationships between fuzzy sets. It is enough to take into account that each fuzzy set can be represented by its crisp α-level sets. Relationship between crisp or fuzzy concepts defined on two related universes can be used in variety of applications, especially in decision making and approximate reasoning.

Acknowledgments. This work was supported by Science and Technology Assistance Agency under the contract No. APVV-0375-06 and by the grant agency VEGA, grant number 1/0090/10.

References

1. Bodjanova, S.: Granules of a Fuzzy Relation. In: 5th International Summer School on Aggregation Operators, Palma, Mallorca, pp. 65–70 (2009)
2. Calvo, T., Kolesarova, A., Komornikova, M., Mesiar, R.: Aggregation Operators: Properties, Classes and Construction Methods. In: Calvo, T., Mayor, G., Mesiar, R. (eds.) Aggregation Operators, New Trends and Applications, pp. 3–104. Physica-Verlag, Heidelberg (2002)
3. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
4. Pawlak, Z., Skowron, A.: Rough Sets: Some Extensions. Information Sciences 177, 28–40 (2007)
5. Polkowski, L.: Rough Sets: Mathematical Foundations. Physica-Verlag, Heidelberg (2002)
6. Li, T.-J., Zhang, W.-X.: Rough Fuzzy Approximations on Two Universes of Discourse. Information Sciences 178, 892–906 (2008)
7. Yao, Y.Y., Wong, S.K.M., Wang, L.S.: A Non-numerical Approach to Uncertain Reasoning. International Journal of General Systems 23, 343–359

8. Yao, Y.Y., Wong, S.K.M., Lin, T.Y.: A Review of Rough Sets Models. In: Lin, T.Y., Cercone, N. (eds.) Rough Sets in Data Mining: Analysis for Imprecise Data, pp. 47–75. Kluwer, Boston (1997)
9. Yao, Y.Y.: Information Granulation and Approximation in a Decision-theoretical Model of Rough Sets. In: Pal, S.K., Polkowski, L., Skowron, A. (eds.) Rough-neural Computing: Techniques for Computing with Words, pp. 491–518. Springer, Berlin (2003)
10. Yao, J.T., Zhang, M.: Feature Selection with Adjustable Criteria. In: Ślęzak, D., Wang, G., Szczuka, M.S., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 204–213. Springer, Heidelberg (2005)
11. Zhong, N., Skowron, A.: A Rough Set-based Knowledge Discovery Process. International Journal of Mathematics and Computer Science 11, 603–619 (2001)
12. Ziarko, W.: Probabilistic Rough Sets. In: Ślęzak, D., Wang, G., Szczuka, M.S., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 283–293. Springer, Heidelberg (2005)

Combined Bayesian Networks and Rough-Granular Approaches for Discovery of Process Models Based on Vehicular Traffic Simulation

Mateusz Adamczyk, Paweł Betliński, and Paweł Gora

Institute of Informatics,
Warsaw University,
Banacha 2, 02-097 Warsaw, Poland
www.mimuw.edu.pl

Abstract. The aim of this paper is to summarize our experiments for discovering process of creation of traffic jams. These experiments were conducted during work on our master theses. We obtained data sets from the vehicular traffic simulator which were used to create a proper ontology based on the domain knowledge. The ontology was used as a schema for hierarchical classifier, which used Bayesian network created by genetic algorithm and rough sets based methods.

Keywords: Granules, ontology, domain knowledge, ontology approximation, hierarchical classifier, rough sets, Bayesian network, time window, process mining, cellular automaton, traffic modelling.

1 Introduction

Discovering knowledge from real-life data is a hard task. Concepts to be learned are vague, complex and often temporal. Their proper understanding requires domain knowledge, which is not present in a given data set.

In this paper we present method for investigating process of creation of traffic jam. First, we obtain data from the vehicular traffic simulator ([2]) then we encode domain knowledge in the ontology which describes hierarchy of vague concepts approximated by classifier. In our work, concepts and their approximations form granule and granules from one level form higher level granules. This schema form hierarchical classifier for predicting creation of traffic jams in the city. First two levels of the hierarchy use rough-based classifiers, and the last one – Bayesian network.

2 Traffic Simulation Framework

Traffic Simulation Framework (TSF) is a program for simulating traffic in cities. Its functionality was described in details in [5].

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 278–287, 2010.

Traffic simulation model is based on well-known model created by K. Nagel and M. Schreckenberg. It describes traffic on straight, one–way road, which is divided into some number of cells. Each cell may contain one car or may be empty. In each tick of discrete clock, cars change their position and speed with carefully chosen rules (see [8] or [5]). This model was broadly examined and generalized (see [4], [14]).

In TSF roads are modelled as a graph, which was created using maps from [12]. All edges are divided into cells which forms a cellular automaton with transition rules which are a generalization of these presented in [8]. TSF introduces support for crossroads and traffic lights. Moreover, each car is an autonomous agent which has start and end points of its journey, path connecting them and preferences regarding its road behaviour (see [5]). During our experiments these parameters were carefully chosen to model traffic during morning hours.

3 Vague Concepts, Granular Computing and Ontologies

Most of the time people use concepts, which are vague. Even when they seem natural, we cannot give their precise definition. Some remedy was found in methods, which laid foundations for granular computing: fuzzy sets (see [16]) and rough sets (see [13]). Unfortunately, simple classifiers based on these methods are not sufficient for real-life problems, where examined concept is ,,far" from given data in sense of some semantic distance.

One of solutions to that problem is a granular computing. Some objects and knowledge about them form so-called ,,granules". Simple granules can be grouped into little more complicated granule of higher level of abstraction. This process can be viewed as a human way of achieving data compression (see [10]).

Definition 1. Ontology *is a specification of a conceptualization ([6]).*

It can be used for modelling semantic layer. Usually it contains concepts, relations between them and other distinctions that are relevant for a given problem. It can be expressed in a natural language in some form of a dictionary, or in a more precise manner: in first order logic or as a graph (see [7]).

4 Proposed Ontology

Ontology that we propose has been described in details in [1]. Concepts in this ontology are organized in levels. They correspond to complexity of streets graph. First level describes one section of street (part of street between intersections). Although section itself is a spatial concept, its state changes over time. The state is characterized by the number of cars on section and their average velocity. Direction of their changes and average values are attributes of time windows — fixed time periods. These attributes determine vertex of section's behavioural graph in the current time window (see 1 (a)).

Street sections are grouped into crossroads (more precisely: parts of crossroads which could impact road lanes in examined direction) — the second level of

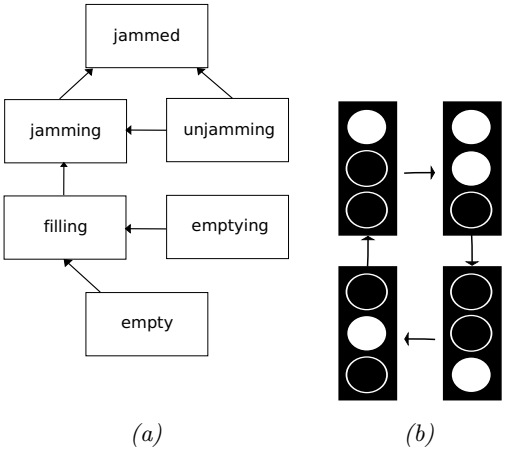


Fig. 1. Graphs used in ontology for describing behaviour: (a) Base graph for behaviour of part of city. Each concept was properly defined for examined case (section, type of crossroad, etc.). „Shortcuts" – edges which corresponds to paths in given graph were omitted for clarity. (b) Graph of traffic lights changes (for section which is entering crossroad in examined direction).

concepts hierarchy. Similarly to street sections, crossroads are observed in time window sequences. Behaviour of sections and state of relation form attributes of behavioural paths. They make possible to identify the state (vertex) of a given crossroad in crossroad's behavioural graph (see 1 (a, b)). Of course, concepts in vertices are dependent on geometry of observed crossroad.

Identified behaviours of examined crossroads form time series describing state of some part of city which were analysed using Bayesian network.

5 A Bayesian Network Classifier Predicting Traffic Jams

In this chapter we will present method of predicting traffic jam on a given street section using Bayesian networks. This method has been designed and implemented by Paweł Betliński (details can be found in [2]). Whole approach is a kind of process mining method (see [3], [11], [15]).

Suppose we have a street section between 2 consecutive crossroads, where we would like to predict a traffic jam. It seems that the main features which impact future traffic on this section are current situation on this section and in sections of direct entry and exit. The example is presented in figure 2 where we consider sections in the depth 1 and 2 from the investigated section.

The whole operation of the program will be illustrated in a situation such as in figure 2 (b), that is, with appeal to a depth 2. The aim is to predict any traffic jam, which appears in the section labelled C. However, we must first define notion of traffic jam, and also character of data flowing into the program from

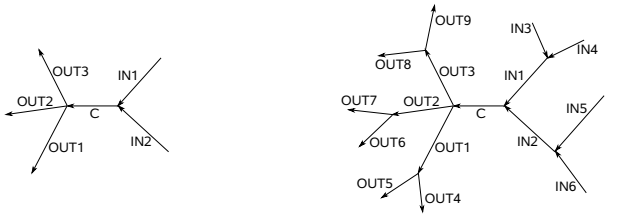


Fig. 2. Example of entry and exit sections to a depth 1 (a) and 2 (b)

simulations. This definitions will be here quite simple and have no connection with ontology area, but they were very useful in program testing.

Time series, which comes to the program after a simulation, is - keeping to our example of figure 2 (b), a data array of 15 attributes: C, $OUT1$, $OUT2$, ..., $OUT9$, $IN1$, $IN2$, ..., $IN6$. k-th array row registers situation of all 15 sections in k-th time segment of simulation - where each segment corresponds in real urban traffic to five minutes. This value of each attribute in k-th row is simply the average speed of all vehicles going through the street section corresponding to the attribute during k-th 'five minutes'. However, for simplicity, such average speed was not recorded directly in a table, but only the digit between 0 and 5 - depending on a segment, in which an average speed is located, namely as in table 3 (a).

So, for example, digit 3 in 17-th row and column corresponding to attribute C means that in 17-th 'five minutes' period of simulation on the street section denoted by C cars were driving with the average speed in range $(30, 40]$.

Construction of a meaningful definition of a traffic jam on a basis of sensory data is a separate issue, and this topic was elaborated in [1]. But for the purpose of the experiments there was created modest definition, stating that a traffic jam on a given street segment occurs when the average speed of vehicles travelling through that part within 'five minutes' of simulation is no more than 10 km/h - which corresponds to digit 0 in the table.

After constructing time series using simulation we placed the following objective: create a classification model, which would enable prediction of traffic jams in the section C, e.g. with 2 steps advance (1 step = 'five minutes'). Our solution is based on Bayesian networks (see [9]).

The approach depends on two basic parameters. The first is a size of time window. Time window is formed by cutting some number of rows (this number is exactly the size of the window) from data table described above. Such a time window is intended to be used as a basis for predicting a traffic jam which may occur on the street section C. The second fundamental parameter is the number of steps ahead of our predictions - i.e. on how many steps before appearance of traffic jam (0 value of attribute C) program should predict this situation.

Velocity segment (km/h)	sign
[0, 10]	0
(10, 20]	1
(20, 30]	2
(30, 40]	3
(40, 50]	4
above 50	5

(a)

d	p	k	S [%]	T [%]
1	2	2	25	51
1	2	4	29	57
1	2	6	33	59
1	1	2	35	62
1	1	4	41	68
1	1	6	46	71
1	0	2	44	70
1	0	4	50	75
1	0	6	54	79
2	2	2	48	73
2	2	4	55	80
2	2	6	59	81
2	1	2	57	77
2	1	4	60	86
2	1	6	64	88
2	0	2	61	82
2	0	4	68	89
2	0	6	69	93

(b)

Fig. 3. (a) Average speed segments. (b) Statistics S and T for all 18 experiments.

For example, using time window of the size 5 and predicting with 1 step in advance mean that the program predicts e.g. value 0 of attribute C in 23-th row based on time window from row 17 to 21. If program predicts with two steps advance, this would mean that it would have predicted this traffic jam based on time window from row 16 to 20, etc.

We would expect some time before a traffic jam in section C characteristic symptoms in a course of time series, so some specific value configuration of time window observed some number of steps before 0 value of attribute C. These characteristic configurations in time window gave the idea of applying Bayesian networks.

As a reminder, Bayesian network learned on a basis of some data table having attributes $X_1, ..., X_a$, $a \in \mathbb{N}$, is a directed, acyclic graph with vertices $X_1, ..., X_a$, along with parameters (assigned to each vertex) defining the appropriate conditional distribution of a given vertex provided its parents. Well-trained network, which means as sparse as it is possible, illustrates in its structure natural relationships in data - for the given vertex its parents in a network are direct causes. It is important also that sparse Bayesian network holds in a compressed manner joint distribution of attributes $X_1, ..., X_a$ (so for some big a it might be still possible to have a Bayesian network describing all distribution, while direct notation of this distribution needs for example $2^a - 1$ numbers in case of binary attributes). Such Bayesian network can be used later as a classifier, predicting value of some decision attribute provided given value of other attributes.

It is not a simple task to obtain a good network structure designed to describe distribution of data. Finding an optimal graph (relative to fixed criteria - for example, a commonly used Bayesian criterion) in a family of all graphs having a vertices becomes impossible task quite quickly with an increase of a. Paweł Betliński designed for this purpose the genetic algorithm, which was later applied to learning Bayesian networks analysing traffic (see [2]). Much emphasis in this genetic algorithm was placed on a careful design of cross-over and mutation operators. We omit description of this algorithm, focusing on the main program.

One of the ideas in our approach is to create a network that simply models distribution of characteristic for subsequent appearance of traffic jam value configurations in time window. This distribution generally can be very complicated, e.g. distribution in 100-dimensional space - if we consider time window of size 5×20, and this would be a huge problem for many other methods, for example clustering.

Now we go to a more formal description of the method. Suppose that k (size of time window) and p (on how many steps before a traffic jam we want to predict it) and $0 < \delta < 1$ are constants. The algorithm consists of two phases:

1. First phase:
   - Repeat m times (where m is a fixed parameter) a single simulation (every time with different parameters) until value of attribute C is 0 for the first time. Sensory data from simulation obtained for a fixed area (e.g. such as on figure 2 (b)) are converted to time series. From each of m generated time series we extract a time window of size k, which appears p steps before first traffic jam - that is first value 0 of attribute C.
   - After completing the above iteration, we have as the result m time windows, which can be now combined into one large data table. In the first row first time window is placed (we create one row from k rows of the time window, writing them one by one in time order), in the second row we put the second time window, and so on. Let's call created table as D_1. It has m rows and n columns, where n denote how many numbers form each time window.
   - We create Bayesian network BN_1 modelling distribution of all rows in table D_1. The aforementioned genetic algorithm finds network structure, and then standard method (counting from table D_1 frequencies expressing conditional distributions) is used to learn parameters of this graph.
2. Second phase:
   - We conduct many simulations, and for each simulation we capture from corresponding time series (for the same area as in the first phase) all time windows of size k, such that its value configuration has probability greater then δ, according to created Bayesian network BN_1. We save such time windows and values of attribute C p steps later after the window.
   - We combine all obtained windows in one table - in each row we place one time window as before, but we add one more (decision) column, which values indicate whether in sector C traffic jam appeared p steps after time window saved in this row: we write 1 - if it appeared, otherwise 0. This table has $n + 1$ columns and will be called D_2.

– We create Bayesian network BN_2 modelling distribution of all rows in table D_2 (we use the same method as in the first phase). BN_2, in conjunction with BN_1, is our final classifier, predicting traffic jam in sector C with p steps in advance.

Let's now consider a sense of this method. Result of the first phase is the Bayesian network BN_1. We know, that it should model distribution of value configurations in time windows (of size k), which occurs p steps before traffic jam in C. But in the first phase we observe only this time windows, after which we have traffic jam p steps later, so we do not have any knowledge about all other windows. By analyzing all windows we see, that some of them, which were considered as a reason of traffic jam, because they occurs often before it, in fact are not any reason, and, e.g., only in 10 percent cases after such window we have traffic jam. So the first phase is like a sieve - it eliminates an excess of time windows leaving only those, which potentially might be traffic jam reason. Value of decision is more balanced in this way - which greatly facilitates task of BN_2.

In the second phase we catch from all potential traffic jam reasons really important rules. We conduct simulations again and catch all time windows which are frequent (value configuration - read from Bayesian network - has probability in BN_1 greater then δ) and cause traffic jams in the future. For example we could set $\delta = 0.01$, and focus in this way on at most 100 most frequent windows appearing before traffic jam - and use later only them to create meaningful rules.

This rules determines Bayesian network BN_2, created on the basis of table D_2, which illustrates consequences of each interesting for BN_1 window (whether we have or not have traffic jam p steps later).

Networks BN_1 and BN_2 create a classification model, which can be used in practice. If we want to predict in step t_0 traffic jam in step t_0+1+p in sector C, then we first count from BN_1 the probability pr of value configuration in time window from step $t_0 - k + 1$ to step t_0. If $\delta \geq pr$ - we predict, that a traffic jam will not appear. Otherwise, we count from BN_2 the probability pr_1 of value configuration in our time window (from step t_0-k+1 to step t_0) combined with traffic jam appearance p steps later (so value 1 of last attribute - like in table D_2), and probability pr_0 of value configuration combined with no traffic jam p steps later (value 0 of last attribute). If $pr_1 > pr_0$ we predict, that traffic jam will appear, otherwise - we predict, that no.

6 Combined Approach

So far we described separate parts of our project, which we are working on. Idea of combining described approaches is very simple and based on granular computing. On every level of complexity one could apply Bayesian networks instead of rough sets based method. Our choice is to apply it on last level. So instead of using table with sensory data as a basis for Bayesian network inference, we simply use now data generated by first two levels of described ontology. This significantly reduces dimension of the problem, so Bayesian networks BN_1 and BN_2 can be learned faster. Moreover, obtained classifier can now reason

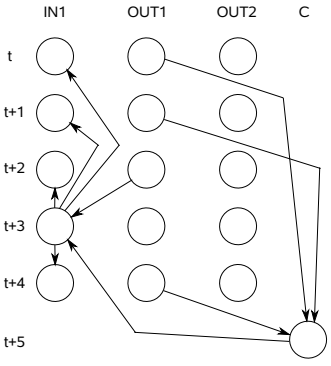


Fig. 4. Result: BN_2 Bayesian network

on higher level concepts, which increases its prediction power. Additionally, its representation can be now easier examined by human.

Figure 4 shows one of graph discovered by our approach. It represents Bayesian network, which describes relations between examined crossroads. As one can notice, state C of Ślasko-Dabrowski bridge is mostly determined by states of crossroad OUT1. This fact is consistent with our observations during simulations. Crossroad OUT2 was always almost empty and chosen by small number of cars, which were leaving the bridge. Additionally, we can see in graph, that the direct reason of state of crossroad IN1 in step t + 3 is situation on the bridge in step t + 5. It seems to be natural too, because cars which are in some step on the bridge were in previous steps at the crossroad IN1.

7 Experiments

Now we present the results of the experiment, in which sector C is Ślasko-Dabrowski bridge, with direction toward the center. For this experiment we used special type of simulation which is called 'Morning in Warsaw' (traffic simulates morning in the city, when people go from home to work).

The whole experiment actually consists of 18 smaller, appointed by choosing any combination of the following parameters:

- depth d of entry and exit sections in an analysis: 1 (illustrated in figure 5 (a)) and 2 (illustrated in figure 5 (b)). For both of these depths there were not taken all the sections for this depth. We selected for analysis only those, which tend to have dense traffic.
- size of a time window $k = 2, 4, 6$
- parameter $p = 0, 1, 2$ (how many steps before traffic jam we want to predict it).

(a) *(b)*

Fig. 5. Area on which program was tested: depth 1 (a) and depth 2 (b)

Common to all experiments were: $m = 1000$ - number of simulations in the first phase (which means that the table D_1 has 1000 rows), and simulations in second phase have been repeated until table D_2 reached 10000 rows. Parameter δ was more complicated: $\delta = (0.05)^{win\_vol}$, where win_vol means number of elements in a time window (height $\times$ width), so δ depends on d and k. When a time window increases (i.e. d or k increases) - δ decreases - this means we allow more time windows for the second phase. That is good, because when win_vol increases there are more possible value configurations of a time window.

In each of the 18 cases learned networks BN_1 and BN_2 were then tested to see how well they predict traffic jams. This test was based on 500 additional simulations. Each simulation was running until this step, when in the section C appears traffic jam (value 0 of attribute C). Results achieved during this test will be described by statistics, which we call S and T. Statistic S is the ratio of number of these traffic jam signals given by the classifier, which were correct, to the total number of traffic jam signals. So in other words statistic S tells us how often the classifier was right telling that in p steps there will appear a traffic jam in C. Statistic T is the ratio of number of those of the 500 traffic jams, which have been properly captured by the network p steps before, to 500 - the total number of traffic jams. In other words T tells us how often real traffic jam was detected in advance of p steps.

Table 3 (b) presents rounded results of the above statistics S and T for all 18 experiments. We summarize them as follows:

- Generally results for T are better than for S, which means that our classifier detects most of real traffic jams, but it also often detects traffic jam, when it actually doesn't appear.
- A deeper analysis with greater monitored area gives better results - in both statistics S and T.
- The larger window size, the better results of S and T.
- The larger parameter p, the more difficult is traffic jam to predict - which seems to be quite natural.

Acknowledgement

The research has been partially supported by the grants N N516 368334 and N N516 077837 from Ministry of Science and Higher Education of the Republic of Poland.

References

1. Adamczyk, M.: Ontologies and their approximations, master thesis, Warsaw University (2009)
2. Betliński, P.: Temporal association rules and their clustering, master thesis, Warsaw University (2009)
3. Bridewell, W., Langley, P., Todorovski, L., Dzeroski, S.: Inductive process modeling. Machine Learning 71, 1–32 (2008)
4. Chowdhury, D., Schadschneider, A.: Self-organization of traffic jams in cities: effects of stochastic dynamics and signal periods. Physical Review E (Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics) 59(2), 1311–1314 (1999)
5. Gora, P.: Traffic Simulation Framework - a cellular automaton-based tool for simulating and investigating real city traffic. In: Recent Advances in Intelligent Information Systems, Warsaw, pp. 642–653 (2009)
6. Gruber, T.: What is an Ontology?, http://www-ksl.stanford.edu/kst/what-is-an-ontology.html
7. Gruber, T.: Ontology, http://tomgruber.org/writing/ontology-definition-2007.htm
8. Nagel, K., Schreckenberg, M.: A cellular automaton model for freeway traffic. Journal de Physique, 2221—2229 (1992)
9. Neapolitan, R.E.: Learning Bayesian Networks. Prentice Hall, Upper Saddle River (2003)
10. Nguyen, H.S., Jankowski, A., Peters, J.F., Skowron, A., Szczuka, M., Stepaniuk, J.: Discovery of Process Models from Data and Domain Knowledge: A Rough-Granular Approach
11. Nguyen, H.S., Skowron, A., Stepaniuk, J.: Discovery of Changes along Trajectories Generated by Process Models Induced from Data and Domain Knowledge. In: Proceedings of the Workshop on Concurrency, Specification and Programming, vol. 3, pp. 350–362 (2008)
12. OpenStreetMap, http://www.openstreetmap.org
13. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177(1), 3–27 (2007)
14. Schadschneider, A.: The Nagel-Schreckenberg model revisited. The European Physical Journal B 10(3), 573–582 (1999)
15. Skowron, A., Suraj, Z.: Discovery of Concurrent Data Models from Experimental Tables: A Rough Set Approach. In: Proceedings of the The First International Conference on Knowledge Discovery and Data Mining, pp. 288–293. AAAI Press, Menlo Park (1995)
16. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 333–353 (1965)

On Scalability of Rough Set Methods

Piotr Kwiatkowski[1], Sinh Hoa Nguyen[2], and Hung Son Nguyen[1]

[1] Institute of Mathematics, Warsaw University,
Banacha 2, 02-097 Warsaw, Poland
[2] Polish-Japanese Institute of Inf. Technology,
Koszykowa 86, 02008, Warszawa, Poland
son@mimuw.edu.pl

Abstract. This paper presents some recent results of the research on the scalability of rough set based classification methods. The proposed solution is based on the close relationship between reduct calculation problem in rough set theory and association rule generation problem. This is a continuation of our previous results (see, e.g. [10] [11]). In this paper, the set of decision rules satisfying the test object is generated directly from the training data set. To make it scalable, we adopted the idea of the FP-growth algorithm for *frequent item-sets* [7], [6]. The experimental results on some benchmark data sets are showing the ability of the proposed solution to process a growing data sets.

Keywords: Data mining, Scalability, Rough set, Lazy learning.

1 Introduction

Classification of new unseen objects is the most important task in data mining. There are many classification approaches like "nearest neighbors", "naive Bayes", "decision tree", "decision rule set", "neural networks" etc. Every classification method has some advantages and disadvantages, hence the choice of classification methods in practical data mining applications depends on different criteria like: accuracy, description clearness, time and memory complexity etc.

This paper is related to the rule-based classification approach based on rough set theory. The typical manner consists of two basic steps: generalization and specification. In generalization step, some set of decision rules is constructed from data as a knowledge base. In the specialization step the set of rules that match a new object (to be classified) is selected and a conflict resolving mechanism will be employed to make the decision for the new object. This approach is quite common in classification methods based on rough set theory (see e.g., [2], [15], [17], [18]). The main reproach to the application of the present above method in data ming is related to the lack of scalability.

The biggest troubles stick in the rule induction step, where the potential number of all rules is exponential. Most of existing heuristics for eager rule induction have at least $O(n^2)$ time complexity, where n is the number of objects in the analyzed data set. However not all of the generated rules are later used in

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 288–297, 2010.

the classification phase. Moreover, the existing algorithms require multiple data scanning. In case of large decision table, the data must be held in a database system and the main problem is to minimize the number SQL queries used in the algorithm.

This paper we adopt the idea of FP-growth algorithms [7], [6] to implement the rough set based lazy classification method. The FP-growth algorithm is known as an efficient and scalable method for frequent pattern discovery from transaction data sets. We present the method called FDP, which is a modification of FP-growth, but is applicable for decision tables. We present the experimental results to confirm the advantages of the proposed method.

2 Basic Notions

In this Section, we recall some well known notions related to rough sets and classification systems.

An *information system* [12] is a pair $\mathbb{A} = (U, A)$, where U is a non-empty, finite set of *objects* and $A = \{a_1, ..., a_k\}$ is a non-empty finite set of *attributes (or features)*, i.e. $a_i : U \to V_{a_i}$ for $i = 1, ..., k$, where V_{a_i} is called *the domain of* a_i. Let $B = \{a_{i_1}, ..., a_{i_j}\} \subset A$, the set $INF_B = V_{a_{i_1}} \times ... \times V_{a_{i_j}}$ is called *information space defined by* B and the function $inf_B : U \to INF_B$ defined by $inf_B(u) = \langle a_{i_1}(u), ..., a_{i_j}(u)\rangle$ is called "B-information map".

Two objects $x, y \in U$ are called indiscernible by attributes from B if $inf_B(x) = inf_B(y)$. It has been shown that the indiscernibility relation, defined by $IND(B) = \{(x, y) : inf_B(x) = inf_B(y)\}$, is the equivalent relation (see [14]). For any $u \in U$, the set $[u]_B = \{x \in U : (x, u) \in IND(B)\}$ is called the indiscernibility class of u relative to B. Many notions in rough set theory can be defined on the basis of such classes. The main subject of rough set theory is concept description, which is the most important challenge in Data Mining. Any *concept* can be associated with the set of elements belonging to this concept. In rough set theory, any concept $X \subset U$ can be described by attributes from $B \subset A$ by the pair of B-lower approximation and B-upper approximation of X, respectively, where

$$\underline{B}X = \{u \in U : [u]_B \subset X\}, \qquad \overline{B}X = \{u \in U : [u]_B \cap X \neq \emptyset\}$$

Any information system of the form $\mathbb{A} = (U, A \cup \{dec\})$ with a distinguished attribute dec is called a *decision table*. The attribute $dec \notin A$ is called the *decision attribute* (or the decision, for short).

The classification problem can be formulated in terms of decision tables. Assume that objects from an universe $\mathbb{X}$ are classified into d classes by a decision function $dec : \mathbb{X} \to V_{dec} = \{1, ..., d\}$ which is unknown for learner. Every object from $\mathbb{X}$ is characterized by attributes from A, but the decision dec is known for objects from a sample set $U \subset \mathbb{X}$ only. The information about function dec is given by decision table $\mathbb{A} = (U, A \cup \{dec\})$. The problem is to construct from $\mathbb{A}$ a function $L_{\mathbb{A}} : INF_A \to V_{dec}$ in such a way that the probability $\mathbf{P}(\{u \in \mathbb{X} : dec(u) = L_{\mathbb{A}}(inf_A(u))\})$ is sufficiently high. The function $L_{\mathbb{A}}$ is

called *decision algorithm* or *classifier* and the methods constructing them from given decision table $\mathbb{A}$ are called *classification methods*.

2.1 Rough Sets and Classification Problem

In this paper, we are dealing with the classification method based on *minimal consistent decision rules*, which is preferred by many Rough Set based classification methods, e.g., [2], [15], [17], [18].

Let $\mathbb{A} = (U, A \cup \{dec\})$ be a decision table and $k \in V_{dec}$. By *decision rule* for k^{th} decision class we mean the expressions of form

$$(a_{i_1} = v_1) \wedge ... \wedge (a_{i_m} = v_m) \Rightarrow (dec = k) \tag{1}$$

where $a_{i_j} \in A$ and $v_j \in V_{a_{i_j}}$. For any decision rule $\mathbf{r}$ of form (1), the set of objects from U satisfying the assumption of $\mathbf{r}$ is called the carrier of $\mathbf{r}$ and is denoted by $[\mathbf{r}]$. By *length* and *support* of decision rule $\mathbf{r}$ we denote the number of descriptors and the number of objects satisfying the assumption of $\mathbf{r}$, i.e. $sup(\mathbf{r}) = |[\mathbf{r}]|$. The *confidence* of decision rule $\mathbf{r}$ is defined by $conf(\mathbf{r}) = \dfrac{|[\mathbf{r}] \cap DEC_k|}{|[\mathbf{r}]|}$. The decision rule $\mathbf{r}$ is called *consistent* with $\mathbb{A}$ if $conf(\mathbf{r}) = 1$. The decision rule $\mathbf{r}$ is called minimal consistent decision rule if it is consistent with $\mathbb{A}$ and any decision rule $\mathbf{r}'$ created from $\mathbf{r}$ by removing one of descriptors from left hand side of $\mathbf{r}$ is not consistent with $\mathbb{A}$.

The set of all minimal consistent decision rules for a given decision table $\mathbb{A}$, denoted by $MinConsRules(\mathbb{A})$, can be found by computing *object oriented reducts* (or local reducts) [8], [2] [17]. This paper is based on the boolean reasoning approach to local reducts [14], [8], [11].

The set $MinConsRules(\mathbb{A})$ can be used as a knowledge base in classification systems. In data mining philosophy, we are interested in extraction of *short and strong* decision rules with *high confidence*. The linguistic features like "short", "strong" or "high confidence" of decision rules can be formulated using of their length, support and confidence. In practice, instead of $MinConsRules(\mathbb{A})$, we are using the set of short, strong, and high accuracy decision rules defined by:

$$RULES(\mathbb{A}, \lambda, \sigma, \alpha) = \left\{ \mathbf{r}\text{: } length(\mathbf{r}) \leq \lambda,\ sup(\mathbf{r}) \geq \sigma \text{ and } conf(\mathbf{r}) \geq \alpha \right\}$$

All heuristics for object oriented reducts can be modified to induce the set $RULES(\mathbb{A}, \lambda, \sigma, \alpha)$ of decision rules.

Discretization of real value attributes is another important task in data mining, particularly for rule based classification methods. Empirical results show that the quality of classification methods depends on the discretization algorithm used in the preprocessing step. In general, discretization is a process of searching for a partition of attribute domains into intervals and unifying the values over each interval. Hence, the discretization problem can be defined as a problem of searching for a relevant set of cuts on the attribute domain.

In rough set theory, the optimal discretization problem has been transformed into a corresponding problem related to reducts of a new decision table [11].

The greedy algorithm for this approach, called MD-heuristic, has been implemented in RSES system. It has been shown that MD-heuristic for discretization is an efficient preprocessing method for rule based classifiers [2].

There are two main classification approaches called eager and lazy. The eager (or laborious) methods induce a generalized model from the input data (the generalization step) and uses extracted model to classify new objects (specialization step). Typical rule based classification methods consist of three phases:

1. **Learning phase:** generates a set of decision rules from a given decision table $\mathbb{A}$.
2. **Rule selection phase:** selects from $RULES(\mathbb{A})$ the set of rules supported by x. We denote this set by $MatchRules(\mathbb{A}, x)$.
3. **Post-processing phase:** makes decision for x using a voting algorithm for decision rules from $MatchRules(\mathbb{A}, x)$

In lazy learning approaches, new objects are classified without generalization step. For example, in kNN (k Nearest Neighbors) method, the decision of new object x can be made by weighting the decision of k nearest neighbors of x. In lazy decision tree methods, we are trying to reconstruct the path $p(x)$ of an "imaginable decision tree" that can be applied for new object x.

3 Scalable Classification Methods in Rough Sets

The scalability means the ability of an algorithm to process a growing input data. In this paper, we are interested in the scalability with respect to the growing number of training examples. The scalability is one of the most advisable properties of data mining algorithm.

Unfortunately, as we recalled before, the time and memory complexity of existing rule induction algorithms can not by applied to very large decision tables. That is why the rule induction methods based on rough sets are criticized due to the lack of scalability. We will show that some of them can be implemented in such a way that they become more scalable in the client-server environment.

The first proposition is related to discretization methods. The idea was based on using "divide and conquer" technique to localize the cut that is very close

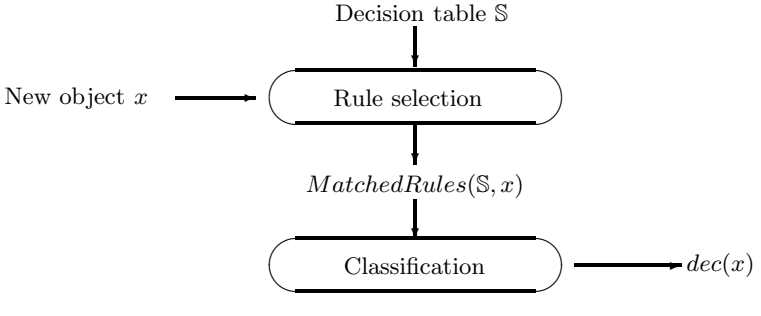


Fig. 1. The lazy rule-based classification system

to the optimal with respect to discernibility measure. It has been shown that it can be done by using only $O(\log n)$ simple SQL queries, where n is the number of objects. This technique has been generalized for other measures.

The second proposition is based on a tricky implementation of lazy rule-based classification approach presented in Fig. 1. In general, lazy methods need more time complexity for the classification step, i.e., the answer time for the question about decision of a new object is longer than in eager classification methods. But lazy classification methods are *well scalable*, i.e., they can be realized for larger decision table using distributed computer system [4], [3], [13], [11].

In other words, we will extract the set of decision rules covering the object x directly from data without explicit rule generation. We show that this diagram can work for the classification method described in previous section using the set of decision rules from $MinRules(\mathbb{S}, \lambda, \sigma, \alpha)$. Formally, the problem is formulated as follows: *given a decision table* $\mathbb{S} = (U, A \cup \{dec\})$ *and a new object* x*, determine the set* $MatchRules(\mathbb{S}, x) = \{\mathbf{r} \in MinRules(\mathbb{S}, \lambda, \sigma, \alpha) : x \text{ satisfies } \mathbf{r}\}$ *of decision rules that can be classified by* x. In [10], we proved that the set $MatchRules(\mathbb{S}, x)$ can be calculated by a modification of *Apriori algorithm* proposed in [1] for frequent item set generation from data bases. Another searching method for $MatchRules(\mathbb{S}, x)$, is based on FP-growth algorithm [7], [6], has been proposed in [9].

The FP-growth algorithm is known as an efficient and scalable method for frequent pattern discovery from transaction data sets. We present the method called FDP, which is a modification of FP-growth, but it is applicable for decision tables. This method consists of the following steps:

- Construction of the data structure called $FDP(x)$ (Frequent Decision Pattern tree). This step requires only two data scanning passes:
  - The first scanning pass is required to calculate the frequencies of descriptors from $inf_A(x)$. After the first data scan, these descriptors are ordered with respect to their frequencies. The low-frequent descriptors are useless in constructing strong decision rules and can be removed. Let $DESC(x)$ be the resulting list of frequent descriptors.
  - In the second scanning pass, each training object u is converted into a list $D(u)$ of frequent descriptors from $DESC(x)$ that occur in $inf_A(u)$, and then we insert the list $D(u)$ into the data structure $FDP(x)$.
- Generation of the set of frequent decision rules from $FDP(x)$ by a recursive procedure. This step does not guarantee the minimality of the obtained rules (some rules are still reducible)
- Insert the obtained rules into a data structure called *the minimal rule tree* – denoted by $MRT(x)$ – to get the set of irreducible decision rules. This data structure can be used to perform different voting strategy.

As we see later, the key concept in this method is the FDP tree structure. In fact, similarly to the original FP-tree, FDP is the prefix tree for the collection of ordered list of descriptors. But, unlike FP-tree, each node in FDP tree consists of four fields: *descriptor_name, support, class_distribution and node_link*, where *descriptor_name* is the name of descriptor, *support* is the number of training objects that contain all descriptors on the path from the root to the current

$\mathbb{A}$	a_1	a_2	a_3	a_4	dec
ID	outlook	temp.	hum.	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no
x	sunny	mild	high	TRUE	?

$\Longrightarrow$

ID	descriptor lists	dec
1	d3, d1	no
2	d3, d4, d1	no
3	d3	yes
4	d3, d2	yes
5		yes
6	d4	no
7	d4	yes
8	d3, d2, d1	no
9	d1	yes
10	d2	yes
11	d2, d4, d1	yes
12	d3, d2, d4	yes
13		yes
14	d3, d2, d4	no

Fig. 2. A decision table $\mathbb{A}$ and test object x

node, *class_distribution* is the detail support for each decision class and *node_link* are used to create list of nodes of the same descriptor.

Example. The detailed definitions and algorithms for this method were described in [9]. Because of the space limitation, we will illustrate the proposed method by the following example.

Let us illustrate our concept for the *golf data set* presented in Figure 2 (left). The test object induces four descriptors: $d1 : a1 = sunny$, $d2 : a_2 = mild$, $d3 : a_3 = high$ and $d4 : a_4 = TRUE$. Thus we can fix the order of descriptors as follow: $DESC(x) = [d3, d2, d4, d1]$. After the first data scan, the training objects can be rewritten as presented in Fig. 2. The corresponding FDP tree for this collection of frequent descriptor lists is shown in Fig. 3. In order to generate decision rules from the FDP tree, one can apply the FDP-growth algorithm which is the modification of FP-growth algorithm [7], [6]. The detail of this step has been described in [9]. In this example, we can obtain the following set of 4 decision rules.

1	(outlook = sunny) $\wedge$ (hum. = high) $\Rightarrow$ $play = no$
2	(outlook = sunny) $\wedge$ (temp. = mild) $\wedge$ (windy = TRUE) $\Rightarrow$ $play = yes$
3	(outlook = sunny) $\wedge$ (temp. = mild) $\wedge$ (hum. = high) $\Rightarrow$ $play = no$
4	(outlook = sunny) $\wedge$ (hum. = high) $\wedge$ (windy = TRUE) $\Rightarrow$ $play = no$

One can see that this is not the set of irreducible decision rules, because, the rules number 3 and 4 are the extensions of rule nr 1. To reduce the set of rules one can use the additional data structure called MRT (minimal rule tree). In fact, MRT is the modification of FPMAX tree, the data structure for extraction of maximal frequent patterns, presented in [5]. The following figure illustrates the resulting MRT tree after inserting all decision rules. After all steps, one can obtain two minimal decision rules presented in Fig. 4.

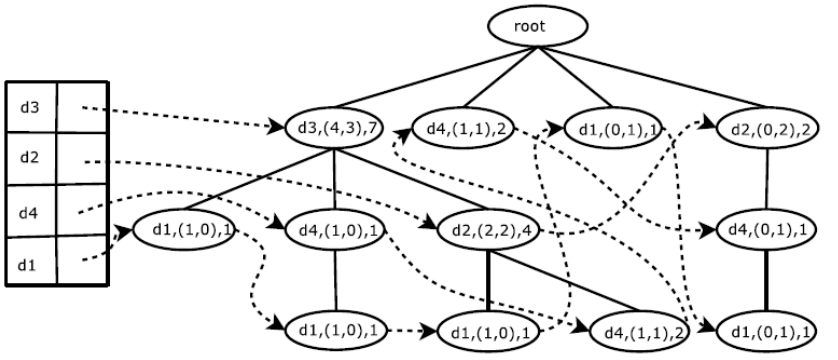
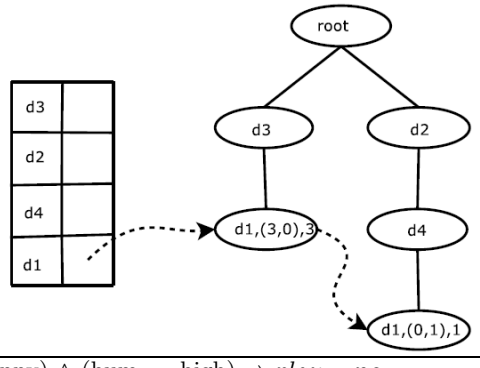


Fig. 3. The FDP tree for the object x from Table 2 and the set of decision rules extracted from the tree



1	(outlook = sunny) $\wedge$ (hum. = high) $\Rightarrow$ $play = no$
2	(outlook = sunny) $\wedge$ (temp. = mild) $\wedge$ (windy = TRUE) $\Rightarrow$ $play = yes$

Fig. 4. The MRT tree for the previous set of decision rules

4 Experimental Results

The FDP-growth algorithm was implemented and tested on data sets from UCI Machine Learning Repository. We compared the accuracy of FDP-growth algorithm with other lazy classifiers: IBk (nearest neighbors classifier) and LBR (Naive Bayes classifier) which are available in WEKA [16]. All experiments were done on PC with dual Processor Athlon X2 4000+ (2 x 2.1GHz) and 4GB RAM.

The first experiment was performed on the *Poker-hand* data set. This data set consists of 10 conditional attributes and 9 decision classes. The training data set consists of 25010 instances, while the test data contain 1000000 instances.

In order to verify the scalability of the proposed solution, we switched the role of this data sets. The experiments were performed on training data sets of different sizes: 10000, 20000, 50000, 100000, 200000, 500000 and 1000000. The accuracy of classifiers were estimated on the sample of 1000 testing instances.

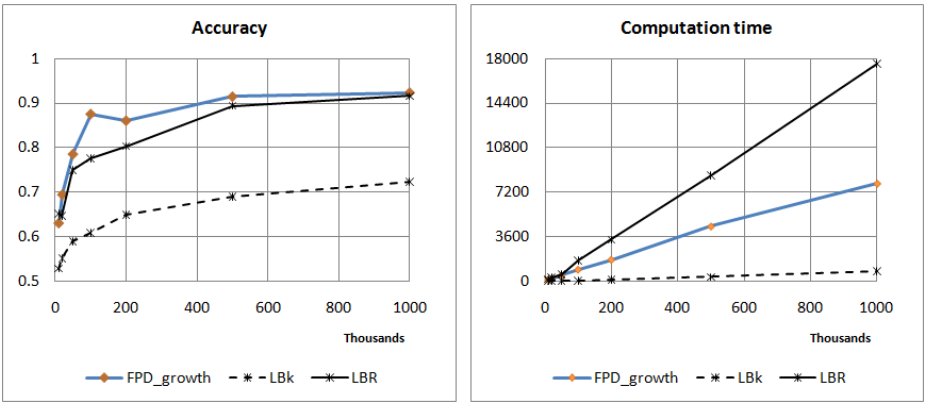


Fig. 5. Comparing the accuracy and computation time of three lazy classifiers for poker-hand data

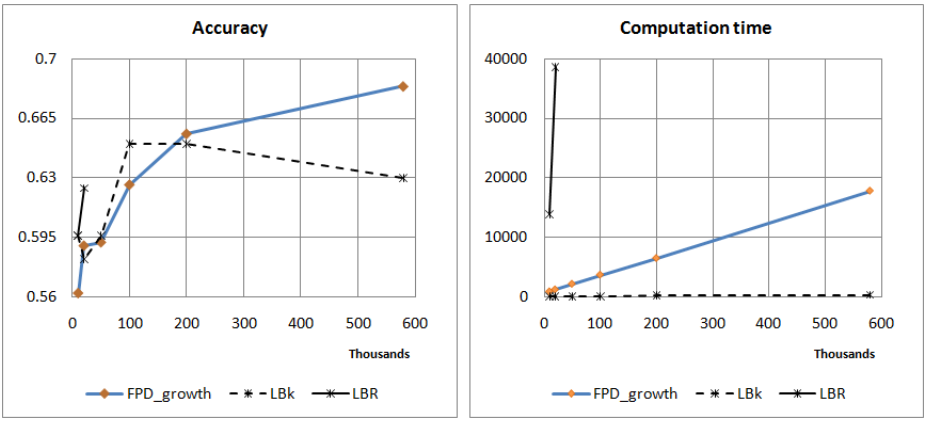


Fig. 6. Comparing the accuracy and computation time of three lazy classifiers for Forest Covertype data

The second experiment has been done on the *Forest Covertype* data set. This data set consists of 581012 objects 54 conditional attributes and 7 decision classes. A sample of 580000 instances was used as the training set. In order to verify the scalability of the proposed solution, the experiments were performed on training data sets of different sizes: 10000, 20000, 50000, 100000, 200000 and 580000. The accuracy of classifiers were estimated on the sample of 500 unseen instances.

The third experiment has been done on the data set called *Pen-Based Recognition of Handwritten Digits*. This is a small data set consisting of 7494 training objects and 3699 test objects, 16 conditional numeric attributes and 10 decision classes. The experiments were performed on the samples of sizes: 7494, 5000,

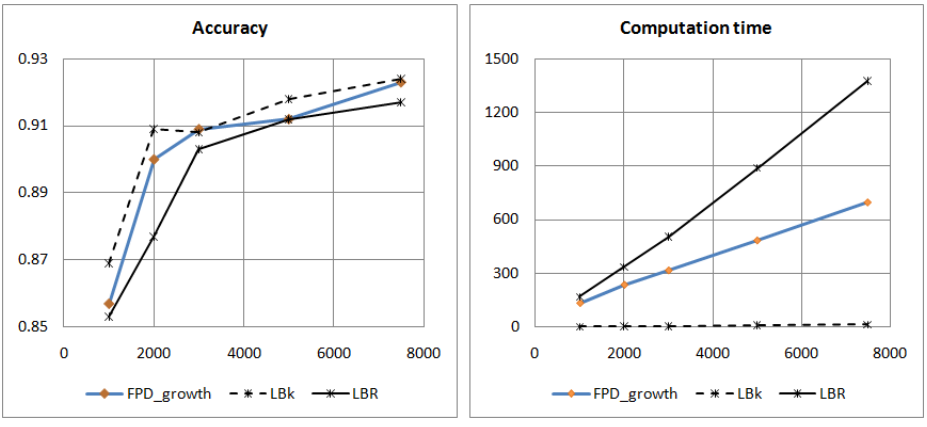


Fig. 7. Comparing the accuracy and computation time of three lazy classifiers for Pen-Based Recognition of Handwritten Digits data

3000, 2000, 1000. The accuracy of classifiers were estimated on the sample of 500 instances from the test data set.

On Fig. 5, Fig. 6 and Fig. 7, we present the detailed results of the described above experiments. The plots of classification accuracy for different training data sizes are presented on the left hand side, while the plots of computation time are presented in the right hand side of these figures.

5 Conclusions

We have presented a scalable lazy classifier which is is a rough set based classifier. We have modified the FP-growth algorithm to calculate the set of minimal decision rules for test objects. The experiment results are showing that the computation time seems to be linearly depends on the size of training set. The proposed method can be easily implemented in the distributed computer system.

If we are constructing the FDP-tree for each training object, we will have an algorithm for object oriented reducts calculation for the decision table. Hence the proposed method can be also applied for eager learning. We plan to use the data structures proposed in this paper to develop efficient incremental learning methods for stream data.

Acknowledgement. The research has been partially supported by grants N N516 368334 and N N516 077837 from Ministry of Science and Higher Education of the Republic of Poland.

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of association rules, Menlo Park, CA, USA. American Association for Artificial Intelligence, pp. 307–328 (1996)

2. Bazan, J.G.: A comparison and non-dynamic rough set method for extracting laws decision tables. In: Polkowski, L., Skowron, A. (eds.) Rough Sets in Knowledge Discovery 1. Methodology and Applications. Studies in Fuzziness and Soft Computing, pp. 321–365. Physica-Verlag, Heidelberg (1998)
3. Bondi, A.B.: Characteristics of scalability and their impact on performance. In: WOSP 2000: Proceedings of the 2nd international workshop on Software and performance, pp. 195–203. ACM, New York (2000)
4. Fayyad, U.M., Haussler, D., Stolorz, P.E.: Mining scientific data. Commun. ACM 39(11), 51–57 (1996)
5. Grahne, G., Zhu, J.: High performance mining of maximal frequent itemsets. In: Proceedings of 6th International Workshop on High Performance Data Mining, HPDM 2003 (2003)
6. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, San Francisco (2000)
7. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Chen, W., Naughton, J., Bernstein, P.A. (eds.) 2000 ACM SIGMOD Intl. Conference on Management of Data, May 2000, pp. 1–12. ACM Press, New York (2000)
8. Komorowski, H.J., Pawlak, Z., Polkowski, L.T., Skowron, A.: Rough Sets: A Tutorial, pp. 3–98. Springer, Singapore (1999)
9. Kwiatkowski, P.: Scalable classification method based on FP-growth algorithm (in Polish). Master's thesis, Warsaw University (2008)
10. Nguyen, H.S.: Scalable classification method based on rough sets. In: Alpigini, J.J., Peters, J.F., Skowron, A., Zhong, N. (eds.) RSCTC 2002. LNCS (LNAI), vol. 2475, pp. 433–440. Springer, Heidelberg (2002)
11. Nguyen, H.S.: Approximate boolean reasoning: Foundations and applications in data mining 4100, 334–506 (2006)
12. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Theory and decision library. D: System theory, knowledge engineering and problem solving, vol. 9. Kluwer Academic Publishers, Dordrecht (1991)
13. Shafer, J.C., Agrawal, R., Mehta, M.: Sprint: A scalable parallel classifier for data mining. In: Vijayaraman, T.M., et al. (eds.) VLDB 1996, Proceedings of 22nd International Conference on Very Large Data Bases, Mumbai, India, September 3-6, pp. 544–555. Morgan Kaufmann, San Francisco (1996)
14. Skowron, A., Rauszer, C.M.: The discernibility matrices and functions in information systems, ch. 3, pp. 331–362. Kluwer Academic Publishers, Dordrecht (1992)
15. Stefanowski, J.: On rough set based approaches to induction of decision rules. In: Polkowski, L., Skowron, A. (eds.) Rough Sets in Knowledge Discovery 1. Methodology and Applications. Studies in Fuzziness and Soft Computing, pp. 500–529. Physica-Verlag, Heidelberg (1998)
16. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
17. Wroblewski, J.: Covering with reducts - a fast algorithm for rule generation. In: Polkowski, L., Skowron, A. (eds.) RSCTC 1998. LNCS (LNAI), vol. 1424, pp. 402–407. Springer, Heidelberg (1998)
18. Ziarko, W.: Rough sets as a methodology for data mining. In: Polkowski, L., Skowron, A. (eds.) Rough Sets in Knowledge Discovery 1. Methodology and Applications. Studies in Fuzziness and Soft Computing, pp. 554–571. Physica-Verlag, Heidelberg (1998)

Interestingness Measures for Association Rules within Groups

Aída Jiménez, Fernando Berzal, and Juan-Carlos Cubero

Dept. Computer Science and Artificial Intelligence,
ETSIIT - University of Granada, 18071, Granada, Spain
{aidajm,jc.cubero,fberzal}@decsai.ugr.es

Abstract. The study of association rules within groups of individuals in a database is interesting to define their characteristics and their behavior. In this paper, we define group association rules and we study interestingness measures for them. These evaluation measures can be used to rank groups of individuals and also rules within each group.

1 Introduction

Association rules have been used to analyze the relationships among the frequent itemsets in transactional and relational databases. Let $I = \{I_1, I_2, ..., I_m\}$ be a set of items. Let D be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Let S be a set of items. A transaction T is said to contain S if and only if $S \subseteq T$ [1]. An **association rule** is an implication of the form $A \Rightarrow C$, where $A \subseteq I$, $C \subseteq$ I , and $A \cap C = \emptyset$.

Databases can naturally contain groups of individuals that share some characteristics [2]. For example, in a census database, we can define groups of individuals according to their sex, their marital status, whether they have children, or even by combining several of such features. *Men who have children* are an example of such a group.

In this paper, we will describe how association rules can be defined in those groups to study the features that individuals in the same group have in common. For example, in the group of *men who have children*, we could find an association rule *lives in suburbs* $\Rightarrow$ *owns 2 car*, which is interpreted as *men who have children* and live in suburbs own 2 cars with some confidence value. If we obtained more rules, we could characterize the *men who have children* group and we could compare their behavior to the behavior of other groups in the database, for example, *men who are single.*

We define a **group** as a set of items $G = \{G_1, G_2, ..., G_n\}$ such that $G \subseteq I$. A **group association rule** $G : A \Rightarrow C$ is an association rule $A \Rightarrow C$ defined over the group G. In other words, a group association rule $G : A \Rightarrow C$ is equivalent to the classical association rule $GA \Rightarrow C$.

In this paper, we will describe how to adapt some of the interestingness measures that have been defined for association rules [3] [4] to group association rules and how these modified measures will help us to rank the different groups in a database in order to highlight the most interesting ones [5].

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 298–307, 2010.

Our paper is organized as follows. In Section 2, we describe some rule evaluation metrics proposed in the literature. Section 3 introduces interestingness measures for group association rules. In Section 4, we explain how to order groups and group association rules within each group. Finally, we end our paper with some conclusions in Section 5.

2 Interestingness Measures for Standard Association Rules

The classical measures used to characterize an association rule are its support and its confidence [6, 1].

Definition 1. *The support of an itemset X in the database D is defined as the percentage of transactions that contain X, i.e.,*

$$supp(X) = P(X).$$

Definition 2. *The rule $A \Rightarrow C$ holds in the transaction set D with **support** s, where s is the percentage of transactions in D that contain $A \cup C$, i.e.,*

$$supp(A \Rightarrow C) = P(A \cup C).$$

Definition 3. *The rule $A \Rightarrow C$ has **confidence** c in the transaction set D, where c is the percentage of transactions in D containing A that also contain C, i.e.,*

$$conf(A \Rightarrow C) = P(C|A) = \frac{supp(A \Rightarrow C)}{supp(A)}$$

Confidence has some drawbacks as we can see in the example shown in Figure 1 where we have a graphical representation of two rules, $A \Rightarrow B$ and $A \Rightarrow C$. In the case of the $A \Rightarrow B$ rule, we have the following support values for the intervening itemsets: supp(A)= 28%, supp(B)=38%, and supp($A \cup B$) = 21%. Therefore, the confidence for the $A \Rightarrow B$ rule is 75%. In the case of the $A \Rightarrow C$ rule, even though the support of the consequent changes (supp(C)=85%), the confidence value of the $A \Rightarrow C$ rule is also 75%.

In the first case, B was present in 38% of the transactions in the database and its presence increases to 75% in transactions where A is also present. In the second case, however, the presence of the A reduces the presence of C, from 85% to 75%. Therefore, the confidence measure does not let us distinguish between these two cases.

In conclusion, confidence does not take into account the support of the rule consequent, hence it is not able to detect negative dependencies between items. Several measures have been proposed in the literature as alternatives to the support and confidence measures [3]. In the following paragraphs, we describe some of them:

Definition 4. *The interest of the rule $A \Rightarrow C$, also known as lift [7], is defined as:*

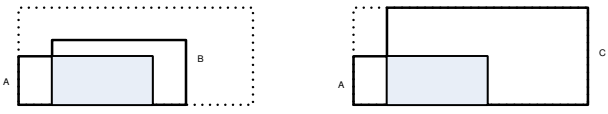


Fig. 1. Graphical depiction of two rules, $A \Rightarrow B$ and $A \Rightarrow C$, both with the same confidence but different consequent support

$$int(A \Rightarrow C) = \frac{supp(A \Rightarrow C)}{supp(A)supp(C)}$$

Interest measures how many times more often A and B occur together than expected if they where statistically independent. Values above 1 indicate positive dependence, while those below 1 indicate negative dependence. The interest of the $A \Rightarrow B$ and $A \Rightarrow C$ rules in the the aforementioned example is $int(A \Rightarrow B)$=4.2 and $int(A \Rightarrow C)$=0.91. Here, $int(A \Rightarrow B) > int(A \Rightarrow C)$, which correspond to our intuition that $A \Rightarrow B$ is more interesting than $A \Rightarrow C$.

Interest measures the degree of dependence between the itemsets. However, it only measures co-occurrence, but not the implication direction because it is a symmetric measure, i.e., $int(A \Rightarrow C)$=$int(C \Rightarrow A)$.

Definition 5. *The conviction[8] of the rule $A \Rightarrow C$ is defined as:*

$$conv(A \Rightarrow C) = \frac{supp(A)supp(\neg C)}{supp(A \cup \neg C)}$$

The advantage of conviction with respect to the confidence measure is that it takes into account both the support of the antecedent and the support of the consequent of the rule. Conviction values in the (0,1) interval mean negative dependence, values above 1 mean positive dependence, and a value of 1 means independence, as happened with the interest measure.

In the example of Figure 1, $supp(\neg B)$=0.62 and $supp(A \cup \neg B) = 0.07$. Therefore, the conviction of the $A \Rightarrow B$ rule is 2.48. In the $A \Rightarrow C$ rule, $supp(\neg C)$=0.15 and $supp(A \cup \neg C) = 0.07$. Therefore, $conv(A \Rightarrow C)$=0.6, which means negative dependence.

Unlike interest, rules that hold 100%, like the *Vietnam veteran* $\Rightarrow$ *more than 5 years old* rule, have the highest possible conviction value of ∞, which is an useful property. If 5% of the people are Vietnam veterans and 90% are more than five years old, then the interest of the *Vietnam veteran* $\Rightarrow$ *more than 5 years old* rule is (0.05)/(0.05)*0.9 = 1.11, sightly above 1, which is the value that would indicate statistical independence [8].

The main drawback of the conviction measure is that is not bounded, i.e., its range is [0,∞]. Therefore, it is difficult to establish a conviction threshold.

Let us now define the gain of a rule as the difference between its confidence and the support of its consequent. Formally,

Definition 6. *The gain of a rule $A \Rightarrow C$ is defined as:*

$$gain(A \Rightarrow C) = conf(A \Rightarrow C) - supp(C).$$

Fig. 2. Graphical examples illustrating the gain (and the certainty factor) of the rules derived from the scenarios represented in Figure 1: $A \Rightarrow B$ (left) and $A \Rightarrow C$ (right)

The gain values for the rules in Figure 1 is $gain\ (A \Rightarrow B)$=0.75-0.38=0.37 and $gain\ (A \Rightarrow C)$=0.75-0.85=-0.10. Figure 2 graphically shows these values. The length of the arrows represents the gain of the rules, i.e., the increase $(A \Rightarrow B)$ or decrease $(A \Rightarrow C)$ in the presence of the consequent given that A is present.

Definition 7. *The certainty factor[9] of a rule $A \Rightarrow C$ is defined as:*

$$CF\ (A \Rightarrow C) = \frac{gain(A \Rightarrow C)}{1 - supp(C)}\ \textit{if}\ gain(A \Rightarrow C) \geq 0,\ \textit{and}$$

$$CF\ (A \Rightarrow C) = \frac{gain(A \Rightarrow C)}{supp(C)}\ \textit{if}\ gain(A \Rightarrow C) < 0.$$

The certainty factor is the gain value normalized into the [-1,1] interval.

The certainty factor is interpreted as a measure of the variation of the probability that C is in a transaction when we consider only those transactions where A is. More specifically, a positive CF measures the decrease of the probability that C is not in a transaction, given that A is.

In the example of Figure 1 a), the CF of the $A \Rightarrow B$ rule is 0.37/(1-0.38)=0.60 while the CF for the $A \Rightarrow C$ rule is -0.10/0.85=-0.12.

3 Interestingness Measures for Group Association Rules

In the following paragraphs, we will explain how to adapt the measures described in Section 2 to group association rules, as well as how these new measures can be useful to evaluate this kind of association rules.

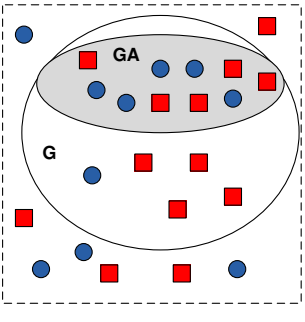


Fig. 3. Graphical representation of a group G

3.1 Group Support

Definition 8. *The support of an itemset X in the group G is the percentage of transactions in G that contain X, i.e.,*

$$supp_G(X) = \frac{P(XG)}{P(G)} = conf(G \Rightarrow X).$$

Figure 3 shows the representation of a group G in an example dataset. The support of *circles (•)* in the group G is $supp_G(\bullet)$= 6/15=0.4.

Definition 9. *The support of the group association rule $G : A \Rightarrow C$ is defined as:*

$$supp_G(A \Rightarrow C) = \frac{P(GAC)}{P(G)} = conf(G \Rightarrow AC).$$

In the previous example, the support of the group association rule $G : A \Rightarrow \bullet$ is $supp_G(A \Rightarrow \bullet) = 5/15 = 0.33$.

3.2 Group Confidence

Definition 10. *The confidence of the group association rule $G : A \Rightarrow C$ is defined as:*

$$conf_G(A \Rightarrow C) = \frac{supp_G(A \Rightarrow C)}{supp_G(A)} = conf(GA \Rightarrow C).$$

The confidence of the rule $G : A \Rightarrow \bullet$ in Figure 3 is $conf_G(A \Rightarrow \bullet) = (5/15) \; / \; (10/15) = 0.5$.

3.3 Group Gain

Definition 11. *The gain of the group association rule $G : A \Rightarrow C$ is defined as:*

$$\begin{aligned} gain_G(A \Rightarrow C) &= conf_G(A \Rightarrow C) - supp_G(C) \\ &= conf(GA \Rightarrow C) - conf(G \Rightarrow C) \end{aligned}$$

The gain represents the difference between the confidence in the presence of the consequent when we know that the antecedent appears in the group, minus the support of the consequent in the group.

In Figure 3, the support of the *circles* in the group G was $supp_G(\bullet) = 6/15 = 0.4$ and the confidence of the $G : A \Rightarrow \bullet$ rule is $conf_G(A \Rightarrow \bullet) = 0.5$. Then, the gain of the rule is $gain_G(A \Rightarrow \bullet) = 0.5\text{-}0.4 = 0.1$. That means that, within the group G, finding a *circle* is 10% more likely when A holds.

Rules with high positive gain values help us to describe subgroups within the group G. For example, in the *men with children* group, the rule *lives in suburbs* ⇒ *2 cars* might have a high positive gain value. That would suggest that, within the *men with children* group, living in suburbs increases the likelihood of owning

2 cars. Therefore, it would be easier to find men with two cars among those how live in suburbs than in the overall group of men who have children.

On the other side, rules with high negative gain values help us to find characteristics that do not define the subgroup. For example, in the *men older than 30* group, the rule *lives downtown* $\Rightarrow$ *owns a car* might have a negative gain. That would suggests that, within the *men older than 30* group, living downtown decrease the likelihood of owning a car. Therefore, it would be more difficult to find men older than 30 who own a car among those that live downtown than in the overall group.

Property 1. *The gain of the* $G : A \Rightarrow C$ *rule is the difference between the gain of the* $GA \Rightarrow C$ *rule and the gain of the* $G \Rightarrow C$ *rule , i.e.,*

$$gain_G(A \Rightarrow C) = gain(GA \Rightarrow C) - gain(G \Rightarrow C)$$

Proof. By Definition 6, $gain(G \Rightarrow C) = conf(G \Rightarrow C) - supp(C)$. Then, we can solve for $conf(G \Rightarrow C)$ as $conf(G \Rightarrow C) = gain(G \Rightarrow C) + supp(C)$. If we replace the $conf(G \Rightarrow C)$ in Definition 11, we obtain $gain_G(A \Rightarrow C) = conf(GA \Rightarrow C) - conf(G \Rightarrow C) = conf(GA \Rightarrow C) - supp(C) - gain(G \Rightarrow C)$. Finally, by Definition 6, $conf(GA \Rightarrow C) - supp(C) = gain(GA \Rightarrow C)$. Therefore, $gain_G(A \Rightarrow C) = gain(GA \Rightarrow C) - gain(G \Rightarrow C)$.

Theorem 2. *The difference between the gain of the* $G : A \Rightarrow C$ *rule in the group* G *and the gain of the* $A \Rightarrow C$ *rule in the database equals the gain of the rule* $A : G \Rightarrow C$ *in the group* A *minus the gain of the* $G \Rightarrow C$ *rule in the database , i.e.,*

$$gain_G(A \Rightarrow C) - gain(A \Rightarrow C) = gain_A(G \Rightarrow C) - gain(G \Rightarrow C).$$

Proof. By Definition 6, $gain(G \Rightarrow C) = conf(G \Rightarrow C) - supp(C)$.

We can isolate $conf(G \Rightarrow C) = gain(G \Rightarrow C) + supp(C)$ and replace it in $gain_G(A \Rightarrow C) = conf(GA \Rightarrow C) - conf(G \Rightarrow C) = conf(GA \Rightarrow C) - (gain(G \Rightarrow C) + supp(C))$.

If we isolate $supp(C)$ from Definition 6 and replace it in the previous expression, we obtain: $gain_G(A \Rightarrow C) = conf(GA \Rightarrow C) - (gain(G \Rightarrow C) + supp(C)) = conf(GA \Rightarrow C) - gain(G \Rightarrow C) - (conf(A \Rightarrow C) - gain(A \Rightarrow C)) = conf(GA \Rightarrow C) - (conf(A \Rightarrow C) - gain(G \Rightarrow C) + gain(A \Rightarrow C))$.

By Definition 11, the first term can be expressed as $conf(GA \Rightarrow C) - conf(A \Rightarrow C) = gain_A(G \Rightarrow C)$. Then, we have $gain_G(A \Rightarrow C) = gain_A(G \Rightarrow C) - gain(G \Rightarrow C) + gain(A \Rightarrow C))$.

Therefore, $gain_G(A \Rightarrow C) - gain(A \Rightarrow C) = gain_A(G \Rightarrow C) - gain(G \Rightarrow C)$.

3.4 Group Gain Normalization

The range of the gain is $[-supp_G(C), 1-supp_G(C)]$. In the followings paragraphs, we propose several ways to normalize the group gain depending on the kind of information we want to highlight. For example, we can normalize the gain into the $[-1, 1]$ interval to obtain a gain factor measure that corresponds to the certainty factor in the general association rule framework.

Group Gain Factor

Definition 12. *The gain factor of the group association rule* $G : A \Rightarrow C$ *is defined as:*

$GF_G\ (A \Rightarrow C) = \frac{gain_G(A \Rightarrow C)}{1 - supp_G(C)}$ *if* $gain_G(A \Rightarrow C) \geq 0$, *and*

$GF_G\ (A \Rightarrow C) = \frac{gain_G(A \Rightarrow C)}{supp_G(C)}$ *if* $gain_G(A \Rightarrow C) < 0$.

In our example, the gain factor of the rule $G : A \Rightarrow \bullet$ is $GF_G(A \Rightarrow \bullet) = 0.1/(1\text{-}0.4) = 0.17$.

This measure is proportional to the group gain. When it is positive, it is also inversely proportional to the value $[1 - supp_G(C)]$. Therefore, all other things being equal, GF will be larger for subgroups of elements that were more common in the group G (i.e., those having a higher $supp_G(C)$). When GF is negative, it is inversely proportional to $supp_G(C)$: it will have a larger absolute value when the subgroup (C) is less frequent in G.

Group Variation

Definition 13. *The variation of a group association rule* $G : A \Rightarrow C$ *is defined as:*

$\delta_G(A \Rightarrow C) = \frac{gain_G(A \Rightarrow C)}{supp_G(C)} = \frac{conf_G(A \Rightarrow C) - supp_G(C)}{supp_G(C)}$

In contrast to the GF, variation is inversely proportional to $supp_G(C)$ when it is positive. It will have a higher value the less frequent C is in G. It should be noted that the variation equals the gain factor when the gain is negative. The variation of the rule $G : A \Rightarrow \bullet$ in Figure 3 is $\delta_G(A \Rightarrow \bullet) = 0.1/(0.4) = 0.25$.

Group Impact

Definition 14. *The impact of the group association rule* $G : A \Rightarrow C$ *is defined as:*

$impact_G(A \Rightarrow C) = supp(GA) * gain_G(A \Rightarrow C)$

The impact of a group association rule represents the number of individuals that are affected by the rule, i.e., the number of individuals that we did not expect to find in the transactions of the group G that contain A (GA) given what we knew about G.

The impact is proportional to $gain_G(A \Rightarrow C)$ and $supp(GA)$. It will be higher for those rules with a high gain and a frequent antecedent A in the group G.

In our example from Figure 3, $impact_G(A \Rightarrow \bullet) = (10)*0.1 = 1$. That should be interpreted as: there is 1 circle that we did not expect to be in GA when we only knew the support of $\bullet$ in G, $supp_G(\bullet) = 0.4$, i.e., we did expect 4 circles in GA but there are 5, actually.

Impact Ratio

Definition 15. *The impact ratio of the group association rule* $G : A \Rightarrow C$ *is defined as:*

$IR_G(A \Rightarrow C) = \frac{impact_G(A \Rightarrow C)}{supp(G)}$

The impact ratio of a group association rule represents the proportion of the impact of the rule in the group G with respect to the size of the group. The impact ratio is $IR_G(A \Rightarrow \bullet) = 1/15 = 0.07$ in the example from Figure 3.

4 Ranking Groups and Group Association Rules

The amount of rules and groups obtained in the rule mining process can be huge, hence it may be difficult to extract useful information from them. All these groups, as well as the rules within them, are obtained in an unsupervised process and we will use the measures we have described to highlight those rules and groups that might be relevant to the user according to several criteria.

In this section, we explain how to rank the groups and the rules within the groups according to their potential interestingness.

4.1 Ranking Rules within a Particular Group

The use of each measure provides us a different ordering among rules. We will choose a measure depending on the kind of information we want to highlight. In this section, we analyze how two rules in a group will have a different relative ordering in a group depending of the measure we use to evaluate them.

Characterizing subgroups within the group. If we are interested in obtaining those rules that characterize subgroups within a group (i.e., rules sharing their consequent), we should use the **gain** measure because a high gain increases our confidence in the presence of the consequent when we know that the antecedent holds.

- If we want to highlight the most frequent subgroups, the **gain factor**, as inversely proportional to the interval $[1 - supp_G(C)]$, should be used.
- If we want to highlight anomalies, the **variation** measure is a better choice since, in contrast to the gain factor, it overweighs those subgroups that have a low support in the group.

Characterizing subgroups within the group using frequent features. If we are interested, not only in the subgroups, but also in using features that are frequent in our database, we should use a measure that takes into account the frequency of the antecedent of the rules, e.g., the **impact** measure.

This measure has the advantage that it has an easy interpretation: it indicates the number of individuals in G that are directly affected by the rule $A \Rightarrow C$, i.e., those individuals that are not expected to be in GA when we only know the overall support of C in the group.

4.2 Ranking Groups within the Database

Apart from the order of the rules within a group, we can establish an ordering relationship among the groups in our database to highlight those groups that

include more interesting rules. For example, this can be useful for analyzing the behavior of groups in our database and studying the features that individuals in the same group have in common.

Impact seems to be a good measure to evaluate the interestingness of the group because it takes into account the number of individuals that are affected by each rule in the group. We should average the impact of the n rules within a given group to indicate the overall interestingness of that group. However, as we have explained in Section 4.1, some rules are more interesting than others. Then, they should not have the same weight.

We can define the weighted impact for the rules in a group using a different interest measure depending on the information we want to highlight. Formally, we define the weighted impact as:

$$\text{Weighted impact}(G) = \frac{\sum_{i=1}^{n} I_G(A \Rightarrow C) \cdot impact_G(A \Rightarrow C)}{\sum_{i=1}^{n} I_G(A \Rightarrow C)}$$

where $I_G(A \Rightarrow C)$ represents one of the interestingness measures described in Section 3 and analyzed in Section 4.1 for each $A \Rightarrow C$ rule in the group G.

Large groups tend to have higher impact values for their rules because the impact depends on the support of the antecedent in the group and it is usually larger in large groups. Therefore, small groups are penalized in the ranking if we use the impact measure. If we also want to take into account the relative size of the groups, we can use the impact ratio measure, which gives us a more balanced ranking. Thus we define a weighted impact ratio measure to rank groups within the database:

$$\text{Weighted IR}(G) = \frac{\sum_{i=1}^{n} I_G(A \Rightarrow C) \cdot IR_G(A \Rightarrow C)}{\sum_{i=1}^{n} I_G(A \Rightarrow C)}$$

5 Conclusions

Databases naturally contain groups of individuals that share some of their features and some aspects of their behavior. In this paper, we have proposed group association rules, which are association rules that are discovered within these groups of individuals.

We have adapted some of the standard interestingness measures for association rules to group association rules and we have also proposed new interestingness measures to evaluate this particular kind of association rules. We have studied the properties of these measures and which ones could be useful for the user depending on the information he is interested in.

Finally, we have proposed some guidelines to rank the groups in a database and the rules within each group depending on the user goals.

Acknowledgements

This work is supported by research project TIN2009-08296.

References

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco (2005)
2. Plasse, M., Niang, N., Saporta, G., Villeminot, A., Leblond, L.: Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Computational Statistics & Data Analysis 52(1), 596–613 (2007)
3. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys 38(3), 9 (2006)
4. Berzal, F., Blanco, I.J., Sánchez, D., Vila, M.A.: Measuring the accuracy and interest of association rules: A new framework. Intelligence Data Analysis 6(3), 221–235 (2002)
5. Bayardo Jr., R.J., Agrawal, R.: Mining the most interesting rules. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (1999)
6. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: 20th International Conference on Very Large Data Bases, pp. 487–499 (1994)
7. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the 1997 ACM SIGMOD international conference on Management of data, vol. 26(2), pp. 265–276 (1997)
8. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the 1997 ACM SIGMOD international conference on Management of data, vol. 26, pp. 255–264 (1997)
9. Shortliffe, E.H., Buchanan, B.G.: A model of inexact reasoning in medicine. Mathematical biosciences 23, 351–379 (1975)

Data Mining in RL-Bags

M. Dolores Ruiz*, Miguel Delgado, and Daniel Sánchez

Dept. Computer Science and Artificial Intelligence, University of Granada,
C/Periodista Daniel Saucedo Aranda s/n , 18071 Granada, Spain
{mdruiz,daniel}@decsai.ugr.es,
mdelgado@ugr.es

Abstract. Many databases in real life involve items with their quantities. This kind of databases can be modeled using the theory of bags or by fuzzy bags if we deal with imprecise properties of objects. We present a general framework for extracting useful knowledge from fuzzy bags or more generally from *RL*-bags, a new type of bag which extends the one of fuzzy bag and preserves the usual crisp properties overall when using the negation. The main contribution is how to deal with the information provided with the *RL*-bags for then mining useful and interesting association rules, as the *RL*-bags involve uncertainty over the quantities associated to the objects.

Keywords: bags, RL-bags, RL-sets, RL-representations, fuzzy rules.

1 Introduction

In this paper, we are going to deal with databases which either contain or can be transformed into bags [12] or fuzzy bags [5]. These databases are special in the sense that a transaction not only contains the information that an item occurs or not, it also specifies the times that an item occurs in each transaction, possibly to a certain degree. This happens very often in the context of market baskets but there are other fields of application as we can see in [7], [9], [8].

In this kind of structures, we may be interested in associations that do not involve items only, but also their frequency of appearance in the bag. For example "most of baskets that contains a lot of bread also contains a lot of milk". As this example shows, these quantities are imprecise very often. Some theories like the theory of fuzzy sets help in dealing with this type of imprecise knowledge [13]. The theory of fuzzy sets has been already used in many situations to represent some real-world domains that are intrinsically fuzzy. But there exist new models to represent imprecise information such as the Restriction Level Representation (RLR) theory which extends that of fuzzy sets and allows us to extend crisp operations to the imprecise case, keeping all the properties of the crisp case, notably those involving negation. On the basis of this theory, we shall present

* We would like to acknowledge support for this work from the Spanish Ministry for Science and Innovation by the project grants TIN2006-15041-C04-01 and TIN2009-08296.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 308–317, 2010.

here the notion of Restriction-Level bag (*RL*-bag for short) that generalizes that of fuzzy bag.

The RLR theory [11] is suitable when we have to operate with vague concepts and we need, or want to keep the ordinary boolean properties. They are also useful when translating crisp procedures as we can see in [6] and for the formulation of properties to the fuzzy case. In our case, we are interested in preserving the usual crisp properties of bags when dealing with uncertainty, specially when mining rules involving the negation of items.

The aim of this work is twofold. First we present the concept of *RL*-bag extending the known concept of fuzzy bag[1]. And second, we present how we can deal with the information provided by the *RL*-bags (the uncertainty and the frequency of items) to obtain meaningful and interesting rules for the user.

The paper is organized as follows: first we review some definitions about *RL*-sets and bags. Then we define the concept of *RL*-bag using the RLR theory. We follow presenting a new approach for mining association rules in *RL*-bags using the approach developed in [6] which extends the measures used for assessing crisp rules to the fuzzy case. Finally, we give some concluding remarks.

2 Representation by Restriction Levels

The basic idea of the RLR theory [11], [10] is that vague properties defined on a set of objects X can be described by a collection of crisp representatives each one being a crisp realization under a certain restriction. The so called restriction levels (RL) are represented by values in the unit interval meaning possible levels of relaxation of the property where 1 corresponds to the most restrictive, 0 means no restriction at all and the restriction level 0.5 is halfway between being totally strict and no strict at all.

Definition 1. *[10] A RL-set Λ is a finite set of restriction levels $\Lambda = \{\alpha_1, \ldots, \alpha_m\}$ verifying that $1 = \alpha_1 > \alpha_2 > \cdots > \alpha_m > \alpha_{m+1} = 0$, $m \geq 1$.*

The *RL*-set of an atomic property represented by means of a fuzzy set A is defined as follows.

Definition 2. *[10] Let be A a fuzzy set defined on the referencial X. Then the RL-set associated to A is given by:*

$$\Lambda_A = \{A(x) \,|\, x \in supp(A)\} \cup \{1\} \tag{1}$$

where $A(x)$ is the membership grade of x to the fuzzy set A, and $supp(.)$ denotes the support of a fuzzy set.

The employed *RL*-set to represent an imprecise property is obtained by the union of the *RL*-sets associated to the atomic properties which define that property.

[1] A fuzzy bag can be seen as a *RL*-bag, but the associated operations between fuzzy bags do not coincide with those defined over *RL*-bags.

A *RL-representation* associated to an imprecise property in X is defined by a pair (Λ, ρ) where Λ is a *RL*-set and $\rho : \Lambda \to \mathcal{P}(X)$ is a function which applies each restriction level into a crisp realization in this level. For example, the *RL*-representation of an imprecise atomic property defined by a fuzzy set A will be the pair (Λ_A, ρ_A), where Λ_A is given by the equation (1) and $\rho_A(\alpha) = A_\alpha = \{x \in X \mid A(x) \geq \alpha\}$ for all $\alpha \in \Lambda_A$.

Given an imprecise property P represented by (Λ_P, ρ_P), the set of crisp representatives of P is the set [10] $\Omega_P = \{\rho_P(\alpha) \mid \alpha \in \Lambda_P\}$.

Definition 3. *[10] Let be (Λ, ρ) a RL-representation with $\Lambda = \{\alpha_1, \ldots, \alpha_m\}$ verifying that $1 = \alpha_1 > \alpha_2 > \cdots > \alpha_m > \alpha_{m+1} = 0$. Let $\alpha \in (0, 1]$ and $\alpha_i, \alpha_{i+1} \in \Lambda$ satisfying that $\alpha_i > \alpha > \alpha_{i+1}$. Then we define*

$$\rho(\alpha) = \rho(\alpha_i). \tag{2}$$

If we look to this definition, this extension for values that there are not in the *RL*-set of the function ρ, is the natural extension if we think in a fuzzy set A and its α-cuts. Using this definition the concept of equivalence between two *RL*-representations is straightforward.

Definition 4. *[10] Let (Λ, ρ) and (Λ', ρ') be two RL-representations on X. We will say that both representations (and the corresponding properties) are* equivalent*, noted by $(\Lambda, \rho) \equiv (\Lambda', \rho')$, if and only if, $\forall \alpha \in (0, 1]$*

$$\rho(\alpha) = \rho(\alpha'). \tag{3}$$

Summarizing, only a finite *RL*-set is necessary for defining a RLR, but the representation extends to any other RL in (0, 1].

The usual boolean operations are extended to RLRs by applying them on the representatives of the same RL of the arguments independently. In particular, we present here the logic operations of disjunction, conjunction and negation. The basic ideas of how they are defined can be found in [10],[11].

Definition 5. *Let P, Q be two imprecise properties with RL-representations (Λ_P, ρ_P), (Λ_Q, ρ_Q). Then, $P \wedge Q$, $P \vee Q$ and $\neg P$ are imprecise properties represented by $(\Lambda_{P \wedge Q}, \rho_{P \wedge Q})$, $(\Lambda_{P \vee Q}, \rho_{P \vee Q})$ and $(\Lambda_{\neg P}, \rho_{\neg P})$ respectively, where $\Lambda_{P \wedge Q} = \Lambda_{P \vee Q} = \Lambda_P \cup \Lambda_Q$, $\Lambda_{\neg P} = \Lambda_P$ and, for all $\alpha \in (0, 1]$,*

$$\begin{aligned} \rho_{P \wedge Q}(\alpha) &= \rho_P(\alpha) \cap \rho_Q(\alpha), \\ \rho_{P \vee Q}(\alpha) &= \rho_P(\alpha) \cup \rho_Q(\alpha), \\ \rho_{\neg P}(\alpha) &= \overline{\rho_P(\alpha)}, \end{aligned} \tag{4}$$

where $\overline{Y}$ is the usual complement of a crisp set Y.

Basic boolean properties that cannot be verified simultaneously by any standard fuzzy set theory (FST) hold simultaneously for RLRs. We want to remark that fuzzy sets are closed with respect to some of these RLR operations in the sense that the corresponding RLR yields the usual nested α-cut representation and hence the result is a fuzzy set. However, this is not true in general when negation is employed.

3 Bags and Bag Databases

Bags, also called multisets, were introduced by R. Yager [12] as set-like algebraic structures where an element can appear more than once. Bags are useful in order to model market basket situations, but they can be used in others contexts. For example, in the real world is usual to find that several individually different objects verify the same properties in a certain context (space, time). In other words, we can say they are instances of the same class. Counting how many objects verify a certain property in a given context seems to be a natural application of bags in practice [5].

There are several definitions of the bag concept [12],[5], but the usual one formalizes a bag B as an application, typically named count function, defined from a set of objects O that maps every object to a non negative integer number. A bag database is defined then as a set of bags, all of them defined over the same set of objects.

When the objects are items, we have bag transactions. An example of such bag transaction[2] could be $B = \{b_1, b_1, b_3, b_3, b_3, b_7, b_7\}$. For brevity, we represent every bag by $B = \{(b_1, q_1), \ldots, (b_k, q_k)\}$ where $b_j \in I$ represents an item and q_j is a non negative integer number for all $1 \leq j \leq k$. Under these conditions, a *bag database* D, is a set of bags $D = \{B_1, \ldots, B_n\}$ where B_i is a bag for every $1 \leq i \leq n$.

4 RL-Bags

In this section we define the new concept of *RL*-Bag using the RLR theory. This formalization will extend the so called fuzzy bags and will provide a good generalization of operations to the crisp case overall when dealing with negation.

The concept of fuzzy bag is presented in [7] as follows:

Definition 6. *Assume O is a set of objects. A fuzzy bag $\tilde{B}$ is a mapping from the cartesian product $O \times [0,1]$ to the set of non-negative integers $\mathbb{N}$ characterized by:*

$$f_{\tilde{B}}(w/b) : [0,1] \times O \longrightarrow \mathbb{N}, \tag{5}$$

where b is an object in O and $w \in [0,1]$ represents its membership grade.

A fuzzy bag $\tilde{B}$ can also be denoted by $\tilde{B} = \{(w_1/b_1, q_1), \ldots, (w_k/b_k, q_k)\}$ where $q_i = f(w_i/b_i) \in \mathbb{N}$ is the count associated to the fuzzy object w_i/b_i.

Example 1. Consider the following two fuzzy bags $\tilde{B}_1$ and $\tilde{B}_2$ defined over the set of objects $O = \{b_1, b_2, b_3\}$:

$$\begin{aligned} \tilde{B}_1 &= \{(0.2/b_1, 2), (0.3/b_2, 3), (0.1/b_3, 4), (0.2/b_3, 5)\} \\ \tilde{B}_2 &= \{(0.4/b_2, 5)\} \end{aligned} \tag{6}$$

Then, $f_{\tilde{B}_1}(0.3/b_2) = 3$, $f_{\tilde{B}_1}(0.2/b_3) = 5$ and $f_{\tilde{B}_2}(0.4/b_2) = 5$ are examples of their associated mappings.

[2] We shall refer in the following to these as simply "bags" for the sake of brevity.

Let $\tilde{B}_\alpha$ be the α-cut of the fuzzy bag $\tilde{B}$. In the same way as fuzzy sets, a fuzzy bag can be represented by its α-cuts via the formula [9]:

$$(f_{\tilde{B}})_\alpha(b) = \sum_{w \geq \alpha} f(w/b). \tag{7}$$

where $(f_{\tilde{B}})_\alpha(b)$ is the number of occurrences of the element b in $\tilde{B}_\alpha$. For instance, in the previous example $\tilde{B}_1$ is represented by its α-cuts as follows:

$$\begin{aligned} (\tilde{B}_1)_{0.1} &= \{(b_1, 2), (b_2, 3), (b_3, 9)\} \\ (\tilde{B}_1)_{0.2} &= \{(b_1, 2), (b_2, 3), (b_3, 5)\} \\ (\tilde{B}_1)_{0.3} &= \{(b_2, 3)\} \end{aligned} \tag{8}$$

where the element $(b_3, 9)$ at the α-cut 0.1 comes from the elements $(0.1/b_3, 4), (0.2/b_3, 5)$. Then, given a fuzzy bag $\tilde{B}$, the following property holds [9]:

$$\forall \alpha, \beta \in (0, 1], \quad \alpha \leq \beta \ \text{ then } \ \tilde{B}_\beta \subseteq \tilde{B}_\alpha. \tag{9}$$

Consequently the α-cuts of a fuzzy bag are nested crisp bags and a fuzzy bag can be represented by the family of all its α-cuts.

There exist other approaches which also generalizes the concept of fuzzy bag using a different formulation [2], [4]. In [4] the main idea is to consider a mathematical correspondence between properties and objects, from which classical bags are obtained as counts. This view enriches the classical view of bags, allowing to define algebraic operations that reduce to classical algebraic and numerical operations on classical bags and fuzzy bags in some particular cases.

In the mentioned papers about fuzzy bags, the set of operations are defined in terms of fuzzy operators. Some of them, such as the complement of a fuzzy bag and the difference between fuzzy bags, suffer from the same problems that the fuzzy operators used to extend them [9]. The restriction level representation theory gives us the tool for avoiding some of these drawbacks by using the operations defined in section 2.

From this point on, we will define the concept of RL-bag using the RLR theory which will allow to manage with fuzzy bags as a set of crisp bags each of them defined over a restriction level.

Definition 7. *Assume that O is a set of objects and $\mathcal{B}(O)$ is the set containing all the possible bags defined over O. A RL-bag, β, will be a pair $(\Lambda_\beta, \rho_\beta)$ where Λ_β is a RL-set and $\rho_\beta : \Lambda_\beta \longrightarrow \mathcal{B}(O)$ is a function which applies each restriction level into a crisp bag defined over the set of objects.*

From this definition it is easy to see that a fuzzy bag is a particular case of RL-bag, but the opposite is not true, because a RL-bag does not necessarily fulfil the property about nested crisp bags when descending the restriction level (or the α-cut in the fuzzy case).

Using Definition 7 it is straightforward to extend the usual logic operators to RL-bags by applying the operators to the crisp bags in each restriction level

Table 1. Examples of RL-bags defined over the set of objects $O = \{b_1, b_2, b_3, b_4\}$

α	$\rho_{\beta 1}$	$\rho_{\beta 2}(\alpha)$
1	$\{(b_3, 2)\}$	$\{(b_1, 1), (b_2, 2)\}$
0.7	$\{(b_1, 2), (b_3, 2)\}$	$\{(b_1, 1), (b_2, 2), (b_3, 1)\}$
0.6	$\{(b_1, 2), (b_2, 1), (b_3, 2)\}$	$\{(b_1, 2), (b_2, 5), (b_3, 1)\}$
0.4	$\{(b_1, 2), (b_2, 1), (b_3, 4)\}$	$\{(b_1, 3), (b_2, 5)\}$

adding to the RL-set the new levels as in definition 5. We want to remark that operations defined over RL-bags following the RLR philosophy do not coincide with those operations defined for fuzzy bags using the fuzzy set theory. So, the real contribution of the RL-bag concept is that of managing fuzzy bags as a set of crisp bags which are satisfied to a certain degree. This generalization of fuzzy bag allows us to operate with fuzzy bags without loosing the crisp properties of bags, and it is also useful when extending crisp measures to the fuzzy case as for instance when mining rules in fuzzy bag databases.

Nevertheless, we are more interested in how the linguistic labels interact with the RL-bags in order to propose a general method for mining association rules from a database constituted by fuzzy bags or RL-bags. The key of our proposal (presented in next section) is how to combine linguistic labels with RL-bags and how to take advantage of the quantities associated to each object for measuring their changes and variation in the whole set of fuzzy bags.

5 Mining Fuzzy Rules in RL-Bags

This section presents a new approach for mining association rules in a database which contains fuzzy bags by using their representation as RL-bags. This kind of databases are very frequent in the field of information retrieval or text mining. For instance, a document d can be seen as a special case of fuzzy bag because we could represent it by a set of triples of the form:

$$d = \{(w_k/i_k, f_k), k = 1, \ldots p\}$$

where $i_k \in I$ is the kth item in the document, $w_k \in [0, 1]$ is the weight or the degree of relevance associated to the item i_k and f_k is the number of appearances of i_k in the document. In these terms, a collection of documents can be considered as a set of fuzzy bags, that is, a fuzzy bag database. Then, it is interesting to develop new methods for mining interesting rules in fuzzy bag databases. In our case we will treat a fuzzy bag as a RL-bag in order to preserve the crisp properties when operating between them.

In our previous work dealing with crisp bags [5], we extract meaningful rules from a crisp bag database considering the user's knowledge about the definition of good representative linguistic terms for the quantities of each item. Our previous approach can be divided in mainly two steps. First, we transform the bag database into a fuzzy transactional one using the linguistic terms defined by

the user and then we extract the fuzzy association rules using the support and certainty factor as proposed in [3].

Following the same idea, we can transform a database which contains RL-bags into a fuzzy one (seen by restriction levels as we will see in Example 2) and then apply some developed tools for mining fuzzy association rules.

Our approach can be summarized in some simple steps as the Algorithm 1 shows.

Algorithm 1. Mining Fuzzy Rules in Fuzzy Bag Databases

Input: Fuzzy bag database (or RL bag database), minsupp, minconf or minCF
Output: Set of fuzzy association rules
1. Database Preprocessing
1.1 The user gives a set of linguistic labels associated to the frequencies of each item.
1.2 Database transformation into a fuzzy database using the previous labels.
2. Mining Process
2.1 Mining Fuzzy Rules in the transformed fuzzy database.
2.1.1 Computation of the associated $4ft$-table in each restriction level.
2.1.2 Computation of the values for the $4ft$-quantifiers associated to support and confidence or certainty factor, extracting those rules that exceed the minimum associated thresholds.

Let $I = \{i_1, \ldots, i_m\}$ be a set of items. We call RL-bag database to a set $D = \{\beta_1, \ldots, \beta_n\}$ of RL-bags defined over I. The first step is to transform D into a new database $\tilde{D}$ formed by a set of crisp bags defined in each restriction level. In this step the participation of the user is necessary in order to define a set of linguistic labels associated to the frequencies of the items appearance. Nevertheless, the previous labels can be defined using a clustering process [1], or simply by defining a partition by means of trapezoidal numbers in the domain $[minf, maxf]$ where the bounds correspond to the minimum value of the frequency (we will use 0 by convention) and to the maximum value of the frequency respectively.

The main idea of this transformation consists in considering the linguistic label as a set of intervals, each one defined in a different restriction level. Then, we see if the frequency $f(w/i)$ associated to the item i at level $\alpha = w$ belongs to the corresponding interval in that restriction level, in that case we will have that the new item $\langle i, label\rangle$ is satisfied in level α (see table 3).

Once we have the fuzzy database defined in terms of restriction levels, we can easily compute for each pair of items X and Y of the previous type ($\langle i, label\rangle$) the associated $4ft$-table, noted by $4ft(\mathcal{M}_\alpha, X, Y, \tilde{D})$, for each RL α as follows [6]:

$\mathcal{M}_{\alpha_i}$	Y	$\neg Y$
X	a_i	b_i
$\neg X$	c_i	d_i

where a_i is the number of transactions in the database $\tilde{D}$ satisfying X and Y at the level α_i, b_i the number of transactions satisfying X and not Y in level α_i,

and so on. The value for the support and for the confidence are extended from the crisp case to the fuzzy case in [6] using the previous $4ft$-table as follows (we also extend the certainty factor in [6] but it is not included here due to the lack of space.)

$$Supp(X \to Y) = \sum_{\alpha_i \in \Lambda_{X \wedge Y}} (\alpha_i - \alpha_{i+1}) \left(\frac{a_i}{a_i + b_i + c_i + d_i} \right) \tag{10}$$

$$Conf(X \to Y) = \sum_{\alpha_i \in \Lambda_{X \wedge Y}} (\alpha_i - \alpha_{i+1}) \left(\frac{a_i}{a_i + b_i} \right) \tag{11}$$

If the $Supp(X \to Y)$ and the $Conf(X \to Y)$ exceed the *minsupp* and *minconf* thresholds, the rule $X \to Y$ will be mined from the database $\tilde{D}$. We want to remark that items X and Y are of the form $\langle i, label \rangle$, so the extracted rules will be of the type: *"most of transactions that contain* ***few*** i_j *also contain* ***a lot*** *of* i_k*"* where we have a relation between the joint occurrence of i_j and i_k and also a relation between their associated frequencies by means of the linguistic labels *few* and *a lot*.

Example 2. Let D_1 be constituted by the following four RL-bags:

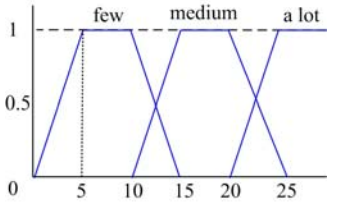
Table 2. D_1

α	β_1	β_2	β_3	β_4
1	$\{(i_1,2),(i_2,1),(i_3,4)\}$	$\{(i_2,1),(i_3,20)\}$	$\{(i_1,2),(i_2,5),(i_3,13)\}$	$\{(i_4,7)\}$
0.8	$\{(i_1,6),(i_2,1),(i_3,5)\}$	$\{(i_1,1),(i_2,1),(i_3,23)\}$	$\{(i_1,2),(i_2,1),(i_3,15)\}$	ϕ
0.5	$\{(i_1,10),(i_2,1),(i_4,4)\}$	$\{(i_1,2),(i_2,1),(i_3,23)\}$	$\{(i_1,2),(i_2,6),(i_3,15)\}$	$\{(i_1,5),(i_4,7)\}$
0.2	$\{(i_1,2),(i_4,6)\}$	$\{(i_1,2),(i_4,12),(i_3,23)\}$	$\{(i_1,2),(i_2,6),(i_3,16)\}$	ϕ

where i_1, i_2, i_3, i_4 are the items associated to the RL-bags $\beta_1, \ldots, \beta_4$. Notice that β_4 does not correspond to a fuzzy bag, but it could come from the negation and/or conjunction of some fuzzy bags.

For every item in $I = \{i_1, i_2, i_3, i_4\}$ we can define a set of linguistic terms related to the frequencies which indicates how to measure the quantity. In Figure 1 there is an example of linguistic labels associated to every item in I. In general, we could have different sets of labels associated to each particular item.

To obtain a new database $\tilde{D}_1$ from D_1 constituted by a set of crisp transactions in each restriction level, we first consider the linguistic labels $L_j \in \{\text{low, medium, a lot}\}$ defined in Figure 1 as a set of intervals in each restriction level, as we show in Figure 1. For that, we consider the same set of restriction levels $\Lambda_{\beta_1} \cup \cdots \cup \Lambda_{\beta_4} = \{1, 0.8, 0.5, 0.2\}$ used in the RL-bag database, and then we take the α-cut of the linguistic label where $\alpha \in \Lambda_{\beta_1} \cup \cdots \cup \Lambda_{\beta_4}$. Then, we check for each restriction level if the quantities associated to the item are in the interval or not. By this procedure, we obtain a new item in the form $\langle i, L_{ij} \rangle$ for each linguistic label as Table 3 shows. In the process, as the linguistic labels are overlapped, one single item can be transformed into two new items, as for instance, at level 0.2 the initial



α	few	medium	a lot
1	$[5, 10]$	$[15, 20]$	$[25, \infty)$
0.8	$[4, 11]$	$[14, 21]$	$[24, \infty)$
0.5	$[2.5, 12.5]$	$[12.5, 22.5]$	$[22.5, \infty)$
0.2	$[1, 14]$	$[11, 22]$	$[21, \infty)$

Fig. 1. Linguistic labels associated to items i_1, i_2, i_3, i_4 and their representation by restriction levels

Table 3. Database $\tilde{D}_1$ resulted from the transformation of D_1 using the set of linguistic labels in Figure 1

α	$\tilde{\beta}_1$	$\tilde{\beta}_2$
1	ϕ	$\{\langle i_3, \text{medium}\rangle\}$
0.8	$\{\langle i_1, \text{few}\rangle, \langle i_3, \text{few}\rangle\}$	ϕ
0.5	$\{\langle i_1, \text{few}\rangle, \langle i_4, \text{few}\rangle\}$	$\{\langle i_3, \text{a lot}\rangle\}$
0.2	$\{\langle i_1, \text{few}\rangle, \langle i_4, \text{few}\rangle\}$	$\{\langle i_1, \text{few}\rangle, \langle i_4, \text{few}\rangle, \langle i_4, \text{medium}\rangle, \langle i_3, \text{a lot}\rangle\}$

α	$\tilde{\beta}_3$	$\tilde{\beta}_4$
1	$\{\langle i_2, \text{few}\rangle\}$	$\{\langle i_4, \text{few}\rangle\}$
0.8	$\{\langle i_3, \text{medium}\rangle\}$	ϕ
0.5	$\{\langle i_2, \text{few}\rangle, \langle i_3, \text{medium}\rangle\}$	$\{\langle i_1, \text{few}\rangle, \langle i_4, \text{few}\rangle\}$
0.2	$\{\langle i_1, \text{few}\rangle, \langle i_2, \text{few}\rangle, \langle i_3, \text{medium}\rangle\}$	ϕ

item $(i_4, 12)$ turn into $\langle i_4, \text{few}\rangle$ and $\langle i_4, \text{medium}\rangle$, because $12 \in [1, 14] \cap [11, 22]$. In the other way round, an initial item can disappear in a certain level if its quantity does not belong to any interval.

Now, we are in the second phase of the process where we have to compute the $4ft$-table, $\mathcal{M}_{\alpha_i}$, associated to every pair of items in each restriction level α_i, and then to check if the support exceeds the imposed threshold, and if so, if confidence also exceeds the $minconf$ threshold. When both conditions are satisfied, we will obtain a fuzzy association rule involving the quantity of items, in this case, rules expressed in terms of the defined linguistic labels.

In this case, we show the $4ft$-table associated to items $X = \langle i_1, \text{few}\rangle$ and $Y = \langle i_4, \text{few}\rangle$ in Table 4 where a_j is the number of transactions in $\tilde{D}_1$ containing both items in level α_j, and so on. Then, we compute the support and confidence measures at every level and we aggregate the values using the formulas (10)

Table 4. $4ft(\mathcal{M}_{\alpha_j}, \langle i_1, \text{few}\rangle, \langle i_4, \text{few}\rangle, \tilde{D}_1)$

$\mathcal{M}_{\alpha_j}$	a_j	b_j	c_j	d_j
$\mathcal{M}_1$	0	0	1	3
$\mathcal{M}_{0.8}$	0	1	0	3
$\mathcal{M}_{0.5}$	2	0	0	2
$\mathcal{M}_{0.2}$	2	1	0	1

and (11) obtaining for instance that $Supp\,(\langle i_1, \text{few}\rangle \rightarrow \langle i_4, \text{few}\rangle) = 0.25$ and $Conf\,(\langle i_1, \text{few}\rangle \rightarrow \langle i_4, \text{few}\rangle) = 0.433$, so if we fix the $minconf$ threshold to 0.7 this rule will not be extracted.

6 Conclusions

The main contributions are the definition of *RL*-bags, a new type of bag which extends the one of fuzzy bag, and the presented framework for extracting useful knowledge from fuzzy bags or more generally from *RL*-bags. In the proposed method, we present how performing some transformations in a *RL*-bag database defining some linguistic labels associated to the quantities, we can achieve interesting association rules from this special type of database.

References

1. Alcalá, R., Alcalá-Fdez, J., Gacto, M.J., Herrera, F.: Genetic learning of membership functions for mining fuzzy association rules. In: Fuzzy Systems Conference, FUZZ-IEEE, pp. 1–6 (2007)
2. Chakrabarty, K., Biswas, R., Nanda, S.: Fuzzy shadows. Fuzzy Sets and Systems 101, 413–421 (1999)
3. Delgado, M., Marín, N., Sánchez, D., Vila, M.A.: Fuzzy association rules: General model and applications. IEEE Transactions on Fuzzy Systems 11(2), 214–225 (2003)
4. Delgado, M., Martín-Bautista, M.J., Sánchez, D., Vila, M.A.: An extended characterization of fuzzy bags. Int. J. of Intelligent Systems 24, 706–721 (2009)
5. Delgado, M., Ruiz, M.D., Sánchez, D.: Pattern extraction from bag databases. Int. J. of Uncert., Fuzziness and Knowledge-Based Systems 16(4), 475–494 (2008)
6. Delgado, M., Ruiz, M.D., Sánchez, D.: A restriction level approach for the representation and evaluation of fuzzy association rules. In: Proceedings of the IFSA-EUSFLAT, Lisbon, Portugal, pp. 1583–1588 (2009)
7. Li, B.: Fuzzy bags and applications. Fuzzy Sets and Systems 34, 61–71 (1990)
8. Rebaïa, A., Martelb, J.-M.: A fuzzy bag approach to choosing the "best" multiattributed potential actions in a multiple judgement and non cardinal data context. Fuzzy Sets and Systems 87(2), 159–166 (1997)
9. Rocacher, D.: On fuzzy bags and their application to flexible querying. Fuzzy Sets and Systems 140, 93–110 (2003)
10. Sánchez, D., Delgado, M., Vila, M.A.: A restriction level approach to the representation of imprecise properties. In: Int. Conference on Information Processing and Management of Uncertainty, Málaga, Spain, pp. 153–159 (2008)
11. Sánchez, D., Delgado, M., Vila, M.A., Chamorro-Martínez, J.: Representation of vagueness via restriction levels. The case of vague quantities. Fuzzy Sets and Systems (2009) (page submitted)
12. Yager, R.R.: On the theory of bags. Int. J. of General Systems 13, 23–37 (1986)
13. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)

Feature Subset Selection for Fuzzy Classification Methods

Marcos E. Cintra[1] and Heloisa A. Camargo[2]

[1] University of Sao Paulo (USP), Mathematics and Computer Science Institute
Sao Carlos, SP - CEP 13566-590 - Brazil
[2] Federal University of São Carlos (UFSCar), Computer Science Department
P.O. Box 676 - CEP 13565-905 - São Carlos, SP
cintra@icmc.usp.br, heloisa@dc.ufscar.br

Abstract. The automatic generation of fuzzy systems have been widely investigated with several proposed approaches in the literature. Since for most methods the generation process complexity increases exponentially with the number of features, a previous feature selection can highly improve the process. Filters, wrappers and embedded methods are used for feature selection. For fuzzy systems it would be desirable to take the fuzzy granulation of the features domains into account for the feature selection process. In this paper a fuzzy wrapper, previously proposed by the authors, and a fuzzy C4.5 decision tree are used to select features. They are compared with three classic filters and the features selected by the original C4.5 decision tree algorithm, as an embedded method. Results using 10 datasets indicate that the use of the fuzzy granulation of features domains is an advantage to select features for the purpose of inducing fuzzy rule bases.

Keywords: Feature selection, fuzzy classification methods, C4.5, Fuzzy C4.5, machine learning.

1 Introduction

Fuzzy systems have been widely used for a variety of tasks, such as classification, optimization and control [1,2]. A system can be defined as a fuzzy system if at least one of its variables is defined in terms of fuzzy sets according to the fuzzy set and fuzzy logic theories, proposed by Loft A. Zadeh [3]. A fuzzy system is usually comprised by a knowledge base and an inference mechanism. The fuzzy data base, which contains the definitions of the features (also named attributes or variables) in terms of fuzzy sets, and the fuzzy rule base, which contains a set of rules defining a given problem, form the knowledge base. The inference mechanism derives the conclusions (or outputs) of the system based on the knowledge base and inputs.

Independently from the adopted approach for the generation of fuzzy rule based systems, feature selection is always an important concern, for it brings the benefits of improving interpretability and reducing the computational cost of

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 318–327, 2010.

the whole process. This issue was tackled with different genetic feature selection based approaches in [4,5]. In [6] a fuzzy neural network is used, and in [7] the authors use a process based on the fuzzy mutual information measure instead.

Filters are a common choice for the task of Feature Subset Selection (FSS). They are simple, fast, and use general characteristics of the data that might suit different datasets and tasks. An approach that considers the aspects of fuzzy logic in the selection process was presented in [8] where an extended definition of the mutual information measure between two fuzzified continuous variables for feature selection was proposed. Another option is the use of embedded methods, such as decision trees, which induce a classifier and select features at the same time. In [9], the authors select features using a crips decision tree before generating fuzzy rule bases. In order to provide a connection between the FSS process and the designed fuzzy system, in [10] we proposed the Fuzzy-Wrapper method, which is a wrapper that uses the Wang & Mendel method [11] to generate fuzzy rule bases as its base algorithm. It produced good results and was significantly better than 4 filters, when compared using 8 datasets. Nevertheless, wrappers have inherent disadvantages related to the cost of the search process and depending on the number of features their use might become unfeasible.

As an alternative to the classic filters and decision trees (used as an embedded method to select features) which do not consider the fuzzy logic to select features, and wrappers, which are costly, in this paper we propose the use of a fuzzy decision tree, based on the classic C4.5 algorithm, to select features. The fuzzy C4.5 is compared to the Fuzzy-Wrapper method, the classic C4.5 decision tree method, and three filters.

The remainder of this paper is organized as follows. In Section 2 the fundamental concepts of fuzzy classification systems are presented. Section 3 describes the main concepts of FSS, including a short description of the Fuzzy-Wrapper method and the fuzzy C4.5 algorithm used in this work. Section 4 presents and discusses the experimental results. Finally, conclusions and future work are presented in Section 5.

2 Fuzzy Classification Systems

Classification is an important task in areas such as pattern recognition, decision making, and data mining, among others. The classification task can be roughly described as: given a set of objects $E = \{e_1, e_2, ..., e_n\}$, also named *examples, cases, or patterns*, which are described by m features, assign a class c_i from a set of classes $C = \{c_1, c_2, ..., c_j\}$ to an object e_p, $e_p = (a_{p_1}, a_{p_2}, ..., a_{p_m})$.

Fuzzy classification systems are rule based fuzzy systems that require the granulation of the features domain by means of fuzzy sets and partitions. The linguistic variables in the antecedent part of the rules represent features, and the consequent part represents a class. A typical fuzzy classification rule can be expressed by

R_k : **IF** X_1 is A_{1l_1} **AND** ... **AND** X_m is A_{ml_m} **THEN** $Class = c_i$

where R_k is the rule identifier, $X_1, ..., X_m$ are the features of the example considered in the problem (represented by linguistic variables), $A_{1l_1}, ..., A_{ml_m}$ are

the linguistic values used to represent the feature values, and $c_i \in C$ is the class. The inference mechanism compares the example to each rule in the fuzzy rule base aiming at determining the class it belongs to.

The classic and general fuzzy reasoning methods are widely used in the literature. Given a set of fuzzy rules (fuzzy rule base) and an input pattern, the classic fuzzy reasoning method classifies this input pattern using the class of the rule with maximum compatibility to the input pattern, while the general fuzzy reasoning method calculates the sum of compatibility degrees for each class and uses the class with highest sum to classify the input pattern. Next section discusses the feature subset selection task.

3 Feature Subset Selection

The task of feature selection aims at finding small subsets of features that describe the dataset at hand as well as or even better than the original set does. The importance of this task lies on the fact that, in practice, the performance of most learning algorithms is affected by the presence of irrelevant and/or redundant features. Besides that, when automatically generating fuzzy rule bases, a particular challenge is the dimensionality problem. Other advantages associated with feature selection are related to reducing the potential hypothesis space by improving data quality, thus increasing the efficiency of the learning algorithm, and enhancing the comprehensibility of the induced classifier [12].

Feature selection methods can be classified into three main categories according to the dependence to the classifier: i)filters; ii) wrappers; iii) embedded methods. These approaches are discussed next.

3.1 Filters

Filters select features before the induction of the classifier. It is a separate process that does not interact with and is independent from the learning algorithm itself. The basic idea of filters is to use general characteristics of the dataset to filter the features before the induction of the classifier takes place.

Some common filters might include the CFS [13], ReliefF [14], and Consistency [15]. CFS stands for Correlation-based Feature Selection and uses the correlation between features to select them. ReliefF ranks the features by their usefulness on distinguishing between very similar examples belonging to different classes and presents an average merit ranking that can be used to select features. Consistency, on the other hand, evaluates the subset of features using the level of consistency between feature values and class values.

These general characteristics used by filters (correlation, distinguishability, consistency, among others) might be appropriate for different domains and tasks. Another advantage of filters is that they are fast and simple.

3.2 Wrappers and the Fuzzy-Wrapper Method

The wrapper approach uses the induction algorithm itself as a black box to evaluate candidate feature subsets, repeating the process on each feature subset until

a stopping criterion is met. This way, wrapper methods take into consideration all the important characteristics of the learning algorithm in the final decision of the FSS process. Wrappers usually use a forward search strategy to avoid the exponential complexity of the search process [16]. However, the computational cost of wrapper methods is prohibitive on large datasets.

In order to take advantage of the fuzzy data base, *i.e.*, the definition of the fuzzy partitions and fuzzy sets that will granulate the features domains in the feature selection process, in [10] we proposed the Fuzzy-Wrapper method, which was further evaluated in [17]. The main idea of this method is to use a wrapper to sequentially generate fuzzy rule bases using the Wang & Mendel (WM) method, combining the features of the dataset. The WM method was originally proposed for regression but can be easily adopted for classification tasks. The best-first heuristic technique is used in order to avoid exponential complexity. The Fuzzy-Wrapper method can be described by the following steps:

1. For a domain with m features, generate all FRBs by the WM method, combining all possible $(m-1)$ features.
2. Remove the feature of the FRB with the best, *i.e.*, lowest error rate, from the dataset, and again generate all FRBs by the WM method combining all possible $(m-2)$ features.
3. Repeat the previous step until only 1 feature remains.
4. Rank the generated FRBs by their error rates.
5. Select the features present in the FRB with the best (lowest) error rate.

Regarding the complexity of the method, since for the WM method the maximum number of fuzzy rules generated is limited to the number of training examples (n), the complexity of the Fuzzy-Wrapper for examples described by m features is $O(n \times m^2)$.

3.3 Embedded Methods for Feature Subset Selection and the Proposed Version of the Fuzzy C4.5

Similarly to wrappers, feature selection by embedded methods is linked with the induction algorithm. However, in this case this link is much stronger than in wrappers, since the feature selection is included in the classifier construction. Embedded methods for FSS induce a classifier that usually do not include all possible features, thus, making it possible to define the set of features included in the classifier as the most relevant ones. A typical example of embedded methods for FSS is decision trees [18].

Decision trees are widely used in machine learning and various algorithms have been proposed for their generation, such as ID3, C4.5, and CART [18,19]. These algorithms generate a tree structure through recursively partitioning the feature space until the whole decision space is completely partitioned into a set of non-overlapping subspaces.

C4.5 is a well-known decision tree algorithm proposed and implemented by Quinlan [18] that uses the information gain and entropy measures when deciding

on the importance of the features, making it possible to select features with the ranking of the generated branches.

Several fuzzy approaches for the induction of decision trees have also been proposed [20], including a fuzzy C4.5 [21]. The fuzzy induction process is very similar to the induction of a classic decision tree. The proposed version of the fuzzy C4.5 algorithm used in this paper applies the same measures of the C4.5 algorithm (entropy and information gain) to decide the importance of the features. However, the features are all defined in terms of fuzzy sets before the induction of the tree. This way, the process can be seen as inducing a tree using only discrete features, since the continuous features are defined in terms of fuzzy sets and the training set is fuzzified before the decision tree induction. Next, we present the main steps of the induction process of the fuzzy tree.

1. Define the fuzzy data base, *i.e.*, the fuzzy granulation of the features domains;
2. Use the fuzzy data base to fuzzify the training set;
3. Calculate the entropy and information gain of each feature to split the training set and generate rules until all features were used or all examples from the training set were classified;
4. Once the tree is induced, it is pruned using the standard pruning rate (25%).

The third step includes all the well-know steps of the classic decision tree algorithms. Notice that the fuzzification of the training data is done before the induction of the tree.

One special issue regarding decision trees is that they can be seen as a set of disjunct rules in which only one rule is fired for a given input example. Regarding fuzzy decision trees, they can be seen as a set of rules that can be fired simultaneously, each one presenting a class and a degree of certainty for a given input pattern to be classified. This characteristic of the fuzzy decision trees allows the use of the fuzzy classic and general reasoning methods, *i.e.*, since more than one rule derived from the decision tree can be fired, an input pattern can be classified with the class of the rule with highest compatibility with the input pattern, or with the class with the highest sum from the set of rules with that given class. Next, we present the experiments and results.

4 Experiments

In order to evaluate the ability of selecting features of the proposed fuzzy C4.5 algorithm, we compared its results with the Fuzzy-Wrapper method. The classic C4.5 decision tree (using the implementation proposed by Quinlan [18]), and three classic filters available at WEKA [22], CFS, ReliefF and Consistency (previously described in Section 3.1) were also tested. 10 benchmark datasets available at the UCI Machine Learning repository [23] were used. In [17], both the classic and general fuzzy reasoning methods were tested. Since the classic reasoning method performed better than the general one, in this paper we decided to adopt only the classic fuzzy reasoning method.

Table 1 summarizes the dataset characteristics giving the total number of examples (Examples), total number of features (Features), including the number of continuous and discrete features in brackets, number of classes (Classes), the majority error (ME), and the number of fuzzy sets for each of the attributes (FS). Examples with missing values were removed.

Regarding the number of fuzzy sets describing each feature, they were defined by the FUZZY-DBD method [24], which chooses the number of fuzzy sets for each feature based on a heuristic process. Specifically, we used triangular shaped fuzzy sets evenly distributed in the features domains.

Table 1. Dataset characteristics

Dataset	Examples	Features	Classes	ME	FS
Credit	653	15(6,9)	2	45.33	2
Cylinder	277	32(19,13)	2	35.74	2
Dermatology	358	34(33,1)	6	68.99	2
Diabetes	769	8(8,0)	2	34.90	2
Glass	220	9(9,0)	7	65.46	7
Heart	270	13(13,0)	2	44.44	2
Ionosphere	351	34(34,0)	2	35.90	3
Segment	210	19(19,0)	7	85.71	3
Vehicle	846	18(18,0)	4	74.23	2
Wine	178	13(13,0)	3	59.74	3

Table 2 presents the original number of features of each dataset (F), the number of features selected by the FUZZY-WRAPPER method (FW), as well as the ones selected by the Fuzzy C4.5 (FC4.5), C4.5, CFS, ReliefF (RF) and Consistency (Cons.) methods. The average number of features is also presented. The best results are light-gray shaded.

Table 2. Original and selected number of features

Dataset	F	FW	FC4.5	C4.5	CFS	RF	Cons.
Credit	15	12	10	9	7	12	13
Cylinder	32	9	24	18	6	16	9
Dermatology	34	17	17	7	16	16	8
Diabetes	8	2	5	6	4	5	8
Glass	9	6	6	9	7	8	7
Heart	13	3	5	12	8	12	10
Ionosphere	34	7	10	14	14	33	7
Segment	19	8	10	12	7	15	9
Vehicle	18	10	11	18	11	18	18
Wine	13	7	8	3	10	13	5
Average	19.5	8.1	10.6	10.8	9	14.8	9.4

Results show that the FUZZY-WRAPPER method was able to select the smallest number of features for 5 out of the 10 datasets, with one tie with FC4.5. The FUZZY-WRAPPER also shows the best average number of selected features. ReliefF, on the other hand, had the worst average of selected features, reducing one of the original number of features for the glass, heart and ionosphere datasets, and no features for the wine and vehicle datasets.

In order to evaluate the quality of the selected sets of features we used the Wang & Mendel (WM) method to generate fuzzy rule bases and calculate their precision. The classification power of these sets of selected features can be seen as an indication of their quality to describe the dataset. It is important to notice that the WM method is a quite simple approach to generate fuzzy rule bases and was used in these experiments because of its simplicity and ease implementation only for preliminary comparisons.

Table 3 presents the error rates and standard deviation (in brackets) for the Fuzzy-Wrapper (FW/WM), fuzzy C4.5 (FC4.5) using the rule base induced by the fuzzy tree, fuzzy C4.5 (FC4.5/WM), C4.5 (C4.5/WM), CFS (CFS/WM), ReliefF (RF/WM), and Consistency (Cons/WM). The suffix WM indicates that the error rates were obtained by the rule bases generated by the Wang & Mendel method. The column division in the middle of the table separates the first three methods (FW/WM, FC4.5, and FC4.5/WM) that make the selection of features using the fuzzyfied data, from the others that use the original data. The best results are light-gray shaded. The last line of the table presents the ranking of the methods according to their error rates.

Notice that the FC4.5 is used as an embedded method, since it induces a fuzzy classifier and select features at the same time. On the other hand, for the FC4.5/WM and the C4.5/WM, the decision tree (fuzzy or not) is used only to select the features which appear in the generated branches. These features are then used by the WM method to generate the fuzzy rule base.

Table 3. Error rates

Dataset	All	FW/WM	FC4.5	FC4.5/WM	C4.5/WM	CFS/WM	RF/WM	Cons/WM
Credit	0.21(0.09)	0.23(0.06)	0.11(0.06)	0.43(0.08)	0.43(0.02)	0.43(0.08)	0.23(0.03)	0.28(0.04)
Cylinder	0.43(0.11)	0.35(0.1)	0.17(0.04)	0.35(0.09)	0.49(0.13)	0.52(0.13)	0.43(0.11)	0.56(0.12)
Derma	0.24(0.08)	0.15(0.06)	0.06(0.05)	0.24(0.08)	0.36(0.04)	0.24(0.11)	0.24(0.08)	0.24(0.08)
Diabetes	0.40(0.07)	0.40(0.04)	0.30(0.05)	0.55(0.07)	0.51(0.06)	0.51(0.08)	0.47(0.06)	0.40(0.05)
Glass	0.42(0.02)	0.37(0.08)	0.25(0.16)	0.65(0.10)	0.50(0.07)	0.54(0.05)	0.42(0.07)	0.48(0.08)
Heart	0.47(0.03)	0.19(0.03)	0.19(0.07)	0.47(0.05)	0.31(0.06)	0.24(0.03)	0.32(0.02)	0.29(0.08)
Iono	0.34(0.05)	0.24(0.07)	0.08(0.06)	0.34(0.03)	0.44(0.04)	0.41(0.08)	0.36(0.09)	0.39(0.10)
Segment	0.25(0.02)	0.21(0.01)	0.12(0.02)	0.26(0.05)	0.26(0.07)	0.27(0.03)	0.26(0.03)	0.34(0.01)
Vehicle	0.76(0.02)	0.36(0.03)	0.37(0.05)	0.61(0.06)	0.43(0.09)	0.52(0.09)	0.43(0.10)	0.43(0.03)
Wine	0.61(0.08)	0.03(0.04)	0.05(0.07)	0.11(0.08)	0.13(0.04)	0.05(0.05)	0.10(0.01)	0.14(0.01)
Ranking	4	2	1	5	8	7	3	6

For this setup, the fuzzy C4.5 classifier (FC4.5) was able to obtain the smallest error rates for 7 out of 10 datasets, with one tie with the Fuzzy-Wrapper method, which obtained the smallest error rates for the remaining datasets. The FC4.5 also performed reasonably well in terms of feature selection. The Fuzzy-Wrapper, on the other hand, was the best in terms of selected features. Nevertheless, it is important to notice that the use of wrappers might not be a reasonable alternative for feature selection for datasets with many features due to its complexity. The features selected by the ReliefF filter produced the third best results, but it is important to bear in mind that ReliefF was the worst approach in terms of selected features.

Other important fact to be taken into consideration is that WM didnot yield good results for any approach, but for the Fuzzy-Wrapper. On the contrary, in many cases the results are worse than the majority error. Nevertheless, the excellent performance of the fuzzy C4.5 both in terms of precision and number of rules (discussed next) can be seen as the required motivation for further experiments and comparisons with other more complex fuzzy rule base inducers, such as genetic approaches.

To test whether there is a difference among the algorithms we used the Friedman test [25] with the null-hypothesis that the performance of all algorithms, assessed in terms of the error rates, are comparable. The last line of Table 3 shows the Friedman ranking. Compared to the Friedman statistics, we could reject the null-hypothesis at 95% of confidence level and then proceed with the Dunn test [25] to compare each algorithm against the others. Results show that the Fuzzy-Wrapper method is better than the CFS method and significantly better than the C4.5 (comparing the rule base generated by the WM method). Results also show that the fuzzy C4.5 using the rule base of the induced trees is significantly better than fuzzy C4.5 and ReliefF, both using the WM fuzzy rule base. Results also show that the fuzzy C4.5, using the decision tree rule base, is extremely better than the C4.5 and CFS, both using the WM fuzzy rule base.

Although the aim of this paper is to assess the methods for feature subset selection, another important aspect to be considered is the interpretability of the resulting rule bases. The Wang & Mendel method generates rules with a conjunction for each input feature in the antecedent of the rules, *i.e.*, if we have m features for a given dataset, each resulting rule will have m conjunctions. Also, the number of rules generated by the WM method is closely connected to the number of features and the number of fuzzy sets describing each feature. On the other hand, with the fuzzy C4.5 method the rules present a variable number of conjunctions in their antecedent and there is no fixed connection between the number of features and the number of resulting rules. This flexibility of the decision trees greatly improves the interpretability of the resulting rule base and is an important advantage when compared to the rule bases generated by the WM method, for instance. In order to allow an overall comparison, for our experiments the number of rules generated by the proposed fuzzy C4.5 algorithm ranged from approximately 5% to 50% of the number of rules generated by the WM method.

5 Conclusion

Fuzzy systems generation have been the focus of a strong research effort in the last years. Depending on the adopted approach, a preselection of features might be essential for the process. Feature selection can also improve the interpretability of the induced rules. The use of filters for this task is a common choice, since they are simple, fast and use general characteristics found in the data that might suit different domains and tasks. Wrappers are also another option which uses the induction algorithm itself to select features, although they are usually

computational costly due to the complexity of the process. A third option is the use of embedded methods, such as decision trees, which induce a classifier and select features at the same time.

This paper proposes a simple version of the Fuzzy C4.5 method to be used for attribute selection, and compares its results with a fuzzy wrapper, 3 filters and the C4.5 classic decision tree algorithm, aiming at proposing fuzzy alternatives to feature selection.

Results show that the fuzzy decision tree is able to select subsets of features with high classification power, being a good alternative to wrapper approaches for datasets with many features. The fuzzy C4.5 algorithm decides the granulation of the features domains in terms of fuzzy sets for the induction process. Furthermore, it is a very simple and fast algorithm, having no restrictions for large datasets described by many features.

As stated before, the WM method is a quite basic approach to generate fuzzy rule bases. This way, we intend to investigate the use of the fuzzy C4.5 method to select features for a genetic generation of fuzzy rule bases. We also intend to investigate the impact of different pruning rates for the generation of the fuzzy C4.5 decision tree.

Acknowledgments

This research was supported by CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico and the São Paulo State Research Foundation - FAPESP (Proc. No. 2007/05390-0), Brazil.

References

1. Cordon, O., Gomide, F.A.C., Herrera, F., Hoffmann, F., Magdalena, L.: Ten years of genetic fuzzy systems: Current framework and new trends. Fuzzy Sets and Systems 141(1), 5–31 (2004)
2. Dumitrescu, D., Lazzerini, B., Jair, L.: Fuzzy Sets and Their Application to Clustering and Training. Int. Series on Computational Intelligence. CBC Press (2000)
3. Zadeh, L.: Fuzzy sets. Information and Control 8, 338–353 (1965)
4. Sanchez, L., Villar, J.R., Couso, I.: Genetic feature selection for fuzzy discretized data. In: Proc. of the IPMU 2008, pp. 1159–1166 (2008)
5. Choi, Y.S., Moon, B.R.: Feature selection in genetic fuzzy discretization for pattern classification problems. IEICE - Trans. Inf. Syst. E90-D(7), 1047–1054 (2007)
6. Li, R.P., Mukaidono, M., Turksen, I.B.: A fuzzy neural network for pattern classification and feature selection. Fuzzy Sets and Systems 130(1), 101–108 (2002)
7. Grande, J., del Rosario Suarez, M., Villar, J.R.: Advances in Soft Computing - A Feature Selection Method Using a Fuzzy Mutual Information Measure, vol. 44. Springer, Heidelberg (2007)
8. Sánchez, L., Suárez, M.R., Villar, J.R., Couso, I.: Mutual information-based feature selection and fuzzy discretization of vague data. International Journal of Approximate Reasoning 49(3), 607–622 (2008)

9. Alonso, J.M., Magdalena, L.: An interpretability-guided modeling process for learning comprehensible fuzzy rule-based classifiers. In: International Conference on Intelligent Systems Design and Applications, pp. 432–437 (2009)
10. Cintra, M.E., de Arruda Camargo, H., Monard, M.C.: Fuzzy feature subset selection using the Wang & Mendel method. In: HIS 2008, vol. 1, pp. 590–595 (2008)
11. Wang, L.: The WM method completed: a flexible fuzzy system approach to data mining. IEEE Transactions on Fuzzy Systems 11, 768–782 (2003)
12. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research 5, 1205–1224 (2004)
13. Hall, M.A.: Correlation-based feature selection for discrete and numeric class Machine Learning. In: Proceedings of the 17th International Conference on Machine Learning, pp. 359–366. Morgan Kaufmann, San Francisco (2000)
14. Kira, K., Rendell, L.: A practical approach to feature selection. International Conference on Machine Learning 1, 368–377 (1992)
15. Liu, H., Setiono, R.: A probabilistic approach to feature selection - a filter solution. In: Proceedings of the 13th International Conference on Machine Learning, vol. 1, pp. 319–327 (1996)
16. Das, S.: Filters, wrappers and a boosting-based hybrid for feature selection. In: Proceedings of the 18th Int. Conf. on Machine Learning, pp. 74–81 (2001)
17. Cintra, M.E., Martin, T.P., Monard, M.C., Camargo, H.A.: Feature subset selection using a fuzzy method. In: International Conference on Intelligent Human-Machine Systems and Cybernetics, vol. 2, pp. 214–217 (2009)
18. Quinlan, J.R.: C4.5 Programs for Machine Learning. Morgan Kaufmann, CA (1988)
19. Quinlan, J.R.: Bagging, boosting and c4.5. In: Proceedings of the 13th Conf. Artificial Intelligence, pp. 725–730 (1996)
20. Olaru, C., Wehenkel, L.: A complete fuzzy decision tree technique. Fuzzy Sets and Systems 138(2), 221–254 (2003)
21. Kazunor, H., Motohide, U., Hiroshi, S., Yuushi, U.: Fuzzy C4.5 for generating fuzzy decision trees and its improvement. Faji Shisutemu Shinpojiumu Koen Ronbunshu 15, 515–518 (1999)
22. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2nd edn. Morgan Kaufmann, San Francisco (2005)
23. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
24. Cintra, M.E., Camargo, H.A., Martin, T.P.: Optimising the fuzzy granulation of attribute domains. In: International Fuzzy Systems Association World Conference, pp. 742–747 (2009)
25. Demšar, J.: Statistical comparison of classifiers over multiple data sets. Journal of Machine Learning Research 7(1), 1–30 (2006)

Restricting the IDM for Classification

Giorgio Corani and Alessio Benavoli

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
CH-6928 Manno, Lugano, Switzerland
{giorgio,alessio}@idsia.ch

Abstract. The *naive credal classifier* (NCC) extends *naive Bayes classifier* (NBC) to imprecise probabilities to robustly deal with the specification of the prior; NCC models a state of ignorance by using a *set* of priors, which is formalized by Walley's *Imprecise Dirichlet Model* (IDM). NCC has been shown to return more robust classification than NBC. However, there are particular situations (which we precisely characterize in the paper) under which the extreme densities included by the IDM force NCC to become very indeterminate, although NBC is able to issue accurately classifications. In this paper, we propose two approaches which overcome this issue, by restricting the set of priors of the IDM . We analyze both approaches theoretically and experimentally.

1 Introduction

The *naive Bayes classifier* (NBC) is often accurate, despite the unrealistic assumption of independence of the features given the class. However, especially on small data sets, NBC can happen to issue *prior-dependent* classifications, i.e., the most probable class varies depending on the adopted prior. This is acceptable if the prior can be carefully elicited to model domain knowledge; otherwise, prior-dependent classifications can be fragile. Usually, NBC is learned using a uniform prior, in the attempt of being *non-informative*. Yet, this solution is hardly satisfactory because the uniform prior models *indifference* rather than ignorance and anyway the choice of any single prior implies some arbitrariness.

The *naive credal classifier* (NCC) extends NBC to imprecise probabilities to robustly deal with the specification of the prior density; NCC models a state of ignorance by using a *set* of priors, which is formalized by Walley's *Imprecise Dirichlet Model* (IDM) (see [2] for a tutorial). IDM satisfies several properties desirable to model prior ignorance, such as the *representation invariance principle* (RIP) and the *likelihood principle* (LP) [2].

NCC turns the set of priors into a set of posteriors by element-wise application of Bayes rule; eventually, it returns all the classes which are *non-dominated*[1] within the set of posteriors. In fact, NCC returns a set of classes when faced with instances whose classification is prior-dependent; it issues weaker but more robust classifications than NBC. We call *determinate* the classifications made of a single class, and *indeterminate* the others.

[1] The definition of dominance is given in Section 2.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 328–337, 2010.

Yet, there are particular situations (which we precisely characterize in the paper) under which the extreme densities included by the IDM force NCC to become very indeterminate; as pointed out in [7, Sec. 6], this behavior is correct in principle, since it shows that the classifications issued by NBC are prior dependent. Yet in such cases the large indeterminacy of NCC is questionable, as it is mostly due to extreme (namely, very skewed) priors; in fact, NCC becomes more determinate if such extreme priors are removed. Moreover, in such cases NBC (learned with uniform prior) achieves good accuracy on the instances indeterminately classified by NCC, which further shows an excessive indeterminacy of NCC. A way to increase the determinacy of NCC in such cases is to remove these extreme densities by restricting the IDM's set of priors by a small amount (an ϵ), in order to remove the boundary. In this paper we propose two approaches to cut off the extreme densities in the IDM; in both cases, the amount of densities removed from IDM is controlled by the parameter $\epsilon > 0$. The value of ϵ determines a trade-off between robustness and informativeness of the issued classifications: increasing ϵ increases informativeness, at a cost of some robustness. The setting $\epsilon = 0$ corresponds to the IDM, which is maximally robust but, at least in such particular cases, leads to a questionable high indeterminacy.

An alternative approach for restricting the credal set by modelling domain knowledge is given in [1], where only those priors that guarantee an improvement of the Mean Squared Error over the Maximum Likelihood Estimator are included in the credal set. In [1] it is also shown that removing extreme densities from the IDM is equivalent to express preferences among subregions of the parameter space; from this viewpoint, the two approaches proposed in this paper are informative; thus, as we show in Sec. 2.1, they cannot satisfy at the same time both RIP and LP.

By experiments, we show under which conditions NCC, learned with Walley's IDM, can become unnecessarily indeterminate; we compare its behavior against that of an alternative credal classifier, CMA (credal model averaging) [3]. Then, we show that NCC becomes more determinate without compromising its reliability, when the two approaches for restricting the IDM are applied.

2 Naive Credal Classifier

NCC models prior near-ignorance by a set of priors; the set is formally defined by using Walley's Imprecise Dirichlet Model (IDM) [6]. NCC updates each prior with the observed likelihood, via element-wise application of Bayes' rule; in this way, NCC turns the set of priors into a set of posteriors. Let us denote the classification variable by C, taking values in the finite set $\mathcal{C}$, where the possible classes are denoted by lower-case letters. We have k features $F_1, \ldots, F_k$ taking generic values $[f_1, \ldots, f_k] = \mathbf{f}$ from the sets $\mathcal{F}_1, \ldots, \mathcal{F}_k$; the features are assumed to be *discrete*. We denote by $\theta_{c,\mathbf{f}}$ the real unknown probability (*chance*) that $(C, F_1, \ldots, F_k)$ equals $(c, \mathbf{f})$, by $\theta_{f_i|c}$ the chance that $F_i = f_i$ conditional on c and by $\theta_{\mathbf{f}|c}$ the chance of $(f_1, \ldots, f_k)$ conditional on c. Let N be the total number of samples; let $n(c)$ and $n(f_i\,|c)$ be the observed frequencies of class

c and of $(f_i\,|c)$. NCC, like NBC, (naively) assumes the independence of the attributes given the class $\theta_{\mathbf{f}|c} = \prod_{i=1}^{k} \theta_{f_i|c}$. The likelihood function is:

$$L(\mathbf{n}|\theta) \propto \prod_{c\in\mathcal{C}} \left[\theta_c^{n(c)} \prod_{i=1}^{k} \prod_{f_i\in\mathcal{F}_i} \theta_{f_i|c}^{n(f_i|c)} \right], \tag{1}$$

where $\mathbf{n}$ denotes the vector of all the above frequencies. Observe that for all c and i, the observations satisfy the *structural constraints* $0 \leq n(f_i\,|c) \leq n(c)$, $\sum_c n(c) = N$ and $\sum_{f_i\in\mathcal{F}_i} n(f_i\,|c) = n(c)$. The prior density is expressed similarly to the likelihood function, except that frequencies $n(\cdot)$ are replaced everywhere by $st(\cdot) - 1$, i.e., the prior is a Dirichlet density with parameters $\alpha(\cdot) = st(\cdot)$. The parameter s is a positive real number which can be regarded as the number of *hidden samples*, in the common interpretation of conjugate Bayesian priors as additional sample units (the number can be fractional, though); the parameters $t(\cdot)$ can be regarded as the proportion of units of the given type; for instance, $t_{c'}$ is the proportion of hidden units having class c' in the hidden samples. Theoretical considerations suggest that s should lie between 1 and 2 [2], while the $t(\cdot)$ are usually set according to the *uniform* prior: $t(c) = \frac{1}{|\mathcal{C}|}$ and $t(a_i|c) = \frac{1}{|\mathcal{C}||\mathcal{F}|}$. By multiplying the prior density and the likelihood function, we obtain a posterior density of the same form as the likelihood, with $n(\cdot)$ replaced by $st(\cdot) + n(\cdot) - 1$. We estimate the posterior joint probability of class and features by taking expectation over the posterior probability of θ, i.e., $P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t})$ equal to:

$$P(c|\mathbf{n}, s, \mathbf{t}) \prod_{i=1}^{k} P(f_i|c, \mathbf{n}, s, \mathbf{t}) = \frac{n(c) + st(c)}{N + s} \prod_{i=1}^{k} \frac{n(f_i\,|c) + st(f_i\,|c)}{n(c) + st(c)}. \tag{2}$$

Equation (2) is the posterior probability densities of class and features returned by NBC. However, the specification of any single prior entails the risk of issuing fragile prior-dependent classifications. Walley's IDM overcomes this problem, by letting the parameters $\mathbf{t}$ vary within intervals instead of being fixed to precise values. In particular, $\mathbf{t}$ vary within the polytope $\mathcal{T}$, defined by the following constraints:

$$\mathcal{T} := \left\{ \sum_{c\in\mathcal{C}} t(c) = 1,\ \sum_{f_i\in\mathcal{F}_i} t(f_i\,|c) = t(c),\ 0 < t(f_i\,|c) < t(c) \quad \forall(i, f_i, c) \right\}. \tag{3}$$

Thus, IDM takes into consideration all the priors densities which belong to the simplex $\mathcal{T}$. Notice that, the above constraints are necessary and sufficient conditions to ensure that all the densities, obtained by letting $\mathbf{t}$ vary in $\mathcal{T}$, are proper. Walley's IDM satisfies the *representation invariance principle* because the uncertainty about any event does not depend on refinements or coarsening of categories; the *likelihood principle*, because posterior inferences depend on the data through the likelihood function only. The specific approach used by NCC [7] to identify the non-dominated classes is called *maximality* [6]. Consider the $1 - 0$ utility functions associated with the actions of choosing class c' or c''.

The family of posterior probabilities $P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t})$ (obtained by letting $\mathbf{t}$ vary in $\mathcal{T}$) are used to determine the lower expected utility of deciding between c' or c''. Class c' dominates c'' if the expected utility w.r.t. $P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t})$ of choosing a class c' over c'' is strictly positive for each $\mathbf{t} \in \mathcal{T}$. In the case of NCC, c' dominates c'' if and only if [7]:

$$\inf_{\mathbf{t}\in\mathcal{T}} \frac{P(c', \mathbf{f}|\mathbf{n}, \mathbf{t}, s)}{P(c'', \mathbf{f}|\mathbf{n}, \mathbf{t}, s)} = \inf_{\mathbf{t}\in\mathcal{T}} \left[\frac{n(c'') + st(c'')}{n(c') + st(c')}\right]^{k-1} \prod_{i=1}^{k} \frac{n(f_i\,|c') + st(f_i\,|c')}{n(f_i\,|c'') + st(f_i\,|c'')} > 1. \tag{4}$$

When faced with a prior-dependent instance, NCC identifies more non-dominated classes and issues an indeterminate classification, thus preserving reliability.

2.1 Restricting IDM

As already observed, by considering all prior Dirichlet densities such that $0 < t(c) < 1$ and $0 < t(f_i\,|c) < t(c)$ for all i and c, IDM excludes the extremes of the simplex $\mathcal{T}$, which correspond to improper densities. This means that the simplex $\mathcal{T}$ is obtained by restricting the set $0 \leq t(c) \leq 1$ and $0 \leq t(f_i\,|c) \leq t(c)$ by an *arbitrary small* ϵ. If $n(f_i\,|c) > 0$ and $n(f_i\,|c) < n(c)$ for each feature i and class c, the posterior densities corresponding to the extremes of the simplex $\mathcal{T}$ are proper for any choice of ϵ. Thus, in this case, we can let ϵ go to zero. This is proved in [7] for the optimization in (4). In particular, it is shown that the infimum of (4) is obtained by letting $t(f_i\,|c') \to 0$ and $t(f_i\,|c'') \to t(c'')$, thus using extreme densities (i.e., $\epsilon = 0$). Then, problem (4) is solved by optimizing on $t(c'')$ only. Since function (4) is convex with respect to $t(c'')$, the minimization can be solved exactly and efficiently.

In some cases, the use of extreme densities in (4) generate what we call the *class problem* and the *feature problem*. The class problem, already observed in [7, Sec. 6], takes place when a class c'' is never observed in the sample; in this case, it is difficult for an alternative class c' to dominate c'': for any value f_i of any feature, there are no data for estimating $P(f_i|c'')$, which therefore under the IDM varies between 0 and 1 and is set to 1 during the minimization. As this behavior repeats for each feature, $P(\mathbf{f}_i|c') \ll P(\mathbf{f}_i|c'')$, thus often preventing an alternative class c' to dominate c''. In fact, c'' will be often identified as non-dominated. The *feature problem* happens instead when there are no observations of one or more values of a certain feature conditional on class c'. In this case, there are no observations for estimating $P(f_i|c')$, which goes to sharp to zero during the solution of (4); this leads to sharp 0 also $P(c', \mathbf{f}|\mathbf{n}, \mathbf{t}, s)$, because of F_i alone, regardless the information coming from all the remaining features. When either the class or the feature problem happen, NCC can get very indeterminate, while at the same time NBC achieves good accuracy on the instances indeterminately classified by NCC; this can be seen as disappointing behavior of NCC.

Note that if $n(c'') = 0$ (class problem) or $n(f_i\,|c') = 0$ (feature problem), the choice of extreme prior densities ($t(c'') = 0$ and, respectively, $t(f_i\,|c') = 0$) lead to improper posteriors. Although, in this case, IDM is still well-defined, we cannot let ϵ go sharp to zero in the optimization in (4). Therefore, the set of

posteriors and the set of non-dominated classes will depend on the choice of ϵ. In this paper, we propose two approaches to remove the extreme densities from $\mathcal{T}$, in order to increase the NCC determinacy.

The first approach uses the following restricted set for $\mathbf{t}$:

$$\mathcal{T}_\epsilon = \left\{ t(c'), t(c'') \geq \epsilon, t(c') + t(c'') = 1, \epsilon \leq t(f_i|c') \leq t(c'), \epsilon \leq t(f_i|c'') \leq t(c'') \right\} \tag{5}$$

where $\epsilon \in (0, 0.5]$; we call NCC_ϵ the resulting classifier. Such an approach is appropriate to deal with the feature problem, as it guarantees $t(f_i|c') \geq \epsilon$ and therefore avoids sharp zeros in the computation of the numerator; however, it should not be too effective against the class problem, as the conditional probabilities at the denominator will nevertheless reach $1 - \epsilon$. Moreover, the credal set of (5) satisfies the RIP, as it is not dependent on the number of categories. However, this comes at a cost. In fact, Eq.(5) requires to adjust the boundary of the credal set on every different pairwise comparison. For instance, when comparing c' with c'', $t(c')$ and $t(c'')$ are lower-bounded by ϵ while $t(c) = 0$ for all the remaining classes; but when comparing c' against c''', $t(c')$ and $t(c''')$ are lower-bounded by ϵ, while $t(c'')$ (and all the remaining $t(c)$) goes to 0 . Therefore, such an approach does not respect the likelihood principle as the set of priors depends on the couple of classes under exam. Moreover, it cannot be guaranteed that if c' dominates c'' and c'' dominates c''', then also c' dominates c''' (*transitivity*), because the pairwise comparisons are in fact performed on credal sets having different boundaries. For this reason, this approach should be used with small values of ϵ, so to enable addressing the feature problem while only minimally perturbing the credal set of the IDM. Using a value of ϵ comprised between 0.01 and 0.1, transitivity has been however always satisfied in our experiments. When the priors are restricted to be in $\mathcal{T}_\epsilon$, the analytical optimization procedure described in [7] remains valid, because the derivatives of function are unchanged compared to [7].

The second approach is based on a ϵ-contamination of the uniform prior of NBC with the set of priors in $\mathcal{T}$, which results in the set:

$$\mathcal{T}_c := \begin{cases} \sum_{c \in \mathcal{C}} t(c) = 1, \quad t(c) \in \left[\epsilon_0 \frac{1}{|\mathcal{C}|}, \epsilon_0 \frac{1}{|\mathcal{C}|} + (1 - \epsilon_0) \right], \\ \sum_{f_i \in \mathcal{F}_i} t(f_i\,|c) = t(c), \quad t(f_i|c) \in \left[\epsilon_i \frac{t(c)}{|\mathcal{F}_i|}, \epsilon_i \frac{t(c)}{|\mathcal{F}_i|} + (1 - \epsilon_i) t(c) \right], \quad \forall (i, c) \end{cases} \tag{6}$$

where the ϵ_0 refers to the class variable, while for each feature a different parameter $\epsilon_i \in (0, 1)$ can be specified. We call NCC_c the resulting classifier, where c stands for "contaminated". This approach, unlike the previous one, satisfies LP (no dependence of the set of priors on the data) but not RIP, since the priors depend on the number of classes (through $1/|\mathcal{C}|$) and number of categories of the features (through $1/|\mathcal{F}_i|$). The minimization problem, has to be numerically approximated because the interval for $t(f_i\,|c)$ depends on $t(c)$ and function (4) is not convex in $t(c)$. When $\epsilon_0 \to 1$ and $\epsilon_i \to 1$ $\forall i$, the set of priors collapses to the uniform prior and thus the classifier coincides with the NBC. Instead, NCC_ϵ

never reduces to a single prior; with $\epsilon = 0.5$ it uses a single prior to compare a couple of classes, but this prior changes with the couple of classes.

3 Credal Model Averaging

Let us consider NBC again: given k features, there are 2^k possible NBCs, each characterized by a different subset of features; we denote by $\mathcal{M}$ the set of such models and by m a generic model. By feature selection, one can identify a single best feature set and then work with a single NBC. An alternative approach is *Bayesian Model Averaging* (BMA), which instead averages over *all* the 2^k different NBCs, the weight assigned to each classifier being proportional to its posterior probability. The joint probability $P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t})$ is obtained by marginalizing m out:

$$P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t}) \propto \sum_{m \in \mathcal{M}} P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t}, m) P(\mathbf{n}|m, s, \mathbf{t}) P(m), \tag{7}$$

where $P(c, \mathbf{f}|\mathbf{n}, s, \mathbf{t}, m)$ is the posterior probability of $c, \mathbf{f}$ computed by m, $P(\mathbf{n}|m, s, \mathbf{t})$ represents the *likelihood* of model m and $P(m)$ the prior probability of m. Dash and Cooper [5] provide an exact and efficient algorithm to compute BMA over 2^k NBCs. This algorithm has been extended to imprecise probabilities in [3], giving rise to credal model averaging (CMA) . In particular, CMA specifies a **set** of prior over the models instead of adopting a single $P(m)$; in fact, CMA imprecisely averages over the 2^k NBCs. CMA is free from both the feature problem and the class problem, as its base classifiers are NBCs.

4 Comparing Credal Classifiers

In order to completely describe the performance of a credal classifier, 4 indicators are necessary: *determinacy*: i.e, the percentage of determinate classifications; *single accuracy*: the accuracy of the classifier when determinate; *set-accuracy*: the accuracy of the classifier when indeterminate; *indeterminate output size*: the average number of classes returned by the classifier when indeterminate. Instead, to compare credal classifiers we adopt two metrics which have been introduced in [4]. We refer to a classifier as *accurate* on a certain instance if its output includes the correct class, regardless how many classes it has returned; we refer to a classifier as *determinate* if its output contains only a single class. The *discounted-accuracy* is: $d\text{-}acc = \frac{1}{n} \sum_{i=1}^{n} (accurate)_i / |Z_i|$, where $(accurate)_i$ is a 0-1 variable, showing whether the classifier is accurate or not on the i-th instance; $|Z_i|$ is the number of classes returned on the i-th instance. Yet, there is no reason for *linearly* discounting the accuracy on the number of returned classes; an alternative non-parametric approach proposed in [4] removes this arbitrariness, being based on a *rank test*. The *rank test* is more robust than *d-acc*, as it does not encode any (arbitrary) functional form for discounting accuracy on the basis of the output size; yet, it uses less pieces of information than d-acc and can be therefore be less sensitive. Overall, a cross-check of both metrics is recommended.

5 Results

We presents results on 45 classification data sets; they are publicly available from the WEKA data sets page.[2] Over each data set, we perform 10 runs of 10-folds cross-validation, namely 100 training/test experiments. Numerical features have been discretized via the entropy-based discretization; within each training-test experiment, the bins are learned on the current training set and then applied unchanged on the current testing set. The comparison of BMA and NBC shows 11 wins for NBC, 25 ties, 9 wins for BMA (over each data sets, the accuracies measured during cross-validation have been compared with a t-test, with $\alpha = 5\%$). There is therefore a balance between the two classifiers. Instead, when considering credal classifiers, CMA clearly dominates NCC: according to the rank test [or the discounted accuracy], there are 23 [26] wins for CMA, 17 [14] ties and 5 [5] wins for NCC[3]. In particular, CMA has much larger determinacy than NCC (on average, 95% vs 76%) and also higher discounted accuracy (0.76 vs 0.70 on average). We must recall that CMA also includes an ϵ parameter, which controls the determinacy of CMA. Yet, even adopting different values of ϵ, CMA remains much more determinate than NCC. The scatter plots of determinacy and discounted accuracy for the two classifiers are in Fig. 1.

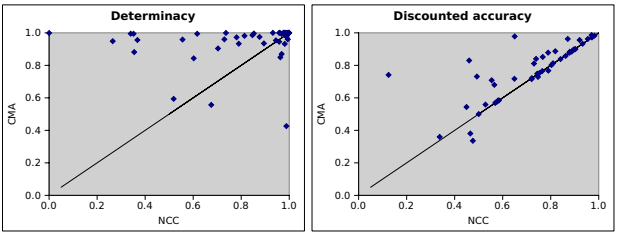


Fig. 1. Scatter plot of determinacy and the discounted-accuracy of CMA and NCC

We focus on three data sets which highlight the consequences of the class and the feature problem; the characteristics of the data sets and the performance of the classifiers are shown in Tab.1. NCC has very low determinacy on these three data sets; this implies that *many* instances are classified in prior-dependent way by NBC. Yet, the classifications issued by NBC using the uniform prior are quite accurate; in particular, they are much more accurate than simply returning the majority class, as shown in Table 1. The model of prior-ignorance which characterizes NCC is indeed theoretically sound, but in these cases its large indeterminacy appears questionable. On squash-stored, NCC suffers from the feature problem: the feature *fruit* has 22 states and requires to estimate 66 parameters for the conditional densities, from only 52 instances; removing this feature increases the NCC determinacy from 31% to 60%. Instead, NCC_ϵ properly deals

[2] http://www.cs.waikato.ac.nz/~ml/weka/index_datasets.html

[3] On each data set, the values discounted accuracy measured for NCC and CMA during cross-validation have been compared via t-test, significance 5%.

Table 1. Comparison of NCC and CMA on three data sets especially difficult for NCC. Majority is the percentage of instances belonging to the most frequent class in the data set.

Data set	N	Feats	$\lvert\mathcal{C}\rvert$	Majority	Accuracy	
					NBC	BMA
primary-tumor	339	17	22	25%	46%	36%
audiology	226	69	24	25%	79%	73%
squash-stored	52	24	3	44%	66%	59%

Data set	Determ.		Disc-acc		NBC accuracy when	
	NCC	CMA	NCC	CMA	NCC det.	NCC ind.
primary-tumor	10%	88%	0.19	0.36	70%	43%
audiology	7%	95%	0.21	0.70	98%	78%
squash-stored	32%	84%	0.49	0.58	70%	63%

with this feature: even using a small ϵ (0.01), determinacy increases from 32% to 42% and discounted-accuracy from 0.48 to 0.57, approaching that of CMA. Moreover single-accuracy (accuracy when determinate) also increases from 70% to 79%, showing that the feature problem prevents NCC to extract useful information from the remaining features. With NCC_c, it is instead necessary to use a larger ϵ (recall that NCC_c lower-bounds the conditional probabilities in the numerator of Eq.(4) by $\frac{\epsilon}{|F_i||\mathcal{C}|}$); for instance, with ϵ=0.1, NCC_c achieves determinacy 36% with discounted-accuracy of 0.53. Moreover, NCC_c too has higher the single-accuracy than NCC. Note that for both NCC_ϵ and NCC_c, adopting increasing ϵ would steadily increase determinacy, as in fact it will reduce the credal set and thus the probability of the instance being prior-dependent. Instead, it cannot be foreseen how the discounted accuracy will vary when ϵ is increased, as this depends on the trade-off between determinacy and accuracy, which cannot be predicted in advance.

The low determinacy of NCC on both audiology and primary-tumor is instead due to the class problem, as several classes are never observed, or observed only once or twice; in fact, removing these classes from the data set largely increases the NCC determinacy. However, NCC_ϵ does not address the class problem, as already pointed out; it is therefore more interesting analyse the behavior of NCC_c. In Fig.2, we show how the main indicators of performance of NCC_c vary with different values of ϵ; for ϵ=1, the classifier corresponds to NBC. This plots highlight the trade-off between robustness and determinacy: increasing ϵ implies higher determinacy, which however comes generally at a cost of some accuracy, both on the instances determinately and indeterminately classified. Domain knowledge can suggest which is a reasonable choice of ϵ. As a last experiment, we have run NCC_ϵ and NCC_c, setting for both $\epsilon = 0.01$, on all the data sets. Overall, both classifiers achieve determinacy and discounted-accuracy

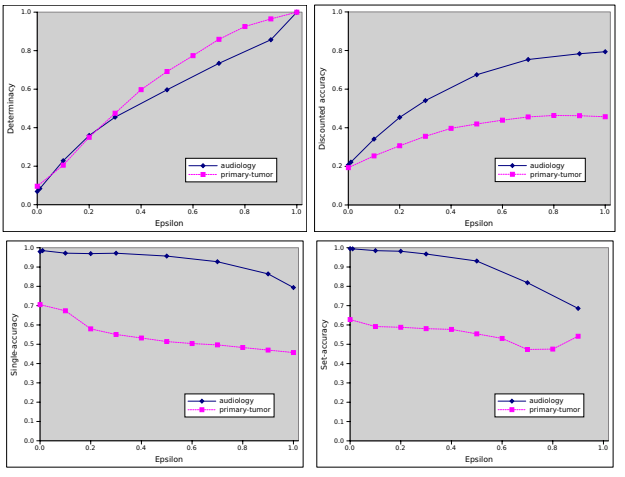


Fig. 2. Sensitivity of NCC_c to the value of ϵ. For ϵ=1, NCC_c corresponds to NBC; therefore, determinacy is 100% and set-accuracy is not measurable.

which is significantly higher than that of NCC, although the impact is generally much lighter than in the three extreme examples previously analyzed.

6 Conclusions

We have presented two approaches to restrict the set of priors of the IDM, in order to overcome the large indeterminacy of NCC, when dealing with what we have called the feature problem and the class problem, discussing advantages and disadvantages of such approaches from a theoretical point of view. Then, by experiments, we have shown that on the data sets where such two problems heavily penalize the NCC determinacy, a small restriction of the set of priors considerably increase the determinacy of the classifier without penalizing its reliability. This is particularly important on real problems, where a trade-off between informativeness and robustness is desirable. As future work, these two approaches could constitute a starting point to design a new classifier, which performs credal model averaging over NCCs characterized by different sets of features. In this case, restricting the imprecision could be a key-issue to manage the quantity of returned indeterminate classifications.

Acknowledgements

This work has been partially supported by the Swiss NSF grants n. 200020-121785/1, 200021-118071/1 and “Hydronet” project. We also thank the referees for their helpful remarks.

References

1. Benavoli, A., de Campos, C.P.: Inference from multinomial data based on a MLE-dominance criterion. In: Proc. of the 10th European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, pp. 22–33. Springer, Heidelberg (2009)
2. Bernard, J.-M.: An introduction to the Imprecise Dirichlet Model for multinomial data. Int. J. of Approximate Reasoning 39(2-3), 123–150 (2005)
3. Corani, G., Zaffalon, M.: Credal Model Averaging: An Extension of Bayesian Model Averaging to Imprecise Probabilities. In: Proc. of the 2008 European Conf. on Machine Learning and Knowledge Discovery in Databases, pp. 257–271 (2008)
4. Corani, G., Zaffalon, M.: Lazy naive credal classifier. In: Proc. of the 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data, pp. 30–37. ACM, New York (2009)
5. Dash, D., Cooper, G.F.: Model Averaging for Prediction with Discrete Bayesian Networks. J. of Machine Learning Research 5, 1177–1203 (2004)
6. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, New York (1991)
7. Zaffalon, M.: Statistical inference of the naive credal classifier. In: ISIPTA 2001: Proc. of the Second Int. Symp. on Imprecise Probabilities and Their Applications, The Netherlands, Shaker, pp. 384–393 (2001)

Estimation of Possibility-Probability Distributions

Balapuwaduge Sumudu Udaya Mendis and Tom D. Gedeon

School of Computer Science, The College of Engineering and Computer Science,
The Australian National University, Australia
{sumudu,tom}@cs.anu.edu.au

Abstract. We demonstrate a theory for evaluating the likelihood of a probability by way of possibility distributions. This theory derives from the standard probability distribution theory by using the possibility to define an arbitrary function whose values are bounded by $[0, 1]$ that represents the confidence that one may have in the outcomes. In other words, when in classic probability theory the probability of an event is represented by an integral of the probability mass over this event, in possibility theory the probability of an event is the integral of the probability mass times the confidence function over the whole space. This theory is then extended in order to define a similar notion to probability distributions, namely Possibility-Probability distributions, which represent, as for probabilities, the possibilities of a calculated probability for a given fuzzy event. In this context, we aim to define an estimation method of such a Possibility-Probability distribution in the case of experimental samples and the corresponding distribution.

Keywords: Possibility, Probability, Probability of Fuzzy Events, and Possibility of Probability.

1 Introduction

When only small samples of data and/or unreliable data is available, first order uncertainty calculations based on ordinary probability is questioned in the literature [1,2]. In this paper, we introduce an enhanced theory for evaluating the likelihood of the probability by way of possibility distributions. Basically, this theory derives from the standard probability distribution theory by using possibility as an an extra confidence measure. The confidence is represented as an arbitrary function whose values lie between 0 and 1, and represent the confidence that one may have in the outcomes. In classic probability theory, the probability of an event is represented by a measurable set, and is the integral of the probability measure over this set. However, in possibility theory the probability of an fuzzy event is the integral of the probability measure weighted by the confidence function over the whole space. This confidence is in general equivalent to the membership function of that particular event.

Next, we extend the method in order to define a similar notion to probability distributions, namely Possibility-Probability Distributions (PPD), which

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 338–347, 2010.

represent, as for probabilities, the possibilities of a calculated probability for a given fuzzy event. In this context, we aim to define an algorithmic estimation of such a possibility distribution in the case of experimental samples and the corresponding distributions.

The earliest concept of the probability of a fuzzy event was introduced in Zadeh [3]. In [4], Zadeh's approach further improved and argued that the probability of a fuzzy event must be a fuzzy number rather than a crisp value. Later, Huang [5] introduced an approach to calculate PPD based on his information diffusion technique. This technique is especially capable of coping with small data samples. Therefore, Huang [6,7] suggested his method of Possibility-Probability calculation was suitable for risk evaluation. Huang and Gedeon [8] extended this method for the use of fuzzy events to calculate Possibility-Probability distributions. However, they did not consider the use of the proper probability calculation of fuzzy events [3]. We aim here to enhance the approach in [8] by using proper probability calculation of fuzzy events. Also, we introduce a further generalized model compared to both approaches in papers [5] and [8].

2 Possibility of Probability

In this section, we first demonstrate the method in [6] for the calculation of PPD. Huang [6] provides the following definition for a Possibility-Probability distribution.

Definition 1. *Let (Ω, φ, P) be a probability space, and P be probability measure. Let the possibility that the probability of x occurring is p be $\pi_x(p)$:*

$$\Pi_p = \{\pi_x(p) \mid x \in \Omega, p \in P\} \tag{1}$$

and is called a possibility-probability distribution.

In definition 1, they consider PPD of non-fuzzy events. They further use the information diffusion method [5] to deal with limited sample data. The mathematical definition of the information distribution is a mapping from a Cartesian product to the unit interval $[0, 1]$ [6,5,7].

Definition 2. *Let $X = \{x_i \mid i = 1, \ldots, n\}$ and $Y = \{y_j \mid j = 1, \ldots, m\}$, then*

$$\mu : X \times Y \to [0, 1] \tag{2}$$

is called an information distribution of X on Y, if $\mu(x, y)$ has the following properties:

(a) $\forall x \in X$, *if* $\exists y \in Y$, *such that* $x = y$, *then* $\mu(x, y) = 1$ *(i.e. μ is reflexive).*
(b) *For* $x \in X$, $\forall y', y'' \in Y$, *if* $\mid y' - x \mid \leq \mid y'' - x \mid$ *then* $\mu(y' - x) \leq \mu(y'' - x)$ *(monotonicity).*
(c) $\sum_{j=1}^{m} \mu(x_i, y_j) = 1, i = 1, \ldots, n$ *(additive).*

μ is called a distribution function of X on Y and can be calculated as follows:

$$\mu(x_i, y_j) = \begin{cases} 1 - \frac{|x_i - y_j|}{\triangle_j} & ; if \mid x_i - y_j \mid \leq \triangle_j \\ 0 & ; if \mid x_i - y_j \mid > \triangle_j \end{cases} \\ = q_{ij} \tag{3}$$

where $\triangle_j = y_{j+1} - y_j$ such that $j = 1, \ldots, m$ and $i = 1, \ldots, n$.

2.1 Interval Set Model

Let $X = \{x_i \mid i = 1, \ldots, n\}$ be a sample, $X \subset \mathbb{R}$, and $Y = \{y_j \mid j = 1, \ldots, m\}$ be a discrete universe of X. Huang [6,7] used the following method to calculate a possibility-probability distribution on intervals,

$$I_j = [y_j - \frac{\triangle_j}{2}, y_j + \frac{\triangle_j}{2}], \ y_j \in Y, \tag{4}$$

with respect to the probability values,

$$p_k = \frac{k}{n}, \ k \in \{1, \ldots, n\} \tag{5}$$

Here k is the number of observations that fall into the interval I_j. From here onwards, $\pi_{I_j}(p_k)$ represents the possibility-probability value of that an event occurs in the interval I_j.

2.2 Joining and Leaving Possibility

Let us take that $x_k \in X$. Next, they assume that the possibility of a probability that includes $x_k \in X$ is based on the distance between center $y_j (\in Y)$ and the data point x_k. If $x_k \in X_{I_j}$[1] (see figure 1) then x_k will leave the interval X_{I_j} with the possibility of,

$$q_{kj}^{-} = \frac{\mid x_k - y_j \mid}{\triangle_j} = 1 - q_{ij} \tag{6}$$

Similarly, if $x_k \notin X_{I_j}$, then x_k will join the interval X_{I_j} with the possibility of,

$$q_{kj}^{+} = 1 - \frac{\mid x_k - y_j \mid}{\triangle_j} = q_{ij} \tag{7}$$

The following figure 1 illustrates the situation. Now take that cardinality of X_{I_j} is n_j. Huang [6] writes that the possibility of probability of an event x occurs in I_j is $\frac{n_j}{n}$ is:

$$\pi_{I_j}(\frac{n_j}{n}) = 1 \tag{8}$$

[1] X_{I_j} is called the interior set of interval I_j [6] that contains all elements of interval I_j.

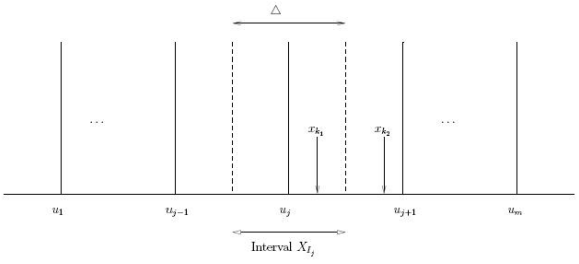


Fig. 1. Intervals on the Universe of Discourse

It is intuitive that $\frac{n_j}{n}$ is the most probable event and thus its possibility is 1. Now, if we assume that $x_k \in X_{I_j}$ may leave the interval I_j when there is a disturbance in the random experiment. Therefore, the possibility that the probability of $x \in I_j$ is $\frac{n_j-1}{n}$ is:

$$\pi_{I_j}(\frac{n_j-1}{n}) = \bigvee_{x_s \in X_{I_j}} q_{sj}^- \tag{9}$$

If two elements leave the interval, we can write that the possibility that the probability of $x \in I_j$ is $\frac{n_j-2}{n}$ is:

$$\pi_{I_j}(\frac{n_j-2}{n}) = \bigvee_{x_{s_1},x_{s_2} \in X_{I_j} | x_{s_1} \neq x_{s_2}} (q_{s_1 j}^- \wedge q_{s_2 j}^-) \tag{10}$$

Similarly, if $x_k \notin X_{I_j}$ may join the interval I_j when there is a disturbance in the random experiment. Therefore, the possibility that the probability of $x \in I_j$ is $\frac{n_j+1}{n}$ is:

$$\pi_{I_j}(\frac{n_j+1}{n}) = \bigvee_{x_s \notin X_{I_j}} q_{sj}^+ \tag{11}$$

Finally, when there are n_j observations in interval I_j, we can write a Possibility-Probability distribution of I_j as follows, [6]:

$$\Pi_{I_j}(p) = \begin{cases} \wedge_{x_s \in X_{I_j}} q_{sj}^- & p = p_0 \\ \vdots & \\ \bigvee_{x_{s_1},x_{s_2} \in X_{I_j} | x_{s_1} \neq x_{s_2}} (q_{s_1 j}^- \wedge q_{s_2 j}^-) & p = p_{n_j-2} \\ \bigvee_{x_s \in X_{I_j}} q_{sj}^- & p = p_{n_j-1} \\ 1 & p = p_{n_j} \\ \bigvee_{x_s \notin X_{I_j}} q_{sj}^+ & p = p_{n_j+1} \\ \bigvee_{x_{s_1},x_{s_2} \notin X_{I_j} | x_{s_1} \neq x_{s_2}} (q_{s_1 j}^+ \wedge q_{s_2 j}^+) & p = p_{n_j+2} \\ \vdots & \\ \wedge_{x_s \notin X_{I_j}} q_{sj}^+ & p = p_n \end{cases} \tag{12}$$

Where $p_{n_j} = \frac{n_j}{n}$.

3 Enhanced Possibility-Probability Calculation Methods

In [8] fuzzy intervals are used instead of the mutually exclusive intervals in [6]. The following method of Possibility-Probability calculation can be found in [8].

Let $A = \{A_1, \ldots, A_m\}$ be m fuzzy events in the space of interest. Here $y_j (\in Y)$ in figure 1 is the middle of the core of the fuzzy set A_j in figure 2. Further, figure 2 illustrates all the fuzzyfied intervals. Let us take that $\mid y_j - y_{j+1} \mid = d$ and

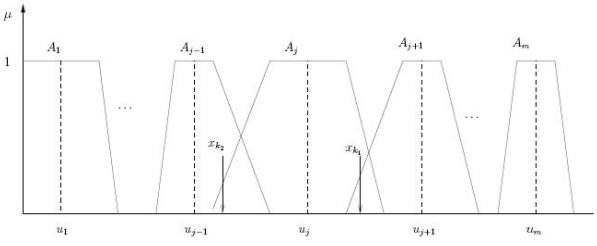


Fig. 2. Intervals on the Universe of Discourse

$X = \{x_1, \ldots, x_n\}$ be the sample space, and these values fall in the interval $[y_j, y_{j+1}]$. Now [8] define S_j and S_{j+1} as follows,

$$S_j = \sum_{i=1}^{n} \mu_{A_j}(x_i) \tag{13}$$

$$S_{j+1} = \sum_{i=1}^{n} \mu_{A_{j+1}}(x_i) \tag{14}$$

$$S = S_j + S_{j+1} \tag{15}$$

Similarly to Huang's method [6], Huang and Gedeon's method [8] also assumes that $\frac{S_j}{S}$ is the maximum possible probability of the event A_j and write

$$\pi_{A_j}(\frac{S_j}{S}) = 1 \tag{16}$$

Now, they consider that the element $x_k \in X$ may leave from fuzzy set A_j, according to the fact that the data point that has the minimum membership to A_j will leave the interval. Therefore, the possibility of probability of A_j occurring being $\frac{S_j - \mu_{A_j}(x_k)}{S}$ as,

$$\pi_{A_j}(\frac{S_j - \mu_{A_j}(x_k)}{S}) = \frac{x_k - y_j}{d} \tag{17}$$

Next if, two points $x_k, x_{k-1} \in X$ may leave the fuzzy set A_j, the the possibility of probability of A_j occurring being $\frac{S_j - \mu_{A_j}(x_k) - \mu_{A_j}(x_{k-1})}{S}$ can be written as

$$\pi_{A_j}(\frac{S_j - \mu_{A_j}(x_k) - \mu_{A_j}(x_{k-1})}{S}) = \frac{x_{k-1} - y_j}{d} \tag{18}$$

where $\mu_{A_j}(x_k) \geq \mu_{A_j}(x_{k-1})$. Conversely, it is possible that some data point will move to the fuzzy set A_j. Suppose that x_t has the minimum membership to A_{j+1} and it is most likely to leave A_{j+1}. Therefore, the possibility of probability of A_j occurring being $\frac{S_j+\mu_{A_{j+1}}(x_t)}{S}$ is,

$$\pi_{A_j}(\frac{S_j + \mu_{A_{j+1}}(x_t)}{S}) = \frac{y_{j+1} - x_t}{d} \tag{19}$$

and the possibility of probability of A_j occurring being $\frac{S_j+\mu_{A_{j+1}}(x_t)+\mu_{A_{j+1}}(x_{+1t})}{S}$ is,

$$\pi_{A_j}(\frac{S_j + \mu_{A_{j+1}}(x_t) + \mu_{A_{j+1}}(x_{t+1})}{S}) = \frac{y_{j+1} - x_{t+1}}{d} \tag{20}$$

where $\mu_{A_{j+1}}(x_t) \geq \mu_{A_{j+1}}(x_{t+1})$. Huang and Gedeon's [8] method can be abstracted into the following form: Let $A = \{A_1, \ldots, A_m\}$ be n fuzzy events in the space of interest. Here $y_j (\in Y)$ is the middle of the core of the fuzzy set A_j. Further, $\mid y_j - y_{j+1} \mid = d$ and $X = \{x_1, \ldots, x_n\}$ be the sample space and the values fall in the interval $[y_j, y_{j+1}]$. Now, let us take that the set $X' = \{x_{k+1}, \ldots, x_{k+h} \mid 0 \leq k < n, 0 < h < n\} \subseteq X$, and the possibility of probability of A_j occurring being $\frac{S_j - \sum_{t=1}^{h} \mu_{A_j}(x_{k+t})}{S}$ as

$$\pi_{A_j}(\frac{S_j - \sum_{t=1}^{h} \mu_{A_j}(x_{k+t})}{S}) = \frac{\mid y_j - \min_{t=1}^{h} x_{k+t} \mid}{d} \tag{21}$$

The possibility of probability of A_j occurring being $\frac{S_j + \sum_{t=1}^{h} \mu_{A_{j+1}}(x_{k+t})}{S}$ is

$$\pi_{A_j}(\frac{S_j + \sum_{t=1}^{h} \mu_{A_{j+1}}(x_{k+t})}{S}) = \frac{\mid y_j - \min_{t=1}^{h} x_{k+t} \mid}{d} \tag{22}$$

4 Possibility-Probability by Employing Probability of Fuzzy Events

Huang's [6,7] method of calculating PPD has serious disadvantages. Firstly, they consider non-fuzzy events to find a possibility of a probability. Therefore, their method can be seen as finding the Possibility-Probability of non-fuzzy events. Secondly, their method assumes that the possibility of an event is proportional to the Euclidian distances between the data points and the centers of those events, and thus lacks a representation of the fuzziness of data against their intervals in the universe.

Huang and Gedeon's [8] measure of the probability of a fuzzy event (eg. probability in equation (16)) is not based on a precise probability measure for a fuzzy event [3]. Further, they consider a special case where all data points in X fall in the interval $[y_j, y_{j+1}]$. In general this not the case, thus equation (15) needs to be improved. In equations (19) and (20), when they write that $\frac{S_j+\mu_{A_{j+1}}(x_t)}{S}$

is the probability of fuzzy set A_i, they violate the concept of the cardinality of a fuzzy set by adding $\mu_{A_{j+1}}(x_t)$ to A_j's cardinality. They also do not consider that A_j's cardinality S_j has already accumulated $\mu_{A_j}(x_t)$. Therefore, they are twice adding x_t's occurrence in the form of $\mu_{A_j}(x_t)$ and $\mu_{A_{j+1}}(x_t)$ to S_j. These are major drawback of Huang and Gedeon's [8] method that we overcome in our methods.

Huang and Gedeon [8] in their first method of finding the possibility of probability also used the same Euclidian distance approach of finding the possibility of their fuzzy events. In their second method, they correctly use fuzzy information, which is the ratio of membership values, to calculate the possibility of probability. However, the second measure is not normalized therefore they need to use an additional scaling function to normalize the data [8].

In this section we provide generalized Possibility-Probability calculation methods that overcome the problems associated with the [6,7] and [8] methods. Additionally, we give a more concrete definition for Possibility-Probability distribution. Unlike Huang's definition in [6], our definition does not consider only a random variable. However, similar to [9], we also use the term "variable", that substitutes for "random variable" in conventional probability theory, considering the fact that a probability measure based on possibility theory can model uncertainty that is caused by more than randomness [1].

Definition 3. *Let $(\mathbb{R}^n, \varphi, P)$ be a probability space, let φ be the σ-field of Borel sets in $\mathbb{R}^n$, and P is a probability measure. Let X be a variable on $\mathbb{R}^n$. Further let A be a fuzzy subset of $\mathbb{R}^n$ in φ. Now, the "$Prob(X\ is\ A)$ takes a value in B" induces a possibility distribution $\Pi_{Prob(X\ is\ A)}$ in P:*

$$\Pi(Prob(X\ is\ A)\ = p) = \pi_{Prob(X\ is\ A)} = \mu_B(p) \tag{23}$$

Where B is a fuzzy sub set in P.

Here $\Pi(Prob(X\ is\ A)\ is\ B)$ or is $\Pi_{Prob(X\ is\ A)}$ denoted the PPD of A. Therefore in short we could also denote this as $\Pi(Prob(A))$ or $\Pi_{Prob(A)}$.

Note 1. As we mentioned earlier, in definition 1, Huang considers PPD of non-fuzzy events. In definition 3, we consider PPD of fuzzy events. Thus, the probability "$P(X\ is\ A)$" in the above 2 equations is the probability of the fuzzy event A denoted by [3]:

Definition 4. *When X is discrete,*

$$P(X\ is\ A) = \sum_i \mu_A(x_i) \times P_X(x_i) \tag{24}$$

Here μ_A is the membership function of the fuzzy set A and P_X denotes the probability distribution function of X.

4.1 Approximation of Possibility of Probability: A Generalized Method

Let us take a situation where we have n number of records of data and a permutation $.(i)$ on each $x \in X$ such that $\mu_{A_j}(x_{(1)}) \geq \mu_{A_j}(x_{(2)}) \geq \ldots \geq \mu_{A_j}(x_{(k)}) \geq \mu_{A_j}(x_{(k+1)}) \ldots \geq \mu_{A_j}(x_{(n)})$ where A_j is the j^{th} fuzzy subset on U (in $\mathbb{R}^n$) and X is a variable on U according to definition (3). Also, for the simplicity of the discussion, in the rest of the paper, we write $\mu_{A_j}(x)$ to denote $\mu_{A_j}(u)$ s.t $u = x$ where $u \in U$ and $x \in X$.

Initial Event. Let k be an integer such that at least $k + 1 \leq n$ and let $X_k = \{x_{(1)}, \ldots, x_{(k)}\}$ be a sub set of X. Further let us take A_j^k as the k^{th} fuzzy event of the fuzzy set A_j. The membership function of the k^{th} fuzzy event A_j^k can be defined as follows:

$$\mu_{A_j^k(x_i)} = \begin{cases} \mu_{A_j(x_i)} \; if \; x_i \in X_k \\ 0 \qquad otherwise \end{cases} \tag{25}$$

Let $n_j^k =\mid X_k \mid = k$ be the cardinality of A_j^k. Now, based on the equation (24), the probability of fuzzy event A_j^k can be calculated as

$$Prob(XisA_j^k) = \sum_{i=1}^{n_j^k} \mu_{A_j^k}(x_i) p(x_i) \tag{26}$$

Now, for the ease of understanding and explaining the leaving and joining possibilities, let us assume that event A_j^k is the starting (initial) event. Similarly to the previous methods the possibility of the initial event is 1.

Leaving Possibility. Next, let the element x_k, which has the lowest membership to the fuzzy event A_j^k, leaves the fuzzy event A_j^k, and let the new resulting fuzzy event state be denoted by A_j^{k-1}. The cardinality of the new fuzzy event A_j^{k-1} is $n_j^{k-1} =\mid X_{k-1} \mid = k - 1$. Based on equation (24), the probability of the new fuzzy event, ie. A_j^{k-1}, can be calculated as follows:

$$Prob(XisA_j^{k-1}) = \sum_{i=1}^{n_j^{k-1}} \mu_{A_j^{k-1}}(x_i) p(x_i) = p_{A_j^{k-1}} \tag{27}$$

Next, the possibility that the probability of A_j^{k-1} occurring being $p_{A_j^{k-1}}$ can be calculated as follows

$$\pi(Prob(XisA_j^{k-1})) = 1 - \max_{i=n_j^{k-1}+1}^{n_j^k} \left[\mu_{A_j^k}(x_i)\right] \tag{28}$$

In general, after $l(\leq k)$ data points leave A_j^k, we can calculate the possibility of the fuzzy event A_j^{k-l} occurring being $p_{A_j^{k-l}}$.

$$\pi(Prob(XisA_j^{k-l})) = 1 - \max_{i=n_j^{k-l}+1}^{n_j^k} \left[\mu_{A_j^k}(x_i)\right] \tag{29}$$

This process will continue until all data points will leave the initial fuzzy event A_j^k.

Note 2. Note that in equation (29) when $l = 0$, it gives the possibility that the probability of A_j occurring being $p_{A_j^k}$ as:

$$\pi(Prob(XisA_j^k)) = 1 - \mu_{A_j^k}(x_{k+1}) = 1 - 0 = 1$$

Joining Possibility. Let us assume that the element x_{k+1} will join the fuzzy event A_j^k, and let the new resulting fuzzy event states be A_j^{k+1}. Now the cardinality of the fuzzy event A_j^{k+1} is $n_j^{k+1} =\mid X_{k+1} \mid= k+1$. Now, based on the equation (24), the probability of the new fuzzy event, ie. A_j^{k+1}, can be calculated as follows:

$$Prob(XisA_j^{k+1}) = \sum_{i=1}^{n_j^{k+1}} \mu_{A_j^{k+1}}(x_i)p(x_i) = p_{A_j^{k+1}} \tag{30}$$

Next, the possibility that the probability of A_j^{k+1} occurring being $p_{A_j^{k+1}}$ can be calculated as follows:

$$\pi(Prob(XisA_j^{k+1})) = \min_{i=n_j^k+1}^{n_j^{k+1}} \left[\mu_{A_j^{k+1}}(x_i)\right] \tag{31}$$

In general, after $l(\leq (n-k))$ data points join A_j^k, we can calculate the possibility of the fuzzy event A_j^{k+l} occurring being $p_{A_j^{k+l}}$.

$$\pi(Prob(XisA_j^{k+l})) = \min_{i=n_j^k+1}^{n_j^{k+l}} \left[\mu_{A_j^{k+l}}(x_i)\right] \tag{32}$$

This process will continue until all possible data points will join the initial fuzzy event A_j^k.

Possibility of Probability. In this subsection, we give a generalized method of calculation of the possibility for the all available probabilities for the event A_j.

$$\begin{aligned} \Pi_{Prob(X\ is\ A_j)} = &\sum_{l=0}^{n_j^k} \frac{1 - \max_{i=n_j^{k-l}+1}^{n_j^k} \left[\mu_{A_j^k}(x_i)\right]}{Prob(X\ is\ A_j^{k-l})} \\ &+ \sum_{l=n_j^k+1}^{n} \frac{\min_{i=n_j^k+1}^{n_j^{k+l}} \left[\mu_{A_j^{k+l}}(x_i)\right]}{Prob(XisA_j^{k+l})} \end{aligned} \tag{33}$$

Here in equation (33), $\sum$ denote an union operation of a fuzzy number and $P(XisA_j^k)$ is given by equation (26).

5 Conclusion

First we have shown the disadvantages of Huang's [6] method. Secondly, we removed the mathematical irregularities in Huang and Gedeon's method [8]. As a result, we provide a generalized method that estimates a PPD from available data. Importantly, the generalized PPD assume the fuzziness of the events that occurs in reality. Our method can be described as a generalized Possibility-Probability calculation method that uses a possibility theory based approach to estimate the likelihood of the reality of the calculated probability. This method is very useful when only a small sample of data is available.

References

1. Darby, J.L.: Estimating terrorist risk with possibility theory. Tech Report of Los-Alamos National Laboratory, University of California, USA, pp. 1–50 (November 2004)
2. Darby, J.L.: Evaluation of risk from acts of terrorism: the adversary/defender model using belief and fuzzy sets. SAND2006-5777, Sandia National Laboratories Tech Report, pp. 1–136 (2006)
3. Zadeh, L.: Probability measures of fuzzy events. Journal Math. Anal. Appl. 23, 421–427 (1968)
4. Yager, R.: A note on probabilities of fuzzy events. Information Sciences 18, 113–122 (1979)
5. Huang, C.: Information diffusion techniques and small sample problem. International Journal of Information Technology and Decision Making 1(2), 229–249 (2002)
6. Huang, C.: An application of calculated fuzzy risk. Information Sciences 142(1-4), 37–56 (2002)
7. Huang, C., Shi, P.: Fuzzy risk and calculation. In: Proc. 18th International Conference of the North American Fuzzy Information Processing Society, New York, June 1999, pp. 10–12 (1999)
8. Huang, Z., Gedeon, T.D.: Information retrieval estimation via fuzzy probability. In: Proceedings of World Automation Congress (WAC), pp. 1–6 (2006)
9. Zadeh, L.: Fuzzy probabilities. World Scientific Series in Advances in Fuzzy Systems, pp. 643–652 (1996)

Bayesian Assaying of GUHA Nuggets

Robert Piché and Esko Turunen*

Department of Mathematics
Tampere University of Technology Tampere, Finland
esko.turunen@tut.fi

Abstract. The General Unary Hypothesis Automaton (GUHA) sifts through large data sets and finds 2×2 contingency tables satisfying certain dependencies expressible by generalized quantifiers that indicate possible relations between attributes. In this work we show how these tables can be further investigated by Bayesian statistical methods. In this way we are able to translate tables that have been discovered by GUHA into verbal statements and probability density plots that are comprehensible to anyone who has a basic understanding of probability.

Keywords: data mining, Bayesian statistics, GUHA, contingency table.

1 Introduction

Introduced in 1966 [1], the General Unary Hypothesis Automaton (GUHA) method is one of the oldest data mining methods; see [2, 5, 6] for a detailed description. The GUHA method is designed for exploratory analysis of large data sets, when the aim is to get orientation in the domain of investigation by analyzing the behaviour of variables, finding interactions among them, etc. The mathematical structure underlying GUHA theory allows software to identify "interesting" features in the data without exhaustive search. There have been several software implementations of GUHA, most notably the freely available *LISp–Miner* developed by Rauch et al. at Prague University of Economics [3].

GUHA is applied to data that can be represented by an $m \times n$ Boolean array, with the 1's or 0's in each column indicating the presence or absence of an observed property among m objects or instances. A Boolean function of the n observed properties can be represented by an m-vector of 1's and 0's, and is called an *attribute*. Any two attributes ϕ and ψ can be summarised by a 2×2 double dichotomy contingency table

	ψ	$\neg\psi$
ϕ	a	b
$\neg\phi$	c	d

where a is the number of instances for which both ϕ and ψ have value 1; the remaining table entries b, c, d are the corresponding counts for $\phi \wedge \neg\psi$, $\neg\phi \wedge \psi$, $\neg\phi \wedge \neg\psi$, and $a+b+c+d = m$. The aim of the exploratory data analysis in GUHA is to identify,

* Corresponding author.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 348-355, 2010.

from all possible 2×2 contingency tables, those tables that indicate the existence of "interesting" relations between attributes: they are defined *true* or *supported by the data*. Relations between attributes that are not true are *false* (not supported by the data). True relations between attributes ϕ and ψ, called *hypotheses*, are the outputs of a GUHA procedure.

Such an analysis is made computationally feasible by restricting attention to relations expressed by *generalised quantifiers*, which satisfy certain monotonicity conditions. Each generalised quantifier has its characteristic truth definition. The LISp–Miner system supports many different generalised quantifiers. In this study we focus on the following four:

Founded implication. For a given fixed strength $0 < p \leq 1$ and support $r > 0$, this relation is true (denoted $\phi \Rightarrow_{p,r} \psi$) if

$$\frac{a}{a+b} \geq p \text{ and } a \geq r$$

and false otherwise. Typically one sets $p \approx 1$, in which case a verbal interpretation of this relation is "most ϕ are ψ".

Founded equivalence. For a given fixed strength $0 < p \leq 1$ and support $r > 0$, this relation is true (denoted $\phi \equiv_{p,r} \psi$) if

$$\frac{a+d}{a+b+c+d} \geq p \text{ and } a \geq r$$

and false otherwise. Typically one sets $p \approx 1$, in which case a verbal interpretation of this relation is "ϕ is almost equivalent to ψ".

Above average. For a given fixed strength $p > 0$ and support $r > 0$, this relation is true (denoted by $\sim^{+}_{p,r}$) if

$$\frac{a}{a+b} \geq \frac{(1+p)(a+c)}{a+b+c+d} \text{ and } a \geq r$$

and false otherwise. Typically one sets $1 + p \gg 1$, in which case a verbal interpretation of this relation is "ψ is much more prevalent among ϕ than in general".

Association. This relation is true (denoted $\phi \sim \psi$) if

$$ad > bc$$

and false otherwise. In statistics, the ratio $\frac{ad}{bc}$ is called the *odds ratio*; a value close to 1 indicates that the attributes ϕ and ψ are independent. Because

$$ad > bc \quad \Leftrightarrow \quad \frac{a}{a+b} > \frac{c}{c+d},$$

a verbal interpretation of this relation is "ψ is more prevalent among ϕ than among $\neg\phi$." Also, because

$$ad > bc \quad \Leftrightarrow \quad \frac{a}{a+c} > \frac{b}{b+d},$$

another verbal interpretation of this relation is "ϕ is more prevalent among ψ than among $\neg\psi$."

After the computer has sifted through a data set and has found all attribute pairs whose 2×2 contingency tables satisfy the truth condition of a given generalised quantifier, the analyst can proceed to a deeper study of the possible relation between the attributes: collect further data collection, apply statistical methods, discuss with subject-domain experts, etc.

A useful starting point in this post-processing is the study and interpretation of the contingency tables themselves. A statistical analysis is attractive because it can provides information about the *uncertainty* associated with inferences. LISp–Miner already contains tools to compute frequentist (e.g. Fisherian) statistical quantities. The results of frequentist statistical analysis (e.g. hypothesis tests), however, often require extensive statistical training and expertise to understand correctly.

In contrast, the results of a Bayesian statistical inference can be understood by anyone with a basic understanding of probability. This is because Bayesian inference produces probability densities of underlying parameters and the *probability* of statements (hypotheses).

In this work we present some Bayesian methods that can be used to evaluate statements that directly correspond to GUHA generalised quantifiers. We focus on the four generalised quantifiers presented above, because the corresponding statistical counterparts are straightforward. We show how to translate tables that have been found by GUHA into verbal statements and probability density plots, that is, we use Bayesian analysis as a GUHA post-processor.

2 Statistical Analysis of Contingency Tables

2.1 Multinomial Model

We start by describing a standard statistical model for 2×2 contingency tables with unconstrained row sums and column sums. Given two attributes ϕ and ψ, there are 4 possible disjoint attribute combinations for every object:

$$x_i \in \{\underbrace{\phi \wedge \psi}_{X_1}, \underbrace{\phi \wedge \neg\psi}_{X_2}, \underbrace{\neg\phi \wedge \psi}_{X_3}, \underbrace{\neg\phi \wedge \neg\psi}_{X_4}\}$$

To each attribute combination X_j one can associate a parameter θ_j that represents its probability of occurrence, given that the parameters $\theta = [\theta_1, \theta_2, \theta_3, \theta_4]$ are known, that is,

$$\mathsf{P}(x_i = X_j \,|\, \theta) = \theta_j \qquad (i \in \{1,\ldots,m\},\ j \in \{1,2,3,4\}).$$

The parameter vector θ is a point in $\mathbb{R}^4$ on the simplex

$$\theta_j \geq 0, \quad \sum_{j=1}^{4} \theta_j = 1.$$

The probability mass function (pmf) for a set of observations $x_1,\ldots,x_m$ that are mutually independent given θ is

$$p(x_{1:m} \mid \theta) = \theta_1^a \cdot \theta_2^b \cdot \theta_3^c \cdot \theta_4^d,$$

with the convention that $0^0 = 1$. The pmf for the corresponding contingency table $y = [a,b,c,d]$ is $y \mid \theta \sim \text{Multinomial}(\theta)$, that is,

$$p(y \mid \theta) \propto \theta_1^a \cdot \theta_2^b \cdot \theta_3^c \cdot \theta_4^d.$$

This is a model of how, given the parameters, the data could have been generated. Statistical inference is the inverse problem: given the data, determine θ. In Bayesian statistics, inference is accomplished using probability theory.

Let $p(\theta)$ be a probability density function (pdf) that describes our state of knowledge about θ before the data is looked at — this is called the *prior* distribution. Then, by Bayes' law, the information provided by the observations improves our state of knowledge, which is now described by the *posterior* distribution

$$p(\theta \mid y) \propto p(y \mid \theta) p(\theta).$$

For the multinomial model, it is convenient to use the prior distribution $\theta \sim \text{Dirichlet}(\alpha', \beta', \gamma', \delta')$, where $\alpha', \beta', \gamma', \delta'$ are positive real values that are chosen to model the prior state of knowledge. The Dirichlet pdf is

$$p(\theta) \propto \theta_1^{\alpha'-1} \cdot \theta_2^{\beta'-1} \cdot \theta_3^{\gamma'-1} \cdot \theta_4^{\delta'-1} \qquad (\theta_j \geq 0, \sum_{j=1}^{4} \theta_j = 1).$$

Small values of $\alpha', \beta', \gamma', \delta'$ give a relatively "vague" distribution, with large dispersion; the density is constant if $\alpha' = \beta' = \gamma' = \delta' = 1$.

With the multinomial model and Dirichlet prior, the posterior pdf is, by Bayes' law,

$$p(\theta \mid y) \propto \theta_1^{a+\alpha'-1} \cdot \theta_2^{b+\beta'-1} \cdot \theta_3^{c+\gamma'-1} \cdot \theta_4^{d+\delta'-1}$$

that is, $\theta \mid y \sim \text{Dirichlet}(\alpha, \beta, \gamma, \delta)$, where $\alpha = \alpha' + a, \beta = \beta' + b, \gamma = \gamma' + c, \delta = \delta' + d$.

This posterior pdf is a complete description of our state of knowledge (and remaining uncertainty about) the parameters of the statistical model.

2.2 Evaluating Hypotheses

In this study, we are interested in using Bayesian methods to evaluate the validity of statements that correspond to generalised quantifiers used in GUHA, in particular, the generalised quantifiers presented in section 1. Consider first how to evaluate the statement "most ϕ are ψ". Recall that, in our statistical model, θ_1 is the proportion of ϕ that are ψ, while θ_2 is the proportion of ϕ that are not ψ. The proportion of ψ among the ϕ is $\frac{\theta_1}{\theta_1+\theta_2}$. Then, given a 2×2 contingency table, there are various ways one could proceed:

1. One could plot the posterior pdf of $\frac{\theta_1}{\theta_1+\theta_2}$. A pdf that has most of its mass at the right end of the interval $[0,1]$ indicates that $\frac{\theta_1}{\theta_1+\theta_2} \approx 1$. The dispersion seen in the graph gives an indication of the uncertainty that remains. The shape of the graph can be summarised by computing some statistics, such as mean and standard deviation. For the Dirichlet prior, the posterior distribution is $\frac{\theta_1}{\theta_1+\theta_2} \,|\, y \sim \text{Beta}(\alpha,\beta)$, for which the mean is $\frac{\alpha}{\alpha+\beta}$ and the variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
2. Suppose we can agree on a numerical value of p for the concept of "most", say $p = 95\%$. One could then compute the value $\mathsf{P}(\frac{\theta_1}{\theta_1+\theta_2} \geq p \,|\, y)$, that is, the posterior probability that the proportion of ψ among the ϕ is $\geq p$. Being a probability, this p value is much easier to understand and interpret than the "significance levels" of classical hypothesis testing.

The statement "ϕ and ψ are almost equivalent" can be analysed as follows. The proportion of $(\phi \wedge \psi) \vee (\neg\phi \wedge \neg\psi)$ is $\theta_1 + \theta_4$, so one plot the posterior pdf of the sum $\theta_1 + \theta_4$ and compute its mean and standard deviation. For the Multinomial-Dirichlet model, $\theta_1 + \theta_4 \,|\, y \sim \text{Beta}(\alpha+\delta, \beta+\gamma)$, for which the mean is $\frac{\alpha+\delta}{A}$ and the variance is $\frac{(\alpha+\delta)(\beta+\gamma)}{A^2(A+1)}$, where $A = \alpha+\beta+\gamma+\delta$. Given a numerical value p to model "almost", one can compute the value $\mathsf{P}(\theta_1 + \theta_4 > p \,|\, y)$.

The statement "ψ is much more prevalent among ϕ than in general" can be analysed as follows. The proportion of ψ in general is $\theta_1 + \theta_3$, while the proportion of ψ that are ϕ is $\frac{\theta_1}{\theta_1+\theta_2}$. One can then proceed to plot the posterior pdf of the ratio $\frac{\theta_1}{\theta_1+\theta_2} / (\theta_1 + \theta_3)$ and compute its mean and standard deviation. Given a numerical value p to model "much more", one can compute the value $\mathsf{P}(\frac{\frac{\theta_1}{\theta_1+\theta_2}}{\theta_1+\theta_3} \geq (1+p) \,|\, y)$.

The statement "ψ is more prevalent among ϕ than among $\neg\phi$" can be analysed similarly. The proportion of ψ that are ϕ is $\frac{\theta_1}{\theta_1+\theta_2}$, the proportion of ψ that are $\neg\phi$ is $\frac{\theta_3}{\theta_3+\theta_4}$. One can plot the posterior pdf of the ratio of these proportions, compute its mean and standard deviation, or compute the value $\mathsf{P}(\frac{\theta_1}{\theta_1+\theta_2} > \frac{\theta_3}{\theta_3+\theta_4} \,|\, y)$. This probability, which is equal to $\mathsf{P}(\frac{\theta_1\theta_4}{\theta_2\theta_3} > 1 \,|\, y)$, is called the *probability of positive association* in [4], where a closed-form formula for it is derived. Note that if $\frac{\theta_1\theta_4}{\theta_2\theta_3} = 1$, then

$$\mathsf{P}(\phi \,|\, \theta)\mathsf{P}(\psi \,|\, \theta) = (\theta_1+\theta_2)\cdot(\theta_1+\theta_3) = \theta_1 \cdot (\theta_1+\theta_2+\theta_3+\theta_4) = \theta_1 = \mathsf{P}(\phi \wedge \psi \,|\, \theta)$$

that is, ϕ and ψ are independent given θ.

3 Example

To illustrate, we use Tjen-Sien Lim's publicly available benchmark data test set [7] from the 1987 National Indonesia Contraceptive Prevalence Survey. These are the responses from interviews of $m = 1473$ married women who were not (as far as they knew) pregnant at the time of interview. The challenge is to learn to predict a woman's contraceptive method from knowledge about her demographic and socio-economic characteristics.

The 10 survey response variables and their types are

Age	integer 16–49
Education	4 categories
Husband's education	4 categories
Number of children borne	integer 0–15
Islamic	binary (yes/no)
Working	binary (yes/no)
Husband's occupation	4 categories
Standard of living	4 categories
Good media exposure	binary (yes/no)
Contraceptive method used	3 categories (None, Long-term, Short-term)

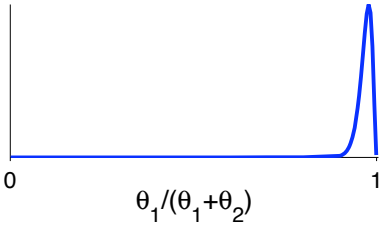
The data was processed into binary form as follows. The three binary variables need no processing. The 3-category variable ("contraceptive method used") is divided into three binary properties, one for each category; each of the four 4-category variables is similarly divided into four binary properties. The age variable is divided into 118 properties: 31 3-year ranges (16–18, 17–19, ..., 47–49), 30 4-year ranges (16–19, ..., 46–49), 29 5-year ranges, and 20 6-year ranges. Similarly, the number-of-children variable is divided into 58 properties: 16 singletons (0, 1,..., 15), 15 two-unit ranges (0–1, 1–2, ..., 14–15), 14 3-unit ranges (0–2, ..., 13–15), and 13 4-unit ranges (0–3, ..., 12–15). Altogether, there were 198 binary properties.

In the first LISp-Miner run, the system was set the task of finding "founded implication" relations $\phi \Rightarrow_{0.95,50} \psi$ with the Contraceptive method properties as ψ and all possible boolean functions of the remaining properties as ϕ. In 7 seconds, after explicitly testing 179 447 tables, 9 contingency tables satisfying the relation were found. One of them was

	ψ	$\neg\psi$
ϕ	92	2
$\neg\phi$	534	842

where ϕ="no children" and ψ ="not using contraceptives". The verbal interpretation of the table is "Most married women without children are not using contraceptives."

Applying the Bayesian model of section 2 to this table, with a vague prior ($\alpha' = \beta' = 1$), the posterior distribution for the proportion of ψ among the ϕ is $\frac{\theta_1}{\theta_1+\theta_2}\,|\,y \sim \text{Beta}(93,3)$. The plot of the pdf shows that most of the probability is concentrated in the right end of the interval.



The mean $\frac{93}{96} \approx 0.969$ and standard deviation ≈ 0.018. Furthermore, we compute

$$\mathrm{P}\left(\frac{\theta_1}{\theta_1+\theta_2} \geq 0.95\,|\,y\right) = \int_{0.95}^{1} \frac{\Gamma(93+3)}{\Gamma(93)\Gamma(3)} t^{92}(1-t)^2 \mathrm{d}t = 0.8595.$$

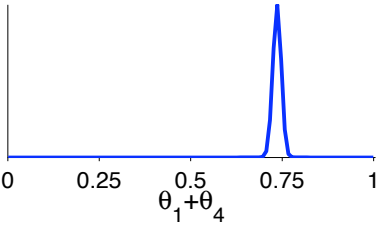
Thus, we are over 85% sure that at least 95% of married women without children are not using contraceptives. These computations require less than 0.01 second on a current laptop, and can be done using publicly available statistical software libraries.

In the second LISp-Miner run, the system was set the task of finding "founded equivalence" relations $\phi \equiv_{0.7,r} \psi$. In 7 seconds, after explicitly testing 106 308 tables, 158 contingency tables satisfying the relation were found. One of them was

	ψ	$\neg\psi$
ϕ	544	355
$\neg\phi$	33	541

where ϕ = "highly educated" and ψ = "husband highly educated". The verbal interpretation of the table is "High education and high education of husband are almost equivalent".

Applying the Bayesian model of section 2 to this table, with a vague prior ($\alpha' = \beta' = \gamma' = \delta' = 1$), the posterior distribution for the proportion of $(\phi \wedge \psi) \vee (\neg\phi \wedge \neg\psi)$ is $\theta_1 + \theta_4 \,|\, y \sim \text{Beta}(1087, 390)$. The plot of the pdf shows that most of the probability is concentrated around the mean, $\mathsf{E}(\theta_1 + \theta_4 \,|\, y) = \frac{1087}{1477} = 0.736$.



The standard deviation is ≈ 0.0115. Furthermore, we compute

$$\mathsf{P}(\theta_1 + \theta_4 \geq 0.70 \,|\, y) = 0.9989.$$

Thus, we are 99.89% sure that high education of husband and wife has at least 70% rate of coincidence. Again, the computing times for this post-processing are negligible.

In the third LISp-Miner run, the system was set the task of finding "above-average" relations $\phi \sim^{+}_{3,15} \psi$. In 3 minutes 17 seconds, after explicitly testing 4 888 398 tables, 14 contingency tables satisfying the relation were found. One of them was

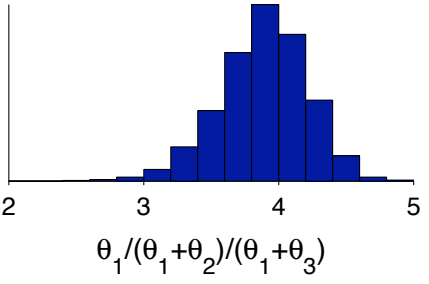
	ψ	$\neg\psi$
ϕ	21	2
$\neg\phi$	312	1138

where ϕ ="Age 37–45 and Children 4 and Husband highly educated and Living standard high", and ψ ="Using long-term contraception method".

Applying the Bayesian model of section 2 to this table, with a vague prior ($\alpha' = \beta' = \gamma' = \delta' = 1$), we obtain the posterior

$$\theta_{1:4} \,|\, y \sim \text{Dirichlet}(22, 3, 313, 1139)$$

To evaluate how much larger is the proportion of ψ among the ϕ than the proportion of ψ in general, we generate 10^4 independent Monte Carlo samples from the posterior and plot the histogram of the proportion ratio $(\frac{\theta_1}{\theta_1+\theta_2})/(\theta_1 + \theta_3)$:



The ratio samples' mean is 3.89 and the standard deviation is 0.33. We note that, although the table satisfies the generalised quantifier for the statement "ψ is over 4 times more prevalent among ϕ than in general", the statistical model indicates that the actual factor may be somewhere between 3 and 4.8. The above Monte Carlo computations and histogram plotting require less than 0.1 second on a current laptop.

The number of samples with ratio larger than 3 is over 9900, so we can say that we are over 99% certain that ψ is at least 3 times more prevalent among ϕ than in general, that is, that the use of long-term contraceptives is at least 3 times more prevalent among rich women aged 37–45 with 4 children and highly educated husband than among married women in general.

4 Conclusions

A GUHA data analysis produces a set of contingency tables, each of which corresponds to a statement asserting the existence of a relation among attributes. In this work we have looked at how the validity of the statement can be evaluated using a statistical analysis of the contingency table. The analysis allows the GUHA user to quantify the statement's uncertainty and thus to present a more complete picture of the information that is coded in the contingency table.

References

[1] Hájek, P., Havel, I., Chytil, M.: The GUHA method of Automatic Hypotheses Determination, Computing, vol. 1, pp. 293–308. Springer, Heidelberg (1966)
[2] Hájek, P., Havránek, T.: Mechanising Hypotheses Formation, Mathematical Foundations for a General Theory. Springer, Berlin (1978)
[3] http://lispminer.vse.cz
[4] Altham, P.M.E.: Exact Bayesian Analysis of a 2×2 Contingency Table, and Fisher's "Exact" Significance Test. J. Royal Stat. Soc., Series B 31(2), 261–269 (1969), http://www.statslab.cam.ac.uk/
[5] Rauch, J., Šimůnek, M.: An Alternative Approach to Mining Association Rules. Computational Intelligence (SCI) 6, 211–231 (2005)
[6] Rauch, J.: Logic of Association Rules. Applied Intelligence 22, 9–28 (2005)
[7] http://b-course.cs.helsinki.fi/obc/cmcexpl.html

Rank Correlation Coefficient Correction by Removing Worst Cases

Martin Krone[1] and Frank Klawonn[1,2]

[1] Department of Computer Science
Ostfalia University of Applied Sciences
Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel, Germany
{ma.krone,f.klawonn}@ostfalia.de
[2] Bioinformatics and Statistics
Helmholtz Centre for Infection Research
Inhoffenstr. 7, D-38124 Braunschweig, Germany

Abstract. Rank correlation can be used to compare two linearly ordered rankings. If the rankings include noise values, the rank correlation coefficient will yield lower values than it actually should. In this paper, we propose an algorithm to remove pairs of values from rankings in order to increase Kendall's tau rank correlation coefficient. The problem itself is motivated from real data in bioinformatics context.

Keywords: Rank correlation coefficient, greedy algorithm, graph algorithms.

1 Introduction

The motivation for the formal problem we will discuss in this paper comes from biological experiments with the bacterium *Pseudomonas aeruginosa*. In these experiments, each of more than 4,000 genes was knocked out and the mutants resulting from the knocked out genes were examined under more than 100 conditions. Conditions are, for instance, different antibiotics in varying concentrations. For each condition we obtain a value, describing the deviation from the "normal condition". Genes that are functionally related are expected to show similar behaviour under the same conditions, especially under those where they should be highly activated (expressed) if they were not knocked out. If we order the deviations from the "normal condition" for each mutant, we can compare these rankings of the conditions. A high correlation between two rankings would be a hint to functionally related genes. However, although many of the genes will play a certain role under almost all conditions, there are some conditions for each gene where it might have no influence. Unfortunately, we do not know which conditions these are for each gene. These conditions can therefore lead to a reduction of the correlation between the genes or mutants.

In order to reduce this effect, we do not consider the correlation with respect to all conditions. For each pair of genes, we are allowed to remove a fixed small number k of conditions, to compute the correlation coefficient. Here we

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 356–364, 2010.

use Kendall's tau rank correlation coefficient [2]. The task is to remove those conditions that lead to the highest increase of the rank correlation coefficient.

The paper is organized as follows. In Sect. 2, we briefly recall Kendall's tau rank correlation coefficient and show how it can be associated with undirected graphs. Section 3 reformulates our problem as a graph problem and discusses the infeasible brute force solution and a greedy approach. An improved greedy algorithm based on a look-ahead strategy is proposed in Sect. 4. Experimental results are provided in Sect. 5 before the final conclusions.

2 Formalization of the Problem

Let x and y be two rankings of length n and both be free of duplicate values (so-called *ties*). Then Kendall's tau rank correlation coefficient can be used to measure the degree of correspondence between x and y. It is defined as

$$\tau = \frac{p_\mathrm{c} - p_\mathrm{d}}{\binom{n}{2}}, \tag{1}$$

where p_c denotes the number of concordant (meaning: in the same order) and p_d the number of discordant (meaning: in the opposite order) among all $\binom{n}{2}$ different pairs. Two pairs (x_i, y_i) and (x_j, y_j) are referred to as concordant if $\operatorname{sgn}(x_i - x_j) = \operatorname{sgn}(y_i - y_j)$ and denoted discordant otherwise. As it is assumed throughout this paper that both rankings are free of ties, every two pairs are either concordant or discordant.

Furthermore, the correspondence between x and y can be represented by an undirected graph by applying the following set of instructions:

1. Create a graph $G = (V, E)$ where V is the set of n nodes labelled $v_1, \ldots, v_n$.
2. For every pair (i, j) s.t. $1 \leq i < j \leq n$, add an undirected edge between v_i and v_j to E if (x_i, y_i) and (x_j, y_j) are discordant.

Theorem 1. *For $-1 \leq \tau \leq 1$, the number of edges in the resulting graph is given by*

$$|E| = \binom{n}{2} \frac{1 - \tau}{2}. \tag{2}$$

Proof. As the number of edges $|E|$ equals the number of discordant pairs and

$$p_\mathrm{c} + p_\mathrm{d} = \binom{n}{2} \tag{3}$$

holds, it follows from (1) that

$$\tau = \frac{\binom{n}{2} - 2|E|}{\binom{n}{2}}. \tag{4}$$

Solving (4) for $|E|$ proves this theorem. □

Thus, the resulting graph is free of edges provided that $\tau = 1$ whereas it will equal the so-called *complete graph* K_n if $\tau = -1$.

3 Reformulation as a Graph Problem

Recall that we seek to delete a fixed constant $k < n$ of conditions for the pair of rankings we want to compare. This set of conditions has to be selected in such a way that as many discordant pairs as possible are removed in order to increase the rank correlation coefficient.

As the correspondence between two rankings can be represented by an undirected graph, an equivalent task is to delete a k-subset of the nodes from the graph such that as many edges as possible are thereby removed. At first glance, finding the best k-subset might seem to be an easily solvable problem. However, as deleting any node from the graph decreases the degree of all its adjacent nodes, this is actually a considerably more difficult task. In fact, this is a variation of the so-called *node-deletion problem* [4]. However, this problem is usually seen with regard to finding a minimum number of nodes whose deletion results in a subgraph that satisfies a given graph-property. As approaches [1] for those kinds of tasks cannot be applied to our specific problem, we will now examine two generic approaches along with their advantages and disadvantages.

3.1 Bruteforce Approach

The most obvious approach to our problem is to determine all $\binom{n}{k}$ subsets of k nodes and select the set whose deletion results in the removal of more edges than any other set. Note that the total number of removed edges does not depend on the order the nodes of a set are deleted in. Thus, it is sufficient to test only one of all $k!$ permutations for each set of k nodes.

While this approach guarantees to find the best set, it will rarely be used in practice unless testing all $\binom{n}{k}$ subsets can be done in reasonable time which will only be possible if both n and $\min(k, n-k)$ are very small. In other cases, one is usually interested in a different approach that requires less steps, but accepts at the same time that less edges in comparison to the bruteforce algorithm might be removed from the graph.

3.2 Using a Greedy Strategy

Greedy algorithms are based on the idea of choosing the local optimum in each step hoping this will lead to the best overall performance. While there are some problems that can efficiently be solved by this approach, such as creating a minimum cost spanning tree [3], greedy algorithms often only find approximate solutions. Note that the definition of the local optimum depends on the specific field of application. As we seek to remove as many edges as possible from the graph by deleting a given number of nodes, a greedy strategy for this problem is to delete the node with the highest degree in each step.

It is important to realize that the strategy described above may fail to remove the maximum number of edges if more than two nodes are deleted from the graph. We show this is indeed true for $k = 3$ by comparing the bruteforce algorithm to this greedy strategy when applied to the graph in Fig. 1 that is based

Table 1. Pair of rankings that can be represented by the graph in Fig. 1

1	2	3	4	5	6	7
5	6	7	1	3	4	2

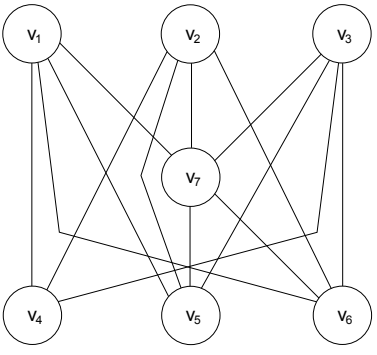


Fig. 1. Example for a graph for which the greedy algorithm removes less edges than the bruteforce approach

on the rankings in Table 1. Initially, the greedy algorithm deletes node v_7 as it has a higher degree than any other node. Consequently, five edges are removed from the graph. Note that all remaining nodes have degree three and the graph is now symmetric with respect to $\{v_1, v_2, v_3\}$ and $\{v_4, v_5, v_6\}$. As each pair of nodes from either set is not adjacent, the greedy algorithm will delete any pair. Thus, a total of eleven nodes are removed from this graph. The bruteforce approach, however, yields a better result. Initially, each pair of nodes of the set $\{v_1, v_2, v_3\}$ is not adjacent and each of these nodes has degree four. Consequently, deleting all nodes of the set $\{v_1, v_2, v_3\}$ in arbitrary order allows the bruteforce algorithm to remove twelve edges from the graph. However, this greedy approach might still be preferred to the bruteforce algorithm as it can easily be implemented and often comes close to the maximum number of edges that can be removed by deleting a certain number of nodes. As this greedy strategy has to find the node having the highest degree in each of k iterations (which can be done in $O(n)$ steps), its overall complexity is $O(kn)$.

4 An Improved Greedy Strategy

Throughout this section, we use S_i to denote the set of nodes that were removed in the first $i-1$ iterations and let $\deg_S(v)$ be the degree of $v \in V$ if all nodes of the set S were removed from the graph.

The greedy strategy that we described in the previous section does not succeed in removing as many edges as possible from a graph in all cases as it does not take into account the number of edges that can be removed in the subsequent

$k-i$ iterations when deciding on the node that will be deleted in the ith step. Thus, in order to improve this greedy strategy for the node-deletion problem, it is necessary to find a way to compute or estimate the number of edges that can be removed in the remaining $k-i$ steps provided that a node v of the set $V \setminus S_i$ was deleted in the ith step. We will now look at three different approaches.

1. One idea is to choose from all $\binom{n-i}{k-i}$ possibilities of selecting $k-i$ nodes from the remaining $n-i$ nodes the set that results in the highest number of removed edges if all its elements are deleted from the current graph. However, this approach basically corresponds to the idea of the bruteforce algorithm described in Sect. 3.1 and will be of limited use for bigger graphs.
2. A different approach is to sort the remaining nodes by their degree in descending order and select the first $k-i$ nodes from the ordered list. Then, approximate the sought-for value by calculating the sum of degrees of the selected nodes. Note that this approach usually overestimates the real number of edges that can be removed in the remaining iterations as it does not consider that some of the selected nodes may be connected by an edge that is consequently counted twice.
3. Moreover, one can also attempt to take into account the effects of deleting v by looking ahead one step in the iteration and apply a second greedy algorithm to determine a lower bound on the number of edges that can be removed in the subsequent iterations. This approach will now be described in more detail.

Assume that the algorithm has completed the first $i-1$ iterations and now has to find the node that will be deleted in the ith step. For each node v of the set $V \setminus S_i$ of the remaining $n-i+1$ nodes, carry out the following steps:

1. Create a copy of the current graph and delete v from this copy.
2. For each of the subsequent $k-i$ iterations, find and delete from the copy the node with the highest degree. Let $h_i(v)$ denote the total number of edges that are removed from the copy in this step.

Finally, set

$$g_i(v) = \deg_{S_i}(v) + h_i(v) \tag{5}$$

to calculate a lower bound on the number of edges that can be removed in the ith step and its subsequent iterations provided that v was deleted in the ith iteration. In order to apply this idea to the node-deletion problem, determine in each iteration $1 \leq i \leq k$ the node $w \in V \setminus S_i$ that maximizes g_i, remove it from the graph along with its adjacent edges and update the set of deleted nodes by setting $S_{i+1} = S_i \cup \{w\}$. A high-level description of this *nested greedy* approach is given in Algorithm 1.

It should be emphasized that $h_i(v)$ provides only a lower bound (but usually a good one provided that $k-i$ is not too large) on the exact number of edges that can be removed in the remaining $k-i$ iterations if v was deleted in the ith step. While an algorithm deleting nodes on the basis of g_i will consequently not remove as many edges as possible in all cases, it will never perform worse than the greedy algorithm described in Sect. 3.2.

Theorem 2. *The nested greedy algorithm never removes less edges than the greedy approach.*

Proof. If $t \in V$ denotes the node having the highest degree in the initial graph, the total number of edges removed by the greedy algorithm equals $g_1(t)$. Algorithm 1 evaluates $g_1(v)$ for all $v \in V$ to calculate a lower bound on the overall number of edges that can be removed under the assumption that v was deleted in the first step and selects the node that maximizes g_1. As $t \in V$, it follows that $g_1(t) \leq \max_{v \in V} g_1(v)$ and the nested greedy algorithm will therefore never remove less edges in comparison to the greedy approach. □

```
1  S ← ∅
2  r ← 0 // Denotes the number of edges removed by this algorithm
3  for i ← 1 to k do
4  |  b ← null
5  |  m ← −1 // Any other value less than zero is fine
6  |  for v ∈ V \ S do
7  |  |  u ← deg_S(v)
8  |  |  S′ ← S ∪ {v} // Create a copy and delete v from this copy
9  |  |  for j ← i + 1 to k do
10 |  |  |  w ← v ∈ V \ S′ : deg_S′(v) ≥ deg_S′(x) ∀ x ∈ V \ S′
11 |  |  |  u ← u + deg_S′(w)
12 |  |  |  S′ ← S′ ∪ {w} // Delete w from the copy
13 |  |  end
14 |  |  if u > m then
15 |  |  |  m ← u
16 |  |  |  b ← v
17 |  |  end
18 |  end
19 |  r ← r + deg_S(b)
20 |  S ← S ∪ {b} // Update graph by deleting b
21 end
```

Algorithm 1. The nested greedy algorithm for the node-deletion problem

The inner loop (lines 1 to 1) of Algorithm 1 that is used to calculate h_i makes this approach more complex in comparison to the greedy algorithm. Recall that the number of nodes that have not been removed from the graph at the beginning of the ith iteration ($1 \leq i \leq k$) equals $n - i + 1$. In each iteration i and for every node $v \in V \setminus S_i$ of the set of the remaining nodes a copy of the graph is created which v is deleted from. Then, $k - i$ iterations (i.e. for $i + 1 \leq j \leq k$) of the inner loop are used to calculate $h_i(v)$. In each of these iterations the node having highest degree among all $n - j + 1$ nodes that remain in the copy of the graph has to be found and deleted from the copy. Thus, the total number of steps can be evaluated as

$$\sum_{i=1}^{k} \left((n - i + 1) \cdot \sum_{j=i+1}^{k} (n - j + 1) \right) = O\left(k^2 n^2 + k^4\right). \tag{6}$$

Recall that a total of k nodes are removed from a graph with n nodes. As k will be much smaller than n in our field of application, the total number of steps is limited by $O\left(k^2n^2\right)$.

Further improvement of Algorithm 1 requires a better estimation of the number of edges that can be removed in the subsequent iterations than the lower bound provided by h_i. We found that increasing the level of nesting the greedy algorithms easily allows for an improved lower bound and will be useful if either the number of nodes in the graph or the number of nodes that one is allowed to remove increases. To illustrate this idea, the entire Algorithm 1 was nested into another greedy algorithm. This new algorithm has running time $O\left(k^3n^3\right)$ as Algorithm 1 will be called $O(n)$ times in each of k iterations and will be denoted *twice nested greedy*. It can be shown using the arguments of the previous proof that this new algorithm will never perform worse than Algorithm 1.

5 Experimental Results

We compared our proposed nested greedy as well as the twice nested greedy algorithm to the bruteforce approach and the greedy algorithm on the basis of three parameters. The first parameter, n, denotes the number of conditions used for the comparison of two rankings. While n was larger than 100 in the experiments with the bacterium Pseudomonas aeruginosa, we used values ranging from 30 to 60 to keep the running times of the algorithms, and in particular of the bruteforce approach, low. The second parameter, denoted τ, represents the value of Kendall's tau rank correlation coefficient for a pair of rankings. Recall that the number of edges in a graph is given by $\binom{n}{2}\frac{1-\tau}{2}$ for $-1 \leq \tau \leq 1$. Our tests included values of 0, 0.25 and 0.5 in order to test the algorithms on graphs of varying sparsity. Finally, as k, the number of conditions that we are allowed to remove to increase the rank correlation coefficient, is unknown, we used two values ranging from 10% to 15% of the value of n.

The tests were performed as follows. For each combination of n and τ, we created a list of 10,000 pairs of rankings with a length of n and a rank correlation coefficient of τ. If this was not possible, for example if $n = 15$ and $\tau = 0$, the pairs of rankings were created in such a way that their rank correlation coefficient is as close to τ as possible. Then, each pair of rankings was used to create an undirected graph using the instructions in Sect. 2. Finally, the algorithms were applied to delete k nodes from each graph in such a way that as many edges are thereby removed. The numbers in the three rightmost columns of Table 2 represent the number of times (out of 10,000) that each of the algorithms removed less edges from a graph in comparison to the bruteforce approach. In order to compare the algorithms on bigger graphs, a slightly different approach was required. For each combination of n and τ, we again created 10,000 pairs of rankings, but excluded the bruteforce approach from the tests as it could not be applied to the resulting graphs in reasonable length of time. As the twice nested greedy algorithm never performs worse than the greedy or the nested greedy approach, all comparisons were no longer done with respect to the bruteforce

Table 2. Results on comparing three different greedy algorithms with respect to the bruteforce approach

n	τ	k	Greedy	Nested Greedy	Twice Nested Greedy
30	0	3	205	7	0
30	0	5	548	31	3
30	0.25	3	149	6	0
30	0.25	5	466	21	0
30	0.5	3	183	3	0
30	0.5	5	479	33	0
40	0	4	291	12	0
40	0	6	482	43	0
40	0.25	4	190	4	0
40	0.25	6	470	44	3
40	0.5	4	227	4	1
40	0.5	6	483	39	1

Table 3. Results on comparing the greedy and the nested greedy approach with respect to the twice nested greedy algorithm

n	τ	k	Greedy	Nested Greedy
50	0	5	269	13
50	0	8	612	49
50	0.25	5	207	9
50	0.25	8	548	36
50	0.5	5	226	8
50	0.5	8	601	48
60	0	6	304	15
60	0	9	567	46
60	0.25	6	269	11
60	0.25	9	566	45
60	0.5	6	261	10
60	0.5	9	559	63

approach, but to the twice nested greedy algorithm. Consequently, the results in Table 3 only provide a valuable indication but do not exactly resemble how well the greedy and the nested greedy algorithm perform on bigger graphs in comparison to the bruteforce approach. We conclude from these results that the greedy algorithm should not be used for the correction of the rank correlation coefficient as it happens quite frequently that this algorithm does not remove as many discordant pairs as possible, even if the number of conditions that one is allowed to remove is small. Our proposed nested greedy algorithm, however, provides a more reliable method and removed less discordant pairs in comparison to the bruteforce approach in only very few cases. While the twice nested greedy algorithm yields results that come even closer to those of the bruteforce approach,

it shall only be used if one accepts the higher running time. However, if both n and $\min(k, n-k)$ are very small, the bruteforce approach might still be used.

6 Conclusions

We have proposed an efficient algorithm to cope with an otherwise infeasible problem. Our basic assumption was that the number k of conditions to be removed is fixed. Future work will include investigations on choosing k automatically based on statistical considerations, i.e. to find the point, when we start to increase the rank correlation coefficient artificially.

References

1. Fujito, T.: A Unified Local Ratio Approximation of Node-Deletion Problems (Extended Abstract). In: Díaz, J. (ed.) ESA 1996. LNCS, vol. 1136, pp. 167–178. Springer, Heidelberg (1996)
2. Kendall, M.G.: A new measure of rank correlation. Biometrika 30, 81–93 (1938)
3. Kruskal, J.B.: On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. Proc. Am. Math. Soc. 7, 48–50 (1956)
4. Yannakakis, M.: Node-and edge-deletion NP-complete problems. In: 10th Annual ACM Symposium on Theory of Computing, pp. 253–264. ACM, New York (1978)

Probabilistic Relational Learning for Medical Diagnosis Based on Ion Mobility Spectrometry*

Marc Finthammer[1], Christoph Beierle[1], Jens Fisseler[1],
Gabriele Kern-Isberner[2], Bülent Möller[2], and Jörg I. Baumbach[3,4]

[1] Dept. of Computer Science, FernUniversität in Hagen
[2] Dept. of Computer Science, Technische Universität Dortmund
[3] Dept. of Metabolomics, Leibniz-Inst. für Analyt. Wissenschaften - ISAS, Dortmund
[4] Dept. Clinical Diagnostics, KIST Europe, Saarbrücken

Abstract. Probabilistic relational modelling and learning is used for the problem of diagnosing lung cancer based on data obtained from peak clusters in ion mobility spectra. Markov Logic Networks and the MLN system Alchemy are employed for various modelling and learning scenarios which are evaluated with respect to ease of use, classification accuracy, and knowledge representation aspects.

1 Introduction

Probabilistic propositional logics have already been studied and used in many applications where uncertainty is inherent in the available information. Exploiting the greater expressive power of first order logic in the combination with probability is the objective of some more recent approaches like Bayesian Logic Programs [3] or Markov Logic Networks (MLNs) [8].

This paper reports on a case study of using probabilistic relational modelling and learning as provided by MLNs and the MLN system Alchemy [4] in the field of biomedical diagnosis. We focus on the early detection of lung cancer by ion mobility spectrometry obtained from the breath a patient exhales [1], a non-invasive diagnostic method which delivers results within a few minutes and can be applied at low costs. To be more precise, we investigated the relationships between the presence of certain peaks in the spectra and the presence of bronchial carcinoma. In order to allow for small variations of peak locations in different spectra, we extracted peak clusters from the data with the help of a multi-level modification of the k-means algorithm. We set up various learning scenarios, and evaluate them with respect to ease of use and classification accuracy. The MLNs prove to return satisfactory results as a logic-based machine learning tool, but also weaknesses with respect to semantical clearness and expressivity become apparent.

* The research reported here was partially supported by the DFG (BE 1700/7-1 and KE 1413/2-1), by Germany's high-tech strategy funds (Project Metabolit-01SF0716) and by the European Union (Project No. 217967). We thank Dr. M. Westhoff (Lung clinic Hemer), Dr. Th. Perl (University Göttingen), B. Bödeker, and B&S Analytik for their successful cooperation, valuable contribution, and general support.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 365–375, 2010.

2 Ion Mobility Spectrometry

In order to determine chemical substances in gaseous analytes, ion mobility spectrometry (IMS) can be used [1]. This method relies on characterizing substances in gases by their ion mobility. Figure 1 illustrates the working principle of an ion mobility spectrometer. After ionisation, ion swarms enter the drift region through an ion shutter. The time needed to pass the drift region is called *drift time*, and the ion mobility is inversely proportional to the drift time. Ion mobility is determined by mapping the drift time to the signal intensity measured at the Faraday plate (cf. Fig. 1). If the gaseous analyte contains various substances, they may reach the Faraday plate at the same time. Therefore, a multi capillary column is used for the pre-separation of different substances [1] so that they enter the spectrometer at different time points, called *retention times*.

Formally, a *drift vector* S is a sequence $S = (z_1, \ldots, z_n)$ of signal intensities z_i measured at the Faraday plate at time points $i \times C$ where C is a fixed time interval. A spectrum M is a sequence $M = (S_0, \ldots, S_m)$ of drift vectors corresponding to distinct retention times $r_{t_0}, \ldots, r_{t_m}$. An IMS spectrum M can be visualized as a heat map on the two dimensions drift time and retention time, while the signal intensity is represented by a colour. One is particularly interested in *peaks* in this heat map as each peak gives information about a particular substance in the analyte. A peak object in a spectrum is characterized by a set of direct or closely related neighbours in the matrix M whose signal value is above a given minimal value; thus, a peak object corresponds to a specific area in the heat map. The determination of peaks in a measurement requires sophisticated processing of the raw spectra (see [1] for details). Peak objects taken from two different measurements that correspond to the same substance occur at corresponding areas in their respective heat maps, and in order to identify such corresponding peaks, they will be mapped to peak clusters.

3 A Modified *k*-Means Algorithm for Peak Clustering

An IMS database $D = \{M_1, \ldots, M_k\}$ is a set of measurements M_i where each measurement is an IMS spectrum $M_i = (S_0^i, \ldots, S_m^i)$. For a spectrum M let $\mathit{peaks}(M)$ denote the set of peaks in M, and $\mathit{peaks}(D) = \cup_{M \in D} \mathit{peaks}(M)$. Then a *peak clustering* for D is a partitioning $PC_1, \ldots, PC_p$ of $\mathit{peaks}(D)$, and $\mathit{peakIndex} : \mathit{peaks}(D) \to \{1, \ldots, p\}$ denotes the mapping that sends each peak P to the index i of the unique *peak cluster* PC_i with $P \in PC_i$. A peak cluster is meant to represent a vague concept of a peak that corresponds to a substance but allows for slight variations in its actual manifestation within a spectrum.

Since in the ideal case each peak in a spectrum corresponds to a different substance, we are only interested in peak clusterings that map all peaks within one and the same measurement to different clusters, i.e., if the restriction $\mathit{peakIndex}_{|\mathit{peaks}(M)}$ is injective for every $M \in D$. Such a clustering for D is called *(measurement) slicing*. In order to obtain a slicing peak clustering for an IMS database D, we extended the well-known k-means algorithm [5] to a *multi-level k-means* approach:

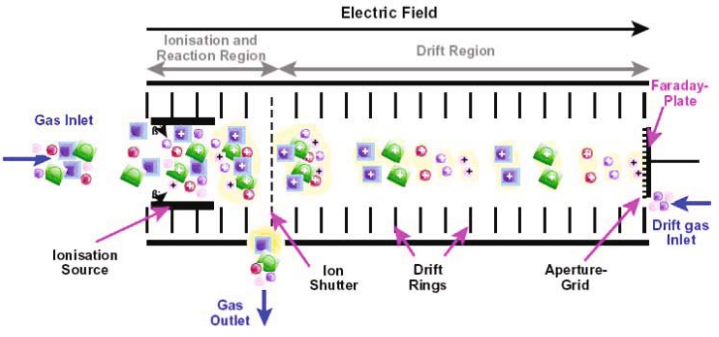


Fig. 1. Schematic overview of an ion mobility spectrometer (from [1])

- First, the k-means algorithm is applied to $peaks(D)$. Each resulting cluster that contains at least two peaks from the same measurement is called *non-slicing* and is processed further on the next level; all other clusters are kept for the final clustering.
- To each non-slicing cluster PC, k-means is applied where the number of clusters to be obtained from PC is set to the maximal number of peaks in PC stemming from the same measurement. This step is applied again on any subsequent level as long as a non-slicing cluster occurs.

It is easy to prove that the multi-level k-means algorithm terminates and yields a slicing peak clustering for every IMS database D. So, the standard k-means algorithm is employed with different values for k that take domain specific information into account. In this way, not a fixed number of clusters results but the clustering is as finely grained as the IMS database requires.

In our case study, we investigated an IMS database consisting of 158 measurements obtained from screening the breath of 158 patients out of which 82 had lung cancer (*bronchial carcinoma, bc*). The idea behind this setting is to support early diagnosis of lung cancer on the basis of the substances a person exhales [1].

Applying the multi-level k-means algorithm to this database yielded a database D_{bc} with 33 peak clusters, in the following referred to by the identifiers $pc0, \ldots, pc32$. For each peak cluster pc, Table 1 shows the probability resp. relative frequency $P(pc)$ that a measurement has a peak belonging to pc, and the conditional probability $P(bc|pc)$ that a measurement having a peak belonging to pc stems from a person having bronchial carcinoma.

In the following sections, we apply methods of probabilistic relational modelling and learning to (a logic representation of) D_{bc}. The logic representation (for convenience, also referred to as D_{bc}) involves atomic formulas of the form $bc(M)$ and $pcInM(pc18, M)$, indicating that measurement M belongs to a person with lung cancer, and that peak cluster $pc18$ occurs in measurement M, respectively.

Table 1. Peak clusters and probabilities calculated for D_{bc} with 158 measurements

pc	$P(pc)$	$P(bc\|pc)$	*pc*	$P(pc)$	$P(bc\|pc)$	*pc*	$P(pc)$	$P(bc\|pc)$
pc0	0.84	0.43	*pc11*	0.04	0.83	*pc22*	0.60	0.40
pc1	0.23	0.68	*pc12*	0.03	0.60	*pc23*	0.65	0.40
pc2	0.93	0.50	*pc13*	0.04	0.86	*pc24*	0.09	0.43
pc3	0.11	0.82	*pc14*	0.89	0.56	*pc25*	0.63	0.56
pc4	0.67	0.55	*pc15*	0.03	0.60	*pc26*	0.53	0.43
pc5	0.39	0.72	*pc16*	0.06	1.00	*pc27*	0.05	0.75
pc6	0.04	0.83	*pc17*	0.51	0.74	*pc28*	0.06	0.33
pc7	0.22	0.91	*pc18*	0.47	0.48	*pc29*	0.25	0.33
pc8	0.05	0.88	*pc19*	0.04	1.00	*pc30*	0.52	0.20
pc9	0.61	0.67	*pc20*	0.03	1.00	*pc31*	0.81	0.51
pc10	0.06	0.20	*pc21*	0.24	0.76	*pc32*	0.24	0.68

4 Relational Probabilistic Learning and Modelling

Inductive Logic Programming (ILP) has been used very successfully in machine learning for inducing first-order hypotheses from examples and background knowledge [6]. For the application scenario described in Sec. 3, we will use the ILP system Aleph [9] for learning bronchial carcinoma diagnosing rules from D_{bc}.

Markov logic [8] establishes a framework which combines Markov networks [7] with first-order logic to handle a broad area of statistical relational learning tasks. The Markov logic syntax complies with first-order logic except that each formula F_i is quantified by an additional (positive or negative) weight value w_i. Semantics are given to sets of Markov logic formulas by a probability distribution over (propositional) possible worlds which is given by a log-linear model defined over weighted ground formulas. The fundamental idea in Markov logic is that first-order formulas are not handled as hard constraints but each formula is more or less softened depending on its weight. These weights induce a kind of priority ordering on the formulas of the knowledge base that determines their respective influence on the probabilities of the log-linear model. A *Markov logic network (MLN)* L is a set of weighted first-order logic formulas together with a set of constants C. The ground Markov network $M_{L,C}$ specifies

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_i w_i n_i(x)\right) \tag{1}$$

as the probability distribution over possible worlds $x \in \mathcal{X}$, where Z is a normalization factor. For each formula F_i, $n_i(x)$ compactly expresses the number of true groundings of F_i in the possible world x. The semantics of L is given by a ground Markov network $M_{L,C}$ constructed from F_i and C [8]. The standard semantics of Markov networks [7] is used for reasoning, e.g. to determine the probabilistic inferences of L (see [8] for details).

Alchemy's inference algorithms allow to calculate (approximately) the conditional probability of a certain ground atom, given some other ground literals as evidence [4]. However, it has to be emphasised that an MLN only allows to

express an if-then-rule as a material implication, not as a conditional probability. Probabilities of material implications may differ largely from conditional probabilities and are known to be quite unintuitive in certain cases. As (1) shows, the weight of an MLN implication is effective for a probability if the implication is logically satisfied, regardless whether both the premise and the consequent hold, or just the premise fails.

5 Learning Classification Rules with ILP and MLNs

In this section, we present different setups to learn MLNs from the data set D_{bc}. Our goal is to calculate the probability that a certain measurement m is from some person with a bronchial carcinoma, given the information for each of the 33 peak clusters whether or not it is contained in measurement m. That is, we want to calculate the conditional probability of $bc(m)$, given the truth values of the literals $pcInM(pc0, m), \ldots, pcInM(pc32, m)$. We use the software package Alchemy [4] which provides several sophisticated algorithms to perform (structure and parameter) learning and inference of MLNs.

We will validate a learned MLN in terms of classification accuracy, defined as the proportion of the correctly predicted (positive and negative) results on the total number of measurements in a testing set. Therefore, we perform a 10-fold cross-validation and determine the average accuracy value of all ten testing sets. Moreover, we will try to elucidate the semantics of MLN-weights by investigating the correlations between them and conditional probabilities.

5.1 Learning Logic Rules with the ILP System Aleph

In our first learning setup, we use the inductive logic programming (ILP) system *Aleph* [9] for learning first-order logic rules from the data set. Besides other parameters, Aleph allows to make detailed specifications about which atoms may appear in the body or head of a rule. We require that the head of a rule must contain the *bc* predicate, and the body must consist of one or more atoms of the *pcInM* predicate, with a constant in the first argument. This way, Aleph is guided to learn classification rules regarding *bc*.

The rules learned with Aleph are displayed in Table 2. We use a compact notation in all tables, where e. g. the atom $pcInM(pc18, M)$ is abbreviated as $Pc18$ and $bc(M)$ as bc. The first seven rules have $bc(M)$ as their consequent, whereas the four other rules have the negated atom $\neg bc(M)$. The premisses of all rules consist of conjunctions of at most three positive *pcInM* literals. From the 33 peak clusters in the data set, there are actually only 18 contained in the rule-set, so the other 15 peak clusters seem to carry no useful information according to the Aleph result. Viewing an implication $p \Rightarrow q$ as a directed rule *if* p *then* q, we can talk of its support (the relative number of database entries satisfying $p \wedge q$) and its confidence (the number of entries satisfying $p \wedge q$ relative to the number of entries satisfying p). Taking a closer look at the support of these rules, it becomes obvious that most rules apply only to a small part of the

Table 2. Rules learned with the ILP system Aleph in compact form. E. g., $Pc5 \wedge Pc8 \Rightarrow bc$ abbreviates $pcInM(pc5, M) \wedge pcInM(pc8, M) \Rightarrow bc(M)$.

#	Rule	Supp.	Conf.	#	Rule	Supp.	Conf.
1	$Pc5 \wedge Pc8 \Rightarrow bc$	0.03	1.00	7	$Pc1 \wedge Pc4 \wedge Pc17 \Rightarrow bc$	0.08	1.00
2	$Pc7 \wedge Pc17 \wedge Pc31 \Rightarrow bc$	0.16	1.00	8	$Pc28 \wedge Pc30 \Rightarrow \neg bc$	0.04	1.00
3	$Pc18 \wedge Pc21 \wedge Pc25 \Rightarrow bc$	0.06	1.00	9	$Pc24 \wedge Pc29 \wedge Pc30 \Rightarrow \neg bc$	0.03	1.00
4	$Pc1 \wedge Pc3 \Rightarrow bc$	0.05	1.00	10	$Pc10 \wedge Pc23 \Rightarrow \neg bc$	0.05	1.00
5	$Pc16 \Rightarrow bc$	0.06	1.00	11	$Pc5 \wedge Pc24 \wedge Pc30 \Rightarrow \neg bc$	0.02	1.00
6	$Pc3 \wedge Pc17 \wedge Pc23 \Rightarrow bc$	0.03	1.00				

data; merely rule 2 applies to some larger amount of measurements. Notice that all rules have confidence 1.0 due to the fact that the whole data set D_{bc} was given to Aleph for learning these rules.

Taking the conditional probabilities from Table 1 into consideration, we can compare the consequent of a rule with the conditional probabilities of the peak clusters in the rule's premise. E. g., for $pcInM(pc5, M)$ and $pcInM(pc8, M)$ – the premise of rule 1 – we have $P(bc|pc5) = 0.72$ and $P(bc|pc8) = 0.88$ in the data set. So each of these peak clusters (by itself) shows a clearly above 0.5 conditional probability and therefore can be considered to be an indicator for bc; and this observation matches the consequent $bc(M)$ of rule 1. Comparing the conditional probabilities of the peak clusters in the premisses of the other rules with their respective consequences shows a similar result: apart from three exceptions, the indication of each peak cluster's conditional probability matches the consequent of the respective rule. That is, a peak cluster with a conditional probability above 0.5 is contained in the premise of a rule whose consequent is $bc(M)$; and a peak cluster with a conditional probability below 0.5 is contained in the premise of a rule whose consequent is $\neg bc(M)$.

Compared to the simple observations from Table 1, the rules in Table 2 seem quite plausible. According to these rules, there are only few dependencies between the 18 peak clusters occurring in the rules, since at most three peak clusters are contained in the same rule and only six of these peak clusters appear two or (at most) three times.

5.2 Learning Weights of Aleph Formulas with Alchemy

In a subsequent step, we take the Aleph implications (Table 2) as logical base structure of an MLN and learn appropriate weights for them from the data set using Alchemy. The resulting weights are represented in Table 3. Evaluating the MLN prediction performance results in an accuracy of 78%.

If we take the implications as if-then-rules, we can determine the conditional probabilities of these rules under the distribution induced by the MLN, i. e. we use Alchemy to calculate the conditional probability of a rule's consequent ground atom given its premise ground atoms as evidence. E. g., for rule 1, Alchemy determines the probability $P(bc(m)|pcInM(pc5, m) \wedge pcInM(pc8, m)) = 0.9800$ in the

Table 3. MLN formulas originated from Aleph rules [78% accuracy]

#	Formula	Wt.	CPr.	#	Formula	Wt.	CPr.
1	$Pc5 \wedge Pc8 \Rightarrow bc$	4.596	0.980	7	$Pc1 \wedge Pc4 \wedge Pc17 \Rightarrow bc$	4.352	0.985
2	$Pc7 \wedge Pc17 \wedge Pc31 \Rightarrow bc$	6.004	0.996	8	$Pc28 \wedge Pc30 \Rightarrow \neg bc$	3.772	0.978
3	$Pc18 \wedge Pc21 \wedge Pc25 \Rightarrow bc$	4.402	0.982	9	$Pc24 \wedge Pc29 \wedge Pc30 \Rightarrow \neg bc$	3.376	0.991
4	$Pc1 \wedge Pc3 \Rightarrow bc$	4.433	0.992	10	$Pc10 \wedge Pc23 \Rightarrow \neg bc$	4.041	0.974
5	$Pc16 \Rightarrow bc$	4.788	0.971	11	$Pc5 \wedge Pc24 \wedge Pc30 \Rightarrow \neg bc$	2.668	0.961
6	$Pc3 \wedge Pc17 \wedge Pc23 \Rightarrow bc$	4.665	0.989				

MLN. The conditional probabilities obtained this way for each rule are stated in the "CPr." column of Table 3. Notice that the conditional probabilities of all rules are not exactly 1.0, as expected, but rather close to it. This is due to the fact that Alchemy performs approximate inference and thereby, as a side-effect, prevents overfitting.

The weights of the MLN formulas are not easy to interpret, since they do not have a clear probabilistic semantics. To put it very simply, the violation of a formula with a relative high weight will lead to a stronger probability decrease (of a possible world) than the violation of a formula with a lower weight, all other things being equal (see equation (1)). So the weight of a formula expresses in a way the relative "importance" or "strength" of a formula within the set of MLN formulas. The learned weights of the eleven formulas are all positive. Taking a look at the support values from Table 2, it is interesting to notice that rule 2 has both the highest support and weight, and that rule 11 has both the lowest support and weight. It has to be kept in mind that the weight of an MLN implication does not reflect (in any way) a conditional probability associated with this implication. This becomes evident by observing that the highest weight is more than twice the lowest, although all conditional probabilities are very similar (close to 1).

5.3 Simple Classification with MLNs

In a further learning setup, we predefine the formula structure of a quite simple MLN: The MLN consists of the 33 implications $pcInM(pc0, M) \Rightarrow bc(M), \ldots,$ $pcInM(pc32, M) \Rightarrow bc(M)$. Since the Alchemy syntax allows to express such "partially grounded" formulas in a compact way, the whole predefined structural Alchemy input merely consists of single line. With this MLN structure, we follow a straightforwardly modelled classification approach: To classify the *bc* state of a measurement, we consider each peak cluster separately, leaving out any connections or dependencies among them. To some extent, this approach resembles Naive Bayes classification, where explicit independence assumptions among classifying attributes are made. The weights learned with Alchemy for the 33 implications are stated in Table 4.

The evaluation of the learned MLN reveals quite a high accuracy of 88%, although the enforced MLN structure lacks any connections between peak

Table 4. Predefined classification formulas and their learned weights [88% accuracy]

Formula	Weight	Formula	Weight	Formula	Weight
$Pc0 \Rightarrow bc$	−1.6771	$Pc11 \Rightarrow bc$	1.0978	$Pc22 \Rightarrow bc$	−0.7165
$Pc1 \Rightarrow bc$	1.1153	$Pc12 \Rightarrow bc$	0.2402	$Pc23 \Rightarrow bc$	−1.7247
$Pc2 \Rightarrow bc$	−0.9009	$Pc13 \Rightarrow bc$	3.0332	$Pc24 \Rightarrow bc$	1.1053
$Pc3 \Rightarrow bc$	2.2471	$Pc14 \Rightarrow bc$	0.4628	$Pc25 \Rightarrow bc$	0.7341
$Pc4 \Rightarrow bc$	2.0735	$Pc15 \Rightarrow bc$	2.3570	$Pc26 \Rightarrow bc$	−0.9295
$Pc5 \Rightarrow bc$	0.5205	$Pc16 \Rightarrow bc$	6.9273	$Pc27 \Rightarrow bc$	1.0999
$Pc6 \Rightarrow bc$	2.2603	$Pc17 \Rightarrow bc$	1.5759	$Pc28 \Rightarrow bc$	0.1661
$Pc7 \Rightarrow bc$	2.6155	$Pc18 \Rightarrow bc$	−0.4797	$Pc29 \Rightarrow bc$	−1.1484
$Pc8 \Rightarrow bc$	3.6903	$Pc19 \Rightarrow bc$	3.5149	$Pc30 \Rightarrow bc$	−3.1752
$Pc9 \Rightarrow bc$	0.6047	$Pc20 \Rightarrow bc$	4.3739	$Pc31 \Rightarrow bc$	0.2722
$Pc10 \Rightarrow bc$	−0.1953	$Pc21 \Rightarrow bc$	2.1169	$Pc32 \Rightarrow bc$	−0.2974

clusters. But the high accuracy suggests that those connections are not of such great importance for classifying the measurements regarding *bc*.

Comparing the learned weights of the implications in Table 4 to the conditional probabilities in Table 1 of the respective peak clusters, shows no clear correlation. At least, it is noticeable, that most implications with a positive weight contain a peak cluster with a conditional probability clearly above 0.5; and that vice versa most implications with a negative weight contain a peak cluster with a conditional probability clearly below 0.5. But since this observations do not hold for peak clusters $pcInM(pc24, M)$, $pcInM(pc28, M)$, and $pcInM(pc32, M)$, it does not allow to draw a direct relation between the weights and the conditional probabilities. It is also peculiar that some peak clusters with an (almost) identical conditional probability belong to implications with significantly different weights (e. g. $pcInM(pc10, M)$ and $pcInM(pc30, M)$, as well as $pcInM(pc12, M)$ and $pcInM(pc15, M)$). Thus, despite its accuracy of 88%, the MLN is somewhat unsatisfactory from a knowledge representation point of view.

5.4 MLN Structure Learning

In our last learning setup, we make use of Alchemy's structure learning feature to learn an MLN from scratch. Alchemy does not allow to make detailed specifications about the formulas to be learned (compared to Aleph in Sec. 5.1), i. e. we cannot impose the requirement that the $pcInM(\_, \_)$ atoms have a constant in the first argument. As a consequence, Alchemy's structure learning algorithm produces no useful results when applied to D_{bc} without any further information. So we modify the relational modelling in some aspect by replacing the binary predicate $pcInM(PC, M)$ by 33 unary predicates $pc0(M), \ldots, pc32(M)$.

The structure (and weight) learning with Alchemy starts from an empty MLN and results in an MLN with 89 formulas (including 34 atomic formulas for all 34 predicates). Table 5 shows 18 of the learned formulas (including those with the highest and lowest overall weights), the other ones are disjunctions over up to

Table 5. Some of the MLN formulas emerged from Alchemy's structure learning [90% accuracy]

#	Formula	Weight	CPr.	Supp.	Conf.
37	$Pc7 \Rightarrow bc$	4.43	1.0000	0.20	0.91
39	$Pc11 \Rightarrow Pc9$	4.82	0.9920	0.04	1.00
44	$Pc17 \wedge Pc28 \Rightarrow Pc21$	5.05	0.9930	0.02	1.00
46	$Pc15 \wedge Pc25 \Rightarrow Pc5$	−4.30	0.0050	0.00	0.00
47	$Pc17 \wedge Pc19 \wedge Pc20 \Rightarrow Pc9$	−8.98	0.0000	0.00	0.00
53	$Pc12 \wedge Pc20 \wedge Pc22 \Rightarrow Pc11$	−8.14	0.0000	0.00	0.00
57	$\neg Pc1 \wedge \neg Pc18 \wedge \neg Pc23 \wedge Pc31 \Rightarrow bc$	6.38	1.0000	0.08	1.00
61	$\neg Pc10 \wedge Pc14 \wedge \neg Pc18 \wedge Pc21 \Rightarrow bc$	7.15	1.0000	0.08	1.00
62	$\neg Pc12 \wedge \neg Pc22 \wedge \neg Pc30 \wedge Pc31 \Rightarrow bc$	7.49	1.0000	0.16	1.00
66	$Pc4 \wedge Pc26 \wedge Pc28 \wedge Pc29 \Rightarrow bc$	−5.62	0.0090	0.00	0.00
68	$\neg Pc9 \wedge \neg Pc13 \wedge \neg Pc16 \wedge Pc23 \wedge \neg Pc29 \Rightarrow \neg bc$	4.01	0.9980	0.16	0.96
70	$Pc1 \wedge Pc3 \wedge \neg Pc15 \wedge \neg Pc23 \wedge Pc26 \Rightarrow \neg bc$	−5.18	0.0000	0.00	0.00
72	$Pc0 \wedge \neg Pc11 \wedge \neg Pc12 \wedge \neg Pc21 \wedge Pc22 \Rightarrow \neg bc$	2.45	0.9210	0.31	0.89
75	$Pc5 \wedge Pc7 \wedge \neg Pc28 \wedge \neg Pc29 \wedge Pc31 \Rightarrow \neg bc$	−2.78	0.0000	0.00	0.00
80	$Pc0 \wedge \neg Pc12 \wedge \neg Pc16 \wedge Pc30 \wedge \neg Pc32 \Rightarrow bc$	−5.55	0.0030	0.01	0.03
81	$\neg Pc6 \wedge \neg Pc13 \wedge \neg Pc28 \wedge Pc31 \wedge Pc32 \Rightarrow \neg bc$	5.61	0.9630	0.05	0.73
82	$\neg Pc3 \wedge \neg Pc4 \wedge Pc25 \wedge \neg Pc28 \wedge \neg Pc32 \Rightarrow \neg bc$	8.77	1.0000	0.07	1.00
89	$\neg Pc3 \wedge \neg Pc11 \wedge Pc13 \wedge \neg Pc17 \wedge \neg Pc31 \Rightarrow \neg bc$	−5.15	0.0000	0.00	0.00

six peak cluster literals. The evaluation of this MLN shows an accuracy of 90%. Compared to the previous results, this MLN models much more connections among the peak clusters and their combined influence regarding $bc(M)$. It has to be noticed that only 13 of the 55 non-atomic formulas involve a bc literal, so the other 42 formulas express connections among the peak clusters regardless of the $bc(M)$ state. The formulas contain positive and negative peak cluster literals as well, whereas the rules in Table 3 contain positive literals only.

As done in Sec. 5.1, we compare the conditional peak cluster probabilities from Table 1 with the premises of implications involving bc literals. Contrary to Sec. 5.1, the comparison shows no clear result: The premises of the implications involve peak cluster literals whose conditional probability goes along with the implication's consequent, as well as peak cluster literals who do not match this pattern. So – compared to previous results – these implications exhibit more complex and subtle connections between the occurrences of peak clusters and the $bc(M)$ state.

Taking the implications as rules, we can determine their support and confidence in the data set: Most of the rules have no high support. The confidence of all rules with positive weight is very high or even 1.0; merely rule 81 shows a somewhat lower confidence. Those rules with a negative weight show (almost) 0.0 confidence. This makes sense considering the impact of a negative weight.

It is not surprising that the weights of this MLN do not suggest any further interpretations or relations, seeing that even the very simple structured MLN from Sec. 5.3 did not allow us to draw any useful conclusions from the weights.

So it can be summed up that this MLN provides a good classification result, but the semantics of the learned weights still remains quite unclear.

6 Conclusions and Future Work

In this paper we illustrated how probabilistic relational modelling and learning can be applied successfully in the biomedical domain to support the diagnosis of lung cancer based on ion mobility spectrometry. We introduced a multi-level k-means clustering algorithm, which we employed to preprocess the IMS input data for logical modelling, and presented three different approaches to learn MLNs from the clustered data set, using the software packages Aleph and Alchemy. In particular, Alchemy proved to be an effective and easy-to-use system for relational probabilistic parameter learning. The validation of the three learned MLNs showed a satisfactory, or good classification accuracy of 78%, 88%, and 90%, respectively. However, from a knowledge representation point of view, the resulting MLNs are not suitable to serve as intelligible knowledge bases. The weight of an MLN formula has no clear semantics, so only for very simple MLNs or in very special cases the impacts of a certain weight can be estimated. Another shortcoming of MLNs is that if-then-rules can only be expressed as material implications. The combination of both these weaknesses makes it hard to interpret the learned MLN formulas and their weights. Our results show that MLN learned from data provides the user with only a very vague and maybe fallacious picture of the domain under consideration. Specifying an MLN by the user from scratch by setting up weights intuitively might yield unexpected, even false results.

As part of our ongoing and future work, we will compare MLN models for IMS data with alternative approaches from the relational machine learning domain, like relational association rules [2] and Bayesian logic programs [3]. Moreover, in order to allow for a more accurate modelling of the application domain, we will take into account also information about intensity and area of peaks which can be extracted from the available data as well.

References

1. Baumbach, J.I., Westhoff, M.: Ion mobility spectometry to detect lung cancer and airway infections. Spectroscopy Europe 18(6), 22–27 (2006)
2. Dehaspe, L., Toironen, H.: Discovery of relational association rules. In: Relational Data Mining, pp. 189–208. Springer, Heidelberg (2000)
3. Kersting, K., De Raedt, L.: Bayesian Logic Programming: Theory and Tool. In: An Introduction to Statistical Relational Learning. MIT Press, Cambridge (2007)
4. Kok, S., Singla, P., Richardson, M., Domingos, P., Sumner, M., Poon, H., Lowd, D., Wang, J.: The Alchemy System for Statistical Relational AI: User Manual. Department of Computer Science and Engineering, University of Washington (2008)
5. MacQueen, J.B.: Some methods of classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)

6. Muggleton, S., De Raedt, L.: Inductive logic programming: Theory and methods. Journal of Logic Programming 19/20, 629–679 (1994)
7. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1998)
8. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning 62(1), 107–136 (2006)
9. Srinivasan, A.: The Aleph Manual (2007), http://www.comlab.ox.ac.uk/activities/machinelearning/Aleph/

Automated Gaussian Smoothing and Peak Detection Based on Repeated Averaging and Properties of a Spectrum's Curvature

Hyung-Won Koh and Lars Hildebrand

Dortmund University of Technology, Computer Science Department,
Chair I Otto-Hahn-Str. 16, 44221 Dortmund, Germany
hyung-won.koh@udo.edu, lars.hildebrand@udo.edu

Abstract. In this paper, we address the two problems of automated smoothing and peak detection in spectral data analysis. We introduce the concept of *triplet significance*, and propose a repeated averaging approach, which is able to find a balance between noise reduction and signal preservation based on properties of a spectrum's curvature. For evaluation purposes, multiple spectra are simulated at different levels of resolution and different distances between peaks for varying amplitudes of uniformly distributed noise. The results empirically show that the proposed methodology outperforms existing approaches based on local maximum detection or the lag-one autocorrelation coefficient.

1 Introduction and Related Work

Spectroscopic experiments commonly provide insights in the atomic or molecular structure and composition of a given sample [1]. Thereby, the pure signal resulting from spectroscopic experiments can often be represented as a superposition of basis functions with a Lorentzian or Gaussian shape. They are known as the *peaks* of a spectrum. For the remainder of this paper, we consider a spectrum to be given as follows:

Definition 1 (Spectrum). *A spectrum is defined as a bivariate data vector* $\mathbf{a} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ *of finite length* n*, with ordered positions* $x_1 < x_2 < \cdots < x_n$ *at equal distances* $\Delta x = x_{i+1} - x_i$*,* $i \in \{1, \ldots, n-1\}$*. We define the respective values* $y_i \in \mathbb{R}$*,* $i \in \{1, \ldots, n\}$ *in a minimalistic manner to be given as*

$$y_i = \sum_{j}^{m} f_j(x_i) + E_i, \tag{1}$$

$$\text{with } f_j(x_i) = A_j \frac{\lambda_j}{\lambda_j^2 + (x_i - \omega_j)^2}. \tag{2}$$

E_i *denotes signal noise at datapoint* x_i*, and* $f_j : \mathbb{R} \to \mathbb{R}$ *represents the* j*-th Lorentz function of the spectrum with horizontal position* ω_j*, with half-width-at-half-height parameter* λ_j*, and with scale* A_j*.*

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 376–385, 2010.

Within this paper, the occurrence of baseline drifts and other perturbations of the signal at lower frequencies are not explicitly considered, based on the assumption that existing preprocessing methods can properly handle these distortion effects, and refer for example to [2] for more detailed information). For reasons of simplicity, we further consider the *standard* case of $\lambda_j = A_j = 1$, $j \in \{1, \ldots, m\}$, resulting in maximal peak heights of $f_j(\omega_j) = 1$.

In the context of automated spectral data analysis, finding an accurate model of the measured signal is a desirable goal to achieve. For this purpose, the problems of *peak picking* and *peak fitting* are commonly solved, namely the identification of the set of underlying basis functions, and subsequently the approximation of the corresponding parameters. Concerning the former problem, peaks can easily be identified by searching for local maxima [3,4,5]. In this way however, two problems arise and need to be handled carefully: Firstly, noise and other distortions of the measurement commonly introduce additional local maxima, leading to an increased number of *false-positive* peaks ("ghost peaks"). Secondly, overlapping peaks are potentially omitted, and the resulting model is likely to be substantially falsified [6].

As reported previously, peaks can alternatively be identified by means of properties of a spectrum's curvature [7]. Empirical studies show that, given a certain degree of smoothness of the considered spectra, peaks can successfully be discriminated from noise based on a scoring of peaks, which basically reflects the length and the degree of the respective curvatures. A positive side-effect from investigating the curvature is that the method can also identify peaks which are not represented as distinctive local maxima but as monotonically increasing (left) or monotonically decreasing (right) "shoulders" of other peaks. For this reason, we stick to this idea and give in the following similar definitions of a peak triplet and its score.

Definition 2 (Peak Triplet). *Given spectrum* $\mathbf{a}$*, and given the second discrete derivative* $\mathbf{a}'' = \{(x_2, y_2''), \ldots, (x_{n-1}, y_{n-1}'')\}$*, with* $y_i'' = y_{i-1} + y_{i+1} - 2\,y_i$*, a peak triplet* p *is defined as a triplet of indices* $p = \{l, m, r\}$*, for which the following holds:*

$$\begin{array}{llll} & y_m'' < 0 & \wedge \quad y_{m-1}'' > y_m'' \leq y_{m+1}'' & \\ \wedge & l < m & \wedge \quad y_{l-1}'' < y_l'' \geq y_{l+1}'' & \wedge \quad y_j'' \geq y_{j+1}'',\ j \in \{l, \ldots, m-1\} \\ \wedge & r > m & \wedge \quad y_{r-1}'' < y_r'' \geq y_{r+1}'' & \wedge \quad y_j'' < y_{j+1}'',\ j \in \{m, \ldots, r-1\} \end{array}$$

Further, we call a pair of peak triplets $(p_1 = \{l_1, m_1, r_1\}, p_2 = \{l_2, m_2, r_2\})$ *to be adjacent, if* $r_1 = l_2$ *holds.*

In other words, a *peak triplet* encapsulates a local minimum of the second derivative by its nearest local maxima in both directions.

Definition 3 (Peak Triplet Score). *The score* $s(p)$ *of a peak triplet* $p = \{l, m, r\}$ *is defined as*

$$s(p) = \sum_{i=l}^{r} z_i, \quad \textit{with } z_i = \begin{cases} -\frac{1}{2} y_i'', & \textit{for} \quad y_i'' < 0 \ \wedge \ \big(i = l \vee i = r\big) \\ -y_i'', & \textit{for} \quad y_i'' < 0 \\ 0 & \textit{else.} \end{cases} \tag{3}$$

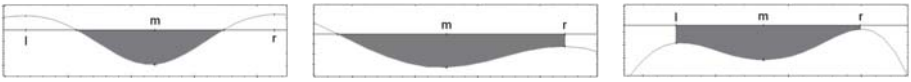


Fig. 1. The second derivative of example peak triplets $\{l, m, r\}$ and the considered area (colored in grey). The horizontal line denotes the zero line.

The score of a peak triplet expresses its likeliness of representing a "real" peak. Negative values of the second derivative are considered exclusively in order to focus on curvatures in clockwise direction only. Fig. 1 shows example peak triplets and the area under consideration.

In extension to [7], we additionally introduce the concept of triplet significance in units of σ as follows:

Definition 4 (Peak Triplet Significance θ). *With Q denoting the set of peak triplets found in blank signal, the significance θ of a peak triplet p with score $s(p)$ is defined as*

$$s(p) = \bar{s} + \theta\,\sigma \quad \Leftrightarrow \quad \theta = \frac{s(p) - \bar{s}}{\sigma}, \tag{4}$$

$$\textit{with} \quad \bar{s} = \frac{1}{|Q|}\sum_{i \in Q} s(p_i) \quad \textit{and} \quad \sigma = \sqrt{\frac{1}{|Q|}\sum_{i \in Q}\big(s(p_i) - \bar{s}\big)^2}$$

as the mean and standard deviation score of all peak triplets in Q, respectively. Further,we call a triplet p to be accepted, if for a given significance threshold δ it holds $\theta \geq \delta$. The set of accepted triplets is in the remainder of this paper denoted as A.

For the purpose of noise reduction, we consider the *mean filter*, which is also known as *moving average* or *box filter* [8]. In particular, we consider repeated averaging with a filter window of length 3, given as

$$smd(y_i) = \frac{1}{3}\sum_{k=i-1}^{i+1} y_{k'}, \quad \text{with } k' = \begin{cases} |k| + 2, & \text{for } k \leq 0, \\ 2n - k, & \text{for } k > n, \\ k, & \text{else.} \end{cases} \tag{5}$$

With the observation that the sum between two consecutive iteration steps only differs by the values y_{i-a} and y_{i+a}, respectively, mean-filtering of n datapoints can be executed in time $O(n + a)$.

Repeated averaging has already been considered decades ago, and is valued for its computational efficiency and easy-to-implement characteristics [9]. In the following we will briefly describe the effects of filtering, and refer for example to [10] for more detailed information.

Repeatedly smoothing by (5) steadily changes the coefficients of the filter window. For example, the first execution of replaces each value y_i by $smd(y_i) = \frac{1}{3}(y_{i-1} + y_i + y_{i+1})$, after the second execution the value at index i is given

as $smd(smd(y_i)) = \frac{1}{9}(y_{i-2} + 2\,y_{i-1} + 3\,y_i + 2\,y_{i+1} + y_{i+2})$, and analogously, $smd(smd(smd(y_i))) = \frac{1}{27}(y_{i-3} + 3\,y_{i-2} + 6\,y_{i-1} + 7\,y_i + 6\,y_{i+1} + 3\,y_{i+2} + y_{i+3})$. In fact, the weights after b times executing (5) equal the *trinomial coefficients*[1] obtained after expansion of $(1 + t + t^2)^b$ (see e.g. [13]). As a consequence of the *central limit theorem*, the coefficients discretely approximate the probability density function of the Gaussian distribution [14]. Mean filtering thus corresponds to a discrete version of convolving the data with a *Gaussian kernel*.

A heuristic non-parametric approach for automated Gaussian smoothing has been proposed in [15], based on the change in the number and the maximal pairwise distance of adjacent local maxima. The shown results though indicate that the proposed method seems to be rather unsuitable for datasets maintaining a higher diversity of peak positions as those considered by the paper.

In [16], an automated smoothing approach is proposed based on the lag-one autocorrelation coefficient ρ_1, given as

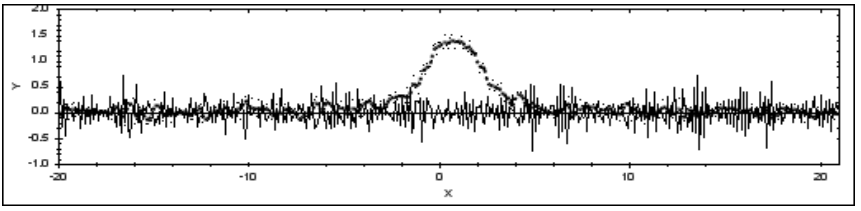
$$\rho_1 = 1 - \frac{1}{2}\frac{\sum_{i=2}(y_i - y_{i-1})^2}{\sum_{i=1} y_i^2}\frac{n}{n-1}, \tag{6}$$

where n stands for the number of real-valued datapoints y_i. In rough summary, the smoothing degree is repeatedly increased, until the lag-one autocorrelation coefficient of the residual, namely the observed spectrum subtracted by the smoothed, is closest to the lag-one autocorrelation coefficient of blank signal. As a drawback, the method tends to excessively smoothen the considered spectra. In the remainder of this paper, an automated method for finding a proper smoothing degree is proposed based on changes in the curvature during the smoothing procedure of a given spectrum.

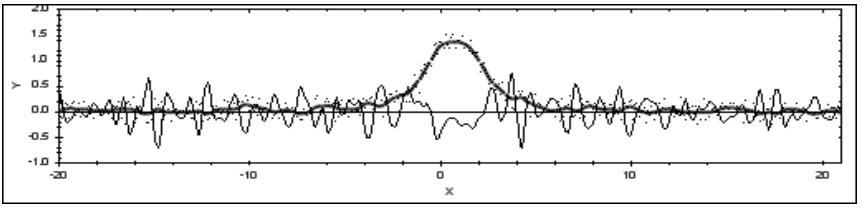
2 Methods

Repeatedly filtering the data by (5) allows to mitigate the impact of noise, but at some point also tends to merge curvatures emerging from distinct Lorentz functions. Figure 2 for example shows the effects of smoothing (thick dotted line) for varying smoothing repeats b of an example spectrum (dotted line) with uniformly distributed noise $U(-0.3, 0.3)$ and significance threshold $\delta = 6.0$. The spectrum is originally given as a sum of two Lorentz functions with parameters $\lambda = A = 1$, and with distance $d = |\omega_2 - \omega_1| = 1.5$ at a resolution $\frac{1}{\Delta x} = 10$. In the beginning ($b = 1$), the second derivative (thin solid line) is highly distorted due to the effects of noise (Fig. 2(a)). Increasing the degree of smoothing by repeatedly executing (5) allows to encapsulate two major clockwise-rotating curvatures of the spectrum as two adjacent peak triplets (Figs. 2(b) - 2(d)). Further smoothing leads to a merge of the two triplets, and only a single peak is observed for $b > 50$

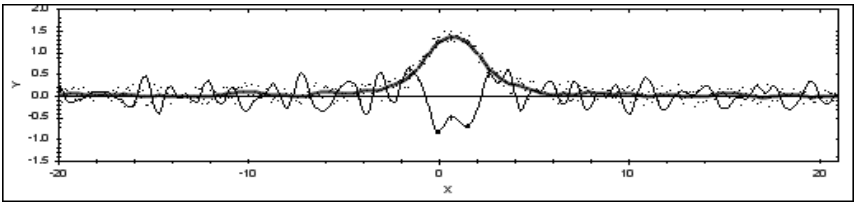
[1] It might be worth mentioning that following [11] no less than Euler found them worthy for a 20-page account [12].



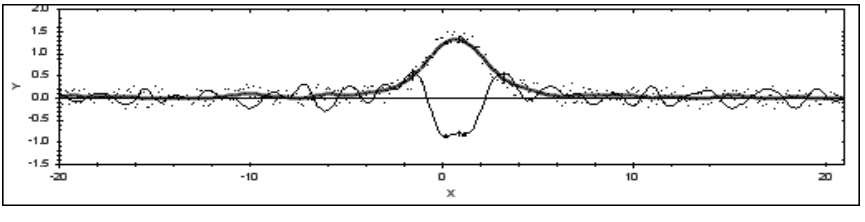
(a) $b = 1$



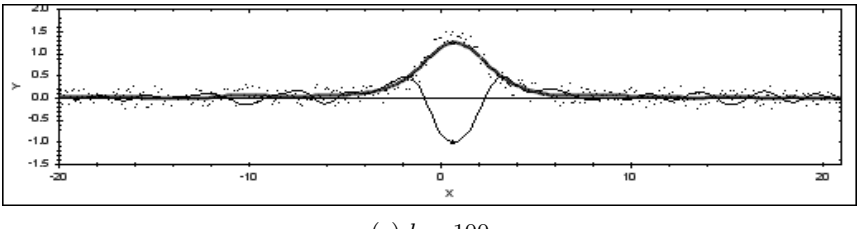
(b) $b = 10$



(c) $b = 25$



(d) $b = 50$



(e) $b = 100$

Fig. 2. Smoothing effects on an example spectrum of two standard Lorentz functions. Shown are the original spectrum (dots) after adding white noise at an amplitude of 0.3, the smoothed spectrum (thick solid line) and the corresponding second derivative (thin solid line). b denotes the number of smoothing steps.

(Fig. 2(e)). The aim thus is to find the number of smoothing repeats, which properly balances the trade-off between noise removal and signal preservation.

For a given significance threshold δ, an intuitive approach for determining a reasonable degree of smoothing is given by maximizing the sum of scores

$$f_{sum}(A) = \sum_{j \in |A|} s(p_j) \tag{7}$$

for the set of accepted triplets A (see Def. 4). However, considering that the score of each triplet itself is given as a sum of second derivative values already, we may presume that merging of two *adjacent* triplets p_1 and p_2 has almost no effect on f_{sum} (compare Def. 3), written as

$$s(p_1) + s(p_2) \approx s(p_{1,2}) \tag{8}$$

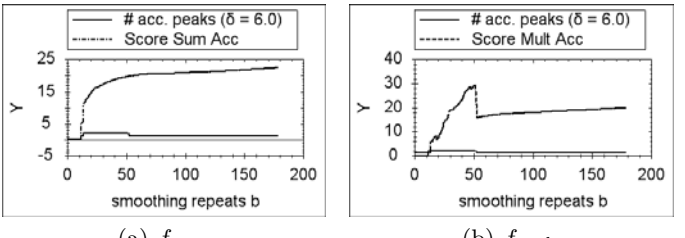
with $p_{1,2}$ denoting the triplet received after merging of p_1 and p_2. To circumvent this problem, we may consider

$$s(p_1) \geq 2 \ \wedge \ s(p_2) \geq 2 \quad \Rightarrow \quad s(p_1)\, s(p_2) \ \geq \ s(p_1) + s(p_2) \approx s(p_{1,2}), \tag{9}$$

and the degree of smoothing can then be found by maximizing

$$f_{mult}(A) = \log\left(\prod_{i=1}^{|A|} \big(2 + s(p_i) - (\bar{s} + \delta\sigma)\big)\right) = \sum_{i=1}^{|A|} \log\big(2 + \Theta_i - \delta\big). \tag{10}$$

δ denotes a predefined significance threshold, and Θ_i denotes the significance of peak triplet p_i (see Def. 4). f_{mult} now allows to prevent significant peak triplets p_i from being merged together, since by definition it holds $\Theta_i \geq \delta$ for all triplets $p_i \in A$. Figure 3 shows the corresponding selection scores f_{sum} and f_{mult} for the merging scenario of Fig. 2. In agreement with the assumption from above, the sum of scores of accepted triplets (f_{sum}, top line in fig. 3(a)) keeps increasing even after the two triplets have merged to a single one ($b > 50$). In contrast, an essential decrease can be observed for f_{mult} (top line in fig. 3(b)). A pseudo-code



(a) f_{sum} (b) f_{mult}

Fig. 3. The respective selection scores f_{sum} and f_{mult} (top lines) of Fig. 2 in comparison with the respective number of selected peaks (bottom line) for increasing smoothing repeats b

Algorithm 1.

Input: Spectrum $\mathbf{a}$, blank signal $\mathbf{a}_{blank}$, significance threshold δ
Output: List of accepted peak triplets A^*

1: $best \leftarrow 0; last \leftarrow \infty; A^* \leftarrow \emptyset;$
2: **while** ($|Q| > 1$ and $|\rho_{1,blank} - \rho_{1,res}| \leq last$) **do**
3: Find set of triplets Q out of blank signal $\mathbf{a}_{blank}$;
4: $last \leftarrow |\rho_{1,blank} - \rho_{1,res}|$;
5: Find accepted peak triplets A on $\mathbf{a}$, given δ;
6: **if** ($f_{mult}(A) \geq best$) **then**
7: $best \leftarrow f_{mult}(A)$; $A^* \leftarrow A$;
8: **end if**
9: Apply (5) on all values in $\mathbf{a}$ and $\mathbf{a}_{blank}$;
10: **end while**
11: **return** A^*;

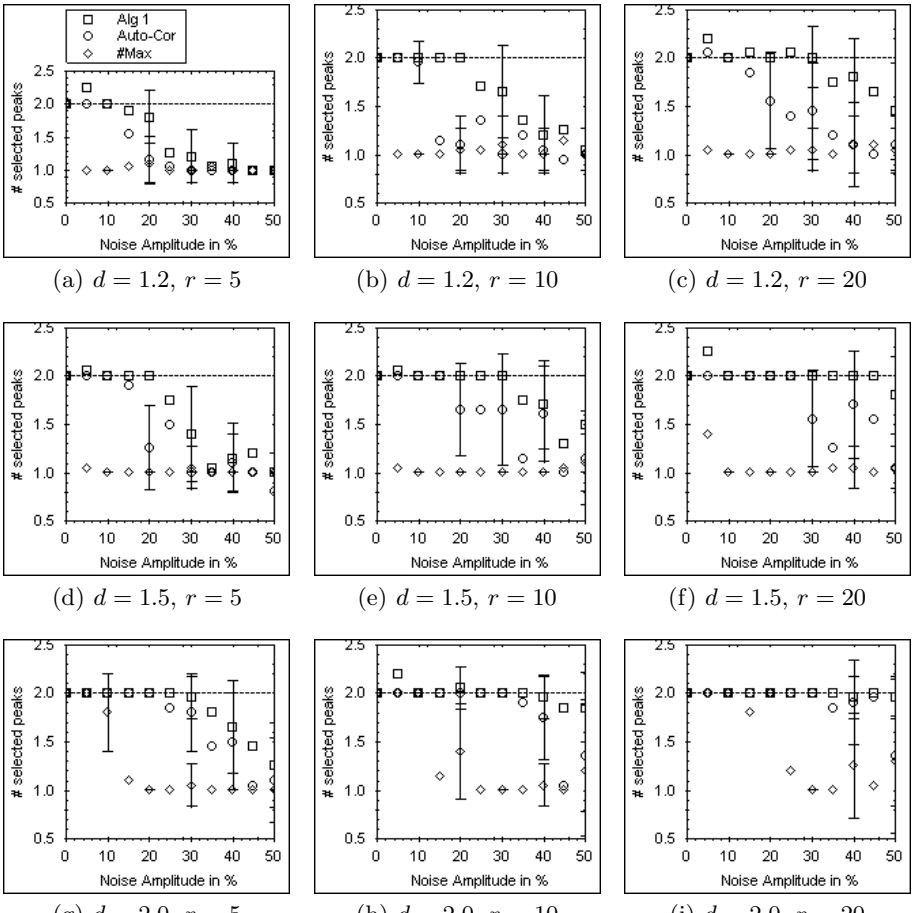
representation for an automated smoothing approach based on f_{mult} is given by Algorithm 1.

In summary, a given spectrum $\mathbf{a}$ is repeatedly smoothed, as long as blank triplets exist, and as long as the autocorrelation coefficient of the residual, $p_{1,res}$, approaches that of blank signal, $p_{1,blank}$. Note that the chosen degree of smoothing by Algorithm 1 is less or equal to that of [16]. With b denoting the number of smoothing repeats needed for the execution of lines 3-11, with n denoting the number of datapoints in $\mathbf{a}$, and with m denoting the number of datapoints in the blank signal, Algorithm 1 has a total worst-case runtime of $O\big(\, b(n + m + |A| + |Q|) \,\big)$.

3 Results

In this section, initial results of Algorithm 1 are presented, based on simulated spectra containing two *standard* Lorentz functions (2) with width and scale parameters $A = \lambda = 1$. With $d = |\omega_2 - \omega_1|$ denoting the distance between the two peaks, and with $r = \frac{1}{\Delta x}$ denoting the resolution of the spectrum, three peak distances $d \in \{1.2, 1.5, 2.0\}$ and three different spectrum resolutions $r \in \{5, 10, 20\}$ are considered, resulting in a total number of nine different smoothing scenarios. Each scenario is sampled 20 times, and uniformly distributed noise $U(-\frac{v}{100}, \frac{v}{100})$ is added to each datapoint, with noise amplitudes v in the range $0 \leq v \leq 50$. Triplets are found within the range $[\omega_1 - 5, \omega_2 + 5]$ out of a total spectral range of $[\omega_1 - 20, \omega_2 + 20]$. All evaluation runs are based on a significance threshold $\delta = 6.0$.

Fig. 4 shows the performance of Algorithm 1 on average out of 20 runs for each of the considered scenarios. The number of accepted peak triplets $|A^*|$ with maximal selection score f_{mult} is denoted as Alg 1, and shown as squares. In addition, the number of accepted triplets $|A|$ found after the last execution of line 5 in Algorithm 1 is denoted as Auto-Cor, and shown as circles. In a sense, these results represent the outcome of the lag-one autocorrelation approach of [16], and thus can be seen as baseline results, compared to which the impact



(a) $d = 1.2$, $r = 5$ (b) $d = 1.2$, $r = 10$ (c) $d = 1.2$, $r = 20$

(d) $d = 1.5$, $r = 5$ (e) $d = 1.5$, $r = 10$ (f) $d = 1.5$, $r = 20$

(g) $d = 2.0$, $r = 5$ (h) $d = 2.0$, $r = 10$ (i) $d = 2.0$, $r = 20$

Fig. 4. Number of accepted peak triplets after the execution of Algorithm 1 for different scenarios and varying noise amplitudes. The length of the error bars equals two times the standard deviation out of 20 runs (see the text for more details).

of f_{mult} can be determined. In addition, for the degree of smoothing chosen by Algorithm 1, the number of local maxima with a maximal value higher than the spectrum average value is denoted as #Max, and shown as diamonds.

The figures generally show that it can be highly beneficial to identify peaks as curvatures of the spectrum rather than as maxima, since both methods *Auto-Cor* and Algorithm 1 leave the spectra in most cases with exactly one maximum after smoothing. Furthermore, it can also generally be observed that an increase in the peak distance d (from top to bottom in all columns) and also an increase in the resolution r (from left to right in all rows) have both a beneficial impact, i.e. the maximal noise amplitude, for which two peaks can still be identified, increases for increasing d or r or both.

An interesting result is given by the fact that Algorithm 1 is capable of identifying both peaks on average for even higher noise amplitudes and even smaller peak distances than *Auto-Cor* in all considered scenarios. Thus, at least for the considered datasets, maximizing (10) apparently comes to a better compromise between noise reduction and signal preservation than minimizing the autocorrelation distance only. The software implementing Algorithm 1 is written in C-Sharp, and is freely available at `http://ls1-www.cs.tu-dortmund.de/~koh`.

4 Summary and Conclusions

In this paper, we have initially proposed an automated smoothing and peak detection method based on repeated averaging and balancing the trade-off between noise reduction and preservation of significant curvatures. By providing a template region of blank signal, only a single peak significance threshold needs to be specified in advance. The time and space complexity of the corresponding algorithm are considerably low, allowing the proposed approach to be considered also for the analysis of high-throughput data.

Empirical studies on simulated datasets show that the number of peaks is found correctly for uniformly distributed noise amplitudes of up to 50% of a peak's maximal value. Furthermore, next to obvious improvements due to increasing distances between the peaks, significant improvements are also observed for increasing the resolution. This is insofar interesting, as typically the observed distances between two adjacent peaks are often substrate specific and inherently dependent on particular properties of a given sample. In contrast, the resolution is commonly limited due to technological boundaries, which will presumably be further improved only as a matter of time.

Of course, observing a correct number does not necessarily imply correct positions of the respective peaks. In addition, it remains to be seen whether similar observations can be made in further studies of more heterogeneous datasets. Further investigations regarding for example different colors of noise, different filter windows to begin with, more heterogeneous datasets concerning varying width and scale parameters of the peak functions, and also regarding the impact on subsequent steps in spectral data analysis, e.g. spectrum modeling or peak alignment, need to be carried out and will be one of the author's ongoing research interests in the near future.

Acknowledgments

This work was partially funded by the European Project Second Generation Locator for Urban Search and Rescue Operations (SGL USaR).

References

1. Ernst, R.R., Bodenhausen, G., Wokaun, A.: Principles of Nuclear Magnetic Resonance in One and Two Dimensions. Oxford University Press, Oxford (1987)
2. Hoch, J.C., Stern, A.S.: NMR data processing. Wiley-Liss., Chichester (1996)

3. Nguyen, N., Huang, H., Oraintara, S., Vo, A.: Peak Detection in Mass Spectrometry by Gabor Filters and Envelope Analysis. Journal of Bioinformatics and Computational Biology 7(3), 547–569 (2009)
4. Nelson, S.J., Brown, T.R.: Peak detection and quantification in nmr spectra using the piqable algorithm. Bul. Magn. Res. 11(3/4), 290–293 (1989)
5. Marshall, R.J.: The determination of peaks in biological waveforms. Computers and Biomedical Research 19(4), 319–329 (1986)
6. Dijkstra, M., Roelofsen, H., Vonk, R.J., Jansen, R.C.: Peak quantification in surface-enhanced laser desorption/inization by using mixture models. Proteomics 6, 5106–5116 (2006)
7. Koh, H.W., Maddula, S., Lambert, J., Hergenröder, R., Hildebrand, L.: An approach to automated frequency-domain feature extraction in nuclear magnetic resonance spectroscopy. Journal of Magnetic Resonance 201(2), 146–156 (2009)
8. Gonzales, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice-Hall, Englewood Cliffs (2002)
9. Faes, T.J.C., Govaerts, H.G., Tenvoorde, B.J., Rompelman, O.: Frequency synthesis of digital filters based on repeatedly applied unweighed moving average operations. Med. & Biol. Eng. & Comput. 32, 698–701 (1994)
10. Cai, L.D.: Some notes on repeated averaging smoothing. In: Pattern Recognition, pp. 597–605 (1988)
11. Andrews, G.E.: Euler's "exemplum memorabile inductionis fallacis" and q-trinomial coefficients. Journal of the American Mathematical Society 3(3), 653–669 (1990)
12. Euler, L.: Observations analyticae. Novi Commentarii Academiae Scientarum Petropolitanae 11, 124–143 (1765); Also in Volume 15 of Opera Omnia, Series 1, Teubner, pp. 50–69
13. Merlini, D., Sprugnoli, R., Verri, M.C.: Some statistics on dyck paths. Journal of Statistical Planning and Inference 101(1-2), 211–227 (2002)
14. Rice, J.: Mathematical Statistics and Data Analysis, 2nd edn. Duxbury Press (1995)
15. Lin, H.C., Wang, L.L., Yang, S.N.: Automatic determination of the spread parameter in gaussian smoothing. Pattern Recognition Letters 17(12), 1247–1252 (1996)
16. Vivo-Truyols, G., Schoenmakers, P.J.: Automatic selection of optimal savitzky-golay smoothing. Analytical Chemistry 78(13), 4598–4608 (2006)

Uncertainty Interval Expression of Measurement: Possibility Maximum Specificity versus Probability Maximum Entropy Principles

Gilles Mauris

LISTIC, University of Savoie, Domaine Universitaire,
BP 80 439, 74 944 Annecy Le Vieux, France
gilles.mauris@univ-savoie.fr

Abstract. This paper pursues previous studies concerning the foundations of a possibility/fuzzy expression of measurement uncertainty. Indeed a possibility distribution can be identified to a family of probability distributions whose dispersion intervals are included in the level cuts of the possibility distribution. The fuzzy inclusion ordering, dubbed specificity ordering, constitutes the basis of a maximal specificity principle for uncertainty expression. We argue that the latter is sounder than the maximal entropy principle to deal with cases of partial or incomplete information, at least in a measurement context. The two approaches are compared on philosophical issues and on some common practical cases.

Keywords: measurement uncertainty, possibility theory, maximum entropy principle, maximum specificity principle, uncertainty intervals.

1 Introduction

Uncertainty interval statement about a quantity under measurement goes back to a very long way [1]. It has taken two main distinct forms [2]. The first form is related to the concept of confidence interval that summarizes the information about a fixed but unknown parameter by an interval (with random extremities) containing the parameter with a specified confidence (between 0 and 100%). The second form is related to the concept of coverage interval (having fixed extremities) defined around the measurement result, which will contain a specified part (between 0 and 100%) of the distribution of the measured values. Commonly, in both forms, a probability level of 95% is considered [3], but there is no compelling scientific reason for this choice. Note that often confusions and difficulties occur with the different kinds of uncertainty intervals that may be used (this is partly caused because their names are formed from a limited number of words such as interval, confidence, coverage). But the different uncertainty intervals are always deduced from dispersion intervals of a random variable X according to an inference method (frequentist, Bayesian, …). In addition, to build uncertainty intervals requires the knowledge of the probability distribution of X. In a lot of situations, the assumption of Gaussian distribution is quite satisfactory (due to the Central Limit Theorem), but it is not always the case, especially when no

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 386–395, 2010.

prior information or a few measurement are available. It is why a lot of authors recommend to use a maximum entropy approach [3], introduced by Jaynes [4], for determining a single probability distribution, the one which has the minimum information content (in the Shannon sense) according to the available partial information, e.g. specified central moments of different orders or specified percentiles.

In previous works, we have proposed a possibility-theory-based alternative approach to the probability ones to express measurement uncertainty [5][6]. This approach considers within one single possibility distribution the whole sets of uncertainty intervals of all the confidence levels between 0 and 1. In this paper, we discuss a maximum specificity principle to build this possibility distribution that discards the maximum entropy principle while satisfying its underlying philosophical foundations, and those of the preceding indifference or insufficient reason principles [1]. The idea is to consider each value of the considered variable as possible while it is not eliminated by the available information which suggests to maximize the probability degrees instead of the Shannon entropy. Note that there exist other different representations than a possibility distribution to deal with partial information about the probability distribution. Among them are probability intervals [7], Ferson's p-boxes [8], Neumaier's clouds [9], but all use a pair of distribution. But the paper is limited to the comparison of the proposed possibility approach to the conventional maximum entropy, and also, it does not consider the lack of missing information under maximum entropy by intervals of second order uncertainty [10]. In section 2, we discuss the philosophical issues of representing partial or incomplete knowledge about measurement distribution either by the probability maximum entropy principle or by the possibility maximum specificity principle. The section 3 presents some practical results obtained by the two approaches on some common cases reflecting different amount of a priori information. Some concluding remarks point out the interest of the possibility approach and some future developments.

2 Philosophical Issues

2.1 Historical Perspectives

The problem of assigning numerical values to probabilities based upon lack of information dates back to the origin of probability [1]. Jakob Bernouilli in Ars Conjectandi in 1713 states that if we are ignorant of the ways an event can occur (and therefore have no reason to believe that one way will occur preferentially compared to another), the event will occur equally likely in any way. But Jakob Bernouilli also argued that this principle of indifference could be used almost exclusively in games of chance, and for others contexts, e.g. judging the risk of death, he advocated another method based on his famous law of large numbers. Laplace in 1813 expressed the indifference principle in the form: equipossible alternatives may be accorded equal probabilities if nothing more is known about the underlying probability distribution. He turned this rule into the cornerstone of a comprehensive theory of probability by stating that this indifference principle relies on symmetry in our belief or judgement in order to obtain numerical values for probabilities. The underlying motivation is, of course, that in this view, the term probability should be understood as a degree of belief and hence the

uniform probability represents exactly the situation where all possible states are equally credible. It was the economist Keynes who renamed the principle of indifference into the principle of insufficient reason in contrast to Leibniz's principle of sufficient reason, and he stressed that it is valid only in the rather special case when there is no knowledge indicating unequal probabilities. The principle of sufficient reason basically says that every fact has a sufficient reason for why it is the way it is and not otherwise.

Since the middle of the 19th century, the principle of insufficient reason had plenty of detractors [11]. One of the first objections is that one cannot derive empirical predictions from a lack of knowledge. Mere ignorance is no ground for any inference whatsoever: *ex nihilo nihil* (nothing comes from nothing, it cannot be that because we are ignorant of the matter we know something about it). Another objection from Bertrand is that when one choose different parameterization for a variable x ranging over a continuum, a probability density that is uniform over x becomes non uniform under a non linear parameter transformation, e.g. $y=x^2$. This conflicts with the intuition that if we are ignorant of x we are also ignorant of y. The last famous objection we mention had been made by Reichenbach in 1935 who claimed that the principle was circular because the only sensible meaning one can give to the word "equipossible" in the Laplace definition is in fact "equiprobable". The principle was also discredited in the first part of the 20th by Fisher, Von Mises, Neuman and Pearson. The revival is due to Jaynes who introduced in 1957 the maximum entropy principle [4] as an extension of the principle of insufficient reason.

2.2 Maximum Entropy Principle

When faced with partial or incomplete probability knowledge such as moments, range, percentile constraints, leading in fact to a family $\mathcal{P}$ of probability distributions and not to a single one, the maximum entropy principle provides a way to select the one having the minimum information content in the Shannon sense. The formulation for the probability density f of a continuous variable having prescribed moments and percentile constraints is [12][13]:

$$f_{\max ent}(x) = \arg\max_{f\in\mathcal{P}}\left(-\int_a^b f(x)\ln(f(x))dx\right)$$

$$\text{subject to } \int_a^b h_i(x)f(x)dx = \mu_i, i=1,...,n \;;\; \int_a^b f(x)dx = 1 \text{ and } f(x)\geq 0$$

where $[a,b]$ is the domain of the variable, $h_i(x)$ is x raised to a certain power for moment constraints or an indicator function for percentile constraints, and the μ_i's are the given values of moments or percentiles of the distribution. The solution of this optimization problem can be obtained by using the method of Lagrange multipliers and leads to the following density:

$$f_{\max ent}(x) = e^{-a_0-1-a_1h_1(x)-a_2h_2(x)-...-ah_n(x)}.$$

When only the normalization and non-negativity density constraints are available, the maximum entropy density is uniform over a bounded domain:

$$f_{\max ent}(x) = e^{-a_0-1} = \frac{1}{b-a}, a \le x \le b .$$

If the first moment μ is available, the maximum entropy density on the non-negative domain is:

$$f_{\max ent}(x) = e^{-a_0-1-a_1 x} = \frac{1}{\mu} e^{-\frac{x}{\mu}}, x \ge 0 .$$

If the first and second moments μ and σ^2 are available over the interval $[-\infty, \infty]$, the maximum entropy density is a Gaussian one:

$$f_{\max ent}(x) = e^{-a_0-1-a_1 x-a_2 x^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty .$$

When percentiles of the distribution are available, the maximum entropy density is a staircase probability density function that satisfies the percentile constraints. Moreover, it is uniform over each interval, and is integrated into a piecewise linear cumulative probability distribution that has the shape of a taut string. But note that the existence a maximum entropy distribution in not guarenteed, even in some simple situations, e.g. for a specified mean μ with a distribution defined on $[-\infty, +\infty]$, and also for the constraints $\mu = 0, \sigma = 1$ and its third moment $EX^3 = 1$ [14].

2.3 Maximum Specificity Principle

A possibility distribution π is a mapping from a set (in the measurement context the set of reals *R*) to the unit interval such that $\pi(x) = 1$ for some *x* belonging to *R* [15]. It generates a set function Π called a possibility measure, and such that $\forall A \subset R, \Pi(A) = \sup_{x \in A} \pi(x)$ and also a necessity measure *N*. The degree of necessity (certainty) of an event A is computed from the degree of possibility of the complementary event $\bar{A}$: $\forall A \subset R, N(A) = 1 - \Pi(\bar{A}) = \inf_{x \notin A}(1 - \pi(x))$.

***Definition*:** a possibility distribution π_1 is called more specific (i.e. more thinner in a broad sense) than π_2 as soon as $\forall x \in R, \pi_1(x) \le \pi_2(x)$ (fuzzy set inclusion).

The more specific π, the more informative it is. If $\pi(x) = 1$ for some *x* and $\pi(y) = 0$ for all $y \ne x$, then π is totally specific (fully precise and certain knowledge), if $\pi(x) = 1$ for all *x* then π is totally non specific (complete ignorance). In fact, a numerical degree of possibility can be viewed as an upper bound to a probability degree [16]. Namely, with every possibility distribution π one can associate a non-empty family of probability measures dominated by the possibility measure: $\mathcal{P}(\pi) = \{P, \forall A \subset R, P(A) \le \Pi(A)\}$. This provides a bridge between probability and

possibility, and there is also a bridge with interval calculus. Indeed, a unimodal numerical possibility distribution may also be viewed as a nested set of coverage intervals, which are the α cuts of $\pi:[\underline{x}_\alpha, \overline{x}_\alpha] = \{x, \pi(x) \geq \alpha\}$. Obviously, the coverage intervals built around the same point x_0 are nested. It has been proven in [6] that stacking the coverage intervals of a probability distribution *F* on top of one another leads to a possibility distribution (denoted π^{x_0} having x_0 as modal value). In fact, in this way, the α-cuts of π^{x_0} are identified with the coverage interval I_β^* of probability level $\beta = 1-\alpha$ around the nominal value x_0.

For a symmetric unimodal (with mode m) probability density, the equivalent possibility distribution is [17]:

$$\pi_F{}^m(m-t) = \pi_F{}^m(m+t) = P(|X-m| \geq t) = 2(1-F(t))$$

In fact, the possibility representations of probability families $\mathcal{P}$ induced by incomplete probabilistic data is clearly related to a bound for the probability $P(|X-m| \geq x)$ when the knowledge of the probability density f associated to the random variable *X* is not available:

$$\pi(x) = \max_{P \in \mathcal{P}} P(|X-m| \geq x)$$

$$\text{subject to } \int_a^b h_i(x) f(x) dx = \mu_i, i = 1, ..., n \; ; \; \int_a^b f(x) dx = 1 \text{ and } f(x) \geq 0$$

Therefore the determination of a maximum specific possibility distribution according to the available information is an optimization problem but different from the one given by the maximum entropy approach. The maximum specificity possibility distribution is also clearly related to probability inequalities. For example, if only the mean μ and the standard deviation σ are known, the possibility distribution can be obtained from the Bienaymé-Chebychev inequality [18]:

$$\pi_{BC}(\mu+t) = \pi_{BC}(\mu-t) = \min(1, \frac{\sigma^2}{t^2}) . \quad (1)$$

2.4 Discussion

The maximum entropy principle is subject to the same objections (ignorance equal to equi-repartition, parameterization influence on ignorance, equi-possible/equi-probable confusion) as the insufficient reason principle. Moreover in the context of measurement uncertainty expression, any maximum information uncertainty has to consider the fact that information of X on a continuous scale exists only on a statement like $x-\varepsilon \leq X \leq x+\varepsilon$ and thus is related to the integral of the probability density, i.e. to the cumulative distribution *F*. Consequently quantities that involve the density in other way, like the entropy, should not form the basis of a method of inference. As claimed by Jaynes, the distribution representing maximum uncertainty must be the

distribution that maximizes the "spread"[4]. The latter has to consider an ordering on the measurement domain; it is not the case of the entropy.

The maximum specificity principle does not suffer from the preceding objections. Indeed when one chooses different parameterization for a variable x ranging over a continuum, a possibility distribution which is rectangular over x remains (by appling Zadeh's propagation principle [15]) rectangular under a parameter transformation, i.e. *ex nihilo nihil.* There is no problem of circularity in the definition since possibility and probability are clearly defined in a separate axiomatic way. The possibility distribution associated to a considered probability distribution is clearly related to the cumulative distribution F, and has thus a measurement meaning. Therefore, the possibility theory offers an interesting alternative approach to the treatment of incomplete knowledge. It proposes a more comprehensive way to express how all information available can properly be taken into account by considering a family of probability distributions instead of a single one. It replaces the *ceteris paribus* probability approach by a *ceteris incognitis* possibility approach using no more information than is available. Indeed, imposing the maximum entropy consists to add information that is not present. Thus the possibility representation is the right maximally noncommittal distribution with regard to missing shape information. In fact, we agree with the objectives of the maximum entropy principle, but we differ in the definition of the information content measure: the lengths of the uncertainty intervals instead of the Shannon entropy. As both approaches provide uncertainty intervals, these intervals can be the basis of practical comparisons between the two approaches as presented in the next section.

3 Some Common Practical Cases of Information Shortage

As discussed above, when the probability density is unknown, some parameters are nevertheless known, e.g. the support, the mode, the mean, the standard deviation. But, knowing one or many of such parameters does not specify the probability density uniquely (it defines a family of probability distributions). In this section we compare the probability maximum entropy and possibility maximum specificity principles from the respective coverage intervals they provide in some common cases.

3.1 Case of Finite Support Distributions

Only the support known

When only the support is known, the maximum entropy probability density is the uniform density and the associated possibility distribution is the uniform possibility distribution. Therefore the two approaches lead to the same uncertainty interval expressions.

Mode known

If the variable is known to be unimodal, then the maximum specific possibility distribution is a triangular possibility distribution with the mode as vertex [19] (see figure 1). In particular if the variable is symmetric then the mode is the middle of the support and the maximal specific possibility distribution is the often used triangular symmetric possibility

distribution [6]. The maximum entropy distribution approach leads also to the same symmetric triangular possibility distribution but only if the mean is imposed to be the middle of the support (it is the case if the symmetry is known). If the mode is known but not the mean, the maximum entropy probability distribution is the uniform distribution. Nevertheless, it is justified to consider the mode for building the uncertainty intervals. Therefore the equivalent possibility distribution is a triangular possibility distribution. Note that the rectangular possibility distribution is the envelope of all the asymmetric triangular possibility distributions with modes ranging along the support. Note also that the knowledge of the mode increases the specificity of the possibility distribution. It is not straightforward to introduce the unimodality constraints in the maximum entropy approach but a way do to it and a few examples are proposed in [20].

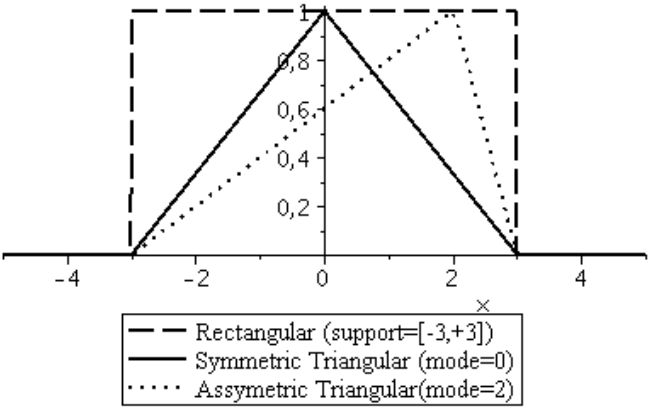


Fig. 1. Possibility distributions from the two principles in the finite support case

3.2 Case of Infinite Support Symmetric Unimodal Distributions

It is the most widely encountered case in the measurement context. If nothing more is specified concerning the distribution, the associated possibility distribution is the totally non specific; the maximal entropy distribution does not exist and the use of an improper probability distribution equals to a constant on the whole universe is recommended [21]. Actually, this improper probability distribution turns out to be a possibility distribution.

Mean and standard deviation known

When the mean μ and the standard deviation σ are known the maximum entropy probability density is a Gaussian density. Note that though the distribution is unimodal, the unimodality property has not been declared in the constraints. Thus adding this important knowledge, does not modify the maximum entropy distribution. Thus the uncertainty is not reduced by the unimodality knowledge contrary to the maximum specificity approach: the Gauss-Winckler inequality [22] leads to a possibility distribution (π_{GW} in eq. 4) rather more specific than the Bienaymé-Chebychev possibility distribution (see figure 2):

$$\pi_{GW}(\mu+t)=\pi_{GW}(\mu-t)=1-\frac{t}{\sqrt{3}\sigma} \quad if \quad t\le\frac{2}{\sqrt{3}}\sigma$$
$$\pi_{GW}(\mu+t)=\pi_{GW}(\mu-t)=\frac{4\sigma^2}{9t^2} \quad if \quad t\ge\frac{2}{\sqrt{3}}\sigma \tag{2}$$

The figure 2 shows the uncertainty intervals obtained from the maximum entropy and maximum specificity principles. The possibility distribution obtained from the Gauss-Winckler includes the one obtained from the Gauss distribution, but they are not very different for probability levels less to 90%.

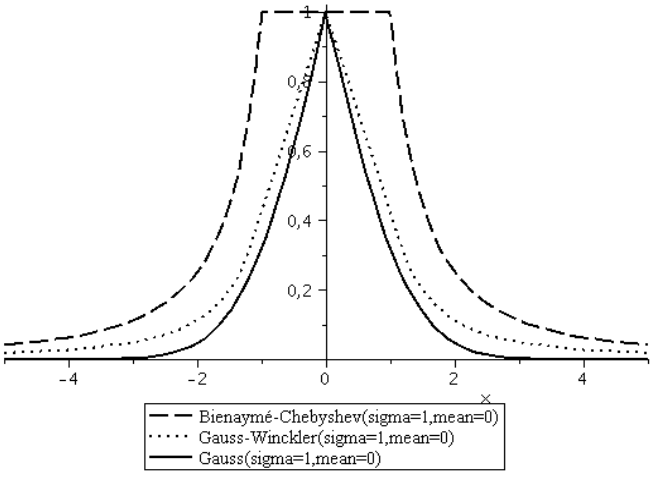


Fig. 2. Possibility distributions from the two principles in the infinite support case

Let us remark that it is possible to deduce the expression of the extremal continuous probability distribution in the specificity order sense, i.e. the one giving the minimum specificity possibility distribution. We obtain first the cumulative distribution by dividing by two the increasing part of the possibility distribution, and reversing and dividing by two the decreasing part. Then we compute the associated density by taking the derivative, thus we obtain:

$$f_{GW}(\mu+t)=f_{GW}(\mu-t)=\frac{1}{2\sqrt{3}\sigma} \quad if \quad t\le\frac{2}{\sqrt{3}}\sigma$$
$$f_{GW}(\mu+t)=f_{GW}(\mu-t)=\frac{4\sigma^2}{9t^3} \quad if \quad t\ge\frac{2}{\sqrt{3}}\sigma$$

Note that this extremal distribution has an infinite standard deviation and thus does not belong to the family from which it is the specificity bound. In the same line, for the non unimodal case, using equation (1), we obtain the following bimodal density:

$$f_{BT}(\mu - t) = f_{BT}(\mu + t) = \frac{\sigma^2}{t^3} \quad if \quad t \geq \sigma$$
$$f_{BT}(\mu + t) = fBT(\mu - t) = 0 \quad if \quad 0 \leq t \leq \sigma$$

These densities can be useful for the inverse transformation from possibility to probability and could be compared to the Smet pignistic transformation [23] developed for a subjective setting.

4 Conclusion

The paper contributions have shown that the probability maximum entropy principle can be replaced by an arguably better (in the sense less informative) possibility maximum specificity principle, at least in a measurement context. Indeed a possibility distribution provides an interesting way to represent different amounts of knowledge about measurement distribution, especially for partial probability information, e.g. a support, a mode, and/or a standard deviation constraint. The comparison of uncertainty intervals resulting from the two uncertainty principles shows that the practical results are quite close in common cases. But the maximum entropy distribution presents some flaws: it does not lead to the maximum uncertainty expression, i.e. the uncertainty intervals are too short (because information that is not available is implicitly added), or it may fail to exist, and the uncertainty can increase with the amount of knowledge. In contrary, the maximum specific possibility distribution always exists, and the more available the information, the more specific it is. In particular the maximum specificity approach makes a good use of the unimodality property that is not the case of the maximum entropy approach. Further developments will deal with how considering other different forms of knowledge in the possibility representation.

References

1. Stigler, S.: The history of statistics, the measurement of uncertainty before 1900. Belknap Press (1986)
2. Cox, M.G., Harris, P.M.: Software support for metrology best practice guide no 6: Uncertainty evaluation, NPL Report (2006)
3. Guide for the expression of uncertainty in measurement, ISO, 2nd edn. (1995); GUM Supplement 1 Propagation of probability distributions using a Monte Carlo method, JCGM, 90 pages (2008)
4. Jaynes, E.T.: Information theory and statistical mechanics. Physical Review 106, 620–630 (1957)
5. Mauris, G., Lasserre, V., Foulloy, L.: Fuzzy modeling of measurement data acquired from physical sensors. IEEE Trans. on Measurement and Instrumentation 49(6), 1201–1205 (2000)
6. Mauris, G., Lasserre, V., Foulloy, L.: A fuzzy approach for the expression of uncertainty in measurement. Int. Journal of Measurement 29(3), 165–177 (2001)
7. De Campos, L., Huete, J., Moral, S.: Probability intervals: a tool for uncertain reasoning. Int. J. Uncertainty Fuzziness Knowledge-Based System 2, 167–196 (1994)

8. Ferson, S., Ginzburg, L., Kreinovich, V., Myers, D., Sentz, K.: Constructing probability boxes and Dempster–Shafer structures. Tech. Report, Sandia Nat.Lab (2003)
9. Neumaier, A.: Clouds, fuzzy sets and probability intervals. Reliable Computing (10), 249–272 (2004)
10. Röder, W., Reucher, E., Kulmann, F.: Features of the expert-System Shell SPIRIT. Logic Journal of the IGPL 14(3), 483–500 (2006)
11. Uffink, J.: Can the maximum entropy principle be explained as a consistency requirement? Studies in History and Philisophy of Modern Physics 26(3), 223–261 (1995)
12. Jaynes, E.T.: Prior probabilities. IEEE Trans. on System Science and Cybernetics 4, 227–241 (1968)
13. Woeger, W.: Probability assignment to systematic deviations by the principle of maximum entropy. IEEE Trans. On Instrumentation and Measurement 36(2), 655–658 (1987)
14. Cover, T., Thomas, J.: Elements of information theory. Wiley, Chichester (2006)
15. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1(1), 3–28 (1978)
16. Dubois, D., Prade, H.: When upper probabilities are possibility measures. Fuzzy Sets and Systems 49, 65–74 (1992)
17. Dubois, D., Foulloy, L., Mauris, G., Prade, H.: Probability-possibility transformations, triangular fuzzy sets and probabilistic inequalities. Reliable Computing 10(4), 273–297 (2004)
18. Mauris, G.: Fuzzy representation of incomplete knowledge about infinite support probability distributions in a measurement context. In: IEEE International Conference on Fuzzy Systems FUZZ-IEEE 2007, CD-ROM, London, UK, July 2007, 6 pages (2007)
19. Baudrit, C., Dubois, D.: Practical representation of incomplete probabilistic information. Computational Statistics & Data Analysis 51(1), 86–108 (2006)
20. Brockett, P.L., Charnes, A., Paick, K.: Information theoretic approach to unimodal density estimation. IEEE Trans. on Information Theory 41(3), 824–829 (1995)
21. Seidenfeld, T.: Entropy and uncertainty. In: Foundations of Statistical Inference, pp. 259–287. D. Reidel Pub. Comp. (1987)
22. Savage, R.: Probability inequality of the Tchebycheff type. Journal of the Research of the NBS 65(3), 211–222 (1961)
23. Smets, P.: Constructing the pignistic probability function in a context of uncertainty Uncertainty. In: Henrion, M., et al. (eds.) Artificial Intelligence, vol. 5, pp. 29–39. North-Holland, Amsterdam (1990)

Lazy Induction of Descriptions Using Two Fuzzy Versions of the Rand Index

Eva Armengol and Àngel García-Cerdaña

IIIA - Artificial Intelligence Research Institute,
CSIC - Spanish Council for Scientific Research,
Campus UAB, 08193 Bellaterra, Catalonia, Spain
{eva,angel}@iiia.csic.es

Abstract. In this paper we introduce an extension of the lazy learning method called *Lazy Induction of Descriptions* (LID). This new version is able to deal with fuzzy cases, i.e., cases described by attributes taking continuous values represented as fuzzy sets. LID classifies new cases based on the relevance of the attributes describing them. This relevance is assessed using a distance measure that compares the correct partition (i.e., the correct classification of cases) with the partitions induced by each one of the attributes. The fuzzy version of LID introduced in this paper uses two fuzzy versions of the Rand index to compare fuzzy partitions: one proposed by Campello and another proposed by Hüllermeier and Rifqi. We experimented with both indexes on data sets from the UCI machine learning repository.

1 Introduction

Case-based reasoning (CBR) is based on the idea that similar problems (cases) have similar solutions. Given a problem to be solved the first step of a CBR method [1] is to retrieve a subset of cases assessed as the most similar to the problem. Depending on the similarity criteria, the subset of retrieved cases will be different and, thus, the solution of the new problem will also be different. Notice that, differently than inductive learning methods (e.g., decision trees), CBR methods are *lazy* in the sense that the problem solving process depends on each new problem. *Lazy Induction of Descriptions* (LID) [3] is a lazy learning method useful for classification tasks. LID retrieves precedents based on the relevance of attributes. This relevance is assessed using a distance measure that compares the correct partition (i.e., the correct classification of cases) with the partitions induced by each one of the attributes.

Although LID is able to deal with relational objects represented as *feature terms* [2], we take here a version of LID that handles objects (cases) represented using *propositional representation*, that is, as a set of pairs attribute-value, where the values are nominal (i.e., they take values in a finite set of values). However, sometimes this representation is not appropriate (for instance to represent people weight, age or some physical measures) being common the necessity to give

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 396–405, 2010.

continuous values to attributes. There are a lot of approaches dealing with attributes taking continuous values. Some of these approaches discretize the continuous values and then they use the usual similarity measures on the discretized values [9,11]. By means of the discretization, continuous values can be handled as nominal. However in this procedure there is a lost of information since the values near to the thresholds of the discretization interval are considered equal but, in fact, they are not. With the goal of reducing such lost of information, other approaches use fuzzy sets to deal with continuous values (see for instance [12]). Previous versions of LID handle cases with attributes having nominal values and, when cases have attributes taking continuous values, they are previously discretized. The distance measure to compare partitions, denoted here by LM, is the one introduced by López de Mántaras in [7]. In this paper we want to analyze the performance of LID when the continuous values of attributes are represented using fuzzy sets. Since the distance LM is not appropriate for this task, it must be replaced by some other measure able to deal with fuzzy partitions.

In this paper we use two fuzzy versions of the Rand index [10]: one proposed by Campello [5], which can compare a fuzzy partition with a crisp one, and another one proposed by Hüllermeier and Rifqi [8], which can compare two fuzzy partitions. In Section 2 we give a brief introduction of LID and the Rand index. In Section 3 we explain the fuzzy version of LID. In Section 4 we show the results of the experiments with fuzzy LID.

2 Lazy Induction of Descriptions

Lazy Induction of Descriptions (LID) is a lazy learning method for classification tasks. LID determines which are the most relevant attributes of a problem and searches in a case base for cases sharing these relevant attributes. The problem is classified when LID finds a set of relevant attributes shared by a subset of cases all of them belonging to the same class. We call *similitude term* the description formed by these relevant features and *discriminatory set* the set of cases satisfying the similitude term.

Given a problem for solving p, the LID algorithm (Fig. 1) initializes D_0 as a description with no attributes, the discriminatory set S_{D_0} as the set of cases satisfying D_0, i.e., all the available cases, and C as the set of solution classes into which the known cases are classified. Let D_i be the current similitude term and S_{D_i} be the set of all the cases satisfying D_i. When the stopping condition of LID is not satisfied, the next step is to select an attribute for specializing D_i. The specialization of D_i is achieved by adding attributes to it. Given a set F of attributes candidate to specialize D_i, the next step of the algorithm is the selection of an attribute $f \in F$. Selecting the most discriminatory attribute in F is heuristically done using a distance (the LM distance in [3]). Such distance is used to compare each partition $\mathcal{P}_f$ induced by an attribute f with the correct partition $\mathcal{P}_c$. The *correct partition* has as many sets as solution classes. Each attribute $f \in F$ induces in the discriminatory set a partition $\mathcal{P}_f$ with as many sets as the number of different values that f takes in the cases. Given a distance

```
Function LID (p, Di, SDi, C)
  if stopping-condition (SDi) then return class (SDi)
      else  fd := Select-attribute (p, SDi, C)
            Di+1 := Add-attribute (fd, Di)
            SDi+1 := Discriminatory-set (Di+1, SDi)
            LID (p, Di+1, SDi+1, C)
  end-if
end-function
```

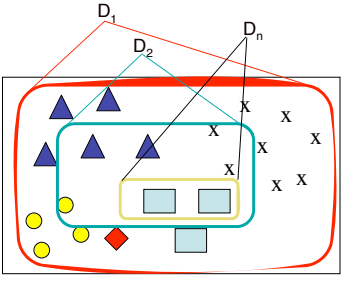


Fig. 1. The LID algorithm. On the right there is the intuitive idea of LID.

Δ and two attributes f and g inducing respectively partitions $\mathcal{P}_f$ and $\mathcal{P}_g$, we say that f is *more discriminatory* than g iff $\Delta(\mathcal{P}_f, \mathcal{P}_c) < \Delta(\mathcal{P}_g, \mathcal{P}_c)$. This means that the partition $\mathcal{P}_f$ is closer to the correct partition than the partition $\mathcal{P}_g$. LID selects the most discriminatory attribute to specialize D_i. Let f_d be the most discriminatory attribute in F. The specialization of D_i defines a new similitude term D_{i+1} by adding to D_i the attribute f_d. The new similitude term $D_{i+1} = D_i \cup \{f_d\}$ is satisfied by a subset of cases in S_{D_i}, namely $S_{D_{i+1}}$. Next, LID is recursively called with $S_{D_{i+1}}$ and D_{i+1}. The recursive call of LID has $S_{D_{i+1}}$ instead of S_{D_i} because the cases that are not satisfied by D_{i+1} will not satisfy any further specialization. Notice that the specialization reduces the discriminatory set at each step, i.e., we get a sequence $S_{D_n} \subset S_{D_{n-1}} \subset \ldots \subset S_{D_0}$. LID has two stopping situations: 1) all the cases in the discriminatory set S_{D_j} belong to the same solution class C_i, or 2) there is no attribute allowing the specialization of the similitude term. When the stopping condition 1) is satisfied, p is classified as belonging to C_i. When the stopping condition 2) is satisfied, S_{D_j} contains cases from several classes; in such situation the *majority criteria* is applied, and p is classified in the class of the majority of cases in S_{D_j}.

Now let us explain how to select the most discriminant attribute using the Rand index [10]. This index is used to compare clusterings being both classical partitions and it takes as basic unit of comparison the way in which two objects are clustered. The situation in which two objects are placed either together in the same cluster in both clusterings, or placed in different clusters in both clusterings, represents a similarity between the clusterings. Conversely, the situation in which two objects are in the same cluster in one clustering and in different clusters in the other, shows a dissimilarity between both clusterings. The Rand index assesses the similarity between clusterings based on the number of equal assignments of pairs of objects normalized by the total number of pairs. Inside LID, the Rand index is used to compare the partitions induced by each one of the attributes describing the objects with the correct partition. Let $X = \{x_1, \ldots, x_n\}$ be a finite set of objects, and let $\mathcal{P} = \{P_1, \ldots, P_k\}$ and $\mathcal{Q} = \{Q_1, \ldots, Q_h\}$ be two partitions of X in k and h sets, respectively. Given two objects x and x' we say that both objects are *paired* in a partition when both objects belong to

```
(define (object :id OBJ-50)
   (Sepallength 7.0)
   (Sepalwidth 3.2)
   (Petallength 4.7)
   (Petalwidth 1.4))
```

```
(define (object :id OBJ-50)
   (Sepallength (define (fuzzy-value)
                        (Value 7.0)
                        (Membership 0 0 1)))
   (Sepalwidth (define (fuzzy-value)
                        (Value 3.2)
                        (Membership 0 1 0)))
   (Petallength (define (fuzzy-value)
                        (Value 4.7)
                        (Membership 0 0.6087 0.3913)))
   (Petalwidth (define (fuzzy-value)
                        (Value 1.4)
                        (Membership  0 1 0))))
```

Fig. 2. On the left there is a propositional representation of an object. On the right there is the representation of the same object extended with the membership vector.

the same set of the partition. Otherwise, we say that both objects are *impaired*. Now let us consider the set $C := \{(x_i, x_j) \in X \times X : 1 \leq i < j \leq n\}$, which can be identified with the set of unordered pairs $\{x, y\}$, with $x, y \in X$. The Rand index between the partitions $\mathcal{P}$ and $\mathcal{Q}$ is defined as follows:

$$R(\mathcal{P}, \mathcal{Q}) = \frac{a + d}{a + b + c + d} \tag{1}$$

where $a = |\{(x, x') \in C : x \text{ and } x' \text{ paired in } \mathcal{P} \text{ and paired in } \mathcal{Q}\}|$,
$b = |\{(x, x') \in C : x \text{ and } x' \text{ paired in } \mathcal{P} \text{ and impaired in } \mathcal{Q}\}|$,
$c = |\{(x, x') \in C : x \text{ and } x' \text{ impaired in } \mathcal{P} \text{ and paired in } \mathcal{Q}\}|$,
$d = |\{(x, x') \in C : x \text{ and } x' \text{ impaired in } \mathcal{P} \text{ and impaired in } \mathcal{Q}\}|$.

Notice that in fact the Rand index gives a measure of the similarity between two partitions. Therefore we say that the attribute f inducing $\mathcal{P}_f$ is *more discriminatory* than the attribute g inducing $\mathcal{P}_g$ iff $1 - R(\mathcal{P}_f, \mathcal{P}_c) < 1 - R(\mathcal{P}_g, \mathcal{P}_c)$.

3 A Fuzzy Version of LID

In this section we explain a fuzzy version of LID using two fuzzifications of the Rand index: the one defined by Campello [5] and another one defined by Hüllermeier and Rifqi [8]. Firstly, we will explain how to represent the fuzzy cases handled by fuzzy LID. The left of Fig. 2 shows an example of an object from the *Iris* data set represented as a set of pairs attribute-value. The right of Fig. 2 shows the fuzzy representation of the same object. Notice that the value of each attribute is an object that has in turn two attributes: Value and Membership. The attribute Value takes the same value v that in the crisp version (for instance, 7.0 in the attribute Sepallength). The attribute Membership takes as value the so-called *membership vector* associated to v, that is, a n-tuple μ, being n the number of fuzzy sets associated to the continuous range of an attribute. Each position i of μ represents the membership of the value v to the corresponding fuzzy set F_i. In the next we will explain how to compute the membership vector.

Given an attribute taking continuous values, let us suppose that the domain expert has given $\alpha_1, \ldots, \alpha_n$ as the thresholds determining the discretization intervals for that attribute. Let α_0 and α_{n+1} be the minimum and maximum respectively of the values that this attribute takes in its range. For each one of the $n+1$ intervals $[\alpha_0, \alpha_1], \ldots, [\alpha_n, \alpha_{n+1}]$ corresponds a trapezoidal fuzzy set defined as follows, where $1 < i < n+1$:

$$F_1(x) = \begin{cases} 1 & \text{when } \alpha_0 \leq x \leq \alpha_1 - \delta_1 \\ \frac{\alpha_1 + \delta_1 - x}{2\delta_1} & \text{when } \alpha_1 - \delta_1 < x < \alpha_1 + \delta_1 \\ 0 & \text{when } \alpha_1 + \delta_1 \leq x \end{cases}$$

$$F_i(x) = \begin{cases} 0 & \text{when } x \leq \alpha_{i-1} - \delta_{i-1} \\ \frac{x - (\alpha_{i-1} - \delta_{i-1})}{2\delta_{i-1}} & \text{when } \alpha_{i-1} - \delta_{i-1} < x < \alpha_{i-1} + \delta_{i-1} \\ 1 & \text{when } \alpha_{i-1} + \delta_{i-1} \leq x \leq \alpha_i - \delta_i \\ \frac{\alpha_i + \delta_i - x}{2\delta_i} & \text{when } \alpha_i - \delta_i < x < \alpha_i + \delta_i \\ 0 & \text{when } \alpha_i + \delta_i \leq x \end{cases}$$

$$F_{n+1}(x) = \begin{cases} 0 & \text{when } x \leq \alpha_n - \delta_n \\ \frac{x - (\alpha_n - \delta_n)}{2\delta_n} & \text{when } \alpha_n - \delta_n < x < \alpha_n + \delta_n \\ 1 & \text{when } \alpha_n + \delta_n \leq x \leq \alpha_{n+1} \end{cases}$$

The parameters δ_i are computed as follows: $\delta_i = p \cdot |\alpha_i - \alpha_{i-1}|$, where the factor p corresponds to a percentage that we can adjust. Figure 3 shows the trapezoidal fuzzy sets defined when $n = 2$. For instance, for the *Iris* data set the values of α_i for the Petallength attribute are: $\alpha_0 = 1$, $\alpha_1 = 2.45$, $\alpha_2 = 4.75$, $\alpha_3 = 6.9$. The value 4.7 taken by the object *obj-50* in the attribute Petallength (Fig. 2) has associated the membership vector $(0, 0.6087, 0.3913)$, meaning that such value belongs to a degree 0 to the fuzzy set F_1 corresponding to the interval $[1, 2.45]$, to a degree 0.6087 to the fuzzy set F_2 corresponding to $[2.45, 4.75]$, and to a degree 0.3913 to the fuzzy set F_3 corresponding to $[4.75, 6.9]$.

In the fuzzy version of LID, the correct partition is the same than in the crisp case since each object belongs to a unique solution class. However, when

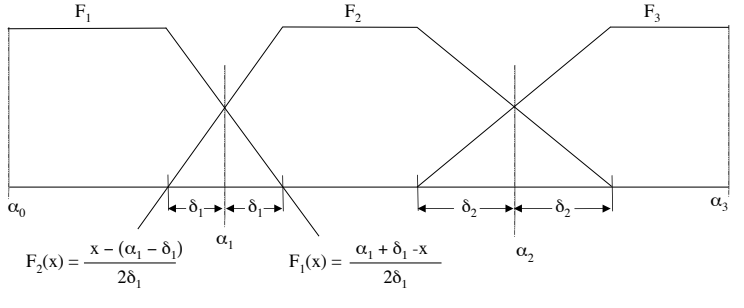


Fig. 3. Trapezoidal fuzzy sets. The values α_1 and α_2 are given by the domain expert as the thresholds of the discretization intervals for a given attribute.

the partitions induced by each attribute are fuzzy, an object can belong (to a certain degree) to more than one partition set. Thus the algorithm of the fuzzy LID is the same explained in Section 2 but using the particular representation for the fuzzy cases and replacing the Rand index by one of its fuzzy versions.

3.1 The Campello's Fuzzy Rand Index

In [5] Campello extends the Rand index to make it feasible to compare fuzzy partitions. To this end, he first rewrite the original formulation of the Rand index in an equivalent form by using basic concepts from set theory. Given the partitions $\mathcal{P}$ and $\mathcal{Q}$ of a set of objects X, and the set C of pairs of elements in X defined in Sec. 2, Campello defines the following subsets of C:

$$V = \{(x, x') : x \text{ and } x' \text{ paired in } \mathcal{P}\},\ W = \{(x, x') : x \text{ and } x' \text{ impaired in } \mathcal{P}\},$$
$$Y = \{(x, x') : x \text{ and } x' \text{ paired in } \mathcal{Q}\},\ Z = \{(x, x') : x \text{ and } x' \text{ impaired in } \mathcal{Q}\}.$$

According to these definitions, the coefficients in Eq. (1) can be rewritten as follows: $a = |V \cap Y|$, $b = |V \cap Z|$, $c = |W \cap Y|$, $d = |W \cap Z|$. When we consider fuzzy partitions, the sets above are fuzzy sets. Let $P_i(x) \in [0, 1]$ be the degree of membership of the object $x \in X$ to the set P_i. Campello defines the fuzzy binary relations V, W, Y and Z on the set C by using the following expressions involving a t-norm $\otimes$ and a t-conorm $\oplus$:

$$V(x, x') = \bigoplus_{i=1}^{k}(P_i(x) \otimes P_i(x')),\quad W(x, x') = \bigoplus_{1 \le i \ne j \le k}(P_i(x) \otimes P_j(x')),$$
$$Y(x, x') = \bigoplus_{i=1}^{h}(Q_i(x) \otimes Q_i(x')),\quad Z(x, x') = \bigoplus_{1 \le i \ne j \le h}(Q_i(x) \otimes Q_j(x')).$$

As it is usually done, Campello takes the intersection of fuzzy binary relations as the t-norm of the membership degrees of the pairs, and he uses the sigma-count principle for defining the fuzzy set cardinality (see [6]). Thus, the coefficients a, b, c, d are obtained as follows:

$$a = |V \bigcap Y| = \sum_{(x,x') \in C}(V(x, x') \otimes Y(x, x'))$$
$$b = |V \bigcap Z| = \sum_{(x,x') \in C}(V(x, x') \otimes Z(x, x'))$$
$$c = |W \bigcap Y| = \sum_{(x,x') \in C}(W(x, x') \otimes Y(x, x'))$$
$$d = |W \bigcap Z| = \sum_{(x,x') \in C}(W(x, x') \otimes Z(x, x'))$$

Then, the fuzzy version of the Rand index is also defined by the equation (1) giving a measure of the similarity between two partitions. Since LID uses a normalized distance measure, we have to take $1 - R(\mathcal{P}, \mathcal{Q})$. The Campello's fuzzy formulation of the Rand index is appropriated to compare a crisp partition with a fuzzy partition. Notice that the correct partition in classification problems is commonly crisp, therefore the use of the distance associated to the Rand index of Campello inside LID is justified. We will denote as CI such distance.

3.2 The Hüllermeier-Rifqi's Fuzzy Rand Index

When CI is used to compare two fuzzy partitions, it presents an important problem since the property of reflexivity is not satisfied. For this reason Hüllermeier and Rifqi proposed in [8] a different fuzzy version for the Rand index which

allows the comparison of two fuzzy partitions and that satisfies all the desirable metric properties. Let us recall their definition.

Given a fuzzy partition $\mathcal{P} = \{P_1, P_2, \ldots, P_k\}$, each object x is characterized by its membership vector $\mathcal{P}(x) = (P_1(x), P_2(x), ..., P_k(x)) \in [0,1]^k$ where $P_i(x)$ is the membership degree of x to the cluster P_i. Given two objects x and x' and two fuzzy partitions $\mathcal{P}$ and $\mathcal{Q}$, the *degree of concordance* of both objects in these partitions is defined by means the expression $1-|E_{\mathcal{P}}(x,x')-E_{\mathcal{Q}}(x,x')|$ where $E_{\mathcal{P}}$ is the fuzzy equivalence relation on X defined by $E_{\mathcal{P}}(x,x') := 1-\|\mathcal{P}(x)-\mathcal{P}(x')\|$ being $\|.\|$ a distance on $[0,1]^k$ yielding values in $[0,1]$. Thus, two objects are equivalent to a degree 1 when both have the same membership degrees in all the sets of the partition. This fuzzy equivalence is used to define the notion of *concordance* as a fuzzy binary relation, which generalizes the crisp binary relation (induced by a crisp partition) defined on the set C of unordered pairs of objects of X using the notions of *paired* and *unpaired*. Then, a distance measure on fuzzy partitions using the *degree of discordance* is defined as $|E_{\mathcal{P}}(x,x') - E_{\mathcal{Q}}(x,x')|$. Thus given a data set X of n elements, and two fuzzy partitions $\mathcal{P}$ and $\mathcal{Q}$ on X, the distance between both partitions is the normalized sum of degrees of discordance:

$$d(\mathcal{P},\mathcal{Q}) = \frac{\sum_{(x,x')\in C} |E_{\mathcal{P}}(x,x') - E_{\mathcal{Q}}(x,x')|}{n(n-1)/2}. \tag{2}$$

Since the Rand index measures similarity, by using the expression $1-d(\mathcal{P},\mathcal{Q})$ we can asses the similarity of two fuzzy partitions $\mathcal{P}$ and $\mathcal{Q}$. In [8] the authors prove that this similarity is a generalization of the Rand index, and they prove also that the distance (2) is a *pseudometric*, i.e., it satisfies the properties of *reflexivity*, *symmetry*, and the *triangular inequality*. Let us recall that a fuzzy partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ is called *normal* if *a*) for each $x \in X$, $P_1(x) + \cdots + P_k(x) = 1$, and *b*) it has a prototypical element, i.e., for every $P_i \in \mathcal{P}$, there exists an $x \in X$ such that $P_i(x) = 1$. In their paper Hüllermeier and Rifqi also show that for normal partitions, and taking the equivalence relation on X defined by

$$E_{\mathcal{P}}(x,x') = 1 - \frac{1}{2}\sum_{i=1}^{k} |P_i(x) - P_i(x')|, \tag{3}$$

the distance defined by the equation (2) is a *metric*, i.e., it also satisfies the property of *separation* ($d(\mathcal{P},\mathcal{Q}) = 0$ implies $\mathcal{P} = \mathcal{Q}$). We have taken this metric as measure of the distance in our experiments. From now on, we will call HR the distance proposed by Hüllermeier and Rifqi using (3).

4 Experiments

We conducted several experiments on four data sets coming from the UCI Repository [4] using the fuzzy versions of the Rand index inside LID. We used four data sets: *iris, heart-statlog, glass* and *thyroids*. For the evaluation of the crisp Rand index we taken the discretization intervals provided by Weka [13], and the same thresholds have been used for defining fuzzy sets. Thus, for instance, for the *Iris* data set, Weka gets the following intervals:

- Attribute Petalwidth: $(\infty, 0.8], (0.8, 1.75], (1.75, \infty)$
- Attribute Petallength: $(\infty, 2.45], (2.45, 4.75], (4.75, \infty)$
- Attribute Sepalwidth: $(\infty, 2.95], (2.95, 3.35], (3.35, \infty)$
- Attribute Sepallength: $(\infty, 5.55], (5.55, 6.15], (6.15, \infty)$

We performed three kinds of experiments: 1) using the crisp Rand index (with the discretization proposed by Weka); 2) using the fuzzyfication proposed by Campello (CI); and 3) using the fuzzyfication proposed by Hüllermeier-Rifqi (HR). The experiments with the crisp Rand index are considered as the baseline of the LID performance. In the fuzzy experiments, to calculate the values δ_i (see Sec. 3) we experimented with $p = 0.05, 0.10, 0.15, 0.20$. Moreover when using the Campello's fuzzyfication we also need to choose a t-norm and a t-conorm. In our experiments we taken the Minimum and the Maximum, respectively.

Table 1 shows the results of LID after seven trials of 10-fold cross-validation. For each index, there are three columns C, I and M corresponding respectively to the percentage of correct, incorrect and multiple answers. LID produces multiple answers when the last similitude term cannot be further specialized and the cases included in its associated discriminatory set belong to several solution classes. In such situation, LID is not able to classify the new problem and, depending on the domain, this can be interpreted as *no solution*. For this reason we counted them separately. We chosen to show the results obtained taking $p = 0.15$ since this is the value producing the least percentage of incorrect classifications. Results obtained with the values 0.05 and 0.10 are not significantly different from those with $p = 0.15$. Worst results are those obtained with $p = 0.20$. The parameter p is a measure of the overlapping degree between two fuzzy sets. In our experiments, the error percentage is not largely influenced by this degree.

It is difficult to extract a clear conclusion about which is the best method since none of them is better than others in all the domains, however the fuzzy versions of LID seems to be better than the crisp version. Our interpretation of this is that the use of fuzzy sets probably supports a more finest classification since, compared with the crisp version, the use of both CR and HR produce a lower percentage of both incorrect and multiple classifications (this happens for all domains except *thyroids*). Thus, when domains have classes with unclear frontiers (i.e., it is difficult to find a discriminant description for them), the use of fuzzy sets can correct these frontiers. Notice also that the percentage of multiple classifications produced by the Rand index is clearly lower than the one produced by both fuzzy versions

Table 1. Percentage of correct classifications (C), incorrect classifications (I) and multiple classifications (M) of LID using the Rand index, CI and HR. Results are the mean of 7 trials of 10-fold cross-validation and they correspond to $p = 0.15$

	Rand			CI			HR		
Data	C	I	M	C	I	M	C	I	M
iris	88.78	5.61	7.89	91.73	5.32	2.95	93.72	1.33	4.95
glass	35.46	9.50	55.04	9.56	6.26	84.18	30.63	13.97	55.40
thyroids	86.56	4.60	8.84	79.15	5.37	15.48	81.19	5.37	13.44
heart-statlog	65.40	16.19	18.41	54.55	14.97	30.48	56.40	16.67	26.93

(except for the *iris* domain). Since the percentage of incorrect classifications is also higher for the Rand index, we conclude that some of the objects that have not been classified using CI and HR (i.e., they produced multiple classifications) have been incorrectly classified using the Rand index. Therefore, the choice of a method has to be done taking into account the characteristics of the application domain. Sometimes it is preferable do not have answer in front of having an incorrect one; however, for some domains, to have more than one answer could be a valuable clue for classifying an object (for instance, when performing knowledge discovery).

Concerning the two fuzzy versions of LID, HR produces a lower percentage of multiple classifications than CI (except for the *iris* domain), however the percentage of incorrect classifications is higher in two of the domains (*glass* and *heart-statlog*). This means that CI is "more sure" in the classification of cases although a lot of times it cannot give a unique classification. Notice that for the *glass* domain the percentage of incorrect classifications is the lowest one; however the percentage of multiple classifications (i.e., no answer) is the highest one. We also conducted some experiments with the *bal* data set, also from the UCI repository, and the results of both HR and CI are not significantly different. Both indexes produce a percentage of incorrect classifications (3.52%) clearly lower than the produced by the Rand index (25.80%). Nevertheless, the Rand index produces a higher percentage of correct answers than the fuzzy indexes (65.53% in front of 60.54%).

Our conclusion is that the difference among the results using crisp and fuzzy indexes is strongly influenced by domain characteristics. Therefore it is necessary to perform an accurate analysis of the application domain (for instance, separability of classes, range of the values, etc.) in order to clearly determine the situations in which an index is better than others.

5 Conclusions and Future Work

We have introduced a new version of the method LID able to deal with fuzzy cases. Thus, cases have attributes taking continuous values which have been represented using fuzzy sets. In the current paper we experimented with LID using two different fuzzyfications of the Rand index one proposed by Campello and the other one proposed by Hüllermeier and Rifqi. From our experiments we concluded that it is difficult to assess a clear judgement about which measure is the best one in terms of classification accuracy. We performed experiments with different overlapping degrees of the fuzzy sets representing the values of the attributes, and we seen that this degrees do not significantly influence accuracy results. Our main conclusion is that the choice among the measures has to be made from an accurate analysis of the characteristics of the application domain.

All measures have a high computational cost, however we plan to experiment with the fuzzy extension proposed by Campello in order to exploit two parameters of the method: the t-norm and the t-conorm. In the experiments we used respectively the Minimum and the Maximum. In the future we plan to experiment with the t-norms of Łukasiewicz and Product and their dual t-conorms.

We also plan to use the similitude term generated by LID as a partial description of the solution classes as we have already done for the crisp version of LID. Now, this similitude term is fuzzy and this opens new opportunities to describe classes. In particular, we are thinking on knowledge discovery processes where the domain experts cannot define clearly the classes. In such domains, a fuzzy description of the classes could be very useful.

Acknowledgments

The authors acknowledge support of the Spanish projects MULOG2 (TIN 2007-68005-C04-01/04), ARINF (TIN 2009-14704-C03-03), Next-CBR (TIN 2009-13692-C03-01), and the grants 2009-SGR-1433/1434 from the Generalitat de Catalunya. The authors also want to thank Pilar Dellunde for her helpful comments and suggestions.

References

1. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications 7(1), 39–59 (1994)
2. Armengol, E., Plaza, E.: Bottom-up induction of feature terms. Machine Learning 41(1), 259–294 (2000)
3. Armengol, E., Plaza, E.: Lazy induction of descriptions for relational case-based learning. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 13–24. Springer, Heidelberg (2001)
4. Asuncion, A., Newman, D.J.: UCI machine learning repository (2007)
5. Campello, R.J.G.B.: A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters 28(7), 833–841 (2007)
6. de Luca, A., Termini, S.: A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. Information and Control 20(4), 301–312 (1972)
7. López de Mántaras, R.: A distance-based attribute selection measure for decision tree induction. Machine Learning 6, 81–92 (1991)
8. Hüllermeier, E., Rifqi, M.: A fuzzy variant of the Rand index for comparing clustering structures. In: IFSA/EUSFLAT Conference, pp. 1294–1298 (2009)
9. Lebowitz, M.: Categorizing numeric information for generalization. Cognitive Science 9(3), 285–308 (1985)
10. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66(336), 846–850 (1971)
11. Schlimmer, J.C.: Learning and representation change. In: Proceedings of the AAAI 1987, pp. 511–515 (1987)
12. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. Journal of Artificial Intelligence Research 6, 1–34 (1997)
13. Witten, I., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S.: Weka: Practical machine learning tools and techniques with java implementations (1999)

Fuzzy Clustering-Based Filter

Luiz F.S. Coletta, Eduardo R. Hruschka, Thiago F. Covoes,
and Ricardo J.G.B. Campello

Computer Science Department
University of São Paulo (USP) at São Carlos, Brazil
{luizfsc,erh,tcovoes,campello}@icmc.usp.br

Abstract. This paper introduces a filter, named FCF (Fuzzy Clustering-based Filter), for removing redundant features, thus making it possible to improve the efficacy and the efficiency of data mining algorithms. FCF is based on the fuzzy partitioning of features into clusters. The number of clusters is automatically estimated from data. After the clustering process, FCF selects a subset of features from the obtained clusters. To do so, we study four different strategies that are based on the information provided by the fuzzy partition matrix. We also show that these strategies can be combined for better performance. Empirical results illustrate the performance of FCF, which in general has obtained competitive results in classification tasks when compared to a related filter that is based on the hard partitioning of features.

1 Introduction

Feature selection involves choosing a subset of the original attributes (features) by eliminating the redundant, uninformative, and noisy ones. This issue has been broadly investigated in supervised learning tasks for which datasets with many features are available, like in text mining and gene expression data analysis. A comprehensive survey of feature selection algorithms is presented in [1]. In brief, there are two fundamentally different approaches for feature selection [2,3]: *wrapper* and *filter*. The former evaluates the subset of selected features using criteria based on the results of learning algorithms that will be ultimately employed, while the latter selects features based on intrinsic properties of the data, being independent of the learning algorithm to be used. Wrappers are often criticized because they require massive amounts of computation [4]. In data mining applications, one usually faces large datasets, and thus methods called filters, which are commonly faster than wrappers, are often more interesting [2]. The readers interested in filtering methods are referred to references [5,4,2,6,7,8] and the bibliography therein. While some filters may involve some kind of transformation of the feature space (e.g., principal component analysis and factor analysis), the present work focuses on finding subsets of features of the original space, mainly because this often allows much simpler and comprehensible results, maintaining the physical interpretation of the selected features.

The feature selection method proposed in this paper is based on the Simplified Silhouette Filter (SSF) [9,10], which removes redundant features by partitioning

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 406–415, 2010.

the original feature set into some hard clusters, from which a subset of features is selected, thus allowing dimensionality reduction. Similarly to SSF, the filter here proposed, named FCF (Fuzzy Clustering-based Filter), also relies on grouping similar features. Contrarily to the approach adopted in SSF, however, FCF uses a fuzzy clustering algorithm — instead of the hard counterpart used by SSF. This characteristic makes FCF more appropriate for detecting clusters with overlapping structures, which are commonly found in real-world applications. Besides using a different clustering algorithm, FCF also differs from SSF in the way features are selected from the induced clusters. In particular, FCF takes advantage of the information contained in the fuzzy partition matrix in order to remove redundant features. To that end, four strategies are investigated. We also show that these strategies can be combined for better performance.

The remainder of this paper is organized as follows. Section 2 describes the SSF algorithm studied in [9,10], whereas Section 3 addresses the proposed FCF. Section 4 analyzes empirical results obtained in 11 datasets by comparing FCF to SSF. Finally, Section 5 concludes the paper and points out some future work.

2 Related Work

In [9,10], the Simplified Silhouette Filter (SSF) was studied and compared to two state-of-art algorithms [6,5]. It has shown competitive results in relation to such algorithms, which were also designed to remove redundant features by clustering them. The filter here proposed (FCF) has been conceived to improve SSF, whose main characteristics are addressed in the sequel.

SSF is based on the partitioning of a set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ of features, $\mathbf{x}_j \in \Re^N$, into a collection $C^X = \{C_1, C_2, \ldots, C_k\}$ of k mutually disjoint subsets of correlated features C_i of X. Features that belong to the same cluster should be more similar (correlated) to each other than features that belong to different clusters. Therefore, it is necessary to devise means of evaluating similarities (in our case, correlations) between features. This problem can be tackled indirectly, i.e., distance measures can be used to quantify dissimilarities (lack of correlation) between features. In this work, a distance measure based on the Pearson correlation coefficient (ρ) — Eq. (1) — is employed.

$$\rho_d(\mathbf{x}_i, \mathbf{x}_j) = 1 - |\rho(\mathbf{x}_i, \mathbf{x}_j)| \tag{1}$$

Attempting at finding a globally optimum solution for clustering problems is usually not computationally feasible [11]. This difficulty has stimulated the search for efficient approximate algorithms. Following this trend, SSF uses a heuristic procedure, based on the simplified silhouette criterion [12], for finding the number of clusters and the corresponding (feature) partitions. To define the simplified silhouette (SS) [12], consider a feature $\mathbf{x}_j$ belonging to cluster C_a. The dissimilarity of $\mathbf{x}_j$ to the medoid of C_a is denoted by $a(j)$. Now let us take into account cluster C_i. The dissimilarity of $\mathbf{x}_j$ to the medoid of C_i will be called $d(\mathbf{x}_j, C_i)$. After computing $d(\mathbf{x}_j, C_i)$ for all clusters $C_i \neq C_a$, the smallest one is selected,

i.e. $b(j) = min\ d(\mathbf{x}_j, C_i)$, $C_i \neq C_a$. This value represents the dissimilarity of $\mathbf{x}_j$ to its neighbor cluster, and the silhouette $s(j)$ is given by:

$$s(j) = \frac{b(j) - a(j)}{max\{a(j), b(j)\}} \tag{2}$$

The higher $s(j)$ the better the assignment of $\mathbf{x}_j$ to a given cluster. In addition, if $s(j)$ is equal to zero, then it is not clear whether the feature should have been assigned to its current cluster or to a neighboring one [13]. Finally, if C_a is a singleton, then $s(j)$ is not defined and the most neutral choice is to set $s(j) = 0$ [14]. The average of $s(j)$, $j = 1, 2, \ldots, M$, can be used as a criterion to assess the quality of a given feature partition. By doing so, the best clustering is achieved when the silhouette value is maximized.

The computation of SS [12] depends only on the achieved partition and not on the adopted clustering algorithm. Thus, SS can be applied to assess partitions (taking into account the number of clusters) obtained by several clustering algorithms. We adopt the well-known k-medoids[1] algorithm to obtain partitions to be evaluated by SS. This algorithm is interrupted as soon as medoids from two consecutive iterations are equal. Roughly speaking, k-medoids is designed to minimize the sum of distances between features and nearest medoids. From the SS criterion viewpoint, good partitions are also obtained when this minimization is suitably performed, as well as when clusters are well separated.

SS is a numeric criterion that allows estimating the number of clusters automatically. Thus, it can provide a way of circumventing an important limitation of k-medoids, namely: k must be determined a priori. In this sense, one can perform multiple runs of k-medoids (for different values of k) and then choose the best available partition, which corresponds to the maximum achieved value for SS. It is also well-known that k-medoids may get stuck at suboptimal solutions for a given k [15]. To alleviate this problem, one can perform multiple runs of k-medoids for a fixed k, as done in the sampling strategy performed by SSF, which is summarized in Algorithm 1.

After running Algorithm 1, two strategies can be used by SSF to select a subset of features from the induced clusters, namely: (SSF-1) For each cluster, its medoid is chosen as a representative feature [9] — by doing so, a subset of k^* features is selected; (SSF-2) Besides selecting the medoid of each cluster, the feature least correlated (less redundant) with the medoid is also selected [10] — thus, two features from each cluster are chosen, resulting in the selection of $2k^*$ features.

3 Fuzzy Clustering-Based Filter (FCF)

The SSF algorithm addressed in Section 2 splits the set of available features into non-overlapping feature clusters. In practice, however, several data sets comprise

[1] The number of clusters, k, will be also denoted further in this paper by c, following the traditional notation adopted in the fuzzy clustering literature.

Algorithm 1. SSF Sampling Strategy [9,10]

Input: Choose the minimum and maximum number of clusters — k_{min} and k_{max} — and a number of initial partitions (n_p) for k-medoids.

```
SSV ⟵ −∞; // Simplified Silhouette Value //
for k ∈ {k_min, ..., k_max} do
    Generate n_p random initial partitions of features into k nonempty clusters;
    Run k-medoids for each initial partition generated in Step 3 and compute
    its corresponding simplified silhouette. Let the best obtained value be BOV;
    if BOV > SSV then
        SSV ⟵ BOV;
        k* ⟵ k;
        Hold the corresponding k* clusters of features;
    end
end
Return SSV and the corresponding k* clusters of features.
```

ill-delineated subsets that cannot be adequately split this way. For instance, there are situations in which the structure existing in the data is characterized by categories that overlap with each other to some degree. In these cases, the use of clustering algorithms that are capable of dealing with such overlapping data clusters is recommended. Fuzzy clustering techniques can naturally cope with this sort of problem since they aim at finding fuzzy clusters to which all the data (in our case, features) belong to some (possibly null) degree. This fact has motivated us to introduce the Fuzzy Clustering-based Filter (FCF), which adopts the fuzzy clustering algorithm known as Fuzzy c-Medoids (FCMdd) [16] to generate clusters of similar (correlated) features.

FCMdd [16] is an adaptation of the popular FCM algorithm [17] to deal with medoids. Let $d(\mathbf{x}_i, \mathbf{x}_j)$ be the dissimilarity between features $\mathbf{x}_i$ and $\mathbf{x}_j$, computed by Eq. (1), and $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_c)$, $\mathbf{v}_i \in X$, be a c-subset of X. Also, let X_c be the set of all c-subsets $\mathbf{V}$ of X. FCMdd aims at minimizing functional (3):

$$Jm(\mathbf{V}; X) = \sum_{j=1}^{M} \sum_{i=1}^{c} \mu_{ij}^{m} \, d(\mathbf{x}_j, \mathbf{v}_i) \tag{3}$$

where the minimization is performed over all $\mathbf{V}$ in X_c, m is the so-called "fuzzifier" whose default value is two, μ_{ij} represents the fuzzy membership of $\mathbf{x}_j$ to cluster i — computed by Eq. (4) — and c is the number of fuzzy clusters.

$$u_{ij} = \left[\sum_{l=1}^{c} \left(\frac{d(\mathbf{x}_j, \mathbf{v}_i)}{d(\mathbf{x}_j, \mathbf{v}_l)} \right)^{2/(m-1)} \right]^{-1} \tag{4}$$

After computing each element μ_{ij} we have a fuzzy partition matrix $\mathbf{U}_{c \times M}$.

In order to automatically estimate the number of clusters (c) from data, FCF essentially uses the sampling strategy summarized in Algorithm 1, except for two important differences: (i) In FCF the FCMdd algorithm [16] substitutes the

k-medoids used by SSF; and (ii) FCF employs a fuzzy version of the silhouette [14] given by [18]:

$$FS = \frac{\sum_{j=1}^{M}(u_{pj} - u_{qj})^{\alpha} s(j)}{\sum_{j=1}^{M}(u_{pj} - u_{qj})^{\alpha}} \quad (5)$$

where μ_{pj} and μ_{qj} are the first and second largest elements of the jth column of the partition matrix $\mathbf{U}$, respectively, α is a weighting coefficient whose default value is 1, and $s(j)$ is the traditional silhouette [14], which is given by Eq. (2), but considering that $a(j)$ is the average distance of $\mathbf{x}_j$ to all other objects belonging to cluster C_a and that $b(j)$ is the average distance of $\mathbf{x}_j$ to all objects belonging to nearest cluster $C_i \neq C_a$.

It is also worth mentioning that the sampling strategy used by SSF (based on multiple runs of k-medoids) outputs a set of hard clusters — Step 11 in Algorithm 1 — whereas FCF performs a post-processing step based on the fuzzy partition matrix corresponding to the best available partition — elected by the FS in (5). To that end, Step 8 of Algorithm 1 has also been modified to hold the respective fuzzy partition matrix, instead of a set of clusters. FCF uses such a partition matrix to perform feature selection according to four strategies, namely:

- Strategy 1: For each cluster, one selects only its medoid, which is the feature most correlated with the other features of the same cluster. The underlying assumption of this strategy is that medoids summarize the information contained in the clusters, in such a way that the remaining features can be deemed redundant. From the partition matrix, the medoid of each cluster i, $i = 1, ..., c$, is the feature j that has the maximum membership value μ_{ij} among $u_{i1}, u_{i2}, ..., u_{iM}$. Thus, this strategy selects c features.
- Strategy 2: Let us assume that the jth feature is classified into the ith fuzzy cluster if the membership μ_{ij} is higher than the membership of this feature to any other fuzzy cluster, i.e., $\mu_{ij} \geq \mu_{qj}$ for every $q \in \{1, ..., c\}, q \neq i$. After classifying features, there may be some of them that are weakly correlated with the medoids. These features can encompass useful information not captured by the medoids and, as such, they are potentially interesting for feature selection purposes. Based on this assumption, Strategy 2 selects, from each cluster i, $i = 1, ..., c$, the feature that has the minimum membership value among those classified into that cluster. This strategy also selects c features.
- Strategy 3: This strategy selects the feature least correlated with all clusters. We define this feature as the one that has the smallest variance in the membership degrees to the clusters. This selection strategy is particularly interesting for situations in which a given feature j has (approximately) equal membership values w.r.t. all clusters, what suggests that the information it holds is not captured by any of them.
- Strategy 4: This is similar to the 3^{rd} strategy, but it only considers pairs of clusters to perform feature selection. More precisely, for every pair of clusters, the feature that presents the smallest variance w.r.t. the respective membership values is selected. Doing so, $c(c-1)/2$ features are selected. The rationale behind this strategy is that features in-between two clusters can subsume information not captured by them.

Features selected by any of the above strategies are included into a subset of features. Obviously these strategies can be combined. For instance, the user may choose to apply only the 1^{st} and the 2^{nd} strategies or, alternatively, he or she may decide to use all of them. From this standpoint, we believe that 5 of such combinations allow deriving FCF variants of particular interest. Two of them (FCF-1 and FCF-12) are motivated by the feature selection criteria adopted by SSF. In particular, FCF-1 uses only the 1^{st} strategy, whereas FCF-12 uses both the 1^{st} and the 2^{nd} strategies. Analogously, we additionally have FCF-13, FCF-14, and, finally, FCF-1234 (all strategies). These FCF variants are experimentally investigated in this paper. Considering computational efficiency, it can be shown that the time complexity of FCF (including any of its feature selection strategies) is estimated as $O(\sum_{c=c_{min}}^{c_{max}} M^2+c^2M)$, where c_{min} and c_{max} are the minimum and maximum number of clusters, respectively.

4 Empirical Evaluation

Eleven datasets were used to assess the performance of the proposed algorithm. Six of them are bioinformatics datasets used by Yeung et al. [19]. These authors created five types of synthetic array datasets with error distributions derived from bioinformatics real data. We used such datasets to test our feature selection method. These datasets (here called Bio1, Bio2, Bio3, Bio4, and Bio5) are composed of 400 genes (instances), described by 20 measurements (features). There are six approximately equal-sized classes in each dataset. In addition, we tested the described feature selection algorithms in a real-world dataset (Yeast Galactose data [19]), which is composed of 20 measurements and 205 genes. In this dataset, the expression patterns reflect four functional categories.

The real-world dataset presented in [20], called Colon Cancer, was also used. This dataset contains 62 samples, whose classes correspond to either tumor or normal biopsies, described by the expression levels of 2,000 genes. The other real-world datasets used in our study are widely known and available at the UCI Machine Learning Repository [21], namely: Ionosphere, Wisconsin Breast Cancer, Pima Indians Diabetes, and Spambase.

In the experiments, we set c_{min}=2, c_{max}=$M/2$, and n_p=20 for clustering features by SSF and FCF. In the experiments with the Colon Cancer dataset, however, we set c_{max}=$\sqrt{M}$ due to its size, for which running experiments with c_{max}=$M/2$ has shown to be computationally prohibitive. From a practical viewpoint, one can consider that these values determine the size of the search space to be assessed, as well as the computational effort to find the corresponding clustering solution. Therefore, domain knowledge, when available, can be incorporated into this approach in order to set those parameters in scenarios in which there are limitations of computational resources.

The quality of each feature subset found by a given filter is here assessed by the generalization capability of the obtained classifier, which is estimated using a 10-fold cross-validation process [22]. We have followed an established methodology to compare different feature selection algorithms. In particular, feature selection

Table 1. Accuracies (%) obtained using NB (standard deviations in parentheses)

Dataset	FCF-1	FCF-12	FCF-13	FCF-14	FCF-1234	SSF-1	SSF-2	Full
Bio1	99.8 (1)	99.8 (1)	99.8 (1)	99.5 (1)	99.5 (1)	**100.0 (0)**	**100.0 (0)**	100.0 (0)
Bio2	96.5 (2)	97.3 (2)	96.5 (2)	96.8 (1)	**97.5 (1)**	96.0 (3)	97.3 (2)	98.8 (1)
Bio3	99.5 (1)	**100.0 (0)**	99.8 (1)	99.8 (1)	**100.0 (0)**	**100.0 (0)**	**100.0 (0)**	100.0 (0)
Bio4	85.8 (3)	96.8 (2)	94.3 (3)	94.3 (3)	96.8 (2)	92.0 (4)	**97.0 (2)**	97.5 (3)
Bio5	99.5 (1)	99.8 (1)	**100.0 (0)**	**100.0 (0)**	99.8 (1)	99.5 (1)	**100.0 (0)**	99.8 (1)
Yeast	92.6 (4)	94.6 (5)	94.6 (5)	**97.1 (2)**	**97.1 (2)**	94.1 (3)	94.6 (4)	97.6 (2)
Wisc	94.1 (2)	94.3 (2)	94.1 (2)	**94.4 (2)**	**94.4 (2)**	93.4 (2)	**94.4 (2)**	96.2 (1)
Pima	69.7 (4)	75.6 (5)	70.8 (5)	**75.8 (5)**	75.6 (5)	65.5 (3)	69.3 (6)	75.6 (5)
Iono	75.2 (7)	83.5 (7)	76.6 (7)	79.2 (10)	82.9 (9)	**86.3 (6)**	82.4 (6)	83.2 (3)
Spam	69.3 (4)	74.2 (3)	69.5 (4)	76.3 (3)	**77.5 (3)**	68.1 (6)	74.0 (4)	79.7 (1)
Colon	46.2 (21)	58.1 (18)	58.3 (20)	58.3 (21)	63.1 (24)	58.3 (19)	**67.9 (15)**	51.9 (20)

Table 2. Accuracies (%) obtained using NN (standard deviations in parentheses)

Dataset	FCF-1	FCF-12	FCF-13	FCF-14	FCF-1234	SSF-1	SSF-2	Full
Bio1	99.8 (1)	**100.0 (0)**	**100.0 (0)**	**100.0 (0)**	**100.0 (0)**	**100.0 (0)**	**100.0 (0)**	100.0 (0)
Bio2	95.8 (2)	**97.3 (1)**	96.8 (2)	96.0 (2)	96.0 (2)	96.3 (2)	96.0 (2)	99.0 (1)
Bio3	99.0 (1)	**100.0 (0)**	99.8 (1)	**100.0 (0)**	**100.0 (0)**	**100.0 (0)**	**100.0 (0)**	100.0 (0)
Bio4	81.5 (4)	95.5 (2)	87.8 (6)	87.8 (6)	95.5 (2)	91.8 (3)	**97.3 (2)**	97.5 (2)
Bio5	99.3 (1)	99.8 (1)	99.8 (1)	99.8 (1)	99.8 (1)	**100.0 (0)**	**100.0 (0)**	100.0 (0)
Yeast	91.6 (5)	94.1 (5)	94.1 (5)	**97.5 (3)**	**97.5 (3)**	90.12 (7)	94.1 (4)	98.1 (3)
Wisc	**93.9 (1)**	93.6 (2)	93.0 (1)	93.7 (2)	93.7 (2)	93.3 (1)	92.9 (2)	95.9 (1)
Pima	62.2 (4)	70.4 (4)	63.0 (5)	**70.6 (4)**	70.4 (4)	64.9 (4)	63.5 (4)	69.8 (5)
Iono	86.0 (7)	88.0 (5)	**88.9 (5)**	88.3 (4)	85.8 (5)	88.3 (4)	88.3 (4)	86.3 (6)
Spam	87.3 (2)	89.2 (2)	87.5 (2)	**90.5 (1)**	90.2 (1)	88.2 (1)	88.9 (2)	90.9 (1)
Colon	48.8 (17)	63.1 (20)	56.7 (24)	72.9 (19)	**81.0 (17)**	75.7 (15)	76.2 (9)	74.3 (10)

has been performed using only the training folds, and classification accuracy has been estimated in the test folds [23]. The same training/test folds were used for all algorithms. In order to provide some reassurance about the validity and non-randomness of the obtained results, we present the results of statistical tests by following the approach proposed by Demšar [24]. In brief, this approach is aimed at comparing multiple algorithms on multiple datasets, and it is based on the use of the well known Friedman test with a corresponding post-hoc test. The Friedman test is a non-parametric counterpart of the well-known ANOVA. If the null hypothesis, which states that the algorithms under study have similar performances, is rejected, then we proceed with the Nemenyi post-hoc test for pair-wise comparisons between algorithms.

Tables 1 and 2 report the average classification results (standard deviations appear within parentheses) for the Naïve Bayes (NB) and Nearest Neighbor (NN) classifiers. These classifiers were chosen to illustrate the relative performance of the feature selection methods evaluated in this work due to their widespread use in practice [25]. We used the NB and NN classifiers available in the Weka System [22], using its default parameters. In Tables 1 and 2, values in bold represent the best accuracies obtained with feature selection in each dataset.

Before delving into the details of some particular results found in each dataset, let us provide an overview of the classification results. Considering the different feature selection strategies for FCF (FCF-1, FCF-12, FCF-13, FCF-14, FCF-1234 — as explained in Section 3), the statistical test procedure used here (Friedman

Table 3. Average Rate (%) of Selected Features

Dataset	FCF-1	FCF-12	FCF-13	FCF-14	FCF-1234	SSF-1	SSF-2
Bio1	18.5	36.5	23.5	37.0	50.0	30.0	60.0
Bio2	15.0	30.0	20.0	30.0	44.5	18.0	36.0
Bio3	14.5	29.0	19.5	28.5	42.5	38.5	77.0
Bio4	10.0	20.0	15.0	15.0	20.0	15.0	30.0
Bio5	18.0	34.5	23.0	32.0	39.0	25.0	50.0
Yeast	15.5	20.5	20.5	28.0	28.0	10.0	15.5
Wisconsin	35.5	50.0	46.6	56.6	56.6	22.22	33.3
Pima	50.0	93.7	62.5	92.5	93.7	37.5	75.0
Ionosphere	13.2	26.1	16.1	34.1	40.8	33.8	63.5
Spambase	48.95	77.89	50.70	89.3	95.2	36.6	64.3
Colon Cancer	0.2	0.3	0.2	0.4	0.6	0.1	0.2

and Nemenyi tests) suggests (with α=5%) that FCF-1 presented worse performance than FCF-(12|14|1234), independently of the classifier used. Comparing FCF-1234 (best FCF variant) with SSF (with its two feature selection strategies, SSF-1 and SSF-2, as explained in Section 2) the statistical procedure indicate difference only between FCF-1234 and SSF-1 (with α=10%) when the NB classifier is used. Although the difference between FCF-1234 and SSF-2 was not statistically significant, the former was capable of obtaining better or equal results than the latter in 64% and 73% of the cases for NB and NN classifiers, respectively.

Taking into account the number of selected features (Table 3), some expected significant differences (with α=5%) were observed, namely: (i) FCF-1 selected less features than FCF-(12|14|1234) and SSF-2; (ii) FCF-13 selected less features than FCF-1234; (iii) SSF-1 selected less features than FCF-1234 and SSF-2.

Now let us shed light on some particular results obtained for specific datasets. In the Bio1 and Bio3 datasets, the classification accuracy obtained using all features (Full - last column) is 100% for both NB and NN. In this scenario, one may wonder if it is possible to remove some redundant features without decreasing classification performance. Our results show that this indeed can occur for practically all the assessed filters. FCF-1 has shown a slightly worse performance than the other filters in Bio3 for NN (accuracy of 99%). However, it is worth mentioning that it provided such a very good classification rate by using the smallest feature subset (with 14.5% of selected features). Therefore, this kind of analysis must be done carefully, for that, depending on the application, a reduction in more than 80% in the number of features (the largest reduction obtained considering all methods) can offer an interesting tradeoff between efficacy and efficiency. The results achieved in Bio2 and Bio5 are very similar to those obtained for Bio1 and Bio3. Considering Bio4, by its turn, all the assessed filters obtained worse accuracies than the use of all available features. FCF presented a significant accuracy reduction, especially using NN. SSF-2 provided the best classification results in this dataset, selecting only 30% of the features. In the Yeast dataset, FCF-14 and FCF-1234 provided the best classification results, using only 28% of the features on average. Similar results were observed in the Spambase and Pima datasets. For Winsconsin, all filters obtained very similar classification results. The same holds for Ionosphere when NN is used, but for NB we can arrive at different conclusions. In particular, SSF-1 presented the

best classification results, selecting only 34% of the features, whereas FCF-12 presented a slightly worse classification performance, but selecting only 26% of the features. Considering Colon Cancer, FCF-1234 and SSF-2 presented the best classification results, selecting less than 1% of the features.

5 Conclusions

This paper introduced the Fuzzy Clustering-based Filter (FCF), which is based on the fuzzy partitioning of features into clusters. FCF allows the elimination of redundant features. It was assessed in eleven classification datasets, using four feature selection strategies proposed in the paper. We have shown that the combination of these strategies lead, in general, to better performance, resulting in a FCF variant here named FCF-1234. This variant also showed improvements in relation to the Simplified Silhouette Filter (SSF) [9,10], which is a related algorithm based on the hard partitioning of features. In particular, although statistically significant differences were not found between FCF-1234 and SSF (selecting two features per cluster), the former presented better or equal classification results in approximately 70% of the performed experiments. The reported results also showed that it is possible to remove redundant features without significantly decreasing classification performance — sometimes even accuracy improvements can be achieved. This fact was clearly observed from the results reported for the Colon Cancer dataset, in which FCF and SSF selected less than 1% of the original features.

Although interesting results have been reported in this paper, there are several issues that can be investigated in the future. For example, a study on different feature selection strategies, taking into account different correlation measures, is a promising future work. Also, a more comprehensive experimental evaluation, comprising more datasets, is in order.

Acknowledgments

We acknowledge the Brazilian Research Agencies CNPq and FAPESP for their financial support to this work.

References

1. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Know. Data Eng. 17(4), 491–502 (2005)
2. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Dordrecht (1998)
3. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)
4. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003)

5. Au, W., Chan, K., Wong, A., Wang, Y.: Attribute clustering for grouping, selection, and classification of gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinformatics 2(2), 83–101 (2005)
6. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell. 24(3), 301–312 (2002)
7. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML1997, pp. 412–420. Morgan Kaufmann Publishers Inc., San Francisco (1997)
8. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. 5, 1205–1224 (2004)
9. Covões, T.F., Hruschka, E.R., de Castro, L.N., dos Santos, A.M.: A cluster-based feature selection approach. In: Corchado, E., Wu, X., Oja, E., Herrero, Á., Baruque, B. (eds.) HAIS 2009. LNCS (LNAI), vol. 5572, pp. 169–176. Springer, Heidelberg (2009)
10. Covões, T.F., Hruschka, E.R.: An experimental study on unsupervised clustering-based feature selection methods. In: ISDA 2009, pp. 993–1000. IEEE Press, Los Alamitos (2009)
11. Arabie, P., Hubert, L.J.: 1. In: An Overview of Combinatorial Data Analysis, pp. 5–64. World Scientific Publishing Company, Singapore (1999)
12. Hruschka, E.R., Campello, R.J.G.B., de Castro, L.N.: Evolving clusters in gene-expression data. Information Sciences 176(13), 1898–1927 (2006)
13. Everitt, B.S.: Cluster Analysis. Edward Arnold and Halsted Press (2001)
14. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley, Chichester (1990)
15. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: a new data clustering algorithm and its applications. Data Mining and Knowledge Discovery 1(2), 141–182 (1997)
16. Krishnapuram, R., Joshi, A., Nasraoui, O., Yi, L.: Low-complexity fuzzy relational clustering algorithms for web mining. IEEE Trans. Fuzzy Syst. 9(4), 595–607 (2001)
17. Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Dordrecht (1981)
18. Campello, R., Hruschka, E.: A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Sets and Systems 157(21), 2858–2875 (2006)
19. Yeung, K., Medvedovic, M., Bumgarner, R.: Clustering gene-expression data with repeated measurements. Genome Biology 4(5), R34 (2003)
20. Alon, U., Barkai, N., Notterman, D., Gishdagger, K., Ybarradagger, S., Mackdagger, D., Levine, A.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Nat. Acad. Sciences USA 96(12), 6745–6750 (1999)
21. Asuncion, A., Newman, D.: UCI machine learning repository (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
22. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)
23. Reunanen, J., Guyon, I., Elisseeff, A.: Overfitting in making comparisons between variable selection methods. J. of Mach. Learn. Res. 3, 1371–1382 (2003)
24. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)
25. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., Mclachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z., Steinbach, M., Hand, D., Steinberg, D.: Top 10 algorithms in data mining. Know. Inform. Systems 14(1), 1–37 (2008)

Fuzzy Classification of Nonconvex Data-Inherent Structures

Arne-Jens Hempel and Steffen F. Bocklisch

Chemnitz University of Technology, D-09107 Chemnitz, Germany
hear@hrz.tu-chemnitz.de

Abstract. The present work dedicates itself to the aggregation of nonconvex data-inherent structures into fuzzy classes. A key feature of this aggregation is its conduction within a closed fuzzy classification framework, being built around a single, generic type of a convex membership function. After a short elaboration concerning this essential building block a novel automated, data-driven design strategy to aggregate complex (nonconvex) data-inherent structures is introduced. The whole aggregation process will be illustrated with the help of an example.

1 Introduction

We are drowning in data but thirsting for knowledge. In order to convert the omnipresent flood of data into valuable and meaningful knowledge it has to be analysed, categorised and represented. However when dealing with real-life data it might be infeasible to derive an analytical model. Reasons for this circumstance are the high complexity of the underlying phenomena, imprecisions (e. g. measuring inaccuracies) or vague knowledge about the phenomena itself (e. g. missing information) [2,7]. In such a case it is assumed that objects (e. g. measurement data) reflect characteristics of the phenomenon in the form of patterns, subsequently referred to as data-inherent structures. The aggregation of those structures into superordinate entities (classes) allows a generalisation, representation and storage of their immanent knowledge [3]. An important aspect within this aggregation process is the representation of occurring uncertainties as supplementary property. One way to take such things into account is a class definition in form of fuzzy sets [9]. In terms of this idea the we will treat the aggregation approach as a fuzzy classification task. In order to present its general idea this paper will refrain from overly-formalised notations in favour of visually comprehensible examples.

2 Fuzzy Pattern Classes

The core component of the classification approach pursued here is a fuzzy set referred to as fuzzy pattern class (FPC). The subsequent sections provide only a necessary survey about its assets. A detailed description (rationale, definition, capabilities and utilisation) can be found in [1,5].

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 416–425, 2010.

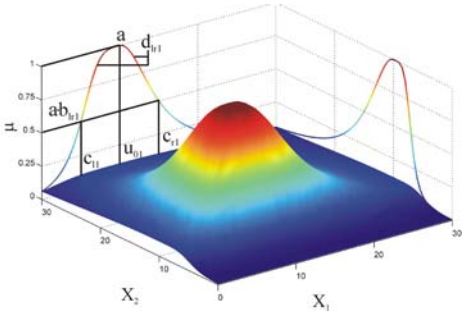


Fig. 1. Membership function and parameters

Definition of a Fuzzy Pattern Class: From a theoretical point a fuzzy pattern class corresponds to an asymmetric, parametric and multivariate membership function (1), defined over its specific universe of discourse (the class space U).

$$\mu(\boldsymbol{x}) = \left(1 + \frac{1}{2N}\sum_{i=1}^{N}[1-\operatorname{sgn}(x_i-u_{0i})]\cdot\left(\frac{1}{b_{\mathrm{l}i}}-1\right)\cdot\left|\frac{x_i-u_{0i}}{c_{\mathrm{l}i}}\right|^{d_{\mathrm{l}i}} + \frac{1}{2N}\sum_{i=1}^{N}[1+\operatorname{sgn}(x_i-u_{0i})]\cdot\left(\frac{1}{b_{\mathrm{r}i}}-1\right)\cdot\left|\frac{x_i-u_{0i}}{c_{\mathrm{r}i}}\right|^{d_{\mathrm{r}i}}\right)^{-1} \quad (1)$$

As Fig. 1 might indicate the multivariate form (1) derives from unidimensional FPC basis functions. Each basis function relates to a specific dimension of U. There it is defined by a set of seven parameters with the following meaning: The left and right class borders c_{l}, c_{r} characterise the support of the membership function in a crisp sense. The parameters $b_{\mathrm{l/r}} \in (0, 1]$ are referred to as border membership, assigning membership values to the class borders $c_{\mathrm{l/r}}$. The fuzziness of a FPC basis function is determined by the parameters $d_{\mathrm{l/r}} \in [2, \infty)$. All FPC basis functions are fuzzy-logically combined into (1) with the help of an N-fold compensatory Hamacher intersection operator, preserving the function concept and its properties. In terms of a PCA the FPC is fit optimal to its supporting data by a supplementary location $\boldsymbol{u_0}$ and rotation $\boldsymbol{\phi}$ of the class space U.

Deduction of Fuzzy Pattern Classes: The deduction of FPCs forms a self-contained part of research establishing two principal approaches [1]. The first way that can be thought of is a definition by expertise, i. e. an expert assigns all FPC parameters based upon task and domain specific knowledge. The second approach is an automatic data-driven method based upon a class labelled set of learning objects. It determines all class parameters in a two step aggregation procedure, see [5]. Figure 2 illustrates the result of this aggregation procedure for a set of 400 objects. Regarding the right hand side of Fig. 2 it needs to be stressed that objects are treated as fuzzy pattern entities. This extension is justified by the fact that every observation (measurement) contains an "elementary fuzziness" c_{e} (e. g. impression of a sensor) [5].

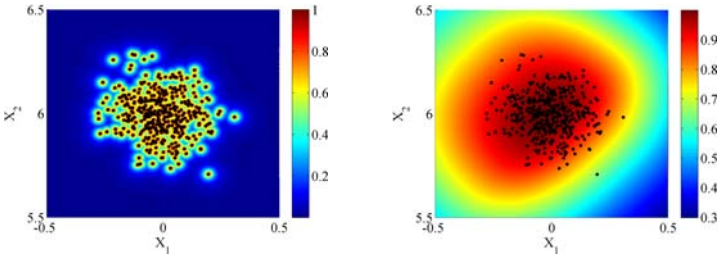


Fig. 2. Aggregation of objects to a fuzzy pattern class

Application and Mode of Operation: For applicative purposes the set of task relevant FPCs is deposited into a so called fuzzy pattern classifier. In operating mode the classifier assigns unknown objects to its class structure. The results of the classification process are the gradual object memberships $\mu^k(\boldsymbol{x})$ to each available class $k = 1, \ldots, K$. All memberships are calculated according to (1) and stored into a vector of memberships $\boldsymbol{\mu} = \left(\mu^1(\boldsymbol{x}), \mu^2(\boldsymbol{x}), ..., \mu^K(\boldsymbol{x})\right)^T$.

Properties of Fuzzy Pattern Classes: Thanks to the side-specific parametric definition FPCs combine a diversity of asymmetric shapes with interpretability and transparency. For the same reason FPCs provide a fair trade-off between data compression, computational cost and generalisation. Together with the data-driven automated aggregation procedure FPCs establish a closed classification framework (from data to FPCs). Besides all advantages FPCs are limited by their convex nature, causing them to form oversimplified classes when it comes to map nonconvex sets of data. In such a case the FPCs envelope the data but they comprise also object-unsupported class space, see Fig. 3.

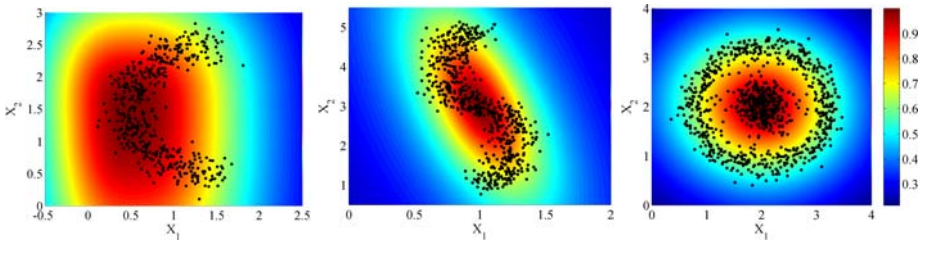


Fig. 3. Examples for nonconvex data-inherent structures and their associated FPCs

Under the condition that the classification framework remains conserved two ways of addressing this major drawback can been though of: Firstly a further segmentation of the data into convex subsets, for example based on cluster analysis [6], secondly a correction of FPCs by complementary fuzzy pattern classes.

3 Concept of Complementary Fuzzy Pattern Classes

The latter and hereafter introduced approach pursues the idea that, the difference of convex sets can be a nonconvex set. In terms of fuzzy logic this means that the negation of a class membership function over its object-unsupported space allows the creation of a nonconvex fuzzy class. Fig. 4 sketches such a negation of FPCs for a C-shaped pattern. According to the approach adopted in this paper, the negating complementary fuzzy pattern classes (CFPC) are defined using (1) ensuring the conservation of the classification framework. As an effect CFPCs exhibit the same properties as FPCs and they can be generated in an automated data-driven manner.

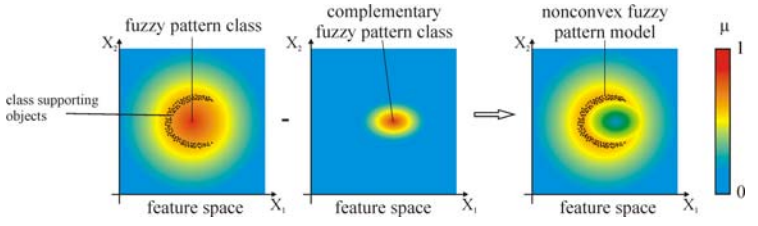


Fig. 4. Nonconvex fuzzy pattern representation via negation

As for the remaining examples in Fig. 3 an S-shape can be corrected by two CFPCs whereas a ring-dot shape can be corrected by double-negating CFPCs.

3.1 Deduction of Complementary Fuzzy Pattern Classes

Given a nonconvex set of learning data the task at hand lies in the determination of the appropriate complementary fuzzy pattern classes. Unfortunately the non-linearity and ambiguity of such a data set render a direct determination of CFPCs in terms of optimisation too costly. Yet a more subtle way to elaborate a data-driven design follows from the automated class generation. Due to the definition of CFPCs it can also be thought of a set of objects (complementary objects) supporting such a type of class, see Fig. 5. Assuming the general case

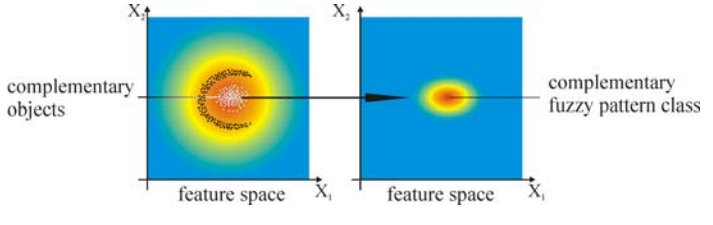


Fig. 5. Complementary objects for the C-shaped problem

it is unlikely to be given a set of complementary objects beforehand. If however such a set is accessible it will be also in accordance with the classification

framework to aggregate it into a CFPC. From this general point of view it becomes obvious that by generating a set of complementary objects a virtue can be made out of the necessity. Since the generation of complementary objects is vital to the subsequent CFPC deduction both aspects have been condensed into the following algorithm, see Fig. 6.

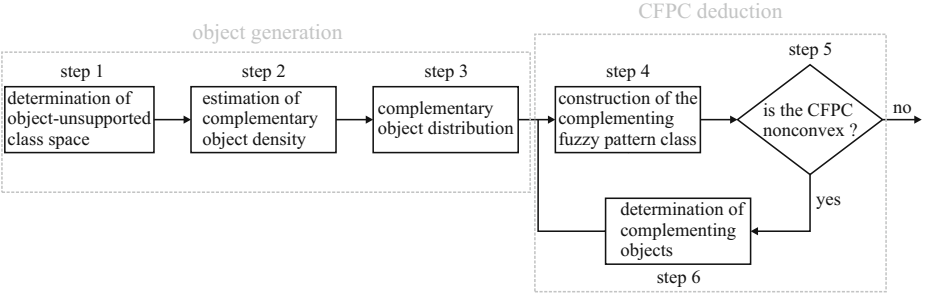


Fig. 6. Design algorithm for complementary fuzzy pattern classes

A variant of this algorithm has been presented in [4]. Scaling exponentially with the dimension of the feature space it is only suitable for low-dimensional cases. The subsequently presented approach addresses the high-dimensional case by perceiving the set of learning objects as a graph over the metric feature space.

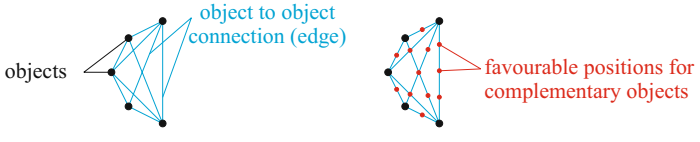


Fig. 7. Idea for a graph based class space exploration

Its leading thought is the distribution of complementary objects alongside object-to-object connections, see Fig. 7. Due to the fact that between M learning objects (vertices) there are $\frac{1}{2}M(M-1)$ object-to-object connections (edges) a graph based approach scales quadratically with the number of objects but independently from the class space dimension.

Step 1 to 3 Generation of Complementary Objects: Since there is little prior information about the shape of a data-inherent structure, the generation process is led by the basic assumptions that: (a) complementary objects are permitted to exist only in the according class space, and (b) complementary objects accumulate only in unsupported areas of the class space.

Step 1: In order to generate complementary objects it is necessary to reveal the object-unsupported areas of the class space. In terms of the underlying graph these areas can be specified by object-to-object connections that are fulfilling two properties. At first such a connection possesses a minimum length of $l_{\min} = 4 \cdot \max(c_e)$, a requirement which is justified by assumption (b). In other words, since objects and complementary objects are considered to be fuzzy entities they must not intersect. All connections of smaller size are discarded, because such an intersection might happen. The second property demands that a connection does not intersect another object within its elementary fuzziness, see right hand side of Fig. 8. This property ensures that connections which are traversing through object-supported class space will be discarded.

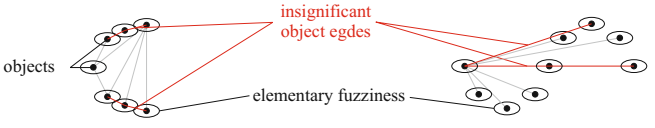


Fig. 8. Exclusion of insignificant object-to-object connections

Both exclusions guarantee that the remaining object-to-object connections perambulate only through the object-unsupported class space. They are referred to as potential connections. According to assumption (b) the complementary objects have to accumulated alongside these potential connections.

Step 2: In order to realise an accumulation of complementary objects on a sound mathematical basis it is necessary to define a density model. Due to the synthetical nature of this approach there are little restrictions for such a model. However to facilitate subsequent tasks it is desirable that this model yields a kind of distance dependent density $g_{\mathrm{cobj}}(s)$. A distance measure that suits such needs is the Euclidean distance s to the nearest learning object.

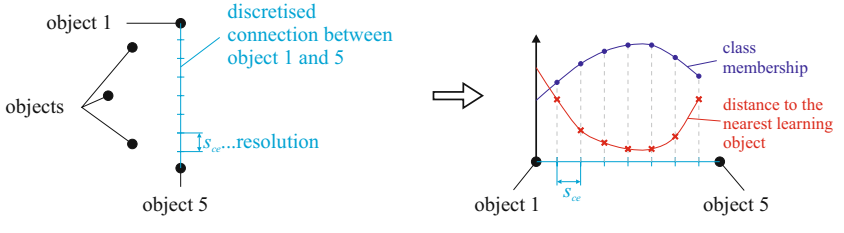


Fig. 9. Determination of distances along object-to-object connections

Along all potential connections s is determined by discretisation according to Fig. 9, yielding a distance for every section. The discretisation resolution s_{ce} follows from the shortest object-to-complementary-object distance $s_{\mathrm{ce}} = 2 \cdot \max(c_e)$.

Based on the available distance measure s the calculation of the complementary object density can be conducted. Among other density models a logistic approach of the following form was found the most promising.

$$g_{\mathrm{cobj}}(s) = \mathrm{rnd}\left(\frac{g_{\max}}{1 + \left(\frac{g_{\max}}{g_{\min}} - 1\right) e^{-g_{\max}(s - s_{\mathrm{ce}})}}\right) \tag{2}$$

It can be easily tuned by the parameters $g_{\max}$ (the maximum density) and (the minimum density) $g_{\min} = g_{\mathrm{cobj}}(s_{\mathrm{ce}})$. The bounds for the minimum density $g_{\min} \in (0, 0.5)$ follow straight from assumption (b). This means that the supremum $g_{\min} = 0.5$ must not be reached, otherwise the rounded density $g_{\mathrm{cobj}}(s)$ would result in complementary objects for distances $s = s_{ce}$. For computational purposes $g_{\min}$ is set to $g_{\min} = 0.49$, realising the steepest slope of (2). The determination of the maximum density $g_{\max} \in [0.5, \infty)$ requires a prespecification of a total number of complementary objects. $g_{\max}$ is fit to this total number with the help of a binary search. After the specification of $g_{\max}$ the logistic density model is applied to the set of all potential connections. In doing so it realises the desired accumulation in form of an integer complementary object equivalent for every section.

Step 3: According to the calculated density the complementary objects are distributed uniformly over their associated sections.

Step 4 to 6 Deduction of the CFPCs: Once the complementary objects have been generated the deduction of a suitable nonconvex FPC classifier can be tackled.

Step 4: All complementary objects will be aggregated to a CFPC.

Step 5: The complementary objects still might form a nonconvex geometry. An application of the aggregation procedure on such a structure results in an oversimplified CFPC μ^{CFPC}. This may suggest that the whole problem was shifted to the complementary objects. This, however, is not the case: According to the generation procedure the sets of complementary and learning objects are disjoint, yet both sets together cover the entire class space. In the case of a nonconvex complementary object geometry the associated CFPC will comprise subsets of learning objects. The question whether such a case is at hand can be perceived as a classification task. Employing our fuzzy classification approach, all learning objects are assigned to the CFPC. If all learning objects yield a low membership, the CFPC would be assumed to be sufficient and deduction is completed. If, on the contrary, learning objects exhibit high degrees of CFPC membership, they would indicate a nonconvex geometry of complementary objects. In such a case the CFPC has to be further specified by a second-level CFPC $\mu^{\mathrm{C^2FPC}}$. Instead of starting over with the whole procedure it follows immediately from complementation property that the learning objects can be seen as complementary objects of the complementary objects. This fortunate circumstance allows to reuse a subset of the already known learning objects as second-level complementary objects.

Step 6: It remains to ascertain which subset of the learning objects is the complementing one. For this purpose assumption (a) can be posed upon the aggregated CFPC. It demands that the second-level complementary objects have to be located within the borders of the CFPC. This is generally satisfied for learning objects with a high degree of CFPC membership. Thanks to the previously conducted classification (step 5) these second-level complementary objects are already know. They are fed back to the algorithm at step 4 and aggregated into the next-level CFPC. In general this CFPC cycle would continue until there is no further complementing object. Since each CFPC is smaller in size compared to its preceding class the disjoint learning and complementary object covering of the class space guarantees the convergence of the algorithm.

3.2 Combination of Complementary Fuzzy Pattern Classes

The overall representation of a data-inherent structure results from a combination of the FPC and its succeeding CFPCs. The combination process is specified according to the deduction process such that if an object does not belong to the CFPC it belongs to the preceding FPC. Moreover the deduction process might imply several levels of negation, which should to be respected in that concatenation. A fuzzy-logical interpretation of this concept derives from the concatenation of the minimum operator and natural complement [8].

$$\mu = \min\left(\mu^{\mathrm{FPC}}, \left(1 - \mu^{\mathrm{CFPC}}\right)\right) \tag{3}$$

4 Example

For the sake of comprehensibility the outlined algorithm will be illustrated with the help of an example. For this purpose it was applied to a ring-dot shaped data structure supported by a set of 900 objects, see Fig. 10.

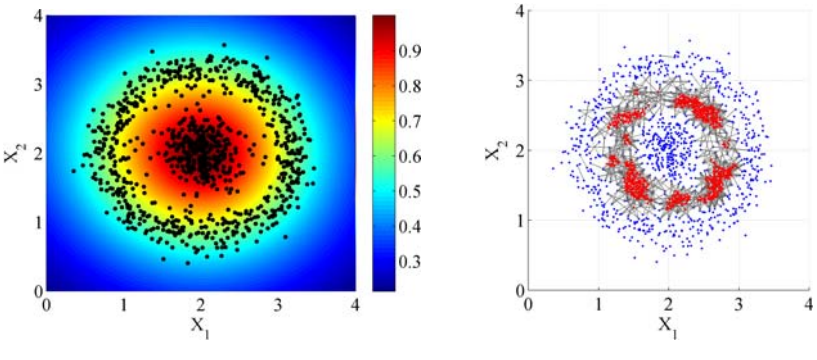


Fig. 10. Left: learning objects and FPC; right: connections and complementary objects

As depicted, the standard FPC μ^{FPC} will assign high memberships to the region in between both object geometries despite the fact that it is object-unsupported. According to Step 1–3 a total of 923 complementary objects have

been generated alongside the grey potential connections. However, the right hand side of Fig. 10 reveals that the complementary objects (red) themselves form a nonconvex (ring like) structure. Consequently their aggregation resulted in an oversimplified CFPC μ^{CFPC} during step 4, a fact that was verified via classification in step 5. Within the selection step 6, all learning objects exceeding a membership $\mu^{\mathrm{CFPC}} > 0.5$ have been assigned as second-level objects. As a matter of fact the classification resulted in high degrees of membership only for the central accumulation of learning objects. Therefore only those central objects have been aggregated into a second-level CFPC $\mu^{\mathrm{C^2FPC}}$. After its aggregation the cycle stopped because the membership of all available complementary objects was sufficiently low ($\mu^{\mathrm{C^2FPC}} < 0.5$). At this point the deduction algorithm for CFPCs stopped and provided three fuzzy pattern classes. All three classes have been combined into a nonconvex fuzzy pattern classifier via (4).

$$\mu = \min\left(\mu^{\mathrm{FPC}}, 1 - \min\left(\mu^{\mathrm{CFPC}}, \left(1 - \mu^{\mathrm{C^2FPC}}\right)\right)\right) \tag{4}$$

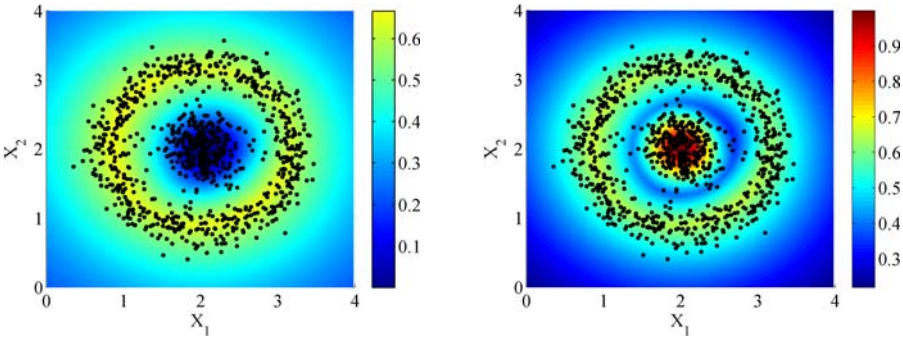


Fig. 11. First and final (second-) level combination of CFPCs

Figure 11 depicts the hierarchical composition of the classifier, together with the supporting objects.

5 Conclusions

This paper is dedicated to a fuzzy and data-driven class aggregation approach for nonconvex data-inherent structures. Its main philosophy is the exclusive application of a fuzzy classification framework being limited due to convexity. To overcome its convex limitations a graph based design strategy relying on so called complementary objects has been elaborated. The key steps within this design strategy are the generation of such complementary objects, their distribution, and the data-driven aggregation into CFPCs. The resulting CFPCs can be applied to negate the preceding FPC realising a nonconvex fuzzy pattern representation. Moreover the mutual negation of FPC and CFPCs implies a hierarchical view to understand such a nonconvex fuzzy class formation. Within

such a hierarchy it is the synergistic effect of mutual negation, shape diversity and interpretability of the of class membership functions that provides a powerful and flexible fuzzy classifier design for almost any shape of data-inherent structures. A possible combination of FPC and CFPCs to a nonconvex overall presentation has been demonstrated with the help of an example. In connection with this example it has been shown that complementary objects have to be constructed only once. Another aspect worth to mention is that the CFPC design works independent from clustering algorithms but features structure capturing. Finally it has to be stressed that the introduced design strategy is an universal approach that can be applied to any unimodal parametric membership function.

Possible applications for such an approach can be found in the areas of engineering geodesy and deformation analysis.

References

1. Bocklisch, S.F.: Prozeßanalyse mit unscharfen Verfahren. Technik, Berlin (1987)
2. Bosc, P., Prade, H.: An introduction to the fuzzy set and possibility theory-based treatment of soft queries and uncertain or imprecise databases. In: Smets, P., Motro, A. (eds.) Uncertainty Management in Information Systems: From Needs to Solutions, pp. 285–324. Kluwer, Dordrecht (1997)
3. Weihs, C., Gaul, W. (eds.): Classification — the Ubiquitous Challenge: Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Dortmund, vol. 28. Springer, Heidelberg (2005)
4. Hempel, A.J., Bocklisch, S.F.: Parametric fuzzy modelling framework for complex data-inherent structures. In: Proceedings of IFSA-EUSFLAT 2009, pp. 885–890 (2009)
5. Hempel, A.J., Bocklisch, S.F.: Fuzzy pattern modelling of data inherent structures based on aggregation of data with heterogeneous fuzziness. In: Rey, G.R., Muneta, L.M. (eds.) Modelling Simulation and Optimization, ch. 28, pp. 637–655. INTECH (2010)
6. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys 31(3), 264–323 (1999)
7. Kwiatkowska, M., Kielan, K., Michalik, K.: A fuzzy-semiotic framework for modeling imprecision in the assessment of depression. In: Proceedings of IFSA-EUSFLAT 2009, pp. 1717–1722 (2009)
8. Rutkowski, L.: Flexible Neuro-fuzzy Systems: Structures, Learning and Performance Evaluation. Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, Norwell (2004)
9. Zadeh, L.A.: Fuzzy sets. Information and Control 8(3), 338–353 (1965)

Fuzzy-Pattern-Classifier Training with Small Data Sets

Uwe Mönks[1], Denis Petker[2], and Volker Lohweg[1]

[1] inIT – Institute Industrial IT, Ostwestfalen-Lippe University of Applied Sciences
Liebigstr. 87, D-32657 Lemgo, Germany
{uwe.moenks,volker.lohweg}@hs-owl.de
[2] OWITA GmbH, Langenbruch 23, D-32657 Lemgo, Germany
denis.petker@owita.de

Abstract. It is likely in real-world applications that only little data is available for training a knowledge-based system. We present a method for automatically training the knowledge-representing membership functions of a Fuzzy-Pattern-Classification system that works also when only little data is available and the universal set is described insufficiently. Actually, this paper presents how the Modified-Fuzzy-Pattern-Classifier's membership functions are trained using probability distribution functions.

Keywords: Fuzzy Logic, Probability Theory, Fuzzy-Pattern-Classification, Machine Learning, Artificial Intelligence, Pattern Recognition.

1 Introduction

In many knowledge-based industrial application there is a necessity to train using a small data set. It is typical that there are less than ten up to some tens of training examples. Having only such a small data set, the description of the underlying universal set, from which these examples are taken, is very vague and connected to a high degree of uncertainty. It was Zadeh [1] who created the basic theory for the nowadays established fuzzy systems, which are suitable for modelling uncertain knowledge using possibility measures. One class of such systems are the *Fuzzy-Pattern-Classifiers (FPC)* introduced by Bocklisch [2] which are widely used in pattern recognition applications for object classification. The basic concept is having a set of *fuzzy membership functions* $\mu : x \rightarrow [0, 1]$ per class which model characteristic features of those classes. These membership functions map an object's feature value $x \in \mathbb{R}$ to the unit interval representing the membership or degree of similarity of x to an ideal class member's feature. All memberships are aggregated subsequently by some *fuzzy aggregation operator*. The object is then assigned to the class having the highest aggregated value.

One established member of the class of Fuzzy-Pattern-Classifiers is Lohweg's *Modified-Fuzzy-Pattern-Classifier (MFPC)* [3,4]. It is widely applied and established in the industry, for instance in printing facilities for checking the print results to give only one example [4,5,6,7]. In these applications, it proved its

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 426–435, 2010.

robustness, performance, and efficiency when implemented in hardware-based solutions.

The MFPC's membership functions are parameterisable unimodal potential functions having at least two degrees of freedom left to the user which demand the application of costly heuristics for finding their values. Mostly, the optimal parameters are not found, resulting in a loss of robustness and therefore deteriorated classification rates.

In this paper we suggest an automatic method of learning the fuzzy membership functions by estimating the data set's probability distribution and deriving the function's parameters automatically from it. The resulting *Probabilistic MFPC (PMFPC)* membership function, extends the MFPC approach to asymmetric membership functions and leaves only one degree of freedom leading to a shorter learning time for obtaining stable and robust classification results.

There exist other approaches in the literature, which go in our direction, but are not applicable here. Rodner and Denzler's approach [8] transfers feature relevance from previous, similar applications to choose the respective features for a new classification task. Our approach is directed to applications where no previous knowledge is available and the features are chosen heuristically. Drobics et al.'s FS-FOIL method is also very promising, but the classification results presented in [9] make use of a bigger training set. Also, the learning approach of finding fuzzy decision rules is different from ours where fuzzy membership functions' shapes are determined.

After having introduced the topic of this paper in this Section, we proceed in Sect. 2 by briefly introducing the Modified-Fuzzy-Pattern-Classifier. In Sect. 3 the new probabilistic parameterisation approach is described. The experiments presented in Sect. 4 return promising results that the incorporation of PMFPC membership functions in fuzzy classification tasks can improve classification results significantly when compared to MFPC. The paper concludes with Sect. 5 and provides an outlook on further research.

2 Modified Fuzzy Pattern Classifier

A hardware optimized derivate of Bocklisch's *Fuzzy-Pattern-Classifier (FPC)* [2] is the *Modified-Fuzzy-Pattern-Classifier (MFPC)*, which can be efficiently implemented as a pattern recognition system on a Field Programmable Gate Array (FPGA), applicable in high-speed industrial applications [4]. Here, its properties shall be briefly introduced. For details, we refer to [4] and [6].

The hardware efficient membership function used for the MFPC is Eichhorn's parameterisable unimodal potential function [10] defined as

$$\mu_{\mathrm{MFPC}}(m, \mathbf{p}) = 2^{-d(m,\mathbf{p})} \in [0,1] \text{ with } d(m, \mathbf{p}) = \left(\frac{|m-S|}{C}\right)^{D}, \tag{1}$$

where $\mathbf{p} = (S, C, D)$ is a parameter vector defining the membership function's properties, namely mean value (S), width (C), and steepness of its edges (D). $d(m, \mathbf{p})$ is the distance measure of the inspected feature m with regard to the

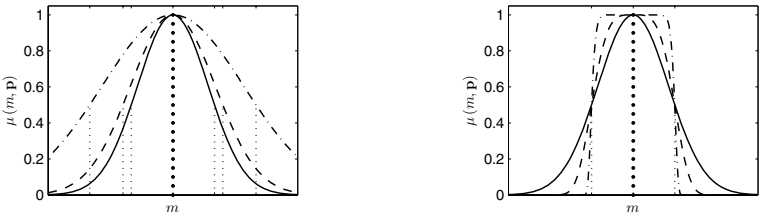


Fig. 1. Sample MFPC membership function at $D = 2$ and $p_{C_e} = 0$ (*solid*). The left and right plots show changes (*dashed* → *dash-dotted*) with increasing p_{C_e} and D, respectively. The vertical *dotted* line shows respective $S \pm C$, the *bold-dotted* line S.

properties of the membership function, i. e. how far is the measured feature m away from its mean value S. A sample MFPC membership function is depicted in Fig. 1.

The MFPC membership function's parameters S and C are obtained automatically during a learning phase after extracting all regarded features m from N typical members of a class by [3] $S = \Delta + m_{\min}, C = (1 + 2p_{c_e}) \cdot \Delta$, where $p_{c_e} \in [0, 1]$ is called percental elementary fuzziness and defines an arbitrary, user-defined width adjustment factor, and where $m_{\max} = \max_{i=1}^{N} m_i$, $m_{\min} = \min_{i=1}^{N} m_i$, $\Delta = \frac{m_{\max} - m_{\min}}{2}$. The integer-valued parameter D is chosen arbitrarily, typically as a power of 2 to keep calculating the distance measure $d(m, \mathbf{p})$ hardware-efficient [3].

The MFPC aggregation of M different features is expressed by

$$h_{\mathrm{MFPC}}(\mathbf{m}, \mathbf{P}) = 2^{-\frac{1}{M}\sum_{i=1}^{M} d_i(m_i, \mathbf{p}_i)}, \text{ with } d_i(m_i, \mathbf{p}_i) = \left(\frac{|m_i - S_i|}{C_i}\right)^{D_i}, \quad (2)$$

where $\mathbf{m}$ is a vector of feature values m_i and $\mathbf{P}$ a matrix of parameter vectors $\mathbf{p}_i$, parameterising each membership function belonging to a feature m_i. It is proved in [6] that the membership functions are aggregated using the well-known *geometric mean* aggregation operator, which is a fuzzy *averaging operator*. Since (2) can be rewritten to

$$h_{\mathrm{MFPC}}(\mathbf{m}, \mathbf{P}) = \left(\prod_{i=1}^{M} 2^{-d_i(m_i, \mathbf{p}_i)}\right)^{\frac{1}{M}} = \left(\prod_{i=1}^{M} \mu_{\mathrm{MFPC},i}(m_i, \mathbf{p}_i)\right)^{\frac{1}{M}},$$

it is possible to use any other fuzzy membership function instead of μ_{MFPC} for existing MFPC applications [6]. μ_{MFPC}'s parameters D and p_{c_e} are not determined automatically and left to the user. An appropriate substitute of μ_{MFPC}, which is parameterised completely automatically (or at least with a smaller number of free parameters) and yields optimal performances, was therefore searched for and found in the *Probabilistic MFPC* membership function. This approach is presented in the following.

3 Probabilistic MFPC Membership Function

To learn a fuzzy membership function automatically, *Random Fuzzy Variables (RFV)* can be applied [11], but this approach has disadvantages towards high-speed real-time applications. The *Probabilistic MFPC (PMFPC)* membership function approach we present here is able to preserve real-time demands (experiments revealed that the parameterisation is executed one order of magnitude faster than the RFV approach) while producing an optimal data set representation by incorporating an estimated probability distribution of the data.

The PMFPC approach is based on a generalised MFPC membership function

$$\mu_{\mathrm{PMFPC}}(m,\mathbf{p}) = 2^{-\mathrm{ld}\left(\frac{1}{B}\right)d(m,\mathbf{p})} \in [0,1] \text{ with } d(m,\mathbf{p}) = \left(\frac{|m-S|}{C}\right)^{D}, \quad (3)$$

where $B \in (0,1]$ is the *class boundary membership* parameter, i. e. defining the membership function's value at $m = S \pm C$: $\mu_{\mathrm{PMFPC}}(S \pm C, \mathbf{p}) = B$. This parameter was actually already introduced by Bocklisch in his Fuzzy-Pattern-Classifier definition [2]. D and B are automatically parameterised in the PMFPC approach. p_{c_e} is yet not automated to preserve the possibility of adjusting the membership function slightly without needing to learn the membership functions from scratch. The algorithms presented in this paper for automatically parameterising parameters D and B are inspired by former approaches: Bocklisch as well as Eichhorn developed algorithms which allow obtaining a value for the (MFPC) potential function's parameter D automatically, based on the used training data set. Bocklisch also proposed an algorithm for the determination of B. For details we refer to [2] and [10]. However, these algorithms yield parameters that do not fulfil the constraints connected with them (cf. Sect. 3.1 and 3.2) in all practical cases. Hence, we propose a probability theory-based alternative described in the following.

3.1 Automatically Parameterising the Steepness of the Edges

Bocklisch formulated constraints for D so the resulting membership function appropriately describes the data set for which the membership function is created [2]. He demands (i) $2 \leq D \leq 20$; (ii) if the objects in the data set are uniformly distributed, the membership function should be sharp-edged ($D = 20$); (iii) in case of an accumulation of objects at the outer boundaries, this distribution is represented by a sharp membership function as well ($D = 20$); and (iv) an inner accumulation of objects should generate a fuzzy membership function, thus $D \to 2$. These constraints are visualised in Fig. 2, showing ten distributed data points X_i and the resulting membership function $\mu(x)$ in accordance to the aforementioned constraints. Bocklisch's and Eichhorn's algorithms adjust D after comparing the actual distribution of objects to a perfect uniform distribution. However, the algorithms tend to change D for every (small) difference between the actual distribution and a perfect uniform distribution. This explains why both algorithms do not fulfil the constraints when applied to random uniform distributions.

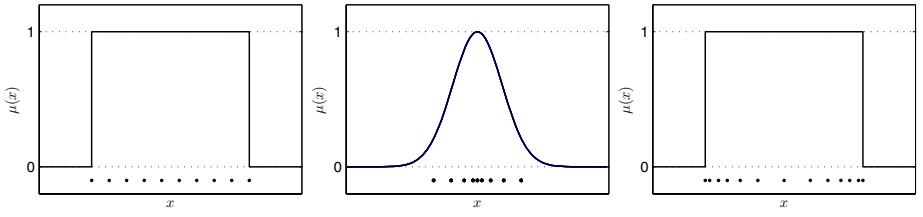


Fig. 2. Distributions of X_i (*bold dots*) and their corresponding membership functions

We actually stick to the idea of adjusting D with respect to the similarity of the actual distribution compared to an artificial, ideal uniform distribution, but we use probability theoretical concepts. Our algorithm basically works as follows: At first, the *empirical cumulative distribution function (ECDF)* of the data set under investigation is determined. Then, the ECDF of an artificial perfect uniform distribution in the range of the actual distribution is determined, too. The similarity between both ECDFs is expressed by its correlation factor which is subsequently mapped to D by a parameterisable function.

Determining the Distributions' Similarity. Consider a sorted vector of n feature values $\mathbf{m} = (m_1, m_2, \ldots, m_n)$ with $m_1 \leq m_2 \leq \ldots \leq m_n$, thus $m_{\min} = m_1$ and $m_{\max} = m_n$. The corresponding ECDF $P_m(x)$ is determined by $P_m(x) = \frac{|\tilde{\mathbf{m}}|}{n}$ with $\tilde{\mathbf{m}} = (m_i | m_i \leq x \ \forall i \in \mathbb{N}_n)$, where $|\mathbf{x}|$ denotes the number of elements in vector $\mathbf{x}$ and $\mathbb{N}_n = [1, 2, \ldots, n]$. The artificial uniform distribution is created by equidistantly distributing n values u_i, hence $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ with $u_i = m_1 + (i-1) \cdot \frac{m_n - m_1}{n-1}$. Its ECDF $P_u(x)$ is determined analogously by substituting $\mathbf{m}$ with $\mathbf{u}$. In the next step, the similarity between both distribution functions is computed by calculating the *correlation factor* [12]

$$c = \frac{\sum_{i=1}^{k} \left(P_m[x_i] - \overline{P_m}\right)\left(P_u[x_i] - \overline{P_u}\right)}{\sqrt{\sum_{i=1}^{k} \left(P_m[x_i] - \overline{P_m}\right)^2 \sum_{i=1}^{k} \left(P_u[x_i] - \overline{P_u}\right)^2}},$$

where $\overline{P_a}$ is the mean value of $P_a(x)$, computed as $\overline{P_a} = \frac{1}{k} \sum_{i=1}^{k} P_a[x_i]$. c's properties can be found in [12]. It is actually the empirical correlation coefficient, demanding sampled data to be determined, necessarily sampled at the same locations x_i. Since $P_m(x)$ cannot be predicted, it seems to be appropriate to sample at k equidistantly spaced locations. k is determined by $k = 10^{\lfloor \log_{10} n \rfloor + 1}$, but at least $k = 50$. This guarantees that the functions are sampled at not less than five times as many sampling points as feature values are available. The equidistant locations are determined as $x_i = m_1 + (i-1) \cdot \frac{m_n - m_1}{k-1} \ \forall i \in \mathbb{N}_k$.

The correlation factor must now be mapped to D while fulfilling Bocklisch's constraints on D. Therefore, the average influence $\overline{\alpha}(D)$ of the parameter D on the membership function μ_{MFPC}, which is the base for μ_{PMFPC}, is investigated to derive a mapping based on it. $\alpha_D(x)$ is determined by taking $\frac{\partial}{\partial D} \mu_{\mathrm{MFPC}}(x, D)$ with $x = \frac{m-S}{C}, x > 0$:

$$\alpha_D(x) = \frac{\partial}{\partial D}\mu_{\mathrm{MFPC}}(x, D) = \frac{\partial}{\partial D}2^{-x^D} = \ln(2)\left(-2^{-x^D}\right)x^D \ln(x).$$

The locations x represent the distance to the membership function's mean value S, hence $x = 0$ is the mean value itself, $x = 1$ is the class boundary $S + C$, $x = 2$ twice the class boundary and so on. The average influence of D on the membership function is evaluated for $-1 \leq x \leq 1$: This interval bears the most valuable information since all feature values of the objects in the training data set are included in this interval, and additionally those of the class members are expected here during the classification process, except from only a typically neglectable number of outliers. Anyway, the range of x must be necessarily bounded since the average influence of D on the membership function, namely $\overline{\alpha}(D) = \frac{1}{x_r - x_l}\int_{x_l}^{x_r} \alpha_D(x)\, dx$, is computing $\alpha_D(x)$'s mean value along x. But since $\lim_{x\to\infty} \alpha_D(x) = 0\ \forall D$, integration of $\alpha_D(x)$ over $\mathbb{R}$ would yield $\overline{\alpha}(D) = 0\ \forall D$, which is not true for that range of x where the majority of objects is present. The mapping of $D : c \to [2, 20]$, which is derived in the following, must take D's average influence into consideration. A graphical representation of $\overline{\alpha}(D)$ is shown in Fig. 3 for the range $2 \leq D \leq 20$, which is actually the only of interest.

Mapping the Distribution's Similarity to the Edge's Steepness. In the general case, the correlation factor c can take values from the interval $[-1, 1]$, but when evaluating distribution functions, the range of values is restricted to $c \in [0, 1]$, which is because probability distribution functions are monotonically increasing. This holds for both distributions, $P_m(x)$ as well as $P_u(x)$. It follows $c \geq 0$. The interpretation of the correlation factor is straight forward. A high value of c means that the distribution $P_m(x)$ is close to a uniform distribution. If $P_m(x)$ actually was a uniform distribution, $c = 1$ since $P_m(x) = P_u(x)$. According to Bocklisch, D should take a high value here. The more $P_m(x)$ differs from a uniform distribution, the more $c \to 0$, the more $D \to 2$. Hence, the mapping function $D(c)$ must necessarily be an increasing function with taking the exponentially decreasing average influence of D on the membership function $\overline{\alpha}(D)$ into consideration (cf. Fig. 3). An appropriate mapping $D : c \to [2, 20]$ is an exponentially increasing function which compensates the changes of μ_{MFPC} with respect to changes of c. While big changes in small c values result in minor

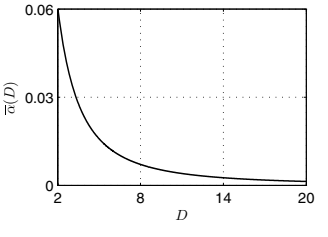


Fig. 3. Average influence of the parameter D on μ_{MFPC} with respect to D

changes of D, implying only a small change of the membership function, D increases rapidly for big correlation factors, not affecting the membership function strongly. We suggest the following heuristically determined exponential function, which achieved promising results during experiments (cf. Sect. 4):

$$D(c) = 19^{c^{2q}} + 1 \;\Rightarrow\; D(c) \in [2, 20], \tag{4}$$

where q is an adjustment parameter. This formulation guarantees that $D \in [2, 20]$ $\forall c$ since $c \in [0, 1]$. Using the adjustment parameter q, D is adjusted with respect to the aggregation operator used to fuse all n membership functions representing each of the n features. Each fuzzy aggregation operator behaves differently. For a fuzzy averaging operator $h(\mathbf{a})$, Dujmović' introduced the objective measure of *global andness* ρ_g (for details cf. [13,6]). Assuming $q = 1$ in the following cases, it can be observed that, when using aggregation operators with a global andness $\rho_g^{h(\mathbf{a})} \to 0$, the aggregated single, n-dimensional membership function is more fuzzy than that one obtained when using an aggregation operator with $\rho_g^{h(\mathbf{a})} \to 1$, where the resulting function is sharp. This behaviour should be compensated by adjusting D in such a way, that the aggregated membership functions have comparable shapes: at some given correlation factor c, D must be increased if ρ_g is high and vice versa. This is achieved by mapping the aggregation operator's global andness to q, hence $q : \rho_g \to \mathbb{R}$. Our suggested solution is a direct mapping of the global andness to the adjustment parameter q, hence $q(\rho_g) = \rho_g \;\Rightarrow\; q \in [0, 1]$. Mapping (4) is now completely defined and consistent with Bocklisch's constraints and our observations regarding the aggregation operator's andness.

3.2 Determining the Class Boundary Membership Parameter

In addition to the determination of D, we present an algorithm to automatically parameterise the class boundary membership B. This parameter is a measure for the membership $\mu_{\mathrm{MFPC}}(m, \mathbf{p})$ at the locations $m \in \{S + C,\, S - C\}$. Typically, the class boundary membership is assigned a value of $B = 0.5$. The algorithm for determining B is based on the algorithm Bocklisch developed [2], but was not adopted as it stands since it has some disadvantages if this algorithm is applied to distributions with a high density especially on the class boundaries. Due to space limitations, this cannot be presented here.

When looking at μ_{MFPC}, the following two constraints on B can be derived: (i) The probability of occurrence is the same for every object in uniform distributions, also on the class boundary. Here, B should have a high value. (ii) For distributions where the density of objects decreases when going towards the class boundaries B should be assigned a small value, since the probability that an object occurs at the boundary is smaller than in the centre.

Hence, for sharp membership functions ($D \to 20$) a high value for B should be assigned, while for fuzzy membership functions ($D \to 2$) the value of B should be low. $B = f(D)$ must have similar properties like $\overline{\alpha}(D)$, meaning B changes quickly where $\overline{\alpha}(D)$ changes quickly and vice versa. We adopted Bocklisch's suitable equation for computing the class boundary membership [2]:

$$B = \frac{1}{1 + \left(\frac{1}{B_{\max}} - 1\right) \cdot \left(\frac{D_{\max}}{D}\right)^{1+\frac{1}{q}}},$$

where $B_{\max} \in (0,1)$ stands for the maximum possible value of B with a proposed value of 0.9, $D_{\max} = 20$ is the maximum possible value of D and q is identical in its meaning and value to q as used in (4).

3.3 An Asymmetric PMFPC Membership Function Formulation

A data set may be represented better if the membership function was formulated asymmetrically instead of symmetrically as is the case with (3). This means

$$\mu_{\mathrm{PMFPC}}(m, \mathbf{p}) = \begin{cases} 2^{-\mathrm{ld}\left(\frac{1}{B_l}\right)\left(\frac{|m-S|}{C_l}\right)^{D_l}}, & m \le S \\ 2^{-\mathrm{ld}\left(\frac{1}{B_r}\right)\left(\frac{|m-S|}{C_r}\right)^{D_r}}, & m > S \end{cases}, \tag{5}$$

where $S = \frac{1}{M}\sum_{i=1}^{M} m_i, m_i \in \mathbf{m}$ is the *arithmetic mean* of all feature values. If S was computed as introduced in (2), the resulting membership function would not describe the underlying feature vector $\mathbf{m}$ appropriately for asymmetrical feature distributions. A new computation method must therefore also be applied to $C_l = S - m_{\min} + p_{C_e} \cdot (m_{\max} - m_{\min})$ and $C_r = m_{\max} - S + p_{C_e} \cdot (m_{\max} - m_{\min})$ due to the change to the asymmetrical formulation. To compute the remaining parameters, the feature vector must be split into the left side feature vector $\mathbf{m}_l = (m_i | m_i \le S)$ and the one for the right side $\mathbf{m}_r = (m_i | m_i \ge S)$ for all $m_i \in \mathbf{m}$. They are determined following the algorithms presented in the preceding Sections, but using only the feature vector for one side to compute this side's respective parameter.

4 Experimental Results on PMFPC

In order to evaluate the classification performance of our probabilistic approach on parameterising the fuzzy membership functions, the same data set is used to learn both the original MFPC membership function μ_{MFPC} and also μ_{PMFPC}. This data set "OCR" (the same as is used in [6]) was compiled in an industrial optical character recognition application and consists of both a training and a test data set. The test data set consists of 746 objects with each 17 features assigned to twelve classes. The dedicated training data set used to learn the membership function comprises 17 images per class, hence 204 images. This represents a typical situation occurring in classification applications, where the training data set from which a robust classifier is to be derived is very small. For details about the data set we refer to [6]. The subsequent classification is executed with different aggregation operators by using the classifier framework presented in [6]. Here, the incorporated aggregation operators are Yager's family of *Ordered Weighted Averaging (OWA)* [14] and Larsen's family of *Andness-directed Importance Weighting Averaging (AIWA)* [15] operators (applied unweighted here)—which both can be

adjusted in their andness degree—and additionally MFPC's original geometric mean (GM). Due to space limitations, we refer to [14] and [15] for the definition of OWA and AIWA operators. As a reference, the data set is also classified using a *Support Vector Machine (SVM)* with a Gaussian radial basis function (RBF). Since SVMs are capable of distinguishing between only two classes, the classification procedure is adjusted to pairwise (or one-against-one) classification according to [16]. Our benchmarking measure is the classification rate $r_+ = \frac{n_+}{N}$, where n_+ is the number of correctly classified objects and N the total number of objects that were evaluated. The best classification rates at a given aggregation operator's andness ρ_g are summarised in the following Table 1, where the best classification rate per group is printed bold.

Table 1. "OCR" classification rates r_+ for each aggregation operator at andness degrees ρ_g with regard to membership function parameters D and p_{C_e}

	Aggregation Operator	μ_{PMFPC}		μ_{MFPC}							
				$D=2$		$D=4$		$D=8$		$D=16$	
ρ_g		p_{C_e}	r_+	p_{C_e}	r_+	p_{C_e}	r_+	p_{C_e}	r_+	p_{C_e}	r_+
0.5000	AIWA	**0.255**	**93.70 %**	0.370	84.58 %	0.355	87.67 %	0.310	92.36 %	0.290	92.90 %
	OWA	**0.255**	**93.70 %**	0.370	84.58 %	0.355	87.67 %	0.310	92.36 %	0.290	92.90 %
0.6000	AIWA	**0.255**	**93.16 %**	0.175	87.13 %	0.205	91.02 %	0.225	92.36 %	0.255	92.23 %
	OWA	**0.255**	**93.57 %**	0.355	84.58 %	0.365	88.47 %	0.320	92.63 %	0.275	92.76 %
0.6368	GM	**0.950**	**84.45 %**	0.155	81.77 %	0.445	82.17 %	0.755	82.44 %	1.000	82.44 %
	AIWA	**0.245**	**91.42 %**	0.135	85.52 %	0.185	90.08 %	0.270	89.81 %	0.315	89.95 %
	OWA	**0.255**	**93.57 %**	0.355	84.72 %	0.355	88.74 %	0.305	92.63 %	0.275	92.76 %
0.7000	AIWA	**1.000**	**83.65 %**	0.420	82.71 %	0.790	82.57 %	0.990	82.31 %	1.000	79.22 %
	OWA	**0.280**	**93.57 %**	0.280	84.85 %	0.310	89.01 %	0.315	92.76 %	0.275	92.63 %

The best classification rates for the "OCR" data set are achieved when the PMFPC membership function is incorporated, which are more than 11 % better than the best incorporating μ_{MFPC}. The Support Vector Machine achieved a best classification rate of $r_+ = 95.04\,\%$ by parameterising its RBF kernel with $\sigma = 5.640$, which is 1.34 % or 10 objects better than the best PMFPC approach.

5 Conclusion and Outlook

Based on the MFPC membership function, we developed and presented a probabilistic parameterisation method, which automatically learns the membership functions based on a given set of training data. This method yields membership functions which outperform any approach using μ_{MFPC} as a fuzzy classifier's membership functions and provides a performance similar to a Support Vector Machine for the evaluated sample data set. Nevertheless, the presented approach is not intended to serve as a SVM substitute, but to show its possible performance compared to a state-of-the-art classification technique while providing robust results for small training data sets and preserving real-time demands as well as hardware-implementability. All results obtained must be seen in the scope of the test case, general statements cannot be derived. Still more data sets need to be classified to see if the trends hold.

Acknowledgement. This work was supported and financed by the German Federal Ministry of Education and Research, Grant-No. 17N2708.

References

1. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy sets and systems 1, 3–28 (1978)
2. Bocklisch, S.F.: Prozeßanalyse mit unscharfen Verfahren. Verlag Technik, Berlin (1987)
3. Lohweg, V.: Ein Beitrag zur effektiven Implementierung adaptiver Spektraltransformationen in applikationsspezifische integrierte Schaltkreise. PhD thesis, Technische Universität Chemnitz (2003)
4. Lohweg, V., Diederichs, C., Müller, D.: Algorithms for hardware-based pattern recognition. EURASIP Journal on Applied Signal Processing 2004(12), 1912–1920 (2004)
5. Niederhöfer, M., Lohweg, V.: Application-based approach for automatic texture defect recognition on synthetic surfaces. In: IEEE International Conference on Emerging Technologies and Factory Automation, vol. (19), pp. 229–232 (2008)
6. Mönks, U., Lohweg, V., Larsen, H.L.: Aggregation Operator Based Fuzzy Pattern Classifier Design. In: Lohweg, V., Niggemann, O. (eds.) Lemgo Series on Industrial Information Technology, vol. 3. IT – Institut Industrial IT, Lemgo (2009)
7. Lohweg, V., Li, R., Türke, T., Willeke, H., Schaede, J.: FPGA-based Multi-sensor Real time Machine Vision for Banknote Printing. In: 21st annual Symposium on IS&T/SPIE Electronc Imaging, San Jose, California USA (2009)
8. Rodner, E., Denzler, J.: Learning with Few Examples by Transferring Feature Relevance. In: Denzler, J., Notni, G., Süße, H. (eds.) DAGM 2009. LNCS, vol. 5748, pp. 252–261. Springer, Heidelberg (2009)
9. Drobics, M., Bodenhofer, U., Peter Klement, E.: FS-FOIL: an inductive learning method for extracting interpretable fuzzy descriptions. International Journal of Approximate Reasoning 32(2-3), 131–152 (2003)
10. Eichhorn, K.: Entwurf und Anwendung von ASICs für musterbasierte Fuzzy-Klassifikationsverfahren am Beispiel eines Schwingungsüberwachungssystems. PhD thesis, Technische Universität Chemnitz (2000)
11. Salicone, S.: Measurement Uncertainty: An Approach via the Mathematical Theory of Evidence. Springer Series in Reliability Engineering. Springer Science+Business Media LLC, Boston (2007)
12. Polyanin, A.D., Manzhirov, A.V.: Handbook of mathematics for engineers and scienctists. Chapman & Hall/CRC, Boca Raton (2007)
13. Dujmović, J.J., Larsen, H.L.: Generalized conjunction/disjunction. International Journal of Approximate Reasoning 46(3), 423–446 (2007)
14. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Transactions on Systems, Man and Cybernetics 18(1), 183–190 (1988)
15. Larsen, H.L.: Efficient Andness-Directed Importance Weighted Averaging Operators. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 11(suppl. 1), 67–82 (2003)
16. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)

Temporal Linguistic Summaries of Time Series Using Fuzzy Logic

Janusz Kacprzyk and Anna Wilbik*

Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
{kacprzyk,wilbik}@ibspan.waw.pl

Abstract. We consider linguistic summaries of time series used for an analysis of the past performance of investment (mutual) funds to help make future investment decisions. We use results from psychology, cognitive sciences and human decision making, which indicate a crucial role of time in the sense that means and ends, like decisions and outcomes, have a varying relevance and impact depending on the time when they occur, notably that what occurs in a more immediate past is more relevant and meaningful that what has occurred earlier. We propose to take into account some of psychological findings related to the importance of time by using different protoforms of linguistic summaries, *temporal linguistic summaries*, a substantial extension of the protoforms employed in our previous works. We consider two types of temporal protoforms exemplified by "Recently, among all segments, most are slowly increasing", and exemplified by "Initially, among all short segments, most are quickly decreasing". We compare them with the traditional ones, and present examples of their use for the analyses of investment funds.

1 Introduction

This paper is a continuation of our previous works (cf. Kacprzyk, Wilbik, Zadrożny [1,2,3] or Kacprzyk, Wilbik [4,5,6]) which deal with the problem of how to effectively and efficiently support a human decision maker in making decisions concerning investments in some financial, notably in investment (mutual) funds. Decision makers are here basically interested in future gains/losses. However, we follow the decision support paradigm, that is, assume primarily the user autonomy and a need to support, not replace, him/her. We are not concerned with forecasting the future daily prices.

This information is related to the history, or past, and this implies some problems. Basically, from our perspective, in all investment decisions the future is what really counts, and the past is irrelevant. But, the past is what we know, and the future is (completely) unknown. Behavior of the human being is to a large extend driven by his/her (already known) past experience. People usually

* Partially supported by the Polish Ministry of Science and Higher Education under Grant No. NN516 4309 33.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 436–445, 2010.

tend to assume that what has happened in the past will also happen (to some, maybe large extent) in the future. By the way, this is the underlying assumption behind the statistical methods too! That attitude clearly implies that the past can be employed to help the human decision maker find a good solution. We follow here this path, i.e. we present a method to subsume the past, to be more specific the past performance of an investment (mutual) fund, by presenting results in a very human consistent way, using natural language statements.

This line of reasoning has often been articulated by many well known investment practitioners, and one can quote here some more relevant opinions. In any information leaflets of investment funds, one may always notice a disclaimer stating that "Past performance is no indication of future returns" which is true. However, on the other hand, for instance, in a well known posting "Past Performance Does Not Predict Future Performance" [7], they state something that may look strange in this context, namely: "... according to an Investment Company Institute study, about 75% of all mutual fund investors mistakenly use short-term past performance as their primary reason for buying a specific fund". But, in an equally well known posting "Past performance is not everything" [8], they state: "... disclaimers apart, as a practice investors continue to make investments based on a scheme's past performance. To make matters worse, fund houses are only too pleased to toe the line by actively advertising the past performance of their schemes leading investors to conclude that it is the single-most important parameter (if not the most important one) to be considered while investing in a mutual fund scheme".

As strange as this may be, we may ask ourselves why it is so. Again, in a well known posting "New Year's Eve: Past performance is no indication of future return" [9], they say "... if there is no correlation between past performance and future return, why are we so drawn to looking at charts and looking at past performance? I believe it is because it is in our nature as human beings ... because we don't know what the future holds, we look toward the past ...".

And, continuing along this line of reasoning, we can find many other examples of similar statements supporting our position. For instance, Myers [10] says: "... Does this mean you should ignore past performance data in selecting a mutual fund? No. But it does mean that you should be wary of how you use that information ... *Lousy performance in the past is indicative of lousy performance in the future...*". And, further: Bogle [11] states: "... there is an important role that past performance can play in helping you to make your fund selections. While you should disregard a single aggregate number showing a fund's past long-term return, you can learn a great deal by studying the *nature of its past returns*. Above all, look for consistency.". In [12], we find: "While past performance does not necessarily predict future returns, it can tell you how volatile a fund has been".

We can quote more, and basically all of them emphasize the importance of looking at the past to help make future decisions, and also generally advocate a more comprehensive look not focused on single values but a very essence of past behavior and returns.

We have followed this line of reasoning in our past papers (cf. Kacprzyk, Wilbik, Zadrożny [1,2,3] or Kacprzyk, Wilbik [4,5,6]), i.e. to try to find a human consistent, fuzzy quantifier based scheme for a linguistic summarization of the past in terms of various aspects of how the time series representing daily quotations of the investment fund(s) behave. However, we have mainly concentrated on a sheer absolute performance, i.e. the time evolution of the quotations themselves. This may be relevant, and sometimes attractive to the users who can see a summary of their gains/loses and their temporal evolution. One can also use a maybe more realistic approach to take into account benchmarks of the particular funds as points of departure which does not change the essence.

Though the use of linguistic data summaries of past performance of the time series representing mutual fund quotations does take into account the importance (or "value") of time, in this paper we will go deeper into this issue by using some results from psychology, cognitive sciences and human decision making. Basically, we will employ some results by Ariely and Zakay [13] who consider the role of time in decision making.

In our case, those psychological analyses will serve the purpose of suggesting, and/or justifying a new types of protoforms of linguistic summaries of time series. Basically, in our recent works (cf. Kacprzyk, Wilbik, Zadrożny [1,2,3] or Kacprzyk, Wilbik [4,5,6]) we have used the following protoforms of the linguistic summaries of times series: "Among all y's, Q are P", exemplified by "among all segments (of the time series) most are slowly increasing", and "Among all R segments, Q are P", exemplified by "among all short segments almost all are quickly decreasing".

However, since in our case the analysis of time series is a highly human focused activity because its very purpose is to provide a human decision maker with some support for making (future) decision, we should take into account some inherent characteristics of time series and their evaluations that are consistent with the human perception of their relevance for the decision making process. One of the crucial aspects in this respect, which will be considered here is the importance of time in the sense that means and ends, like decisions and outcomes, have a carrying relevance and impact depending on the time moment when they occur. Basically, in virtually all cases what occurs in a more immediate past is more relevant and meaningful that what has occurred earlier. This temporal relationships change both the decisions and their evaluation as has been shown in psychology (cf. Ariely and Zakay [13] or Rachlin [14]). Among many approaches one can mention, for instance, a so called *temporal construal theory* by Liberman and Trope [15] who have shown that options are evaluated differently depending on time instants they come into question. They introduce the two main characteristics of options: desirability, which refers to long time wishes or intentions that are far away of their implementation of a decision option, and feasibility, which refers to a short term, close to the implementation characteristics. One can mention other works concerned with similar issues. It should be noted that this fact has already been reflected in (dynamic, or multistage) decision making and control models in which discounting is widely used.

In our context, we propose to take into account some of those psychological findings related to the importance of time by using different protoforms of linguistic summaries of times series, called *temporal linguistic summaries*. We consider two types of temporal protoforms: "E_T among all y's Q are P", exemplified by "Recently, among all segments, most are slowly increasing", and "E_T among all Ry's Q are P", exemplified by "Initially, among all short segments, most are quickly decreasing"; they both go beyond the classic Zadeh's protoforms. We will present formally those new temporal protoforms, compare them with the traditional ones, and present examples of their use.

2 Linguistic Summaries of Time Series

A linguistic summary of data (database) is a (usually short) sentence (or a few sentences), that captures the very essence of the data, that is numeric, large and because of its size not comprehensible for human users. We use here Yager's [16] basic approach, and a linguistic summary includes: a summarizer P, a quantity in agreement Q, i.e. a linguistic quantifier, truth (validity) $\mathcal{T}$ of the summary and optionally, a qualifier R. Thus, basically the core of a linguistic summary is a linguistically quantified proposition in the sense of Zadeh [17] which may be written, respectively as

$$Qy's \text{ are } P \qquad\qquad QRy's \text{ are } P \tag{1}$$

They may be exemplified, respectively by: "*Most* of employees earn *low* salary", $\mathcal{T}$=0.7, or "*Most* of *young* employees earn *low* salary", $\mathcal{T}$=0.82.

In our approach we focus on trends, linear segments extracted from the time series, obtained via using a piecewise linear segmentation method (cf. [18,19]). We consider the following three features of (global) trends in time series: (1) dynamics of change, (2) duration, and (3) variability. By *dynamics of change* we understand the speed of change of the consecutive values of time series. It may be described by the slope of a line representing the trend, represented by a linguistic variable. *Duration* is the length of a single trend, and is also represented by a linguistic variable. *Variability* describes how "spread out" a group of data is. We compute it as a weighted average of values taken by some measures used in statistics: (1) the range, (2) the interquartile range (IQR), (3) the variance, (4) the standard deviation, and (5) the mean absolute deviation (MAD). This is also treated as a linguistic variable.

For practical reasons for all we use a fuzzy granulation (cf. Bathyrshin at al. [20,21]) to represent the values by a small set of linguistic labels as, e.g.: increasing, slowly increasing, constant, slowly decreasing, decreasing, which are equated with fuzzy sets.

For clarity and convenience we employ Zadeh's [22] protoforms for dealing with linguistic summaries [23]. A protoform is defined as a more or less abstract prototype (template) of a linguistically quantified proposition. We have two types of protoforms of linguistic summaries of trends:

– a simple (short) form:

$$\text{Among all segments, } Q \text{ are } P \tag{2}$$

e.g.: "Among all segments, *most* are *slowly increasing*".
– an extended form:

$$\text{Among all } R \text{ segments, } Q \text{ are } P \tag{3}$$

e.g.: "Among all *short* segments, *most* are *slowly increasing*".

The quality of linguistic summaries can be evaluated in many different ways (cf. [4,5,6]). However the basic criterion is the truth value (a degree of truth or validity), introduced by Yager in [16]. It describes the degree of truth (from $[0, 1]$) to which a linguistically quantified proposition equated with a linguistic summary is true.

Using Zadeh's calculus of linguistically quantified propositions [17] it is calculated in dynamic context using the same formulas as in the static case. Thus, the truth value is calculated for the simple and extended form as, respectively:

$$\mathcal{T}(\text{Among all } y\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_P(y_i)\right) \tag{4}$$

$$\mathcal{T}(\text{Among all } Ry\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n}\mu_R(y_i)\wedge\mu_P(y_i)}{\sum_{i=1}^{n}\mu_R(y_i)}\right) \tag{5}$$

where $\wedge$ is the minimum operation (or, for instance, a t-norm, cf. Kacprzyk, Wilbik and Zadrożny [24]). It seems that the minimum operation is a good choice since it can be easily interpreted and the numerical values correspond to the intuition.

3 Temporal Protoforms

We can extend our protoforms given in (2) and (3) by adding a temporal expression E_T like: "recently", "in the very beginning" or "in May 2010", "initially", ... The temporal protoforms have the following forms:
– a simple (short) form:

$$E_T \text{ among all segments, } Q \text{ are } P \tag{6}$$

e.g.: "*Recently* among all segments, *most* are *slowly increasing*".
– an extended form:

$$E_T \text{ among all } R \text{ segments, } Q \text{ are } P \tag{7}$$

e.g.: "*Initially* among all *short* segments, *most* are *slowly increasing*".

To evaluate the quality of those summaries, the most important quality criterion is again the truth value. The computation of truth values of temporal

summaries is very similar to the previous case. We only need to consider temporal expression as an additional external qualifier, as the temporal expression limit the universe of interest, only to the trends (segments) that occur on the time axis described by a fuzzy set modeling expression E_T. We compute the proportion of segments in which "trend is P" and occurred in E_T to those that occurred in E_T. Next we compute the degree to which this proportion is Q.

Truth value of the simple temporal protoform (6) is computed as:

$$\mathcal{T}(E_T \text{ among all } y\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n} \mu_{E_T}(y_i) \wedge \mu_P(y_i)}{\sum_{i=1}^{n} \mu_{E_T}(y_i)}\right) \tag{8}$$

where $\mu_{E_T}(y_i)$ is degree to which a trend (segment) occurs during the time span described by E_T. Similarly we compute the truth of the extended temporal protoform (7) as:

$$\mathcal{T}(E_T \text{Among all } Ry\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n} \mu_{E_T}(y_i) \wedge \mu_R(y_i) \wedge \mu_P(y_i)}{\sum_{i=1}^{n} \mu_{E_T}(y_i) \wedge \mu_R(y_i)}\right) \tag{9}$$

A natural question emerges, how to compute $\mu_{E_T}(y_i)$. Let $\mu_{E_T}(t)$ be a membership function of a fuzzy set representing a linguistic variable E_T. We assume that the considered time span is normalized, i.e. $t \in [0, 1]$, the first observation is made for $t = 0$ and the last for $t = 1$. Let us consider a segment y_i, starting at time a and terminating at time b, $0 \leq a < b \leq 1$. Then

$$\mu_{E_T}(y_i) = \frac{1}{b-a}\int_a^b \mu_{E_T}(t)dt \tag{10}$$

and we can interpret this value as the average membership degree of E_T in $[a, b]$.

For clarity, in this paper – in which we introduce a new concept of temporal protoforms – only the truth value is introduced. In many cases this is not sufficient to differentiate between the summaries obtained and to improve the accuracy of summary evaluation we have to use various quality criteria exemplified by the degree of specificity, degree of appropriateness or measure of informativeness. This issue will be considered in subsequent papers.

4 Numerical Results

The method proposed was tested on data on quotations of an investment (mutual) fund that invests at least 50% of assets in shares listed at the Warsaw Stock Exchange.

Data shown in Figure 1 were collected from January 2002 until the December 2009 with the value of one share equal to PLN 12.06 in the beginning of the period to PLN 35.82 at the end of the time span considered (PLN stands for the Polish Zloty). The minimal value recorded was PLN 9.35 while the maximal one during this period was PLN 57.85. The biggest daily increase was equal to PLN 2.32, while the biggest daily decrease was equal to PLN 3.46. We illustrate the

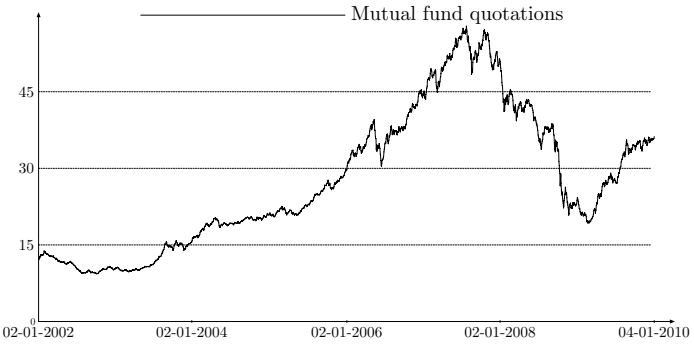


Fig. 1. Daily quotations of an investment fund in question

method proposed by analyzing the absolute performance of a given investment fund, and not against benchmarks, for illustrativeness.

We obtain 362 extracted trends, with the shortest of 1 time unit only, and the longest – 71 time units. We assume 3 labels only for each attribute: short, medium and long for duration, increasing, constant and decreasing for dynamics and low, moderate and high for variability. The use of linguistic values in the summaries is clearly a reflection of a natural information granulation.

The summaries as presented in Section 2 are presented in Table 1. They are ordered according to the truth value.

Table 1. Linguistic summaries

linguistic summary	$\mathcal{T}$
Among all low-variability y's, most are short	1.0000
Among all decreasing y's, almost all are short	1.0000
Among all increasing y's, almost all are short	1.0000
Among all increasing y's, most are low-variability	1.0000
Among all medium y's, most are constant	1.0000
Among all short and decreasing y's, most are low-variability	0.9834
Among all increasing y's, most are short and low-variability	0.9819
Among all y's, most are short	0.9718
Among all y's, most are low-variability	0.9056
Among all moderate-variability y's, most are short	0.8492
Among all constant y's, most are low-variability	0.8085
Among all medium and constant y's, most are low-variability	0.7785
Among all short and increasing y's, almost all are low-variability	0.7657
Among all medium and low-variability y's, almost all are constant	0.7601
Among all medium y's, most are low-variability	0.7353

The temporal linguistic summaries describing the situation after the crisis begun (i.e. after September 2007) are shown in Table 2.

Table 2. Temporal linguistic summaries describing the situation after the crisis begun

linguistic summary	$\mathcal{T}$
After the crisis started among all y's, almost all are short	1.0000
After the crisis started among all short y's, most are low-variability	1.0000
After the crisis started among all low-variability y's, almost all are short	1.0000
After the crisis started among all decreasing y's, almost all are short	1.0000
After the crisis started among all short and decreasing y's, most are low-variability	0.9710
After the crisis started among all y's, most are low-variability	0.9706
After the crisis started among all y's, most are short and low-variability	0.8826
After the crisis started among all moderate-variability y's, almost all are short	0.8150
After the crisis started among all increasing y's, almost all are short and low-variability	0.7465
After the crisis started among all constant y's, almost all are short	0.7175

We may notice some differences between those descriptions. Summaries of the whole time series use more often the word increasing than those describing the crisis and latter time. Also if we compare length of the summaries, e.g. using the following two summaries:

- Among all y's, most are short, $\mathcal{T} = 0.9718$
- After the crisis started among all y's, almost all are short $\mathcal{T}$=1.000

we notice that more short trends (segments) appear in the crisis time. In case of variability, we can't see much difference, as the variability is generally low. Solution here could be adjusting the fuzzy sets describing the linguistic variables: low, moderate and high or the use of other more detailed granulation.

Table 3. Temporal linguistic summaries for the initial stage

linguistic summary	$\mathcal{T}$
Initially among all y's, almost all are constant	1.0000
Initially among all low-variability y's, almost all are constant	1.0000
Initially among all long y's, most are constant	1.0000
Initially among all medium y's, most are constant	1.0000
Initially among all moderate-variability y's, almost all are constant	1.0000
Initially among all long and low-variability y's, most are constant	1.0000
Initially among all medium and low-variability y's, most are constant	1.0000
Initially among all long and moderate-variability y's, most are constant	1.0000
Initially among all short y's, most are constant	1.0000
Initially among all long y's, most are constant and low-variability	0.9501
Initially among all medium y's, most are constant and low-variability	0.8213
Initially among all y's, most are low-variability	0.8161
Initially among all y's, most are constant and low-variability	0.8024
Initially among all y's, majority are long	0.7089
Initially among all y's, majority are long and constant	0.7089
Initially among all constant y's, majority are medium	0.7022

For comparison, in Table 3 we show summaries from the initial stage of the mutual fund, i.e. more or less two first years.

At the first glance we may see that those descriptions are different. In the initial stage many trends were long, this word does not appears in the description from the beginning of the crisis . Also the word constant occurs very often, and there is no word about any increases or decreases. The only common thing between those summaries is their low variability, but this can be caused by a very coarse, three values granulation.

5 Concluding Remarks

Using results from from psychology, cognitive sciences and human decision making, which clearly indicate a crucial role of time in the sense that means and ends, like decisions and outcomes, have a varying relevance and impact depending on the time moment when they occur, notably that what occurs in a more immediate past is more relevant and meaningful that what has occurred earlier, we proposed different protoforms of linguistic summaries of times series, *temporal linguistic summaries*, a substantial extension of the protoforms employed in our previous works. We considered two types of temporal protoforms: "E_T among all y's Q are P", exemplified by "Recently, among all segments, most are slowly increasing", and "E_T among all Ry's Q are P", exemplified by "Initially, among all short segments, most are quickly decreasing". We formally presented those new temporal protoforms, compared them with the traditional ones, and showed examples of their use for the analyses of investment funds.

As a possible future direction, we intend to more explicitly relate our model to other findings in psychology and cognitive science.

References

1. Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic summarization of trends: a fuzzy logic based approach. In: Proceedings of the 11th International Conference Information Processing and Management of Uncertainty in Knowledge-based Systems, pp. 2166–2172 (2006)
2. Kacprzyk, J., Wilbik, A., Zadrożny, S.: On some types of linguistic summaries of time series. In: Proceedings of the 3rd International IEEE Conference "Intelligent Systems", pp. 373–378. IEEE Press, Los Alamitos (2006)
3. Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets and Systems 159(12), 1485–1499 (2008)
4. Kacprzyk, J., Wilbik, A.: An extended, specificity based approach to linguistic summarization of time series. In: Proceedings of the 12th International Conference Information Processing and Management of Uncertainty in Knowledge-based Systems, pp. 551–559 (2008)
5. Kacprzyk, J., Wilbik, A.: A new insight into the linguistic summarization of time series via a degree of support: Elimination of infrequent patterns. In: Dubois, et al. (eds.) Soft Methods for Handling Variability and Imprecision, pp. 393–400. Springer, Heidelberg (2008)

6. Kacprzyk, J., Wilbik, A.: Linguistic summarization of time series using linguistic quantifiers: augmenting the analysis by a degree of fuzziness. In: Proceedings of 2008 IEEE World Congress on Computational Intelligence, pp. 1146–1153. IEEE Press, Los Alamitos (2008)
7. Past performance does not predict future performance, http://www.freemoneyfinance.com/2007/01/past_performanc.html
8. Past performance is not everything, www.personalfn.com/detail.asp?date=9/1/2007
9. New year's eve:past performance is no indication of future return stockcasting, blogspot.com/2005/12/new-years-evepast-performance-is-no.html
10. Myers, R.: Using past performance to pick mutual funds. Nation's Business (1997), findarticles.com/p/articles/mi_m1154/is_n10_v85/ai_19856416
11. Bogle, J.C.: Common Sense on Mutual Funds: New Imperatives for the Intelligent Investor. Wiley, New York (1999)
12. U.S. Securities and Exchange Commission: Mutual fund investing: Look at more than a fund's past performance, www.sec.gov/investor/pubs/mfperform.htm
13. Ariely, D., Zakay, D.: A timely account of the role of duration in decision making. Acta Psychologica 108(2), 187–207 (2001)
14. Rachlin, H.: Judgment, Decision, and Choice: A Cognitive/Behavioral Synthesis. W.H. Freeman & Company, New York (1989)
15. Liberman, N., Trope, Y.: The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory. Journal of Personality and Social Psychology 75, 5–18 (1998)
16. Yager, R.R.: A new approach to the summarization of data. Information Sciences 28, 69–86 (1982)
17. Zadeh, L.A.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Systems 9(2), 111–127 (1983)
18. Keogh, E., Chu, S., Hart, D., Pazzani, M.: An online algorithm for segmenting time series. In: Proceedings of the 2001 IEEE International Conference on Data Mining (2001)
19. Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting time series: A survey and novel approach. In: Last, M., Kandel, A., Bunke, H. (eds.) Data Mining in Time Series Databases. World Scientific Publishing, Singapore (2004)
20. Batyrshin, I.: On granular derivatives and the solution of a granular initial value problem. International Journal Applied Mathematics and Computer Science 12(3), 403–410 (2002)
21. Batyrshin, I., Sheremetov, L.: Perception based functions in qualitative forecasting. In: Batyrshin, et al. (eds.) Perception-based Data Mining and Decision Making in Economics and Finance. Springer, Heidelberg (2006)
22. Zadeh, L.A.: A prototype-centered approach to adding deduction capabilities to search engines – the concept of a protoform. In: Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS 2002), pp. 523–525 (2002)
23. Kacprzyk, J., Zadrożny, S.: Linguistic database summaries and their protoforms: toward natural language based knowledge discovery tools. Information Sciences 173, 281–304 (2005)
24. Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic summarization of time series under different granulation of describing features. In: Kryszkiewicz, M., Peters, J.F., Rybiński, H., Skowron, A. (eds.) RSEISP 2007. LNCS (LNAI), vol. 4585, pp. 230–240. Springer, Heidelberg (2007)

A Comparison of Five Fuzzy Rand Indices

Derek T. Anderson[1], James C. Bezdek[1], James M. Keller[1], and Mihail Popescu[2]

[1] Electrical and Computer Engineering Department, University of Missouri Columbia, MO, 65211, USA
[2] Informatics Institute, University of Missouri Columbia, MO, 65211, USA
dtaxtd@mail.missouri.edu, jcbezdek@gmail.com, {kellerj,popescum}@missouri.edu

Abstract. Five papers have appeared in the last three years that propose different fuzzy generalizations of Rand's classical comparison index for crisp clustering algorithms. We review the five generalizations, compare their complexities, and then give two numerical examples to compare their performance. Our extension (for the pairwise agreements) is $\boldsymbol{O}(n)$, while the other four generalizations are $\boldsymbol{O}(n^2)$.

Keywords: Cluster validity, Rand index, Fuzzy Rand Index.

1 Introduction

Let $O=\{o_1,\ldots,o_n\}$ denote n objects (fish, cigars, motorcycles, beers, etc.). When each object in O is represented by a (column) vector $\mathbf{x}$, the set $X=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}\subset\Re^p$ is an *object data representation* of O. When each object in $o_i \in O$ has a *physical label*, O is a set of *labeled data*; otherwise, O is unlabeled. Let integer c denote the number of classes, $1 < c < n$. Clustering in unlabeled data is the assignment of one of four types of labels to each object in O. The label vectors of the objects are the columns of c-partitions of O, which are sets of (cn) values $\{u_{ik}\}$ that can be conveniently arrayed as $(c \times n)$ matrices, say $U = [u_{ik}]$. The three sets are:

$$M_{pcn}=\left\{U\in\Re^{cn}: u_{ik}\in[0,1]\,\forall\, i,k;\ 0<\sum_{k=1}^{n}u_{ik}\ \forall\, i\right\}=\text{possibilistic c-partitions;} \quad (1a)$$

$$M_{fcn}=\left\{U\in M_{pcn}:\sum_{i=1}^{c}u_{ik}=1\,\forall\, k\right\}=\text{fuzzy or probabilistic c-partitions;} \quad (1b)$$

$$M_{hcn}=\left\{U\in M_{fcn}: u_{ik}\in\{0,1\}\,\forall i,k\right\}=\text{crisp or hard c-partitions.} \quad (1c)$$

It is convenient to have a single name for the set M_{pcn}–M_{hcn}, which contains the fuzzy, probabilistic, and possibilistic c-partitions of O. We call M_{pcn}–M_{hcn} the *soft c-partitions* of O. Clustering algorithms map $X\subset\Re^p$ or $R\subset\Re^{nn}\mapsto M_{pcn}$. Let $CP = \{U_i: 1 \le i \le N\}$ denote N different *candidate partitions* of a fixed object set O that may arise as a result of clustering (X or R) with one algorithm at various values of its

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 446–454, 2010.

parameters; or more generally, with different algorithms, each with its own parameters. Which $U \in CP$ best explains and represents the (unknown) structure in O? This article is about one method for answering this question. Many other methods are nicely discussed in [1-4].

One group of methods for this problem use *comparison indices*, s(U,V). There are various ways to use such indices [5, 6]. The only application we consider in this note is when U is an algorithmically obtained partition, and V is a *reference partition* that purports to represent the "true cluster structure" in O. In this case s(U,V) measures the extent to which U's in CP recover or retrieve the "true" clusters in O, and hence, the sizes of U and V are equal.

We never have an external reference partition in a real clustering situation which, by definition, involves unlabeled data. So, why do this at all? Well, the only way you can evaluate *any* clustering algorithm before using it in a real situation is to see how well it recovers "true but unknown" reference partitions. If nothing else, good recovery rates on data with "known" cluster structure at least provide some psychological reassurance that the clustering algorithm *can* sometimes recover "good clusters".

2 Comparison Indices and the Contingency Table for (U,V)

Let $U \in M_{hrn}$ and $V \in M_{hcn}$ be crisp partitions of O. U and V need not possess the same number of clusters, $r \neq c$. The four classical combinations for pairs of objects from $O \times O$ in clusters of U and V are: (i) paired in U and V; (ii) not paired in U nor in V; (iii) paired in V but not in U; and (iv) paired in U but not in V [6, p. 194]. The comparison of U to V with a similarity measure s begins with the $r \times c$ contingency matrix $N = UV^T$ shown in Table 1 that contains counts of the number of occurrences of each of the four types over the n(n-1)/2 distinct, unordered pairs in $O \times O$. Entry n_{ij} is the number of objects common to classes U_i and V_j.

Table 1. The contingency matrix N

Partition V: V_j = row j of V

	Class	V_1	V_2	...	V_c	Sums
Partition U U_i = row i of U	U_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\bullet}$
	U_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	U_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\bullet}$
	Sums	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet c}$	$n_{\bullet\bullet} = n$

$N = UV^T$

The building blocks of many similarity measures for s(U,V) are the four equations (2a)-(2d). These four equations simply count the number of occurrences amongst the n(n-1)/2 pairs of each of the four types of unordered pairs.

$$a = \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(n_{ij}-1);\ \text{number paired in U and V;} \tag{2a}$$

$$d = \frac{1}{2}\left(n^2 + \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}^2 - \left(\sum_{i=1}^{r} n_{i\bullet}^2 + \sum_{j=1}^{c} n_{\bullet j}^2\right)\right);\ \text{number paired in neither U nor V;} \tag{2b}$$

$$b = \frac{1}{2}\left(\sum_{j=1}^{c} n_{\bullet j}^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}^2\right);\ \text{number paired in V, not U;} \tag{2c}$$

$$c = \frac{1}{2}\left(\sum_{i=1}^{r} n_{i\bullet}^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}^2\right);\ \text{number paired in U, not V.} \tag{2d}$$

The sums (a+d) and (b+c) are usually interpreted, respectively, as (the total number of) *agreements* and *disagreements* between U and V. Anderson et al. [17] tabulate a [non-exhaustive] list of 14 coefficients that have been proposed for s(U,V) based on functions of a, b, c and d; Sokal and Sneath [8] list many others. In this note, the only index we consider is Rand's index, the classical form of which is

$$s_r(U,V) = (a+d)/(a+b+c+d). \tag{3}$$

3 Generalizing s_r(U,V) When U and/or V Are Soft Partitions

Rand's index first appeared in Sokal and Michener in 1958, where it was called a simple matching coefficient [8]. Rand reintroduced this function in 1971 [5], and the literature has consistently referred to it as "Rand's Index" since then. The resurgence of Rand's index in bioinformatics [9-12] has renewed interest in generalizing it, along with some of the other comparison indices based on the elements in Table 1, to various non-crisp cases. Specifically, we mention the papers (in chronological order) of Campello [13, 2007], Frigui et al. [14, 2007], Brower [15, 2009], Hullermeier and Rifqi [16, 2009], and Anderson et al. [17, 2010]. All of these papers generalize the Rand index to the case of U and/or V being fuzzy partitions of the n objects. Next, we briefly review the method used to generalize (3) in each of these five articles.

3.1 Campello [13]

Campello presents a method for fuzzifying the indices of Rand, Jaccard, Fowlkes-Mallow, Hubert and (one version of) the adjusted Rand. Campello's scheme is based on writing equations (2) in an equivalent form using (cardinalities of) intersections of

the crisp subsets of O×O corresponding to each of the four totals, and then replacing the crisp sets with fuzzy ones. Campello's generalization of equation (2a) is:

$$\underbrace{a}_{\text{Campello}} = \sum_{j=1}^{i-1}\sum_{i=2}^{n}\left(\left(\bigvee_{k=1}^{r}\left(u_{ki}\wedge u_{kj}\right)\right)\wedge\left(\bigvee_{k=1}^{c}\left(v_{ki}\wedge v_{kj}\right)\right)\right); \tag{4}$$

3.2 Frigui et al. [14]

These authors present generalizations of the Rand, Jaccard, Fowlkes-Mallow and Hubert indices. They address only the special case where U and V are both $c\times n$ partitions of O and V is a crisp reference partition. Instead of using the elements from Table 1 to compute equations (2), they first convert U and V into $n\times n$ coincidence matrices, $U^{*} = U^{T}U;\ V^{*} = V^{T}V$. Frigui et al.'s generalization of equation (2a) is:

$$\underbrace{a}_{\text{Frigui etal.}} = \sum_{j=2}^{n}\sum_{k=1}^{j-1}\left(\sum_{i=1}^{c}u_{ij}u_{ik}\right)\left(\sum_{i=1}^{c}v_{ij}v_{ik}\right); \tag{5}$$

If product is used for t-norm and sum is used in place of t-conorm, then Frigui's approach is in effect Campello's [15]. However, Campello's approach is more general. It applies to the cases of fuzzy, probabilistic, and possibilistic U and/or V.

3.3 Brouwer [15]

Brouwer discusses another generalization of the Rand, a (third variant of) the adjusted Rand, and Jaccard's index. His approach is also based on formulating the two potentially very large $n\times n$ (bonding) matrices, $U^{*} = U^{T}U;\ V^{*} = V^{T}V$. However, instead of dot product for constructing bonding matrix terms, he instead uses cosine correlation, i.e. $u^{*}_{i,j} = \cos(\mu_i^j)$, where $\cos(\mu_i^j)$ is the angle between vectors $\mathbf{U}_i^T$ and $\mathbf{U}^j$. In the sequel, $\mathbf{A}_i$ and $\mathbf{A}^j$ denote the vectors corresponding to the i-th row and j-th column of any matrix A, and $\langle \mathbf{A}_i, \mathbf{A}^j \rangle$ is the dot product of these two vectors. Brouwer's generalization of equation (2a) is:

$$\underbrace{a}_{\text{Brouwer}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\langle U_i^T, U^j\rangle\langle V_i^T, V^j\rangle}{\|U_i^T\|\|U^j\|\,\|V_i^T\|\|V^j\|}\right)}{2} - \frac{n}{2} = \frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\cos(\mu_i^j)\cos(v_i^j)\right)\right) - n}{2}. \tag{6}$$

3.4 Hullermeier and Rifqi [16]

These authors consider only the Rand index. They argue that Campello's fuzzy Rand index is in some sense defective because it is not a metric. They do not formulate their index in terms of equations (2a)-(2d). Instead, their generalization is guided by the fact that Rand's index counts the number of paired agreements (a+d) divided by the total number of possible pairs (a+b+c+d), and this leads them to a direct generalization of the Rand index:

$$s_{FRHR}(U,V)=1-\left[\sum_{j=i+1}^{n}\sum_{i=1}^{n-1}\left|\left\|V^i-V^j\right\|-\left\|U^i-U^j\right\|\right|\Big/\binom{n}{2}\right]. \quad (7)$$

3.5 Anderson et al. [17]

This paper provides generalizations for 14 comparison indices. It begins by forming the contingency matrix as $N = UV^T$. Anderson et al. note that a modification is needed to accommodate the case when U and/or V are possibilistic. In the possibilistic case, we can have $\sum_{i=1}^{r} n_{i\bullet} > n$, or $\sum_{j=1}^{c} n_{\bullet j} > n$. One or both of the terms $\sum_{i=1}^{r} n_{i\bullet}^2$ and $\sum_{j=1}^{c} n_{\bullet j}^2$ can make d at (2b) relatively large and negative. Depending on b at (2c) and c at (2d), the (soft) Rand index can result in $s_r(U,V) < 0$ or >1. To remedy this, they scale N with $\phi = n/\sum_{i=1}^{r} n_{i\bullet}$ or $\varphi = n/\sum_{j=1}^{c} n_{\bullet j}$. Since $\sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet j}$, $\varphi = \phi$. These authors base their generalization on $N^* = \phi UV^T = [n/\sum_{i=1}^{r} n_{i\bullet}]UV^T$. An advantage of this scaling is that when U and V are crisp, fuzzy or probabilistic partitions, ϕ=1, thus $N^* = N = UV^T$. This shows that *ANY* index based on only the elements of Table 1 will reduce to the original index when U and V are both crisp partitions of the n objects; and otherwise, they will be valid soft generalizations of those indices. Moreover, in the case of possibilistic partitions, the normalization produces index values in the range [0,1]. Anderson et al.'s generalization of (2a) using $N^* = \phi UV^T$ is:

$$\underbrace{a}_{\text{Anderson,et al.}} = \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{n\left\langle U_i,(V^T)^j\right\rangle}{\sum_{k=1}^{r}\sum_{p=1}^{c}\left\langle U_k,(V^T)^p\right\rangle}\right)\left(\left(\frac{n\left\langle U_i,(V^T)^j\right\rangle}{\sum_{k=1}^{r}\sum_{p=1}^{c}\left\langle U_k,(V^T)^p\right\rangle}\right)-1\right). \quad (8)$$

4 Examples

The following two examples compare the five fuzzy Rand indices in Sections 3.1-3.5. Candidate fuzzy partitions are generated by the fuzzy c-means (FCM, [1]) algorithm using the *fcm* function from the MATLAB Fuzzy Logic Toolbox with c = 2, 3,…,10, m = 2, maximum number of iterations MAXIT = 100, objective function error EPS=1e-5 and random partition initialization. Scatterplots of the two data sets, X1 and X2, are shown in Figure 1.

In data set X1, each cluster of 500 points is a sample from a mixture of c=6 equiprobable Gaussian distributions in two dimensions. Means of the six component densities are: $[20\ 20]^T$, $[1\ 40]^T$, $[20\ 70]^T$, $[40\ 6]^T$, $[40\ 30]^T$, and $[60\ 50]^T$, and the common covariance matrix was $\Sigma=1.7I_6$. These clusters are fairly compact and well-separated. We expect the best partition in CP to occur at c = 6. The reference partition V is the crisp 6x3000 partition with six "diagonal" blocks of 500 1's in each row.

Data set X2 is three well separated parallel line clusters of size 200 each. Samples are generated according to $\vec{c}+\alpha(\vec{d}-\vec{c})$, where $\alpha \in [0,1]$ is a uniformly distributed random number and $(\vec{c},\vec{d})$ are line segment endpoints. Clusters in X2 include: $([1\ 1]^T$, $[1\ 10]^T)$, $([6\ 2]^T, [6\ 6]^T)$, and $([10\ 0]^T, [10\ 12]^T)$. We expect the best partition in CP to occur at c = 3. The reference partition V is the crisp 3x600 partition with three "diagonal" blocks of 200 1's in each row.

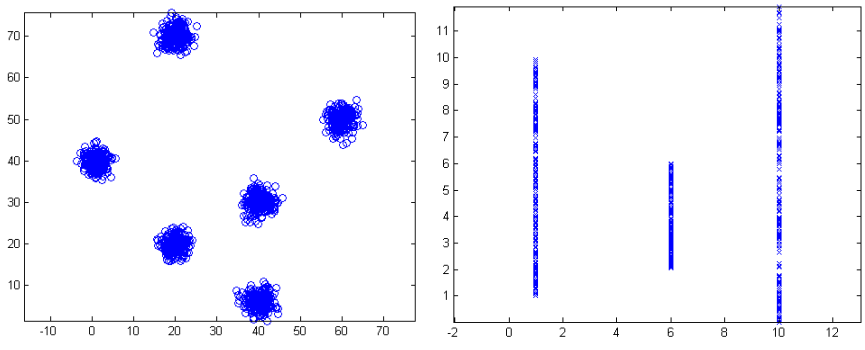


Fig. 1. Gaussian (X1) and parallel line (X2) data sets used for the two examples

Figure 2 shows graphs of the five Fuzzy Rand indices for terminal FCM partitions on X1 as c varies from 2 to 10, so there are 9 candidate partitions in CP. This graph shows two things: first, the five indices are indeed different; and second, they have similar values on this well behaved data set. All five indices have clear maximums at c=6 which points to the most preferable partition in CP. Figure 2 might tempt you to

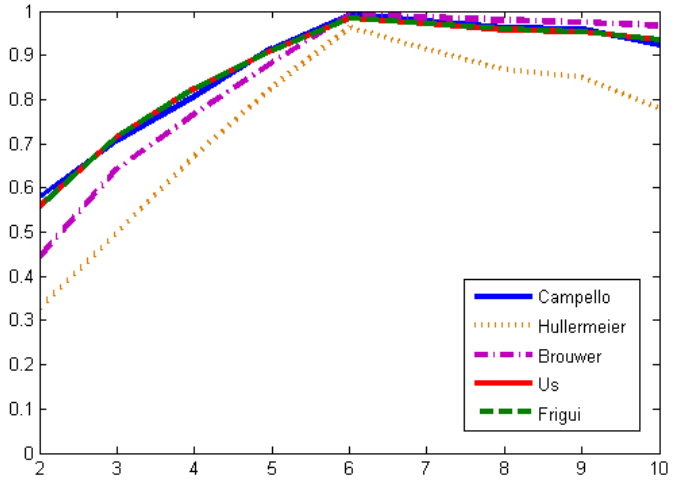


Fig. 2. Comparison of the five fuzzy Rand indices on FCM

conjecture that Hullermeier and Rifki is bounded above by the other four indices, but we have not attempted a proof of this. The other indices all cross each other.

Figure 3 shows graphs of the five Fuzzy Rand indices for terminal FCM partitions on X2 as c varies from 2 to 10. The graph again shows that the five indices are indeed different and they have similar values on this data set. We expect FCM to fail on this example. The crisp V for c = 3 will NOT match well with the FCM c = 3 partition. As expected, no indices have a clear maximum at c = 3. Hullermeier and Rifki have a maximum at c = 4 and the others are at c = 5. Also, Hullermeier and Rifki appear to again be bounded above by the other four indices, which all cross each other.

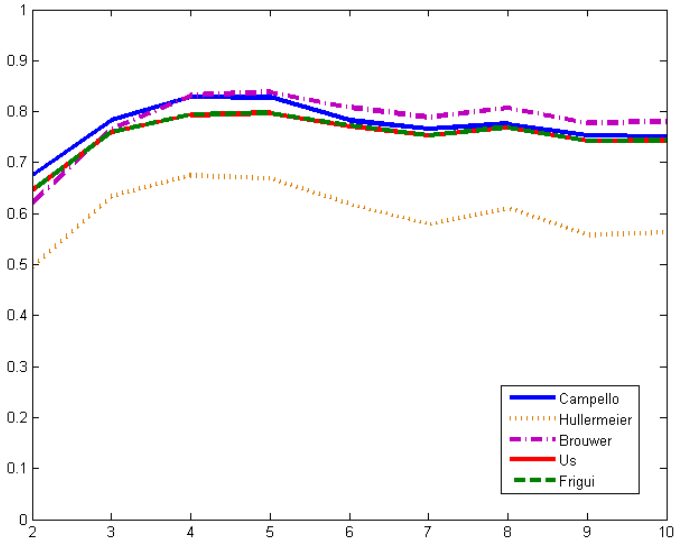


Fig. 3. Comparison of the five fuzzy Rand indices on FCM

5 Computational Complexity

Assuming similar cost for different operations, the cost of evaluation of formula (2a) for all but the index of Hullermeier and Rifqi (who form the fuzzy Rand index directly) are reported in Table 2. The example in the last column of Table 2 shows that in terms of computational costs, Anderson et al.'s method is (at least 3 and at best 5) orders of magnitude less than the other generalizations of Rand's index. Computation of the Rand index at (3) involves calculation of all four equations, (2a)-(2d). Combining the factors as in (3) uses only addition and subtraction, and will cost all methods equally. Hence, we can extend the results of Table 2 from just equation (2a) to equation (3) without loss.

Table 2. Computational complexity for the five fuzzy Rand indices

Method	Computational Complexity		n=1000, r = c = 5
Anderson et al. [17] for (2a)	**O**(2rcn+3rc)	**O**(n)	50,075
Brouwer [15] for (2a) (assuming cosine as a single operation)	**O**($2n^2$+n+2)	**O**(n^2)	2,001,002
Brouwer [15] for (2a) (using dot product and magnitude form)	**O**($4n^2$+ rn^2+ cn^2-n+3rn+3cn+2)	**O**(n^2)	14,029,002
Campello [13] for (2a)	**O**(rn^2+cn^2+(n-n^2)/2-rn-cn)	**O**(n^2)	9,490,500
Hullermeier and Rifqi [16] (for the fuzzy Rand)	**O**(($3rn^2$+$3cn^2$-3rn-3cn)/2+5)	**O**(n^2)	14,985,005
Frigui et al. [14] for (2a)	**O**(rn^2+cn^2+(n-n^2)/2-rn-cn)	**O**(n^2)	9,490,500

6 Discussion and Conclusions

We compared our generalization of the classical Rand index with four other fuzzy generalizations of it both experimentally, and in terms of computational complexity. Our extension of the Rand index is $\boldsymbol{O}(n)$, while the other four are all $\boldsymbol{O}(n^2)$. More examples using different types of data, algorithms and other indices that involve comparing partitions appear in [17]. The advantage of using $N^* = \phi UV^T$ is that this formulation directly generalizes *all* indices that depend only on equations (2) to every combination of (U, V). There are sixteen possible pair types according as each of U, V are crisp, fuzzy, probabilistic or possibilistic, so we have, for example, 16 Rand indices, 16 Jaccard indices, and so on. Each formula there is recovered when U and V are crisp, i.e., these are true generalizations to every case - by definition.

The use of comparison indices for validation of clustering algorithms has the significant advantage of being independent of the correspondence problem for comparing clustering solutions to known reference partitions. When U is soft, one approach to retrieval assessment is to first harden any soft partition U. Then the hardened version of U, say H(U), defines the function $s_e(H(U),V) = 1 - \sum_{k=1}^{n} \left\| [H(U)]^k - V^k \right\|_1 \Big/ 2n,$ which counts the number of label matches. This comparison method is *similar* to assessment by s(U,V), but before using s_e(H(U),V), we *must* register the reference clusters to their algorithmic counterparts. This complication is avoided by s(U,V), because the indices in Table 2 depend only on the values in Table 1; double sums, row sums, or column sums of the entries of $N=UV^T$. Consequently, crisp comparison indices such as Rand's index are *independent of the correspondence problem* which plagues evaluation of retrieval success for soft clustering algorithms by the "harden and count" method represented by s_c(H(U), V). This important advantage for the comparison index method remains true even when $r \neq c$ and the resubstitution error rate cannot even be computed!

Acknowledgments

Derek Anderson is a pre-doctoral biomedical informatics research fellow funded by the National Library of Medicine (T15 LM07089).

References

1. Bezdek, J.C., Keller, J.M., Krishnapuram, R., Pal, N.R.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer, Norwell (1999)
2. Jain, A., Dubes, R.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
3. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press, NY (2006)
4. Hoppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis. Wiley and Sons, Chichester (1999)
5. Rand, W.M.: Objective criteria for the evaluation of clustering methods. JASA 66(336), 846–850 (1971)
6. Hubert, L.J., Arabie, P.: Comparing partitions. J. Classification 2, 193–218 (1985)
7. Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy - The Principles and Practice of Numerical Classification. W. H. Freeman, San Francisco (1973)
8. Sokal, R.R., Michener, C.D.: A Statistical Method for Evaluating Systematic Relationships. The University of Kansas Scientific Bulletin 38, 1409–1438 (1958)
9. Duan, F., Zhang, H.: Correcting the loss of cell-cycle synchrony in clustering analysis of microarray data using weights. Bioinformatics 20(11), 1766–1771 (2004)
10. Thalamuthu, A., Mukhopadhyay, I., Zheng, X., Tseng, G.: Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics 22(19), 2405–2412 (2006)
11. Wong, D.S.V., Wong, F.K., Wood, G.R.: A multi-stage approach to clustering and imputation of gene expression profiles. Bioinformatics 23(8), 998–1005 (2007)
12. Yu, Z., Wong, H.S., Wang, H.: Graph-based consensus clustering for class discovery from gene expression data. Bioinformatics 23(21), 2888–2896 (2007)
13. Campello, R.J.G.B.: A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Patt. Recog. Letters 28, 833–841 (2007)
14. Frigui, H., Hwang, C., Rhee, F.: Clustering and aggregation of relational data with applications to image database categorization. Pattern Recognition (40), 3053–3068 (2007)
15. Brower, R.K.: Extending the Rand, adjusted Rand, and Jaccard indices to fuzzy partitions. J. Intell. Inf. Systems 32, 213–235 (2009)
16. Hullermeier, E., Rifqi, M.: A fuzzy variant of the Rand index for comparing clustering structures. In: Proc. IFSA, Lisbon, Portugal, pp. 1–6 (2009)
17. Anderson, D., Bezdek, J.C., Keller, J.M., Popescu, M.: Comparing soft partitions. IEEE Trans. Fuzzy Systems (2010) (in review)

Identifying the Risk of Attribute Disclosure by Mining Fuzzy Rules*

Irene Díaz[1], José Ranilla[1], Luis J. Rodríguez-Muniz[2], and Luigi Troiano[3]

[1] University of Oviedo, Department of Computer Science, Gijón, Spain
{sirene,ranilla}@uniovi.es
[2] University of Oviedo, Department of Statistics and O.R., Gijón, Spain
luisj@uniovi.es
[3] University of Sannio, Department of Engineering, Benevento, Italy
troiano@unisannio.it

Abstract. In this paper we address the problem of controlling the disclosure of sensible information by inferring them by the other attributes made public. This threat to privacy is commonly known as prediction or attribute disclosure. Our approach is based on identifying those rules able to link sensitive information to the other attributes being released. In particular, the method presented in this paper is based on mining fuzzy rules. The fuzzy approach is compared to (crisp) decision trees in order to highlight pros and cons of it.

1 Introduction

In order to provide a richer set of data to analyze, statistical agencies and offices release information regarding individuals, companies and other organizations. If availability of microdata makes possible to investigate trends and relationships more accurately, on the other side it poses relevant concerns regarding the risk of revealing sensitive information about the respondents. Indeed, publishing aggregate or individual data carries always the risk that individuals or organizations could be identified and confidential information about them could be released. Therefore, on one side there is a need of providing information in order to perform statistical analysis, whereas on the other it is necessary that some relevant information is not revealed.

Statistical Disclosure Control (SDC) aims at releasing statistical records while protecting confidentiality of information at the same time. Among the different threats, there is the possibility that some sensitive information can be obtained by other known data regarding some entity. In this case, the risk is that hidden information is inferred using public information as premise. Discovering a link between hidden and public information is possible can help SDC to prevent such a risk.

* Authors acknowledge financial support by Grant MTM2008-01519 from Ministry of Science and Innovation and Grant TIN2007-61273 from Ministry of Education and Science, Government of Spain.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 455–464, 2010.

In this paper we propose to use rule extraction as means to identify possible paths that if own by an intruder would be able to disclose sensitive information. The reminder is organized as follows: Section 2 is devoted to some preliminaries regarding information disclosure; Section 3 presents Data Mining as means to prevent disclosure of sensitive information; Section 4 presents experimental results; Section 5 draws conclusions and future directions worth being investigated.

2 Information Disclosure

Information disclosure has place when an entity (i.e. a person or an organization) is able to learn something regarding another entity by released microdata sets. For example, illness regarding patients could be released via medical databases, or competitors' financial figures by business databases.

Microdata attributes of interest for statistical disclosure control can refer to respondent identity (key attributes), or to relevant information (sensitive attributes). In order to preserve the respondent's privacy, the direct linkage between key and sensitive attributes is hidden by SDC. This process is known as data anonymization. However, an intruder can still attack data anonymization by reconstructing the original link with respect to some records.

In particular, there are two types of disclosure associated to microdata [13]: (i) identity disclosure when the entity is (re-)associated to some sensitive data in an anonymized database; (2) prediction disclosure when some sensitive data is inferred by the other attributes for some known entity. The first is also known as *re-identification*, the second as *attribute disclosure*.

Different metrics for measuring the level of privacy guaranteed by SDC have been proposed over the time. Among them, k-anonymity [11], l-diversity [4], p-sensitive [12] and t-closeness [3] each of these metrics is able to drive data anonymization with respect to same aspect, but all of them share the common idea that having more records within a group associable to an entity enforce privacy protection.

However privacy should be related to the extent some information can be considered sensitive. For instance, disclosing that incomes are within a given range, can be considered as much as sensitive than more precise information. This case is known in literature as similarity attack.

Therefore diversification, obtaining by altering the initial information, does not necessarily lead to a stronger privacy protection. Even masking or removing a sensitive attribute could be not enough to avoid attribute disclosure.

The aim of this paper is to show evidence that, even if there is no correlation between data, it is still possible to find a link, although approximated, between public and sensitive variables. The simpler this link is, the most likely it can be discovered or known by intruder, representing thus a threat to no-disclosure of sensitive information.

3 Data Mining

Domingo [1] establishes the connections between data mining and statistical disclosure. The problem in attribute disclosure is basically finding an inferential path from released attributes to sensitive information. Such a path can be due to background knowledge. Mining rules, able to reconstruct the hidden linkage from given patterns of the other attributes, can put into evidence that if some knowledge is discovered by intruders, this can be used to break privacy protections. In addition, similarity is inherently a fuzzy concept.

Data mining approaches can be seen as a way of knowledge discovery which is essential for solving problems. Data mining techniques build a model to predict or classify a problem like an expert. In this sense, data mining techniques may infer relationships allowing us to study the problem of Information Disclosure.

There exists a huge number of machine learning approaches to cope with the problem of classifying an example. Some of them, as Neural Networks or Support Vector Machines (SVM) are very effective and efficient but the model they build is little informative for the Statistical Disclosure problem (for example, SVM provides the weights of the support vectors) [2].

On the other hand, Literature reports a considerable number of ID3-based systems [5] and several fuzzy versions of decision trees [6].

4 Attribute Disclosure by Mining Fuzzy Rules

In this paper we use a data mining strategy based on building a decision tree which provides the corresponding *If-Then* rules with fuzzy information in the *If* part. This approach has two main advantages: first, the induction rules provide us with some information about the most sensible attributes, and, second, as the input information is fuzzified the disclosure risk is supposed to be decreased.

To provide the fuzzy rules, it is used a system based on $C4.5$ [7], the so-called ARNI, and its fuzzy extension, the so-called *FArni*. But despite of using *Information gain* as in $C4.5$, both systems ARNI and *FArni* use a measure called *Imputity Level* (*IL*) for determining the quality of the rules induced from examples [10]. *IL* [8] explicitly takes into account not only the probability of success p, but also the difficulty of attaining that amount of examples of class C. Later, once the fuzzy decision tree is induced, *FArni* returns compact fuzzy rule sets after applying a pruning process inherited from ARNI and Fan [9]. *FArni* is presented in detail in [10].

5 Experimental Results

The experiments in this paper were carried out with the TARRAGONA Data Set, which is a real data set comprising figures of 834 companies in the Tarragona area. Data correspond to year 1995. For each company, 13 quantitative variables are given: Fixed assets, Current assets, Treasury, Uncommitted funds, Paid-up capital, Short-term debt, Sales, Labor costs, Depreciation, and Operating profit,

Financial outcome, Gross profit, Net profit, which are considered in this research as target classes. Experimental results, after an initial analysis, show that this approach is feasible and able to emerge background knowledge in preserving data confidentiality.

Let's check the behavior of a fuzzy learner in identifying the disclosure risk over the TARRAGONA Data Set.

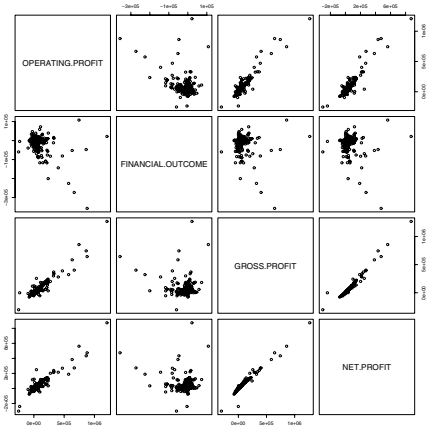


Fig. 1. Pairwise scatter-plots of sensible variables

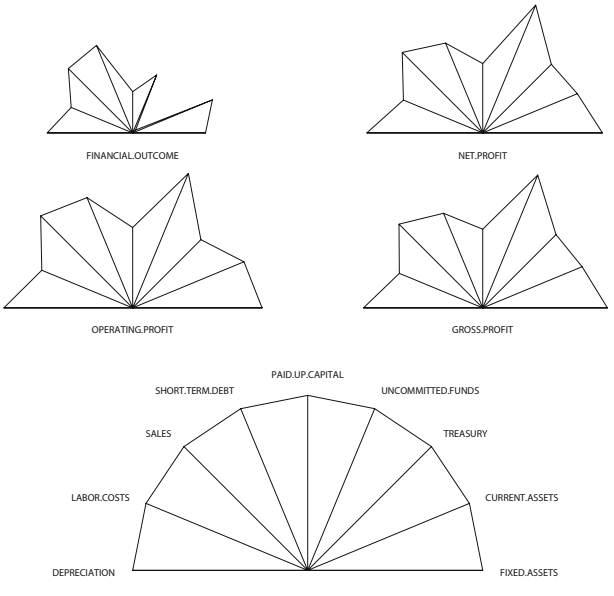


Fig. 2. Radar chart of absolute correlation between public and target variables

In order to determine the attributes to be protected we have performed a preliminary selection. Thus, we have selected four variables: Financial outcome, Gross profit, Net profit and Operating outcome; this selection is motivated by the informative relevance of the values of these variables. But, after performing a statistical study of the correlations among these four variables we have obtained strong (in some case, obvious) dependencies. In particular, Gross profit, Net profit and Operating outcome are strongly correlated and can be considered to be independent on Financial outcome, as outlined in Fig.1.

Moreover, when analysing the degree in which each of these variables is depending on the rest, we can see how the behaviour of Net profit, Gross profit and Operating outcome is, more ore less similar. In Figure 2 we represent these dependencies in a radar chart (attributes on each radar axis are shown in the lower graphic). We can observe how the fore mentioned three variables have very similar behavior. Therefore, in a second step, we have choose only Net profit and Financial outcome as the attributes to be protected.

Finally, there is no strong correlation between variables being released, as depicted by pairwise scatter-plots in Fig.3. Therefore we can assume the variables used for inference as independent.

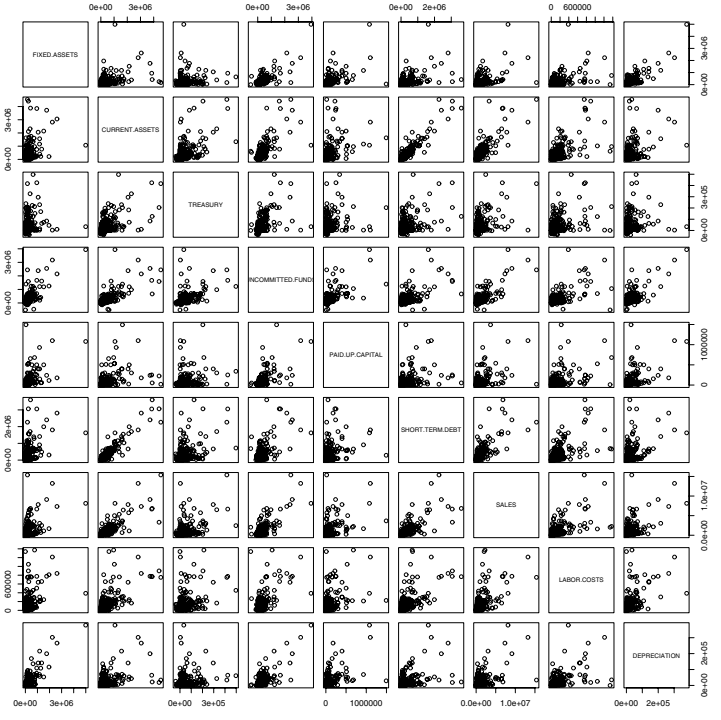


Fig. 3. Pairwise scatter-plots of released variables

The low correlation between sensitive variables, as chosen in our experimentation, leads to the need of adopting an alternative approach to statistics in identifying threats to attribute disclosure.

Therefore four classifiers will be built, one for each objective variable. As in many cases of practical interest it is not important that precise values of sensible attributes are disclosed but rather the ranges they belong to, discovering treats of disclosing sensitive information can be regarded as a classification problem, where an intruder could be able to associate non-sensitive data to a class of sensitive information.

Once we have selected the variables to protect, we want to measure how strong is this protection if we attack the data with a data mining rule generator as *FArni*. Since this method predicts only one class at once, we have constructed one classifier for each variable. Therefore, two learners will be built, one for Net profit and the other one for Financial outcome. On the other hand, since *FArni* needs a class with discrete domain, we have discretized each variable using quintiles as breaking points and then splitting the examples in five classes named A, B, C, D, and E.

FArni constructs a fuzzy domain for each attribute. We have selected five trapezoidal fuzzy sets labelled *very low*, *low*, *medium*, *high* and *very high* for every domain. Each trapezoidal fuzzy set has been constructed assuming that the 1-level set contains the fourth part of the total number of cases with non-zero membership to the considered fuzzy set.

Obtained fuzzy rules are shown in the following subsections. We also compare these results to the ones obtained by using the correspondent crisp rule generator.

5.1 Fuzzy Rules Obtained for Disclosing Net Profit

- Net Profit is A *if*:
  - *P*aidup-capital is high *and* *U*ncommitted-funds is low
- Net Profit is B *if*:
  - *U*ncommitted-funds is very low *and* *D*epreciation is low *and* *T*reasury is low
- Net Profit is C *if*:
  - *T*reasury is very low *and* *L*abor-costs is very low *and* *D*epreciation is very low *and* *P*aidup-capital is low *and* *F*ixed-assets is very low *and* *U*ncommitted-funds is low
- Net Profit is D *if*:
  - *L*abor-costs is high *P*aidup-capital is low *and* *U*ncommitted-funds is low
- Net Profit is E *if*:
  - *U*ncommitted-funds is very high
  - *D*epreciation is low *and* *F*ixed-assets is low *P*aidup-capital is very low *and* *T*reasury is very low

- *L*abor-costs is very high
- *U*ncommitted-funds is medium
- *S*hortterm-Debt is medium *and* *U*ncommitted-funds is low
- *L*abor-costs is low *and* *P*aidup-capital is low
- *P*aidup-capital is low *and* *F*ixed-assets is low
- *D*epreciation is low *and* *P*aidup-capital is low
- *T*reasury is medium

5.2 Crisp Rules Obtained for Disclosing Net Profit

– Net Profit is *A* *if*:

- *U*ncommitted-funds $\leq$ 12091 *and* *T*reasury $\leq$ 5545
- *U*ncommitted-funds $\leq$ 15407 *and* *T*reasury $\leq$ 7347 *and* *L*abor-costs $\leq$ 46958
- *U*ncommitted-funds $\leq$ 15407 *and* *S*ales $\leq$ 115535

– Net Profit is *B* *if*:

- *D*epreciation $\leq$ 1997 *and* *U*ncommitted-funds $\leq$ 15407 *and* *P*aidup-capital $\leq$ 2000 *and* 115535 $<$ *S*ales $\leq$ 204889

– Net Profit is *C* *if*:

- 18431 $<$ *U*ncommitted-funds $\leq$ 50769 *and* 2734 $<$ *D*epreciation $\leq$ 4496 *and* 9800 ¡*P*aidup-capital $\leq$ 20250 *and* C*urrent-assets* $\leq$ 76155

– Net Profit is *D* *if*:

- 50769 $<$ *U*ncommitted-funds $\leq$ 145424 *and* *L*abor-costs $>$ 26297 *and* *D*epreciation $>$ 1997 *and* *T*reasury $>$ 2683 *and* *S*ales $\leq$ 547947

– Net Profit is *E* *if*:

- *S*ales $>$ 1.27169e+06 *and* *D*epreciation $>$ 5536
- *U*ncommitted-funds $>$ 145424

5.3 Fuzzy Rules Obtained for Disclosing Financial Outcome

– Financial Outcome is *A* *if*:

- *S*ales is medium *and* *U*ncommitted-funds is low *and* *T*reasury is very low
- *S*hortterm-debt is very high *and* *U*ncommitted-funds is low
- C*urrent-assets* is very high *and* *T*reasury is very high
- *S*hortterm-debt is medium *and* *T*reasury is medium
- *S*hortterm-debt is low *and* C*urrent-assets* is medium *and* *U*ncommitted-funds is low
- *D*epreciation is medium *and* *U*ncommitted-funds is low
- C*urrent-assets* is medium *and* *T*reasury is low
- *F*ixed-assets is low *and* *D*epreciation is low *and* *P*aidup-capital is very low *and* *L*abor-costs is very low
- *L*abor-costs is very high *and* *T*reasury is very low

- C*urrent-assets* is high
- *P*aidup-capital is high *and U*ncommitted-funds is low
- *L*abor-costs is medium *and T*reasury is medium *and U*ncommitted-funds is low

– Financial Outcome is *B if:*

- *T*reasury is high *and U*ncommitted-funds is low

– Financial Outcome is *C if:*

- *P*aidup-capital is medium *and T*reasury is very low *and S*ales is very low *and U*ncommitted-funds is low *and L*abor-costs is low

– Financial Outcome is *D if:*

- *L*abor-costs is very low *and P*aidup-capital is very low *and F*ixed-assets is very low *and D*epreciation is low *and U*ncommitted-funds is low *and T*reasury is low

– Financial Outcome is *E if:*

- *S*ales is medium *and U*ncommitted-funds is high *and* C*urrent-assets* is medium
- C*urrent-assets* is low *and L*abor-costs is high *and D*epreciation is very low
- *T*reasury is medium *and* C*urrent-assets* is very high
- *T*reasury is medium *and L*abor-costs is very high
- C*urrent-assets* is low *and T*reasury is very high

5.4 Crisp Rules Obtained for Disclosing Financial Outcome

– Financial Outcome is *A if:*

- *S*hortterm-debt > 143049
- *F*ixed-assets > 146919 *and D*epreciation > 5969
- *F*ixed-assets > 109215 *and D*epreciation > 9711
- C*urrent-assets* > 178945 *and F*ixed-assets > 11723

– Financial Outcome is *B if*:

- $124587 <$ *S*ales ≤ 318796 *and S*hortterm-debt > 70266 *and T*reasury ≤ 6404 *and P*aidup-capital > 3600 *and P*aidup-capital ≤ 38000 *and* C*urrent-assets* ≤ 104984 *and F*ixed-assets ≤ 68152

– Financial Outcome is *C if*:

- *U*ncommitted-funds ≤ 28411 *and T*reasury ≤ 5776 *and S*hortterm-debt > 21546 *and S*hortterm-debt ≤ 43227 *and D*epreciation > 586 *and D*epreciation ≤ 2072 *and P*aidup-capital > 800 *and L*abor-costs > 17249

– Financial Outcome is *D if*:

- *L*abor-costs ≤ 17249 *and S*hortterm-debt ≤ 43227 *and S*ales ≤ 158049 *and F*ixed-assets ≤ 6133 *and P*aidup-capital ≤ 8500 *and T*reasury ≤ 10121

- C*urrent-assets* $\leq$ 39986 *and* *S*hortterm-debt $\leq$ 43227 *and* *U*ncommitted-funds $\leq$ 28411

– Financial Outcome is *E if*:

- *T*reasury $>$ 13592 *and* *U*ncommitted-funds $>$ 22238 *and* *L*abor-costs $\leq$ 23039 *and* C*urrent-assets* $\leq$ 104984
- *T*reasury $>$ 6564 *and* *U*ncommitted-funds $>$ 29125 *and* *S*hortterm-debt $\leq$ 31511
- *T*reasury $>$ 5776 *and* *S*hortterm-debt $\leq=$ 43227 *and* *F*ixed-assets $\leq$ 6068 *and* *P*aidup-capital $\leq$ 6200 *and* *L*abor-costs $>$ 17249
- *T*reasury $>$ 11565 *and* *U*ncommitted-funds $>$ 59198 *and* C*urrent-assets* $>$ 104984 *and* *S*hortterm-debt $\leq$ 143049 *and* *F*ixed-assets $\leq$ 70563 *and* *P*aidup-capital $\leq$ 38000 *and* *S*ales $\leq$ 707711

5.5 Experimentation Outcomes

In general, rules mined by FArni (fuzzy) are simpler than those obtained by ARNI (crisp). The threat of disclosing attributes is related to the possibility of being aware or building such links as background knowledge. Simplicity of rules derives from the number of predicates involved in the antecedents and by interpretability of them. In both cases, fuzzy rules resulted less structured and easier to understand. As the risk of attribute disclosure is higher in the case of simpler rules, the fuzzy approach looks able to find a higher number of threats.

6 Conclusions and Future Directions

In this paper we investigated the application of fuzzy rules mining as a means for discovering conditions able to infer sensitive information, also known as attribute disclosure, although approximated. Experimental results are encouraging. Fuzzy rules are generally simpler and easier to interpret than other approaches, based on decision trees for example. In the future we aim to study at which extent rules can be generalized, and what is the role of background knowledge in determining logical connections between data.

References

1. Domingo-Ferrer, J., Torra, V.: On the connections between statistical disclosure control for microdata and some artificial intelligence tools. Inf. Sci. Inf. Comput. Sci. 151, 153–170 (2003)
2. Joachims, T.: Making large-scale support vector machine learning practical. In: Smola, A., Scholkopf, B., Burges, C. (eds.) Advances in Kernel Methods: Support Vector Machines. MIT Press, Cambridge (1998)
3. Li, N., Li, T.: t-closeness: Privacy beyond k-anonymity and?-diversity. In: Proceedings of IEEE International Conference on Data Engineering (2007)

4. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: 22nd IEEE International Conference on Data Engineering (2006)
5. Pal, N.R., Chakraborty, S.: Fuzzy rule extraction from id3-type decision trees for real data. IEEE Transactions on Systems, Man, and Cybernetics, Part B 31(5), 745–754 (2001)
6. Pham, D.T., Bigot, S., Dimov, S.S.: Rules-f: a fuzzy inductive learning algorithm. In: Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, pp. 1433–1444 (2006)
7. Quinlan, R.: c4.5: Programs of machine learning (1993)
8. Ranilla, J., Luaces, O., Bahamonde, A.: A heuristic for learning decision trees and pruning them into classification rules. AICom (Artificial Intelligence Communication) 16(2) (2003) (in press)
9. Ranilla, J., Antonio, B.: Fan: Finding accurate inductions. International Journal of Human Computer Studies 56(4), 445–474 (2002)
10. Ranilla, J., Rodriguez-Muniz, L.J.: A heuristic approach to learning rules from fuzzy databases. IEEE Intelligent Systems 22(2), 62–68 (2007)
11. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13, 1010–1027 (2001)
12. Sun, X., Wang, H., Li, J., Ross, D.: Achieving p-sensitive k-anonymity via anatomy. In: Proceedings of the 2009 IEEE International Conference on e-Business Engineering, Washington, DC, USA, pp. 199–205. IEEE Computer Society, Los Alamitos (2009)
13. Willenborg, L., de Waal, T.: Statistical Disclosure Control in Practice. Springer, Heidelberg (1996)

Explicit Descriptions of Associative Sugeno Integrals

Miguel Couceiro and Jean-Luc Marichal

University of Luxembourg, Mathematics Research Unit
6, rue Richard Coudenhove-Kalergi
L–1359 Luxembourg, G.-D. Luxembourg
miguel.couceiro@uni.lu, jean-luc.marichal@uni.lu

Abstract. The associativity property, usually defined for binary functions, can be generalized to functions of a given fixed arity $n \geq 1$ as well as to functions of multiple arities. In this paper, we investigate these two generalizations in the case of Sugeno integrals over bounded distributive lattices and present explicit descriptions of the corresponding associative functions. We also show that, in this case, both generalizations of associativity are essentially the same.

Keywords: bounded distributive lattice, Sugeno integral, associativity, idempotency, functional equation.

1 Introduction

Let X be an arbitrary nonempty set. Throughout this paper, we regard vectors $\mathbf{x}$ in X^n as n-strings over X. The 0-string or *empty* string is denoted by ε so that $X^0 = \{\varepsilon\}$. We denote by X^* the set of all strings over X, that is, $X^* = \bigcup_{n\in\mathbb{N}} X^n$. Moreover, we consider X^* endowed with concatenation for which we adopt the juxtaposition notation. For instance, if $\mathbf{x} \in X^n$, $y \in X$, and $\mathbf{z} \in X^m$, then $\mathbf{x}y\mathbf{z} \in X^{n+1+m}$. Furthermore, for $\mathbf{x} \in X^m$, we use the short-hand notation $\mathbf{x}^n = \mathbf{x}\cdots\mathbf{x} \in X^{n\times m}$. In the sequel, we will be interested both in functions of a given fixed arity (i.e., functions $f\colon X^n \to X$) as well as in functions defined on X^*, that is, of the form $g\colon X^* \to X$. Given a function $g\colon X^* \to X$, we denote by g_n the restriction of g to X^n, i.e. $g_n := g|_{X^n}$. In this way, each function $g\colon X^* \to X$ can be regarded as a family $(g_n)_{n\in\mathbb{N}}$ of functions $g_n\colon X^n \to X$. We convey that g_0 is defined by $g_0(\varepsilon) = \varepsilon$.

In this paper, we are interested in the associativity property, traditionally considered on binary functions. Recall that a function $f\colon X^2 \to X$ is said to be *associative* if $f(f(xy)z) = f(xf(yz))$ for every $x, y, z \in X$. The importance of this notion is made clear by its natural interpretation. Essentially, it expresses the fact that the order in which variables are bracketed is not relevant. This algebraic property was extended to functions $f\colon X^n \to X$, $n \geq 1$, as well as to functions $g\colon X^* \to X$ in somewhat different ways.

A function $f\colon X^n \to X$ is said to be *associative* if, for every $\mathbf{xz}, \mathbf{x}'\mathbf{z}' \in X^{n-1}$ and every $\mathbf{y}, \mathbf{y}' \in X^n$ such that $\mathbf{xyz} = \mathbf{x}'\mathbf{y}'\mathbf{z}'$, we have $f(\mathbf{x}f(\mathbf{y})\mathbf{z}) = f(\mathbf{x}'f(\mathbf{y}')\mathbf{z}')$.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 465–470, 2010.

This generalization of associativity to n-ary functions goes back to Dörnte [6] and led to the generalization of groups to n-groups (polyadic goups).[1] In a somewhat different context, this notion has been recently used to completely classify closed intervals made of equational classes of Boolean functions; see [2].

On a different setting, associativity can be generalized to functions on X^* as follows. We say that a function $g\colon X^* \to X$ is *associative* if, for every $\mathbf{xyz}, \mathbf{x'y'z'} \in X^*$ such that $\mathbf{xyz} = \mathbf{x'y'z'}$, we have $g(\mathbf{x}g(\mathbf{y})\mathbf{z}) = g(\mathbf{x'}g(\mathbf{y'})\mathbf{z'})$. Alternative formulations of this definition appeared in the theory of aggregation functions, where the arity is not always fixed; see for instance [1,14,16,17].

In general, the latter definition is more restrictive on the components g_n of $g\colon X^* \to X$. For instance, the ternary real function $f(xyz) = x - y + z$ is associative but cannot be the ternary component of an associative function $g\colon \mathbb{R}^* \to \mathbb{R}$. Indeed, the equations

$$g_2(g_2(xy)z) = g_2(xg_2(yz)) = x - y + z \tag{1}$$

have no solution, for otherwise we would have $y = g_2(g_2(y0)0)$ and hence

$$g_2(xy) = g_2(xg_2(g_2(y0)0)) = g_2(g_2(xg_2(y0))0) = x - g_2(y0).$$

This would imply $g_2(xy) = x - y$, which contradicts (1).

In this paper we show that, in the case of Sugeno integrals on bounded distributive lattices, the two notions of associativity are essentially the same. More precisely, given a bounded distributive lattice L, we have that a Sugeno integral $f\colon L^n \to L$ is associative if and only if it is the n-ary component of some associative function $g\colon L^* \to L$; see Corollary 7. This paper is organized as follows: in Sect. 2 we provide some preliminary results, which are then used in Sect. 3 to obtain explicit descriptions of those associative Sugeno integrals; see Theorems 4 and 6.

2 Preliminary Results

The following proposition provides useful reformulations of associativity of functions $g\colon X^* \to X$.

Proposition 1. *Let $g\colon X^* \to X$ be a function. The following assertions are equivalent:*

(i) g is associative.
(ii) For every $\mathbf{xyz} \in X^*$, we have $g(\mathbf{x}g(\mathbf{y})\mathbf{z}) = g(\mathbf{xyz})$.
(iii) For every $\mathbf{xy} \in X^*$, we have $g(g(\mathbf{x})g(\mathbf{y})) = g(\mathbf{xy})$.

Remark 2. (i) Associativity of functions $g\colon X^* \to X$ was defined in [16] and [17] as in assertions *(iii)* and *(ii)* of Proposition 1, respectively. For a recent reference, see [14].

[1] The first extensive study on polyadic groups was due to Post [20]. This study was followed by several contributions towards the classification and description of n-groups and similar "super-associative" structures; to mention a few, see [7,8,9,11,12,15,19].

(ii) As observed in [1], associative functions $g\colon X^* \to X$ are completely determined by their unary and binary components. Indeed, for every $n \in \mathbb{N}$, $n > 2$, and every $x_1, \ldots, x_n \in X$, we have

$$g(x_1 \cdots x_n) = g_2(g_2(\cdots g_2(g_2(x_1x_2)x_3)\cdots)x_n).$$

3 Associative Sugeno Integrals

Let L be a bounded distributive lattice, with 0 and 1 as bottom and top elements. A (*lattice*) *polynomial function* is any mapping $f\colon L^n \to L$ which can be obtained as combinations of projections and constant functions using the lattice operations $\wedge$ and $\vee$. Our interest in these lattice polynomial functions comes from the fact that, as observed in [18], (*discrete*) *Sugeno integrals* can be regarded as idempotent lattice polynomial functions, that is, polynomial functions satisfying $f(x^n) = x$ for every $x \in X$. This view has several appealing aspects, in particular, concerning normal form representations of Sugeno integrals. Indeed, as shown by Goodstein [13], polynomial functions on bounded distributive lattices coincide exactly with those functions representable in disjunctive normal form (DNF).

More precisely, for $I \subseteq [n] = \{1, \ldots, n\}$, let $\mathbf{e}_I \in \{0,1\}^n$ be the characteristic vector of I and let $\alpha_f\colon 2^{[n]} \to L$ be the function given by $\alpha_f(I) = f(\mathbf{e}_I)$. Then

$$f(\mathbf{x}) = \bigvee_{I \subseteq [n]} \big(\alpha_f(I) \wedge \bigwedge_{i \in I} x_i\big). \tag{2}$$

Thus, a function $f\colon L^n \to L$ is a Sugeno integral integral if and only if f fulfills (2) with $\alpha_f(\emptyset) = 0$ and $\alpha_f([n]) = 1$. For further background, see [3,4].

Theorem 3 ([4]). *A function $f\colon L^n \to L$ is a Sugeno integral if and only if it is idempotent and satisfies*

$$f(\mathbf{x}y\mathbf{z}) = \operatorname{med}\big(f(\mathbf{x}0\mathbf{z}), y, f(\mathbf{x}1\mathbf{z})\big), \quad \textit{for every } \mathbf{x}y\mathbf{z} \in L^n. \tag{3}$$

The following theorem is an immediate consequence of Theorem 6 in [5] and it restricts the disjunctive normal form of n-ary Sugeno integrals.

Theorem 4. *Let $f\colon L^n \to L$ be a Sugeno integral. If f is associative, then*

$$f(\mathbf{x}) = (b_n \wedge x_1) \vee \Big(\bigvee_{i=1}^{n} (b_n \wedge c_n \wedge x_i)\Big) \vee (c_n \wedge x_n) \vee \bigwedge_{i=1}^{n} x_i, \tag{4}$$

where $b_n = f(10^{n-1})$ and $c_n = f(0^{n-1}1)$.

Remark 5. (i) We observe that equation (4) can be rewritten in a more symmetric way as

$$f(\mathbf{x}) = (b_n \wedge x_1) \vee \operatorname{med}\Big(\bigwedge_{i=1}^{n} x_i, b_n \wedge c_n, \bigvee_{i=1}^{n} x_i\Big) \vee (c_n \wedge x_n).$$

This formula reduces to $f(\mathbf{x}) = \operatorname{med}\big(\bigwedge_{i=1}^{n} x_i, b_n, \bigvee_{i=1}^{n} x_i\big)$ as soon as f is a symmetric function (i.e., invariant under permutation of its variables).

(ii) A *term function* $f\colon L^n \to L$ is a Sugeno integral satisfying $\alpha_f(I) \in \{0,1\}$ for every $I \subseteq [n]$. By Theorem 4, the only associative term functions $f\colon L^n \to L$ are $\mathbf{x} \mapsto x_1$, $\mathbf{x} \mapsto x_n$, $\mathbf{x} \mapsto \bigwedge_{i=1}^n x_i$, and $\mathbf{x} \mapsto \bigvee_{i=1}^n x_i$.

We say that a function $g\colon L^* \to L$ is a *Sugeno integral* if every g_n, $n \geq 1$, is a Sugeno integral. The following theorem yields a description of associative Sugeno integrals $g\colon L^* \to L$. For a generalization to polynomial functions $g\colon L^* \to L$, see Theorem 7 in [5].

Theorem 6. *A Sugeno integral $g\colon L^* \to L$ is associative if and only if $g_1(x) = x$ and, for $n \geq 2$,*

$$g_n(\mathbf{x}) = (b_2 \wedge x_1) \vee \Big(\bigvee_{i=1}^n (b_2 \wedge c_2 \wedge x_i) \Big) \vee (c_2 \wedge x_n) \vee \bigwedge_{i=1}^n x_i, \tag{5}$$

where $b_2 = g_2(10)$ and $c_2 = g_2(01)$.

Proof. Sufficiency can be verified by making use of Proposition 1.

To verify that the conditions are necessary, note that since each g_n is associative, by Theorem 4, each g_n has the form (4) with $b_n = g_n(10^{n-1})$ and $c_n = g_n(0^{n-1}1)$. By associativity and Theorem 3, for every $n \geq 3$,

$$g_n(10^{n-1}) = g_2(g_{n-1}(10^{n-2})0) = \mathrm{med}(0, g_{n-1}(10^{n-2}), g_2(10)).$$

By reasoning recursively, one can see that $b_n = b_2$. Similarly, one can verify that $c_n = c_2$, for every $n \geq 3$. □

Even though associativity for functions $g\colon L^* \to L$ seems more restrictive on their components g_n than associativity for functions of a given fixed arity, from Theorems 4 and 6 it follows that associativity for Sugeno integrals $f\colon L^n \to L$ naturally extends componentwise to Sugeno integrals $g\colon L^* \to L$.

Corollary 7. *Let $f\colon L^n \to L$ be a Sugeno integral. Then f is associative if and only if there is an associative Sugeno integral $g\colon L^* \to L$ such that $g_n = f$.*

Proof. Clearly, the condition is sufficient. Conversely, if f is associative, then by Theorem 4

$$f(\mathbf{x}) = (b_n \wedge x_1) \vee \Big(\bigvee_{i=1}^n (b_n \wedge c_n \wedge x_i) \Big) \vee (c_n \wedge x_n) \vee \bigwedge_{i=1}^n x_i,$$

where $b_n = f(10^{n-1})$ and $c_n = f(0^{n-1}1)$. Let $g\colon L^* \to L$ be the Sugeno integral such that $g_1(x) = x$ and, for $m \geq 2$,

$$g_m(\mathbf{x}) = (b \wedge x_1) \vee \Big(\bigvee_{i=1}^m (b \wedge c \wedge x_i) \Big) \vee (c \wedge x_n) \vee \bigwedge_{i=1}^m x_i,$$

where $b = f(10^{n-1})$ and $c = f(0^{n-1}1)$. Clearly, $g_n = f$ and by Theorem 6 we have that g is associative. □

Remark 8. (i) The case when L is a connected order topological space was considered by Fodor [10] who obtained an explicit description of those nondecreasing binary functions which are idempotent, continuous, and associative.

(ii) Many associative functions $g\colon L^* \to L$ have been investigated in aggregation theory in the special case when L is the real unit interval $[0,1]$; see, e.g., [14]. To give an example, there are only four such associative functions whose n-ary restrictions are nondecreasing and stable under interval scale transformations (i.e., g_n commutes with unary positive affine functions), namely, $\mathbf{x} \mapsto x_1$, $\mathbf{x} \mapsto x_n$, $\mathbf{x} \mapsto \bigwedge_{i=1}^n x_i$, and $\mathbf{x} \mapsto \bigvee_{i=1}^n x_i$. In particular, an associative function $g\colon [0,1]^* \to [0,1]$ whose n-ary restrictions are *discrete Choquet integrals* necessarily reduces to one of these four functions.

References

1. Beliakov, G., Pradera, A., Calvo, T.: Aggregation Functions: A Guide for Practitioners. Studies in Fuziness and Soft Computing. Springer, Berlin (2007)
2. Couceiro, M.: On the Lattice of Equational Classes of Boolean Functions and its Closed Intervals. J. Mult.-Valued Logic Soft Comput. 14(1-2), 81–104 (2008)
3. Couceiro, M., Marichal, J.-L.: Polynomial Functions over Bounded Distributive Lattices, http://arxiv.org/abs/0901.4888
4. Couceiro, M., Marichal, J.-L.: Characterizations of Discrete Sugeno Integrals as Polynomial Functions over Distributive Lattices. Fuzzy Sets and Systems 161, 694–707 (2009)
5. Couceiro, M., Marichal, J.-L.: Associative Polynomial Functions over Bounded Distributive Lattices. Order (to appear), http://arxiv.org/abs/0902.2323
6. Dörnte, W.: Untersuchengen über einen Verallgemeinerten Gruppenbegriff. Math. Z. 29, 1–19 (1928)
7. Dudek, W.A.: Varieties of Polyadic Groups. Filomat 9, 657–674 (1995)
8. Dudek, W.A.: On Some Old and New Problems in n-ary Groups. Quasigroups and Related Systems 8, 15–36 (2001)
9. Dudek, W.A., Glazek, K., Gleichgewicht, B.: A Note on the Axiom of n-Groups. Coll. Math. Soc. J. Bolyai 29, 195–202 (1977)
10. Fodor, J.C.: An Extension of Fung-Fu's Theorem. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 4(3), 235–243 (1996)
11. Glazek, K.: Bibliography of n-Groups (Polyadic Groups) and Some Group-Like n-ary Systems. In: Proc. of the Symposium on n-ary Structures, Macedonian Academy of Sciences and Arts, Skopje, pp. 253–289 (1982)
12. Glazek, K., Gleichgewicht, B.: Remarks on n-Groups as Abstract Algebras. Colloq. Math. 17, 209–219 (1967)
13. Goodstein, R.L.: The Solution of Equations in a Lattice. Proc. Roy. Soc. Edinburgh Sect. A 67, 231–242 (1965/1967)
14. Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E.: Aggregation Functions. Encyclopedia of Mathematics and Its Applications, vol. 127. Cambridge University Press, Cambridge (2009)
15. Hosszú, M.: On the Explicit Form of n-Group Operations. Publ. Math. Debrecen 10, 88–92 (1963)
16. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms. Trends in Logic – Studia Logica Library, vol. 8. Kluwer Academic Publishers, Dordrecht (2000)

17. Marichal, J.-L.: Aggregation Operators for Multicriteria Decision Aid. Ph.D. thesis, Institute of Mathematics, University of Liège, Liège, Belgium (December 1998)
18. Marichal, J.-L.: Weighted Lattice Polynomials. Discrete Math. 309(4), 814–820 (2009)
19. Monk, J.D., Sioson, F.M.: On the General Theory of m-Groups. Fund. Math. 72, 233–244 (1971)
20. Post, E.L.: Polyadic Groups. Trans. Amer. Math. Soc. 48, 208–350 (1940)

Continuity of Choquet Integrals of Supermodular Capacities*

Nobusumi Sagara

Faculty of Economics, Hosei University,
4342, Aihara, Machida, Tokyo 194–0298, Japan
nsagara@hosei.ac.jp

Abstract. We prove the continuity of the Choquet integral of supermodular capacities, in L^∞ with respect to the weak*-topology, employing a useful relationship between convex games and their Choquet integrals. The main result is applied to generalized fair division problems, and the existence of Pareto optimal α-allocations is demonstrated for the case of nonadditive measures.

Keywords: Supermodularity, Choquet integral, Weak*-continuity, Capacity, Core, Fair allocation, Pareto optimality, Envy-freeness.

MSC 2000: Primary: 28B05, 28E10; secondary: 91A12.

1 Preliminary

Let $(\Omega, \mathscr{F})$ be a measurable space, where $\mathscr{F}$ is a σ-algebra of subsets of a nonempty set Ω. Throughout this paper, a *set function* is a real-valued function on $\mathscr{F}$ that vanishes at the empty set.

A set function ν is *monotone* if $\nu(A) \leq \nu(B)$ for every $A, B \in \mathscr{F}$ with $A \subset B$; ν is *supermodular* (or *convex*) if $\nu(A) + \nu(B) \leq \nu(A \cup B) + \nu(A \cap B)$ for every $A, B \in \mathscr{F}$. A supermodular set function is monotone if and only if it is nonnegative.

A set function ν is *bounded* if $\sup_{A \in \mathscr{F}} |\nu(A)| < \infty$. A monotone set function is bounded. A set function ν is of *bounded variation* if $\|\nu\| := \sup \sum_{i=1}^{k} |\nu(A_i) - \nu(A_{i-1})|$ is finite, where the supremum is taken over all finite chains $\emptyset = A_0 \subset A_1 \subset \cdots \subset A_k = \Omega$ in $\mathscr{F}$.

Given a set function ν, an element $N \in \mathscr{F}$ is *ν-null* if $\nu(A \cup N) = \nu(A)$ for every $A \in \mathscr{F}$. If $N \in \mathscr{F}$ is ν-null, then $\nu(N) = 0$; ν is *null-additive* if $A \cap N = \emptyset$ and $\nu(N) = 0$ imply $\nu(A \cup N) = \nu(A)$.

A set function ν is *absolutely continuous* with respect to a set function μ if every μ-null set is ν-null; ν is *equivalent* to μ if an element in $\mathscr{F}$ is ν-null if and only if it is μ-null.

* This research is supported by a Grant-in-Aid for Scientific Research (No. 21530277) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 471–479, 2010.

A set function ν is *strictly monotone* if $A \subset B$ with ν-nonnull $B \setminus A$ implies that $\nu(A) < \nu(B)$.

A set function ν is *continuous from above* at A if $\nu(A_k) \to \nu(A)$ for every sequence $\{A_k\}$ in $\mathscr{F}$ with $A_k \downarrow A$; ν is *continuous from below* at A if $\nu(A_k) \to \nu(A)$ for every sequence $\{A_k\}$ in $\mathscr{F}$ with $A_k \uparrow A$; ν is *continuous* if it is both continuous above and continuous below at every element in $\mathscr{F}$. A continuous monotone set function is called a *capacity* (or a *fuzzy measure*).

For later use, we employ the following result.

Proposition 1.1. (i) *For a null-additive monotone set function ν, an element $N \in \mathscr{F}$ is ν-null if and only if $\nu(N) = 0$.*
(ii) *A bounded supermodular set function is of bounded variation.*

(For a proof of condition (i), see [10, Theorem 2.1]; for that of condition (ii), see [8].)

Let $B(\Omega, \mathscr{F})$ be the space of bounded measurable functions on Ω with the supremum norm. The *Choquet integral* $\hat{\nu} : B(\Omega, \mathscr{F}) \to \mathbb{R}$ of a set function ν is defined by an improper Riemann integral of the form

$$\hat{\nu}(f) = \int_0^{+\infty} \nu(f \geq t)dt + \int_{-\infty}^{0} [\nu(f \geq t) - \nu(\Omega)]dt,$$

where $\nu(f \geq t)$ denotes the value of ν at the measurable set $\{\omega \in \Omega \mid f(\omega) \geq t\}$. Note that this integral exists whenever ν is of bounded variation (see [10, Theorem 7.21]); $\nu(A) = \hat{\nu}(\chi_A)$ for every $A \in \mathscr{F}$, where χ_A is the characteristic function of $A \in \mathscr{F}$.

The next result is due to [2].

Proposition 1.2. *For every set function ν of bounded variation, the following conditions are equivalent:*

(i) *ν is supermodular on $\mathscr{F}$;*
(ii) *$\hat{\nu}$ is concave on $B(\Omega, \mathscr{F})$;*
(iii) *$\hat{\nu}$ is supermodular on $B(\Omega, \mathscr{F})$, i.e., $\hat{\nu}(f) + \hat{\nu}(g) \leq \hat{\nu}(f \vee g) + \hat{\nu}(f \wedge g)$ for every $f, g \in B(\Omega, \mathscr{F})$;*
(iv) *$\hat{\nu}$ is superadditive on $B(\Omega, \mathscr{F})$, i.e., $\hat{\nu}(f) + \hat{\nu}(g) \leq \hat{\nu}(f + g)$ for every $f, g \in B(\Omega, \mathscr{F})$.*

(For a proof, see [8].)

Let $ba(\Omega, \mathscr{F})$ be the space of finitely additive set functions on $\mathscr{F}$ of bounded variation with the total variation norm, which is the dual space of $B(\Omega, \mathscr{F})$ (see [6, Theorem IV.5.1]), with the corresponding duality denoted by $\langle f, \lambda \rangle$ for $f \in B(\Omega, \mathscr{F})$ and $\lambda \in ba(\Omega, \mathscr{F})$. The space $ba(\Omega, \mathscr{F})$ is endowed with the weak*-topology.

For a set function ν, define the subset $\mathscr{C}(\nu)$ of $ba(\Omega, \mathscr{F})$ by

$$\mathscr{C}(\nu) = \{\lambda \in ba(\Omega, \mathscr{F}) \mid \nu \leq \lambda \text{ and } \lambda(\Omega) = \nu(\Omega)\}.$$

The set $\mathscr{C}(\nu)$ is called the *core* of ν in cooperative game theory. Note that $\mathscr{C}(\nu)$ is (possibly empty) weak*-compact in $ba(\Omega, \mathscr{F})$ because it is bounded in the total variation norm and weak*-closed (see [6, Corollary V.4.3]).

The following conditions on supermodular set functions indicate a profound relationship between their cores and Choquet integrals.

Proposition 1.3. *For every bounded supermodular capacity ν, the following conditions are satisfied:*

(i) *$\mathscr{C}(\nu)$ is nonempty.*
(ii) *Every element in $\mathscr{C}(\nu)$ is countably additive.*
(iii) *There exists a finite measure μ such that*

$$\lim_{\mu(A)\to 0} \sup\{\lambda(A) \mid \lambda \in \mathscr{C}(\nu)\} = 0.$$

(iv) *ν is equivalent to any control measure[1] for $\mathscr{C}(\nu)$.*
(v) *$-\hat{\nu}$ is the support function of $\mathscr{C}(\nu)$, i.e.,*

$$\hat{\nu}(f) = \min\{\langle f, \lambda\rangle \mid \lambda \in \mathscr{C}(\nu)\} \quad \textit{for every } f \in B(\Omega, \mathscr{F}).$$

Condition (i) is due to [7] and [15]; Conditions (ii) and (iii) are attributed to [15]; Condition (iv) is proven by [8]; Condition (v) is demonstrated by [16], which implies that ν is *exact*, i.e.,

$$\nu(A) = \min\{\lambda(A) \mid \lambda \in \mathscr{C}(\nu)\} \quad \text{for every } A \in \mathscr{F}. \tag{1.1}$$

All of these assertions have been proven in full generality by [8].

2 Main Result

Let $L^\infty(\Omega, \mathscr{F}, \mu)$ be the space of μ-essentially bounded functions on Ω, with which the weak*-topology $\sigma(L^\infty, L^1)$ is endowed. While $\hat{\nu}$ is Lipschitz of rank $\|\nu\|$ on $B(\Omega, \mathscr{F})$ if ν is of bounded variation (see [8]), the following theorem, the main result of this paper, states that the continuity of $\hat{\nu}$ can be strengthened to the weak*-continuity on $L^\infty(\Omega, \mathscr{F}, \mu)$ when ν is a supermodular capacity.

Theorem 2.1. *For every bounded set function ν, the following conditions are equivalent.*

(i) *ν is a supermodular capacity;*
(ii) *ν is of bounded variation and for every control measure μ for $\mathscr{C}(\nu)$, the Choquet integral $\hat{\nu} : B(\Omega, \mathscr{F}) \to \mathbb{R}$ of ν has a unique extension to $L^\infty(\Omega, \mathscr{F}, \mu)$ on which $\hat{\nu}$ is concave and weak*-continuous.*

[1] A finite measure μ satisfying condition (iii) is called a *control measure* for $\mathscr{C}(\nu)$.

Proof. (i) ⇒ (ii): Choose any control measure μ for $\mathscr{C}(\nu)$. Since every element in $\mathscr{C}(\nu)$ is absolutely continuous with respect to μ, we have that if $\mu(N) = 0$, then $\lambda(N) = 0$ for every $\lambda \in \mathscr{C}(\nu)$. By (1.1), we have $\nu(A \cup N) = \min_{\lambda \in \mathscr{C}(\nu)} \lambda(A \cup N) = \min_{\lambda \in \mathscr{C}(\nu)} \lambda(A) = \nu(A)$ for every $A \in \mathscr{F}$, which demonstrates that every μ-null set is ν-null. Take any $f \in L^\infty(\Omega, \mathscr{F}, \mu)$. Then $f = g$ a.e. for some $g \in B(\Omega, \mathscr{F})$. Since

$$(f \geq t) = [(f \geq t) \cap (f = g)] \cup [(f \geq t) \cap (f \neq g)]$$

for every $t \in \mathbb{R}$ and the sets $(f \geq t) \cap (f \neq g)$ and $(g \geq t) \cap (f \neq g)$ are μ-null, we have $\nu(f \geq t) = \nu(g \geq t)$. Hence, $\hat{\nu}(f) = \hat{\nu}(g)$ and the value $\hat{\nu}(f)$ is well defined because ν is of bounded variation by Proposition 1.1. It follows from this argument that the Choquet integral $\hat{\nu}$ defined on $B(\Omega, \mathscr{F})$ has a unique extension to $L^\infty(\Omega, \mathscr{F}, \mu)$ (which we do not relabel).

The concavity of $\hat{\nu}$ on $L^\infty(\Omega, \mathscr{F}, \mu)$ follows easily from Proposition 1.2 or from the observation that $\hat{\nu}$ is the pointwise minimum of the family of linear functionals $f \mapsto \langle f, \lambda \rangle$ on $B(\Omega, \mathscr{F})$ over $\lambda \in \mathscr{C}(\nu)$.

From Proposition 1.1(ii), it follows that ν is of bounded variation.

Let $\{f_\alpha\} \subset L^\infty(\Omega, \mathscr{F}, \mu)$ be a convergent net with $f_\alpha \to f$ in $\sigma(L^\infty, L^1)$. Then $\{f_\alpha\}$ is norm bounded in view of its relative weak*-compactness (see [6, Corollary V.4.3]). Since for each α there exists an element $\lambda_\alpha \in \mathscr{C}(\nu)$ such that $\hat{\nu}(f_\alpha) = \langle f_\alpha, \lambda_\alpha \rangle$ by Proposition 1.3(v) and $\mathscr{C}(\nu)$ is weak*-compact in $ba(\Omega, \mathscr{F})$, we can extract a weak*-convergent subnet $\{\lambda_\alpha\}$ (which we do not relabel) with $\lambda_\alpha \to \lambda_* \in \mathscr{C}(\nu)$. We thus have

$$\begin{aligned} |\langle f_\alpha, \lambda_\alpha \rangle - \langle f, \lambda_* \rangle| &= |\langle f_\alpha - f, \lambda_* \rangle + \langle f_\alpha, \lambda_\alpha - \lambda_* \rangle| \\ &\leq \left| \int (f_\alpha - f) d\lambda_* \right| + \|f_\alpha\|_\infty \left| \int \chi_\Omega d\lambda_\alpha - \int \chi_\Omega d\lambda_* \right| \\ &\leq \left| \int (f_\alpha - f) g_* d\mu \right| + C \left| \int \chi_\Omega d\lambda_\alpha - \int \chi_\Omega d\lambda_* \right| \to 0, \end{aligned}$$

where $g_* = \frac{d\lambda_*}{d\mu} \in L^1(\Omega, \mathscr{F}, \mu)$ and the constant C is such that $\|f_\alpha\|_\infty \leq C$ for each α. Therefore, $\hat{\nu}(f_\alpha) \to \langle f, \lambda_* \rangle$. Choose any $\lambda \in \mathscr{C}(\nu)$. We then have

$$\langle f_\alpha, \lambda_\alpha \rangle = \hat{\nu}(f_\alpha) \leq \langle f_\alpha, \lambda \rangle = \int f_\alpha g d\mu \quad \text{for each } \alpha,$$

where $g = \frac{d\lambda}{d\mu} \in L^1(\Omega, \mathscr{F}, \mu)$. Taking the limit for this inequality yields $\langle f, \lambda_* \rangle \leq \langle f, \lambda \rangle$ for every $\lambda \in \mathscr{C}(\nu)$. Therefore, $\hat{\nu}(f) = \min_{\lambda \in \mathscr{C}(\nu)} \langle f, \lambda \rangle = \langle f, \lambda_* \rangle$, and hence $\hat{\nu}(f_\alpha) \to \hat{\nu}(f)$.

(ii) ⇒ (i): Let μ be a control measure stated in condition (ii). Since the concavity of $\hat{\nu}$ is equivalent to the supermodularity of ν by Proposition 1.2, it suffices to show that ν is continuous.

To this end, let $A_k \uparrow A$ in $\mathscr{F}$. Then $\{\chi_{A_k}\}$ is a bounded sequence in $L^\infty(\Omega, \mathscr{F}, \mu)$ with $\chi_{A_k} \uparrow \chi_A$ a.e. For every $f \in L^1(\Omega, \mathscr{F}, \mu)$, we have $\chi_{A_k} f \to \chi_A f$ a.e. with $|\chi_{A_k} f| \leq |f|$ for each k. Then from Lebesgue's dominated convergence theorem,

$\int \chi_{A_k} f d\mu \to \int \chi_A f d\mu$ for every $f \in L^1(\Omega, \mathscr{F}, \mu)$, which implies $\chi_{A_k} \to \chi_A$ in $\sigma(L^\infty, L^1)$. Therefore, $\nu(A_k) = \hat{\nu}(\chi_{A_k}) \to \hat{\nu}(\chi_A) = \nu(A)$, and hence ν is continuous from below at every $A \in \mathscr{F}$.

The verification of the continuity from above is similar.

3 Application to Fair Allocation Problems

The fair division of a cake among a finite number of players is formulated as the partitioning of a measurable space $(\Omega, \mathscr{F})$. Here, the cake Ω (nonempty set) is a metaphor for a divisible heterogeneous commodity and the σ-algebra $\mathscr{F}$ of subsets of Ω describes a possible collection of pieces of the cake.

There are n players, indexed by $i = 1, \ldots, n$, whose preference on $\mathscr{F}$ is given by a *utility function* $\nu_i : \mathscr{F} \to \mathbb{R}$, in terms of which the inequality $\nu_i(A) \geq \nu_i(B)$ means that A is at least as good as B for player i. A *partition* of Ω is an n-tuple $(A_1, \ldots, A_n)$ of mutually disjoint elements $A_1, \ldots, A_n$ in $\mathscr{F}$ whose union is Ω, where each A_i is a piece of the cake given to player i. The primitive of a fair division problem consists of $\langle (\Omega, \mathscr{F}), \{\nu_i\}_{i=1}^n \rangle$.

Let $B(\Omega, \mathscr{F}; \mathbb{R}^n)$ be the space of $\mathbb{R}^n$-valued bounded measurable functions on Ω with the sup norm, where its generic element is denoted coordinately by $(f_1, \ldots, f_n)$. An element $(f_1, \ldots, f_n) \in B(\Omega, \mathscr{F}; \mathbb{R}^n)$ is an *allocation* of Ω if $\sum_{i=1}^n f_i = 1$ and $f_1, \ldots, f_n \geq 0$. The set of allocations of Ω is denoted by $\mathscr{A}$. Note that an n-tuple of measurable sets $(A_1, \ldots, A_n)$ is a partition of Ω if and only if $\sum_{i=1}^n \chi_{A_i} = 1$. The primitive of a fair allocation problem corresponding to that of the fair division problem $\langle (\Omega, \mathscr{F}), \{\nu_i\}_{i=1}^n \rangle$ is $\langle B(\Omega, \mathscr{F}), \{\hat{\nu}_i\}_{i=1}^n \rangle$, where $\hat{\nu}_i$ is the Choquet integral of ν_i.

Definition 3.1. *An allocation $(f_1, \ldots, f_n)$ is:*

(i) *Envy free if $\hat{\nu}_i(f_j) \leq \hat{\nu}_i(f_i)$ for each $i, j = 1, \ldots, n$.*
(ii) *Weakly Pareto optimal if there exists no allocation $(g_1, \ldots, g_n)$ such that $\hat{\nu}_i(f_i) < \hat{\nu}_i(g_i)$ for each $i = 1, \ldots, n$.*
(iii) *Pareto optimal if there exists no allocation $(g_1, \ldots, g_n)$ such that $\hat{\nu}_i(f_i) \leq \hat{\nu}_i(g_i)$ for each $i = 1, \ldots, n$ and $\hat{\nu}_j(f_j) < \hat{\nu}_j(g_j)$ for some j.*

The following result is a partial generalization of [1] and [18], the former demonstrated the existence of Pareto optimal envy-free allocations in the fair division problem and the latter showed that of Pareto optimal envy-free partitions in the fair division problem, both of whom assumed that each ν_i is a nonatomic finite measure.

Theorem 3.1. *If ν_i is a strictly monotone, null-additive, supermodular capacity for each $i = 1, \ldots, n$, then there exists a Pareto optimal envy-free allocation.*

4 Proof of Theorem 3.1

Under the hypothesis of Theorem 3.1, every element in $\mathscr{C}(\nu_i)$ is a finite measure and there exists a control measure μ_i for $\mathscr{C}(\nu_i)$ by Proposition 1.3. Let μ be

the finite measure given by $\mu = \sum_{i=1}^{n} \mu_i$. Then μ is a control measure for each $\mathscr{C}(\nu_i)$. From Theorem 2.1, each $\hat{\nu}_i$ defined on $B(\Omega, \mathscr{F})$ has a unique extension to $L^\infty(\Omega, \mathscr{F}, \mu)$ on which $\hat{\nu}_i$ is concave and weak*-continuous on $L^\infty(\Omega, \mathscr{F}, \mu)$.

We say that a function $T : L^\infty(\Omega, \mathscr{F}, \mu) \to \mathbb{R}$ is *strictly monotone* if $f \leq g$ and $f \neq g$ implies that $T(f) < T(g)$.

Lemma 4.1. *If ν_i is a strictly monotone, null-additive, supermodular capacity, then $\hat{\nu}_i$ is strictly monotone on $L^\infty(\Omega, \mathscr{F}, \mu)$.*

Proof. By the strict monotonicity of ν_i, every element in $\mathscr{C}(\nu_i)$ is nonzero and nonnegative. Suppose that the inequality $f \leq g$ with $f \neq g$ in $L^\infty(\Omega, \mathscr{F}, \mu)$ holds. Then there exists some $A \in \mathscr{F}$ with $\mu(A) > 0$ on which $f < g$. Since ν_i is equivalent to μ by Proposition 1.3(iv), the set A is ν_i-nonnull, which implies that $0 < \nu_i(A) \leq \lambda(A)$ for every $\lambda \in \mathscr{C}(\nu_i)$ by Proposition 1.1(i). Hence, $\hat{\nu}_i(f) = \min_{\lambda \in \mathscr{C}(\nu_i)} \langle f, \lambda \rangle < \min_{\lambda \in \mathscr{C}(\nu_i)} \langle g, \lambda \rangle = \hat{\nu}_i(g)$ by Proposition 1.3(v) and the weak*-compactness of $\mathscr{C}(\nu_i)$.

In the definition of allocations, we may adopt $L^\infty(\Omega, \mathscr{F}, \mu; \mathbb{R}^n)$ in place of $B(\Omega, \mathscr{F}; \mathbb{R}^n)$ because such a replacement makes the possible utility values of players along the allocations unchanged by Theorem 2.1.

Lemma 4.2. *Suppose that ν_i is strictly monotone, null-additive, supermodular capacity for each i. Then, an allocation is Pareto optimal if and only if it is weakly Pareto optimal.*

Proof. It is evident that Pareto optimality implies weak Pareto optimality. We show the converse implication. Let $(f_1, \ldots, f_n)$ be a weakly Pareto optimal allocation in $L^\infty(\Omega, \mathscr{F}, \mu; \mathbb{R}^n)$. Suppose that $(f_1, \ldots, f_n)$ is not Pareto optimal. There then exists an allocation $(g_1, \ldots, g_n)$ in $L^\infty(\Omega, \mathscr{F}, \mu; \mathbb{R}^n)$ such that $\hat{\nu}_i(f_i) \leq \hat{\nu}_i(g_i)$ for each i and $\hat{\nu}_j(f_j) < \hat{\nu}_j(g_j)$ for some j. Thus, there exists some $A \in \mathscr{F}$ with $\mu(A) > 0$ on which g_j is positive. By the weak*-continuity of $\hat{\nu}_j$, there exists some $\varepsilon > 0$ such that $\hat{\nu}_j(f_j) < \hat{\nu}_j((1-\varepsilon)g_j)$. Let $h_i = g_i + \frac{\varepsilon g_j}{n-1}$ for $i \neq j$ and $h_j = (1-\varepsilon)g_j$. Then the resulting allocation $(h_1, \ldots, h_n)$ satisfies $\hat{\nu}_i(f_i) < \hat{\nu}_i(h_i)$ for each i by the strict monotonicity of $\hat{\nu}_i$. This contradicts the weak Pareto optimality of $(f_1, \ldots, f_n)$.

As mentioned above, the set $\mathscr{A}$ of allocations can be defined in L^∞ without loss of generality by

$$\mathscr{A} = \left\{ (f_1, \ldots, f_n) \in L^\infty(\Omega, \mathscr{F}, \mu; \mathbb{R}^n) \mid \sum_{i=1}^{n} f_i = 1,\ f_1, \ldots, f_n \geq 0 \right\}.$$

Note that $\mathscr{A}$ is weak*-compact in $L^\infty(\Omega, \mathscr{F}, \mu; \mathbb{R}^m)$ because it is bounded in the essential sup norm and weak*-closed (see [6, Corollary V.4.3]). Define the utility possibility set by

$$V = \{(\hat{\nu}_1(f_1), \ldots, \hat{\nu}_n(f_n)) \in \mathbb{R}^n \mid (f_1, \ldots, f_n) \in \mathscr{A}\}.$$

It follows from the weak*-continuity of $\hat{\nu}_i$ and the weak*-compactness of $\mathscr{A}$ that V is compact in $\mathbb{R}^n$. Let $\mathscr{A}^*$ be the set of Pareto optimal allocations and define the *Pareto frontier* of V by

$$V^* = \{(\hat{\nu}_1(f_1), \ldots, \hat{\nu}_n(f_n)) \in \mathbb{R}^n \mid (f_1, \ldots, f_n) \in \mathscr{A}^*\}.$$

By Lemma 4.2, V^* is nonempty and included in the boundary of V.

Lemma 4.3 ([9]). *There exists a homeomorphism φ from Δ^{n-1} to V_n^* such that $\varphi(s) = \rho(s)s$ for $s \in \Delta^{n-1}$ with $\rho : \Delta^{n-1} \to (0, \infty)$ a continuous function.*

To prove Theorem 3.1, the following simple observation by [17] plays an important role for establishing Pareto optimal envy-free allocations.

Lemma 4.4 ([17]). *For every Pareto optimal allocation $(f_1, \ldots, f_n)$ there exists some j such that $\hat{\nu}_i(f_j) \leq \hat{\nu}_i(f_i)$ for each $i = 1, \ldots, n$.*

Proof. Arbitrarily take a Pareto optimal allocation $(f_1, \ldots, f_n)$. Suppose, to the contrary, that for each j there exists some $\pi(j) \in \{1, \ldots, n\}$ such that $\hat{\nu}_{\pi(j)}(f_{\pi(j)}) < \hat{\nu}_{\pi(j)}(f_j)$. Then the map π from $\{1, \ldots, n\}$ into itself defined by $j \mapsto \pi(j)$ satisfies $\pi(j) \neq j$ for each j. We thus have $\pi^s(j) \neq \pi^{s+1}(j)$ and $\hat{\nu}_{\pi^{s+1}(j)}(f_{\pi^{s+1}(j)}) < \hat{\nu}_{\pi^{s+1}(j)}(f_{\pi^s(j)})$ for every $s = 0, 1, \ldots$, where π^s is the s-th iteration of π with π^0 the identity map on $\{1, \ldots, n\}$. If $\pi^s(j)$ is distinct from $\pi^t(j)$ for every $s \neq t$, then $\{\pi^s(j)\}_{s=0}^{\infty}$ constitutes an infinite sequence of positive integers, which is obviously impossible. Therefore, for some integers $s > t \geq 0$, we have $\pi^s(j) = \pi(j)^{s-t}$. Let $i_0 = \pi^s(j), i_1 = \pi^{s-1}(j), \ldots, i_t = \pi^{s-t}(j)$ and $I = \{i_0, \ldots, i_t\}$. It is evident that $\hat{\nu}_{i_0}(f_{i_0}) < \hat{\nu}_{i_0}(f_{i_1}), \ldots, \hat{\nu}_{i_{t-1}}(f_{i_{t-1}}) < \hat{\nu}_{i_{t-1}}(f_{i_t})$, and $\hat{\nu}_{i_t}(f_{i_t}) < \hat{\nu}_{i_t}(f_{i_0})$. Define the allocation $(g_1, \ldots, g_n)$ by

$$g_i = \begin{cases} f_{i_{k+1}} & \text{if } i = i_k \text{ with } 1 \leq k \leq t-1, \\ f_{i_0} & \text{if } i = i_t, \\ f_i & \text{if } i \notin I. \end{cases}$$

It is obvious that the resulting allocation $(g_1, \ldots, g_n)$ satisfies $\hat{\nu}_i(f_i) < \hat{\nu}_i(g_i)$ for each $i \in I$ and $\hat{\nu}_i(g_i) = \hat{\nu}_i(f_i)$ for each $i \notin I$, and this contradicts the Pareto optimality of $(f_1, \ldots, f_n)$.

Lemma 4.5 ([14]). *Let $\Delta_i = \{(\alpha_1, \ldots, \alpha_n) \in \Delta^{n-1} \mid \alpha_i = 0\}$ for each $i = 1, \ldots, n$. If the collection $\{C_1, \ldots, C_n\}$ is a closed covering of Δ^{n-1} satisfying $\Delta_i \subset C_i$ for each i, then $\bigcap_{i=1}^n C_i \neq \emptyset$.*

Proof (Proof of Theorem 3.1). Let ψ be the homeomorphism from V^* to Δ^{n-1} defined by $\psi = \varphi^{-1}$ where φ is the homeomorphism from Lemma 4.3. For each $j = 1, \ldots, n$, define the sets $\mathscr{C}_j, C_j$ and D_j by

$$\begin{aligned} \mathscr{C}_j &= \{(f_1, \ldots, f_n) \in \mathscr{A}^* \mid \hat{\nu}_i(f_j) \leq \hat{\nu}_i(f_i)\ \forall i = 1, \ldots, n\}, \\ C_j &= \{(\hat{\nu}_1(f_1), \ldots, \hat{\nu}_n(f_n)) \in \mathbb{R}^n \mid (f_1, \ldots, f_n) \in \mathscr{C}_j\}, \\ D_j &= \{(x_1, \ldots, x_n) \in V^* \mid x_j = 0\}. \end{aligned}$$

As $\bigcup_{j=1}^{n} \mathscr{C}_j = \mathscr{A}^*$ by Lemma 4.4, the collection $\{\psi(C_1), \ldots, \psi(C_n)\}$ of sets in $\mathbb{R}^n$ is a closed covering of Δ^{n-1}. It is easy to verify that D_j is nonempty. Indeed, a solution to the maximization problem

$$\max \left\{ \sum_{i=1}^{n} \alpha_i x_i \mid (x_1, \ldots, x_n) \in V \right\}$$

with $\alpha \in \Delta^{n-1}$ and $\alpha_j = 0$ belongs to D_j in view of the strict monotonicity and continuity for each $\hat{\nu}_i$. Note that $\psi(D_j) \subset \psi(C_j)$ for each j. Because ψ is given by

$$V^* \ni (x_1, \ldots, x_n) = \rho(s)s \mapsto s \in \Delta^{n-1},$$

we have $\psi(D_j) = \Delta_j$ for each j. Therefore, by Lemma 4.5, there exists some $s \in \bigcap_{j=1}^{n} \psi(C_j) = \varphi^{-1}(\bigcap_{j=1}^{n} C_j)$. Then for some $(f_1, \ldots, f_n) \in \mathscr{A}^*$, we have $(\hat{\nu}_1(f_1), \ldots, \hat{\nu}_n(f_n)) = \rho(s)s \in \bigcap_{j=1}^{n} C_j$. Suppose that $(f_1, \ldots, f_n) \notin \mathscr{C}_j$ for some j. We then have $(\hat{\nu}_1(f_1), \ldots, \hat{\nu}_n(f_n)) \notin C_j$, a contradiction. Therefore, $(f_1, \ldots, f_n) \in \bigcap_{j=1}^{n} \mathscr{C}_j$. By construction, it is obvious that the allocation $(f_1, \ldots, f_n)$ is Pareto optimal and envy free.

References

[1] Akin, E.: Vilfred Pareto cuts the cake. J. Math. Econom. 24, 23–44 (1995)
[2] Choquet, G.: Theory of capacities. Ann. Inst. Fourier (Grenoble) 5, 131–295 (1955)
[3] Dall'Aglio, M.: The Dubins-Spanier optimization problems in fair division theory. J. Comput. Anal. Appl. 130, 17–40 (2001)
[4] Dall'Aglio, M., Maccheroni, F.: Disputed lands. Games Econom. Behav. 66, 57–77 (2009)
[5] Dubins, L.E., Spanier, E.H.: How to cut a cake fairly. Amer. Math. Monthly 68, 1–17 (1961)
[6] Dunford, N., Schwartz, J.T.: Linear Operators, Part I: General Theory. John Wiley & Sons, New York (1958)
[7] Kelley, J.L.: Measures on Boolean algebras. Pacific J. Math. 9, 1165–1177 (1959)
[8] Marinacci, M., Montrucchio, L.: Introduction to the mathematics of ambiguity. In: Gilboa, I. (ed.) Uncertainty in Economic Theory, Routledge, New York, pp. 46–107 (2004)
[9] Mas-Colell, A.: The price equilibrium existence problem in topological vector lattices. Econometrica 54, 1039–1053 (1986)
[10] Pap, E.: Null-Additive Set Functions. Kluwer Academic Publishers, Dordrecht (1995)
[11] Sagara, N.: A characterization of α-maximin solutions of fair division problems. Math. Social Sci. 55, 273–280 (2008)
[12] Sagara, N.: Fair division problems with nonadditive evaluations: Existence of solutions and representation of preference orderings. In: Varela, J., Acuña, S. (eds.) Handbook of Optimization Theory: Decision Analysis and Application. Nova Science Publishers, Hauppauge (2010) (forthcoming)
[13] Sagara, N., Vlach, M.: Convexity of the lower partition range of a concave vector measure. Adv. Math. Econ. 13, 155–160 (2010)

[14] Scarf, H.: The approximation of fixed points of a continuous mapping. SIAM J. Appl. Math. 15, 1328–1343 (1967)
[15] Schmeidler, D.: Cores of exact games, I. J. Math. Anal. Appl. 40, 214–225 (1972)
[16] Schmeidler, D.: Integral representation without additivity. Proc. Amer. Math. Soc. 97, 255–261 (1986)
[17] Varian, H.R.: Equity, envy, and efficiency. J. Econom. Theory 9, 63–91 (1974)
[18] Weller, D.: Fair division of a measurable space. J. Math. Econom. 14, 5–17 (1985)

Inclusion-Exclusion Integral and Its Application to Subjective Video Quality Estimation

Aoi Honda and Jun Okamoto

Department of Systems Design and Informatics, Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka 820-8502, Japan
aoi@ces.kyutech.ac.jp
NTT Service Integration Laboratories, NTT Corporation
okamoto.jun@lab.ntt.co.jp

Abstract. A new integral for non-additive measure encompassing the Choquet integral is proposed. We construct a model using it for multi-criteria decision making problem and verify its feasibility applying to the subjective quality estimation of video delivery services.

Keywords: Choquet integral, non-additive measure, multi-criteria decision making, inclusion-exclusion integral.

1 Introduction

The monotone measure, which is also called a capacity[1] or a fuzzy measure [2,3], is not additive so that we cannot use the Lebesgue integral as a integral with respect to monotone measure. Several integrals have been proposed for non-additive measure but the Choquet integral is the most popular and used in applications of non-additive measure because it has some preferable or useful properties. Then general classes of integral, which contains the Choquet integral and also the Sugeno integral[2,3], are discussed by several authors [4,5,6,7]. In this paper, we propose a new integral, called inclusion-exclusion integral, with respect to non-additive measure. This integral encompasses the Choquet integral. We also construct a new "objective" subjective evaluation model using it. In section 4 we apply our new model to the subjective evaluation problem of video delivery services qulaity using real large data. After that we discuss the conventional method of this field that can be interpreted as our inclusion-exclusion integral.

2 Preliminaries

Throughout the paper, the whole set is denoted by $\Omega := \{1, 2, \ldots, n\}$ and 2^Ω denotes the power set of Ω. The minimum and the maximum operation in 2^Ω are denoted by $\wedge$ and $\vee$, respectively. For a set A, the number of elements of A is denoted by $|A|$.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 480–489, 2010.

Definition 1 (monotone measure). *A set function $v : 2^\Omega \to [0,1]$ is called a monotone measure if it satisfies*
1. $v(\emptyset) = 0, v(\Omega) = 1$, *and*
2. $v(A) \leq v(B)$ *whenever* $A \subseteq B, A, B \in 2^\Omega$.

Definition 2 (Choquet integral). *Let v be a monotone measure defined on 2^Ω, and f a non-negative function on Ω. The Choquet integral of f w.r.t. v is defined by*

$$(C)\int f\ dv := \sum_{i=1}^{n}\big(f(\sigma(i)) - f(\sigma(i-1))\big)v(\{\sigma(i),\ldots,\sigma(n)\}),$$

where σ is a permutation on Ω such that $f(\sigma(1)) \leq \cdots \leq f(\sigma(n))$ and $f(\sigma(0)) := 0$.

Definition 3 (Möbius transform). *Let v be a set function on 2^Ω. The Möbius transform of v, denoted by m^v, is defined by*

$$m^v(A) := \sum_{B\subseteq A}(-1)^{|A\setminus B|}v(B) \tag{1}$$

for any $A \in 2^\Omega$. And v and m^v are one-to-one correspondence with

$$v(A) = \sum_{B\subseteq A} m^v(B) \tag{2}$$

for any $A \in 2^\Omega$.

Proposition 1 ([8]). *The Choquet integral of v is represented with Möbius transform of v by*

$$(C)\int f dv = \sum_{A\subseteq\Omega}\left(\bigwedge_{i\in A} f(i)\right) m^v(A). \tag{3}$$

Definition 4 (k-additive measure). *Let v be a monotone measure on 2^Ω and k a positive integer. A monotone measure v which satisfies $m^v(A) = 0$ whenever $|A| > k$ is called k-additive measure.*

A k-additive measure admits an interpretation that this measure has interactions only among k criteria. The proposition below shows that the Choquet integral is represented as a linear combination of the Möbius transforms of v.

Definition 5 (T-norm [9,10,11] etc.). *If a binary operation $\otimes : [0,1] \times [0,1] \to [0,1]$ satisfies*
1. $0 \otimes 0 = 0, x \otimes 1 = x$ *for any* $x > 0$,
2. $x \leq y$ *implies* $x \otimes z \leq y \otimes z$,
3. $x \otimes y = y \otimes x$ *and*
4. $x \otimes (y \otimes z)) = ((x \otimes y) \otimes z)$,

then $\otimes$ is called a T-norm.

By *4* in Definition 5, $\otimes$ is extended to $\otimes : [0,1]^n \to [0,1]$.

3 Inclusion-Exclusion Integral

The inclusion-exclusion integral is defined as follows.

Definition 6 (inclusion-exclusion integral). *Let v be a positive monotone measure defined on 2^Ω, $\otimes : \cup_{k=1}^{n}[0,1]^k \to [0,1]$ be a T-norm and let f be a non-negative function on Ω satisfying $f(i) \leq 1$ for $i \in \Omega$. The inclusion-exclusion integral (IE integral) is defined by*

$$\otimes \int f \; dv := \sum_{A \subseteq \Omega} \left\{ \sum_{B \supseteq A} \left((-1)^{|B \setminus A|} \bigotimes_{i \in B} f(i) \right) \right\} v(A), \tag{4}$$

where $\otimes_{i \in B} w_i = \otimes\{w_i \mid i \in B\}, B \subseteq \Omega$.

We named it after the inclusion-exclusion formula. Various inclusion-exclusion integrals can be introduced using various T-norms.

Example 1. Adopting the multiplication as the T-norm in inclusion-exclusion integral, we have

$$(M) \int f \; dv := \sum_{A \subseteq \Omega} \left\{ \sum_{B \supseteq A} \left((-1)^{|B \setminus A|} \prod_{i \in B} f(i) \right) \right\} v(A).$$

We call it multiple inclusion-exclusion integral (MIE integral).

Proposition 2. *Let v be a positive monotone measure defined on 2^Ω. Let f and g be non-negative functions on Ω dominated by 1.*
(i) $f \leq g$ implies

$$(M) \int f \; dv \leq (M) \int g \; dv.$$

(ii) If v is a classical measure, i.e., v is additive, then

$$(M) \int f \; dv = \int f \; dv,$$

where the right-hand side is the Lebesgue integral.
(iii)

$$(M) \int \chi_A \; dv = v(A),$$

for any $A \in 2^\Omega$, where χ_A is the characteristic function of the set $A \in 2^X$.

Proposition 3. *Let v be a positive monotone measure defined on 2^Ω. We have for every pair (f, g) of non-negative functions on Ω dominated by 1. We have*

$$(M) \int (f + g) \; dv \leq (M) \int f \; dv + (M) \int g \; dv$$

if and only if

$$v(A \cup B) + v(A \cap B) \leq v(A) + v(B)$$

and

$$(M)\int (f+g)\ dv \geq (M)\int f\ dv + (M)\int g\ dv$$

if and only if

$$v(A \cup B) + v(A \cap B) \geq v(A) + v(B).$$

The Choquet integral also have properties of Proposition 3 and 4.

Proposition 4. *The inclusion-exclusion integral is represented with Möbius transform of v by*

$$\otimes \int f\ dv = \sum_{A \subseteq \Omega} \left(\bigotimes_{i \in A} f(i) \right) m^v(A). \tag{5}$$

Hereafter we use the above form as the IE integral instead of (4).

Corollary 1. *The Choquet integral is one of inclusion-exclusion integrals, whose T-norm is the minimum operation. In other words, the Choquet integral of v is represented by*

$$(C)\int f\ dv := \sum_{A \subseteq \Omega} \left\{ \sum_{B \supseteq A} \left((-1)^{|B \setminus A|} \bigwedge_{i \in B} f(i) \right) \right\} v(A).$$

4 Application to Subjective Quality Estimation of Video Delivery Services

We analyze the large amount of real data of human subjective evaluations on a product. We regard that this product has n criteria, $X_1, X_2, \ldots, X_n$, determining its performance which means $\Omega = \{1, 2, \ldots, n\} := \{X_1, X_2, \ldots, X_n\}$. The data concerning these criteria are obtained by a survey with large amount of sample products whose values for $X_1, X_2, \ldots, X_n$ are slight different from others. A datum is $(n+1)$-tuple which consists of n explanatory objective values $x_1, x_2, \ldots, x_n$ for $X_1, X_2, \ldots, X_n$ and one subjective value y which is intuitive evaluation of the product. In other words, data are $(x_1^1, \ldots, x_i^1, \ldots, x_n^1, y^1), \ldots, (x_1^j, \ldots, x_i^j, \ldots, x_n^j, y^j), \ldots, (x_1^m, \ldots, x_i^m, \ldots, x_n^m, y^m)$ and the last one of $(n+1)$-tuple is the intuitive evaluation of the product. The format of each datum is as Table 1. Since the scale of values depend on each criteria, we shall normalize them in $[0, 1]$ before analyzing with

$$f^j(X_i) = \frac{x_i^j - \min\limits_j (x_i^j)}{\max\limits_j (x_i^j) - \min\limits_j (x_i^j)}.$$

Table 1. Crude data

No.	x_1	x_2	x_3	x_4	x_5	evaluation y
1	44.46	0.93	0.35	0.77	1.00	4.79
2	41.24	1.38	0.42	1.36	1.00	4.58
3	39.06	1.33	0.45	1.91	1.00	4.54
⋮	⋮	⋮	⋮	⋮	⋮	⋮
j	36.49	0.72	0.35	2.77	6.85	4.04
⋮	⋮	⋮	⋮	⋮	⋮	⋮
m	42.58	0.00	0.00	100.00	69.96	3.42

For a criterion such that the smaller values make the product the better evaluation, we reverse the value in $[0, 1]$:

$$f^j(X_i) = 1 - \frac{x_i^j - \min\limits_j(x_i^j)}{\max\limits_j(x_i^j) - \min\limits_j(x_i^j)}. \tag{6}$$

n-tuple data normalised in this way correspond to $f(X_1), f(X_2), \ldots, f(X_n)$.

Table 2. Normalized data

No.	$f(X_1)$	$f(X_2)$	$f(X_3)$	$f(X_4)$	$f(X_5)$	y
1	0.8272	0.7362	0.7127	0.9974	1.00	4.79
2	0.7270	0.6057	0.6751	0.9915	1.00	4.58
3	0.6591	0.6208	0.6751	0.9915	1.00	4.54
⋮	⋮	⋮	⋮	⋮	⋮	⋮
j	0.5789	0.7952	0.7821	0.9850	0.6735	4.04
⋮	⋮	⋮	⋮	⋮	⋮	⋮
m	0.7688	1.00	0.9771	0.00	0.8233	3.42

We shall consider that n objective values are input values and one subjective value is output value.

We use real data of subjective quality on video delivery services. We regard that the quality has five criteria:

X_1 := overall noize,
X_2 := degradation caused by block distortion,
X_3 := degradation associated with blurring,
X_4 := local spatial degradation,
X_5 := freeze degradation,

determining its performance. These data are obtained by a survey with large amount, which is 1139 samples whose values of X_1, X_2, X_3, X_4 and X_5 are slight different from others. A datum consists of five explanatory objective values for X_1, X_2, X_3, X_4, X_5, and one subjective value which is perceptual quality of the video. Table 3 and 4 are data that we used. The smaller values of X_2, X_3, X_4 and X_5 are better so that we reverse them using (6).

4.1 Multiple Linear Regression Model

The multiple linear regression models are widely used for multi-criteria decision making problems. The multiple linear regression model is given by

$$y = \sum_{i=1}^{n} f(X_i)a_i + e.$$

In the case applying this model, we estimate coefficients $a_1, \ldots, a_n$ and e by data using, in general, the method of least squares. Each a_i means the degree of importance of i-th criterion X_i. If we don't need them, we need not have to normarize $f(X_i)$.

The advantages of this model are following:

1. The linearity of the model enables us to identify the model easily, in other words, we can estimate parameters $a_1, \ldots, a_n$ and e easily.
2. There are many tools or software for this model.

The problem of this model is following:

1. There exist several unexplainable cases by this model. One reasons for it is assumed that this model cannot handle the interactions between criteria.

4.2 Inclusion-Exclusion Integral Model

In this model, it is assumed that the degrees of importance of criteria have interactions with other criteria. Let $N = \{1, \ldots, n\}$ be a set of criterion and v be a monotone measure. The evaluation for f and that of the case of $n = 5$ are given by

$$y = \sum_{A \subseteq N} \left(\bigotimes_{i \in A} f(X_i) \right) m^v(X_i) + e.$$

If it is needed, we can obtain v by m^v with (2). Obtained v is not necessarily monotone.

We adopt the minimum operation and the multiplication as the T-norm in the inclusion-exclusion integral, in other words, we use the Choquet integral and the MIE integral. The advantages of this model are following:

1. This model can express interactions among criteria, so that it can be applied to various cases more than the multiple linear regression model.

2. The form of the integration is linear, so that we can use tools or software for the multiple linear regression model.

The problem of this model is as following:

1. In the case of the Choquet Integral, it can be difficult to retain independency among values. For example, in an extreme case that all data satisfies $f(1) < f(2) < \cdots < f(n)$, then the values $f(1), f(1) \wedge f(2), \ldots, f(1) \wedge \cdots \wedge f(n)$ are completely same. In the case that there are some data depending on other data, we can not estimate parameters m^v at all.
2. The model consists of 2^n parameters, which are $v(A)$ for $2^N \setminus \{\emptyset\}$ and e. The more n increases, the more the number of parameters explodes. We have to prepare large amount of data corresponding to the number of parameters. In other words, the more the number of parameters increases, the more we need large amount data. For example, in general, it is known that to use 16 parameters more than 1000 data is needed.

We can reduce the second problem utilizing k-additive measure as v. Using k-additive measure make the number of parameters decrease from 2^n to $\sum_{j=1}^{k} {}_n\mathrm{C}_j + 1$. It also makes retaining independency less difficult, so that problem 1 is reduced. k-additive model is given by

$$y = \sum_{\substack{A \subseteq N \\ |A| \leq k}} \left(\bigotimes_{i \in A} f(X_i) \right) m^v(X_i) + e.$$

Now we have 5 criteria and adopting $k = 2$ we can decrease from 32 parameters to 16 parameters.

Remark that parameters are not restricted, so that we cannot guarantee obtained v by m^v is monotone.

5 Results and Discussion

We have examined three models as follows:

1. the usual multiple linear regression model (Section 4.1),
2. the Choquet integral model with 2-additive measure (Section 4.2) and
3. the MIE integral model with 2-additive measure (Section 4.2).

We have estimated parameters of each model and obtained the squared multiple correlation coefficient R^2, the root mean squared error(RMSE) and Akaike's information criterion(AIC). The squared multiple correlation coefficient R^2 is the square of the correlation coefficient of real data values and estimated values:

$$R^2 := \left(\frac{\sum_{i=1}^{m} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{m} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{m} (\hat{y}_i - \bar{\hat{y}})^2}} \right)^2,$$

where y_i is the true value, $\hat{y}_i$ is the estimate of y_i, $\bar{y_i}$ is the mean of $y_i, i = 1, \ldots, m$ and $m = 1139$. It means that the estimation is better if R is closer to 1. The RMSE is calculated by

$$\mathrm{RMSE} := \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}.$$

The AIC is calculated by

$$\mathrm{AIC} = m\left(\log\left(\frac{2\pi}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2\right) + 1\right) + 2(p+1),$$

where p is the number of parameters. It means that the estimation is better if its AIC is smaller. In general, the more number of parameters, the model fit data the more. In such case the AIC enables us to compare models whose number of parameters are different.

Table 3. Squared correlation coefficients and root mean squared errors

model	p	R^2	RMSE	AIC
multiple linear regression model	6	0.567	0.664	2781.95
Choquet integral model	16	0.663	0.464	2393.74
MIE model	16	0.702	0.435	2320.23

The results are shown in Table 3. Results of analysis indicate the validity of our proposed method comparing with the multiple linear regression model.

On the other hand, a model for subjective video quality has been recommended by Telecommunication Standardizaion Sector of International Telecommunication Union (ITU-T) as J.247 Annex A [12]:

$$\begin{aligned} y &= a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5\log_{10}(x_5) \\ &\quad + a_6(a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4)\log_{10}(x_5) + e \\ &= a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5' \\ &\quad + (a_1a_6)x_1x_5' + (a_2a_6)x_2x_5' + (a_3a_6)x_3x_5' + (a_4a_6)x_4x_5' + e, \end{aligned}$$

where $x_5' := \log_{10}(x_5)$. This model specialize in subjective video quality estimation. We can regard it as the MIE integral with 2-additive measure considering that coefficients of $x_2x_3, x_2x_4, \ldots, x_3x_4$ are vanished. Values of X_5 largely concentrate to a small interval so that translating them by the increasing function $\varphi(x) = \log_{10} x$ is reasonable for the sake of dispersing them. If we make other values translation adequately, we obtain better estimation [4,13].

Using x_5' instead of x_5, we obtain the following results.

In also this case, IE integral models are better than conventional models.

Table 4. squared correlation coefficients and root mean squared errors

model	p	R^2	RMSE	AIC
multiple linear regression model	6	0.651	0.596	2658.89
Choquet integral model	16	0.733	0.522	2527.89
MIE model	16	0.723	0.532	2549.51
ITU-T R. J.247 Annex A	7	0.712	0.542	2552.72

6 Conclusions

We have examined the validity of the proposed model applying to the estimation for subjective evaluation with large reliable data. We may go on from the results of the examination to the conclusion that our proposed model is valid for analyzing subjective evaluation problems. Adopting other T-norms, for example parametrized families of T-norms, we can obtain better model to meet each problem.

The proposed IE integral model is based on the non-additive set function and it can be represented as a formula of a linear combination by the Möbius transform, so that we can take advantage of the multiple linear regression analysis. Using the inverse Möbius Transform, we obtain non-additive measure, which is not necessarily monotone, therefore we can know not only degrees of importance of each item but also more information from results of the analysis. For example, interactions between explanatory values.

The future plans are to studying properties of the inclusion-exclusion integral and to verify its feasibility more detail particularly applying this model to various data and also investigating other conventional models of various subjective evaluation problems which can be interpreted as inclusion-exclusion integral models.

References

1. Choquet, G.: Theory of capacities. Ann. Inst. Fourier 5, 131–295 (1953)
2. Sugeno, M.: Theory of fuzzy integrals and its applications. PhD Thesis, Tokyo Institute of Technology (1974)
3. Sugeno, M.: Fuzzy measures and fuzzy integrals–a survey. In: Gupta, M.M., Saridis, G.N., Gaines, B.R. (eds.) Fuzzy automata and decision processes, pp. 89–102. North Holland, Amsterdam (1977)
4. Murofushi, T., Sugeno, M.: Fuzzy t-conorm integrals with respect to fuzzy measures: generalization of Sugeno integral and Choquet integral. Fuzzy Sets and Systems 42, 57–71 (1991)
5. Mesiar, R.: Choquet-like integral. J. Math Anal. Appl. 194, 477–488 (1995)
6. Benvenuti, P., Vivona, D.: General theory of the fuzzy integral. Mathware and soft computing 3, 199–209 (1996)
7. Klement, E.P., Kolesarova, A., Mesiar, R.: Lipschitz continuity of discrete universal integrals based on copuras. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 18(1), 39–52 (2010)

8. Grabisch, M., Labreuche, C.: A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. A Quarterly Journal of Operations Research 6, 1–44 (2008)
9. Schweizer, B., Sklar, A.: Statistical metric spaces. Pacific. J. Math. 10, 313–334 (1960)
10. Klement, E.P.: Some mathematical aspects of fuzzy sets: Triangular norms, fuzzy logics, and generalized measures. Fuzzy Sets and Systems 90, 133–140 (1997)
11. Mesiar, R., Pap, E.: Different interpretations of triangular norms and related operations. Fuzzy Sets and Systems 96, 183–189 (1998)
12. ITU-T Recommendation J.247: Objective perceptual multimedia video quality measurement in the presence of a full reference (2008)
13. Mori, D., Honda, A., Uchida, M., Okazaki, Y.: Quality evaluations of network services using a non-additive set function. In: Proceedings of Modeling Decisions for Artificial Intelligence, CD-ROM (2008)

Fuzzy Measure Spaces Generated by Fuzzy Sets*

Antonín Dvořák and Michal Holčapek

University of Ostrava, Institute for Research and Applications of Fuzzy Modeling,
30. dubna 22, 70103 Ostrava, Czech Republic
antonin.dvorak@osu.cz, michal.holcapek@osu.cz
http://irafm.osu.cz

Abstract. In this paper we study fuzzy measures which are defined on algebras of fuzzy subsets of a given fuzzy set. Subsequently, we use these fuzzy measures to define a fuzzy integral. We show basic properties of these fuzzy measures and integrals. We are motivated by our research on fuzzy quantifiers, where this fuzzy integral is used.

1 Introduction and Motivation

In this paper we study fuzzy measures which are defined on algebras of fuzzy subsets of a given fuzzy set. We are motivated by our study of fuzzy quantifiers [1,2] which generalizes the theory of generalized quantifiers [3,4] and follows a research line started in [5]. In [2], we define fuzzy quantifiers of type $\langle 1 \rangle$ determined by fuzzy measures, where fuzzy measurable and measure spaces and fuzzy measures were defined on algebras of fuzzy subsets of a given crisp set M. Fuzzy quantifiers were then defined as integrals of fuzzy sets.

When we tried to generalize these quantifiers to type $\langle 1, 1 \rangle$ (and possibly higher), we found out that it would be advantageous to define fuzzy measure space on an algebra of fuzzy subsets of a given *fuzzy* set A. Then, the measure of this fuzzy set A is always equal to $\top$ (or 1 depending on which structure of truth values we use). This generating fuzzy set A is a model of the *restriction* part of $\langle 1, 1 \rangle$ quantifier. For example, in "Many Swedes are blond", "Swedes" (or "to be a Swede") is the restriction and "blond" (or "to be blond") is the scope of quantifier "many". It is then natural to consider the restriction as the domain of objects to which the scope applies. Hence, it is natural to consider it as a (fuzzy) set which should have the maximal possible measure.

Fuzzy measures and integrals ([6], see also [7,8]) are important tools allowing us to compare classical or fuzzy sets with respect to their size. Usually, fuzzy measures are set functions defined on some algebra of sets which are monotone with respect to inclusion and which assign zero to the empty set. In this contribution, fuzzy measures are defined on algebras of fuzzy sets (measure spaces) and, generally, they attain values from a complete residuated lattice $\mathbf{L}$.

A fuzzy integral is then defined using these fuzzy measure spaces and some of its properties are investigated. This integral can be used for the definition of

* This paper has been supported by the grant IAA108270901 of the GA AV ČR.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 490–499, 2010.

(models of) fuzzy quantifiers. Contrary to usual definitions of fuzzy integrals, these integrals can be used to integrate all fuzzy sets, that means, also fuzzy sets which are not standardly measurable with respect to the used fuzzy measure space. This enables us to introduce fuzzy quantifiers over spaces of all fuzzy sets and not only over spaces of all measurable fuzzy sets (cf. [9]).

Let us remark that, besides these fuzzy measure spaces it is also possible to define *complementary fuzzy measure spaces* on a fuzzy set A. We start with the same fuzzy measurable space and define *complementary fuzzy measure* in such a way that the measure of the empty set is $\top$ and the measure of A is $\bot$ (see analogous definition in [2], Sect. 3.1). If the structure of truth values is a complete MV-algebra, then it is possible to define a complementary fuzzy measure by means of a fuzzy measure and vice versa. Integrals on these complementary fuzzy measure spaces can be used for defining quantifiers like "no" "at most half", etc. Due to limited space, we are not investigating this type of fuzzy measures further in this paper.

This paper is structured as follows: Sect. 2 brings necessary preliminaries on residuated lattices and fuzzy sets. In Sect. 3 we introduce our concept of fuzzy measurable spaces, define fuzzy measures on these spaces and show some of their properties. Further, we define isomorphisms of fuzzy measurable spaces and of fuzzy measure spaces. To have these isomorphisms defined is vital for the investigation of isomorphism-invariant fuzzy quantifiers. In Sect. 4 we define fuzzy integral on fuzzy measure space generated by a given fuzzy set A and present its basic properties. Finally, in Sect. 5 we comment on our results and provide hints about directions of our further research. Due to a limited space, we are not providing full proofs of theorems in this paper.

2 Preliminaries

2.1 Structures of Truth Values

In this paper, we suppose that the structure of truth values is a *complete residuated lattice* (see e.g. [10]), i.e., an algebra $\mathbf{L} = \langle L, \wedge, \vee, \to, \otimes, \bot, \top \rangle$ with four binary operations and two constants such that $\langle L, \wedge, \vee, \bot, \top \rangle$ is a complete lattice, where $\bot$ is the least element and $\top$ is the greatest element of L, respectively, $\langle L, \otimes, \top \rangle$ is a commutative monoid (i.e., $\otimes$ is associative, commutative and the identity $a \otimes \top = a$ holds for any $a \in L$) and the adjointness property is satisfied, i.e.

$$a \leq b \to c \quad \text{iff} \quad a \otimes b \leq c \tag{1}$$

holds for each $a, b, c \in L$, where $\leq$ denotes the corresponding lattice ordering. A residuated lattice is *divisible*, if $a \otimes (a \to b) = a \wedge b$ holds for arbitrary $a, b \in L$, and satisfies the *law of double negation*, if $(a \to \bot) \to \bot = a$ holds for any $a \in L$. A divisible residuated lattice satisfying the law of double negation is called an *MV-algebra*. For other information about residuated lattices we refer to [10,11].

Example 1. It is easy to prove (see e.g. [12]) that the algebra

$$\mathbf{L}_{\mathrm{T}} = \langle [0,1], \min, \max, T, \to_T, 0, 1 \rangle,$$

where T is a left continuous t-norm and $a \to_T b = \bigvee \{c \in [0,1] \mid T(a,c) \leq b\}$ defines the residuum, is a complete residuated lattice. Moreover, if T is the Łukasiewicz t-norm, i.e., $T(a,b) = \max(a+b-1, 0)$ for all $a, b \in [0,1]$, then $\mathbf{L}_{\mathrm{T}}$ is a complete MV-algebra called a *Łukasiewicz algebra (on* $[0,1]$*)*.

Let us define the following additional operations for all $a, b \in L$:

$$a \leftrightarrow b = (a \to b) \wedge (b \to a), \qquad (\textit{biresiduum})$$
$$\neg a = a \to \bot. \qquad (\textit{negation})$$

2.2 Fuzzy Sets

Let $\mathbf{L} = \langle L, \wedge, \vee, \to, \otimes, \bot, \top \rangle$ be a complete residuated lattice and M be a universe of discourse (possibly empty). A mapping $A : M \to L$ is called a *fuzzy set on* M.[1] A value $A(m)$ is called a *membership degree of* m *in the fuzzy set* A. The set of all fuzzy sets on M is denoted by $\mathcal{F}(M)$. Obviously, if $M = \emptyset$, then the empty mapping $\emptyset$ is the unique fuzzy set on $\emptyset$ and thus $\mathcal{F}(\emptyset) = \{\emptyset\}$. A fuzzy set A on M is called *crisp*, if there is a subset X of M such that $A = 1_X$, where 1_X denotes the characteristic function of X. Particularly, $1_\emptyset$ denotes the empty fuzzy set on M, i.e., $1_\emptyset(m) = \bot$ for any $m \in M$. This convention will be also kept for $M = \emptyset$. The set of all crisp fuzzy sets on M is denoted by $\mathcal{P}(M)$. A fuzzy set A is *constant*, if there is $c \in L$ such that $A(m) = c$ for any $m \in M$. For simplicity, a constant fuzzy set is denoted by the corresponding element of L, e.g., a, b, c.[2] Let us denote $\mathrm{Supp}(A) = \{m \mid m \in M \ \&\ A(m) > \bot\}$ and $\mathrm{core}(A) = \{m \mid m \in M \ \&\ A(m) = \top\}$, the *support* and *core* of a fuzzy set A, respectively. Obviously, $\mathrm{Supp}(1_X) = \mathrm{core}(1_X) = X$ for any crisp fuzzy set. A fuzzy set A is called *normal*, if $\mathrm{core}(A) \neq \emptyset$. Let $A \in \mathcal{F}(M)$ and Z be a set. Then $A \restriction Z$ denotes the restriction of $A : M \to L$ to Z.

Let $\{A_i \mid i \in I\}$ be a non-empty family of fuzzy sets on M. Then the *union of* A_i is defined by

$$\left(\bigcup_{i \in I} A_i\right)(m) = \bigvee_{i \in I} A_i(m) \tag{2}$$

[1] In many papers (see e.g. [10]), a mapping $A : M \to L$ is called $\mathbf{L}$-fuzzy set or $\mathbf{L}$-fuzzy subset on M. Since we will always deal with a fixed complete residuated lattice in the following text, we suppose that the denotation "fuzzy set" without a reference to the considered residuated lattice is sufficient.

[2] We suppose that the meaning of this symbol will be unmistakable from the context, that is, it should be clear when an element of L is considered and when a constant fuzzy set is assumed.

for any $m \in M$ and the *intersection of* A_i is defined by

$$\left(\bigcap_{i\in I} A_i\right)(m) = \bigwedge_{i\in I} A_i(m) \tag{3}$$

for any $m \in M$. Let A, B be fuzzy sets on M. The *difference* of A and B is a fuzzy set $A \setminus B$ on M defined by

$$(A \setminus B)(m) = A(m) \wedge (B(m) \to \bot) \tag{4}$$

for any $m \in M$ and the *complement* of A is a fuzzy set $\overline{A} = 1_M \setminus A$. Finally, an extension of the operations $\otimes$ and $\to$ on L to the operations on $\mathcal{F}(M)$ is given by

$$(A \otimes B)(m) = A(m) \otimes B(m) \quad \text{and} \quad (A \to B)(m) = A(m) \to B(m) \tag{5}$$

for any $A, B \in \mathcal{F}(M)$ and $m \in M$, respectively.

Since we will deal with fuzzy sets over different universes, let us introduce the ordering relation between fuzzy sets on arbitrary universes. We say that a fuzzy set A on M is a *fuzzy subset of* a fuzzy set B on M' and denote by $A \subseteq B$, if $A(m) \leq B(m)$ for any $m \in \mathrm{Supp}(A)$. The set of all fuzzy subsets of A on M is denoted by $\mathcal{F}(A)$. Thus

$$\mathcal{F}(A) = \{B \mid B \in \mathcal{F}(M) \text{ and } B \subseteq A\}. \tag{6}$$

Further, we say that a fuzzy set A on M is *equal* to a fuzzy set B on M' and denote by $A = B$, if $A \subseteq B$ and $B \subseteq A$. Obviously, if A and B are fuzzy sets on the same universe M, then the proposed relations coincide with the common definitions of ordering and equality relations.

Let $f : M \to M'$ be a mapping. A mapping $f^{\to} : \mathcal{F}_{\mathbf{L}}(M) \to \mathcal{F}_{\mathbf{L}}(M')$ defined by $f^{\to}(A)(m) = \bigvee_{m' \in f^{-1}(m)} A(m')$ is called the *fuzzy extension* of the mapping f. Obviously, if f is a bijective mapping, then $f^{\to}(A)(f(m)) = A(m)$ for any $m \in M$.

3 Fuzzy Measures

In the following, we will consider algebras of fuzzy sets as a base for defining fuzzy measures of fuzzy sets.

Definition 1. *Let A be a non-empty fuzzy set on M. A subset $\mathcal{F}$ of $\mathcal{F}(A)$ is an* algebra of fuzzy sets on A, *if the following conditions are satisfied*

(i) $1_\emptyset, A \in \mathcal{F}$,
(ii) if $X \in \mathcal{F}$, *then* $A \setminus X \in \mathcal{F}$,
(iii) if $X, Y \in \mathcal{F}$, *then* $X \cup Y \in \mathcal{F}$.

We denote $\mathbf{Alg}(A)$ the set of all algebras of fuzzy sets on A. A pair $(A, \mathcal{F})$ is called a *fuzzy measurable space* (on A), if $\mathcal{F}$ is an algebra of fuzzy sets on A. Let $(A, \mathcal{F})$ be a fuzzy measurable space and $X \in \mathcal{F}(A)$. We say that X is $\mathcal{F}$*-measurable*, if $X \in \mathcal{F}$.

Remark 1. In order to refer to the universe of discourse M of A, we will sometimes write $\mathrm{Dom}(A)$ instead of M.

Remark 2. For a non-empty set M, one checks easily that the proposed definition of algebra of fuzzy sets on 1_M coincides with the standard definition of algebra of fuzzy sets on M (cf. [6]).

Example 2. $\{1_\emptyset, A\}$ and $\mathcal{F}(A)$ are trivial examples of algebras of fuzzy sets on A. The set of all crisp subsets of a fuzzy set A is not, in general, an algebra of fuzzy sets on A.

Contrary to the algebra of sets, the conditions of the algebra of fuzzy sets does not ensure that the intersection of two $\mathcal{F}$-measurable fuzzy sets is also $\mathcal{F}$-measurable fuzzy set, i.e., in general, there can be $A, B \in \mathcal{F}$ with $A \cap B \notin \mathcal{F}$. Therefore, we define the following type of algebra.

Definition 2. *We say that an algebra $\mathcal{F}$ of fuzzy sets on A is* closed under intersections*, if $C \cap D \in \mathcal{F}$, whenever $C, D \in \mathcal{F}$.*

Theorem 1. *Let* $\mathbf{L}$ *be a complete residuated lattice satisfying the law of double negation and $\overline{A} = 1_\emptyset$. Then each algebra of fuzzy sets on A is closed under intersections.*

Let us introduce the concept of fuzzy measure as follows. The definition is a modification of the definition of a normed measure with respect to truth values (see e.g. [7,8]).

Definition 3. *Let $(A, \mathcal{F})$ be a fuzzy measurable space. A mapping $\mu : \mathcal{F} \to L$ is called a* fuzzy measure *on $(A, \mathcal{F})$, if*

(i) $\mu(1_\emptyset) = \bot$ *and* $\mu(A) = \top$,
(ii) if $B, C \in \mathcal{F}$ *such that* $B \subseteq C$, *then* $\mu(B) \leq \mu(C)$.

A triplet $(A, \mathcal{F}, \mu)$ is called the *fuzzy measure space*, if $(A, \mathcal{F})$ is a fuzzy measurable space and μ is a fuzzy measure on $(A, \mathcal{F})$. We denote $\mathbf{Fms}(M)$ the class of all fuzzy measurable spaces defined on a non-empty universe M, i.e., a fuzzy measure space $(A, \mathcal{F}, \mu)$ belongs to $\mathbf{Fms}(M)$, if $A \in \mathcal{F}(M)$. It is easy to see that $\mathbf{Fms}(M) \subseteq \mathbf{Fms}(M')$, whenever $M \subseteq M'$.

In the following text, a fuzzy measure space $(A, \mathcal{F}, \mu)$ may be shortly denoted by the bold symbol $\mathbf{A}$, i.e. $\mathbf{A} = (A, \mathcal{F}, \mu)$.

Sometimes, it is useful to extend a fuzzy measure on a new fuzzy measurable space, where we may have a problem to define it directly. One of the well-known approaches for measures is to establish an inner (or outer) measure, when the values of inner measure are obtained approximatively using known values for

fuzzy subsets (or supersets). In the following part, we will do it for the fuzzy case.

Let $(A, \mathcal{F})$ be a fuzzy measurable space and $X \in \mathcal{F}(M)$. Denote $\mathcal{F}_X$ the set of all $\mathcal{F}$-measurable sets which are contained in X, i.e.,

$$\mathcal{F}_X = \{B \mid B \in \mathcal{F} \text{ and } B \subseteq X\}. \tag{7}$$

Note that $1_\emptyset \in \mathcal{F}_X$ for each $X \in \mathcal{F}(M)$ and if X is an $\mathcal{F}$-measurable set, then also $X \in \mathcal{F}_X$. If $X = A$, then we will write only $\mathcal{F}$ instead of $\mathcal{F}(A)$. The following theorem introduces *inner fuzzy measure on* $(M, \mathcal{F}')$ *determined by fuzzy measure* μ on some fuzzy measurable space $(M, \mathcal{F})$, where $\mathcal{F} \subseteq \mathcal{F}'$.

Theorem 2. *Let* $(A, \mathcal{F}, \mu)$ *be a fuzzy measure space and* $\mathcal{F}' \in \mathbf{Alg}(A)$ *such that* $\mathcal{F} \subseteq \mathcal{F}'$. *A mapping* $\mu^* : \mathcal{F}' \to L$ *defined by*

$$\mu^*(X) = \bigvee_{B \in \mathcal{F}_X} \mu(B) \tag{8}$$

is a fuzzy measure on the fuzzy measurable space $(A, \mathcal{F}')$. *Moreover,* μ^* *and* μ *coincide on* $\mathcal{F}$. *We say that* μ^* *is the* inner fuzzy measure on $(A, \mathcal{F}')$ determined by μ.

Example 3. Let $(A, \mathcal{F}, \mu)$ be the fuzzy measure space where $\mathcal{F} = \{1_\emptyset, A\}$ and $\mathcal{F}' = \mathcal{F}(A)$ (the set of all fuzzy subsets of A). Then

$$\mu^*(X) = \bigvee_{B \in \mathcal{F}_X} \mu(B) = \begin{cases} \top, & X = A, \\ \bot, & \text{otherwise}, \end{cases}$$

since $\mathcal{F}_X$ is either $\{1_\emptyset\}$ or $\{1_\emptyset, A\} = \mathcal{F}$, $\mu(1_\emptyset) = \bot$ and $\mu(A) = \top$.

In the following part we will define an isomorphism between fuzzy measure spaces. These definitions are vital for the investigation of *isomorphism invariant* fuzzy quantifiers (ISOM).

Definition 4. *Let* $(A, \mathcal{F})$ *and* $(B, \mathcal{G})$ *be fuzzy measurable spaces. We say that a mapping* $g : \mathcal{F} \to \mathcal{G}$ *is an* isomorphism between $(A, \mathcal{F})$ and $(B, \mathcal{G})$, *if*

(i) g *is a bijective mapping with* $g(1_\emptyset) = 1_\emptyset$,
(ii) $g(X \cup Y) = g(X) \cup g(Y)$ *and* $g(A \setminus X) = B \setminus g(X)$ *hold for any* $X, Y \in \mathcal{F}$,
(iii) there exists a bijective mapping $f : \mathrm{Dom}(A) \to \mathrm{Dom}(B)$ *with* $X(m) = g(X)(f(m))$ *for any* $X \in \mathcal{F}$ *and* $m \in \mathrm{Dom}(A)$.

Theorem 3. *Let* $(A, \mathcal{F})$, $(B, \mathcal{G})$ *be fuzzy measurable spaces and* $g : \mathcal{F} \to \mathcal{G}$ *be a surjective mapping. Then* g *is an isomorphism between* $(A, \mathcal{F})$ *and* $(B, \mathcal{G})$ *if and only if there exists a bijective mapping* $f : \mathrm{Dom}(A) \to \mathrm{Dom}(B)$ *such that* $g = f^{\to}$.

Definition 5. *Let* $(A, \mathcal{F})$ *and* $(B, \mathcal{G})$ *be fuzzy measurable spaces. We say that a mapping* $g : \mathcal{F} \to \mathcal{G}$ *is an* isomorphism between $(A, \mathcal{F}, \mu)$ and $(B, \mathcal{G}, \mu')$, *if*

(i) g *is an isomorphism between* $(A, \mathcal{F})$ *and* $(B, \mathcal{G})$,
(ii) $\mu(X) = \mu'(g(X))$ *for any* $X \in \mathcal{F}$.

The following lemma formulated for a fuzzy measure space is a straightforward consequence of the definition of isomorphism of fuzzy measure spaces.

Lemma 1. *Let* $(A_1, \mathcal{F}_1, \mu_1)$ *and* $(A_2, \mathcal{F}_2, \mu_2)$ *be isomorphic fuzzy measure spaces. If we put* $B_i = A_i \restriction \mathrm{Supp}(A_i)$, $\mathcal{G}_i = \{X_i \restriction \mathrm{Supp}(A_i) \mid X_i \in \mathcal{F}_i\}$ *and* $\mu'_i(X \restriction \mathrm{Supp}(A)) = \mu_i(X)$ *for* $i = 1, 2$, *then the triplets* $(B_1, \mathcal{G}_1, \mu'_1)$ *and* $(B_2, \mathcal{G}_2, \mu'_2)$ *are isomorphic fuzzy measure spaces.*

If g is an isomorphism between fuzzy measure spaces $\mathbf{A} = (A, \mathcal{F}, \mu)$ and $\mathbf{B} = (B, \mathcal{G}, \mu')$, then we will write $g(A, \mathcal{F}, \mu) = (B, \mathcal{G}, \mu')$ or shortly $g(\mathbf{A}) = \mathbf{B}$.

We say that a system $\mathcal{A}$ of fuzzy measure spaces from $\mathbf{Fms}(M)$ is a *closed under isomorphisms system of fuzzy measure spaces in* $\mathbf{Fms}(M)$, if it holds that if $\mathbf{A} \in \mathcal{A}$ and $\mathbf{B} \in \mathbf{Fms}(M)$ are isomorphic, then $\mathbf{B} \in \mathcal{A}$. In the following text, for simplicity, we will omit the term "under isomorphisms" in "closed under isomorphisms" and say only "closed system of fuzzy measure spaces in $\mathbf{Fms}(M)$". Note that there are closed systems of fuzzy measure spaces containing non-isomorphic fuzzy measure spaces. If a system $\mathcal{A}$ of mutually isomorphic fuzzy measure spaces in $\mathbf{Fms}(M)$ is closed, then we say that $\mathcal{A}$ is *closed system of mutually isomorphic fuzzy measure spaces in* $\mathbf{Fms}(M)$. Obviously, each closed system is a union of closed systems of mutually isomorphic fuzzy measure spaces.

Lemma 2. *A system* $\mathcal{A}$ *of fuzzy measure spaces in* $\mathbf{Fms}(M)$ *is closed if and only if* $f^{\rightarrow}(\mathbf{A}) \in \mathcal{A}$ *for any* $\mathbf{A} \in \mathcal{A}$ *and any permutation*[3] f *on* $\mathrm{Dom}(A)$.

An important class of fuzzy measure spaces which will be used in our theory of fuzzy quantifiers, namely when we will study so-called *permutation invariance* of fuzzy quantifiers, covers the fuzzy measure spaces being invariant under automorphisms in the following sense.

Definition 6. *We say that* $(A, \mathcal{F}, \mu)$ *is a* cardinal *fuzzy measure space, if*

(i) if $X \in \mathcal{F}$, *then* $f^{\rightarrow}(X) \in \mathcal{F}$,
(ii) $\mu(X) = \mu(f^{\rightarrow}(X))$

hold for any $X \in \mathcal{F}$ *and for any permutation* f *on* $\mathrm{Dom}(A)$.

Remark 3. The denotation "cardinal" in the previous definition means that the measures are invariant under the same cardinality of fuzzy sets, where we can say that two fuzzy sets X, Y have the same cardinality, if there exists a permutation f on $\mathrm{Dom}(A)$ such that $f^{\rightarrow}(X) = Y$.

Lemma 3. *A set* $\{\mathbf{A}\}$ *forms a closed system of fuzzy measure spaces in* $\mathbf{Fms}(A)$ *if and only if* $\mathbf{A}$ *is a cardinal fuzzy measure space.*

Lemma 4. *If* $\mathbf{A}$ *is a cardinal fuzzy measure space, then* A *is a constant fuzzy set.*

[3] A *permutation* on a set Z is a bijective mapping from Z to itself.

Example 4. Trivial examples of closed systems of mutually isomorphic fuzzy measure spaces in $\mathbf{Fms}(M)$ are a trivial fuzzy measure space $(A, \{\emptyset, A\}, \mu)$, where $A \in \mathcal{F}(M)$, and a fuzzy measure space $(A, \mathcal{F}(A), \mu)$, where $\mu(A) = \mu(B)$ whenever there exists a permutation f on $\mathrm{Dom}(A)$ with $f^{\rightarrow}(A) = B$. Further examples can be obtained from the second fuzzy measure space in such a way that $\mathcal{F}(A)$ is replaced by $\mathcal{F}$ of all fuzzy sets from $\mathcal{F}(A)$ different from the constant fuzzy sets taking a value $c \in]\bot, a[$ (if $c \neq \neg\neg c$) or values $c, \neg c \in]\bot, a[$ (if $c = \neg\neg c$), where $a = A(m)$ for any $m \in \mathrm{Dom}(A)$.[4]

Lemma 5. *Let $\mathcal{A} = \{\mathbf{A}_i \mid i \in I\}$ be a closed system of fuzzy measure spaces in $\mathbf{Fms}(A)$. Put $\mathcal{F}$ the least algebra containing $\bigcup_{i\in I} \mathcal{F}_i$ and*

$$\mu(X) = \bigvee_{i\in I} \bigvee_{Y\in\mathcal{F}_i(X)} \mu_i(Y). \tag{9}$$

Then $\mathbf{A} = (A, \mathcal{F}, \mu)$ is a cardinal fuzzy measure space.

4 $\odot$-Fuzzy Integral

In this part, we will introduce a type of fuzzy integral that can be defined on an arbitrary fuzzy measure space $(A, \mathcal{F}, \mu)$. The integrated functions are fuzzy sets on $\mathrm{Dom}(A)$. For a convention with the classical measure theory, we will prefer, in this part, the denotation f, g for the integrated functions instead of X, Y, nevertheless, we will deal with them as with fuzzy sets. For example, $f \cap g$ denotes the intersection of fuzzy sets. This integral is defined over a general operation $\odot$ which substitutes one of the operations $\wedge$ and $\otimes$, i.e., $\odot \in \{\wedge, \otimes\}$. This integral will be used as a basis for the definition of fuzzy quantifiers of type $\langle 1, 1\rangle$ determined by fuzzy measures.

Definition 7. *Let $(A, \mathcal{F}, \mu)$ be a fuzzy measure space with $M = \mathrm{Dom}(A)$, $f : M \to L$ and X be an $\mathcal{F}$-measurable fuzzy set. The* $\odot$-fuzzy integral of f on X *is given by*

$$\int_X^{\odot} f \, d\mu = \bigvee_{Y\in\mathcal{F}_X\setminus\{1_\emptyset\}} \bigwedge_{m\in\mathrm{Supp}(Y)} (f(m) \odot \mu(Y)). \tag{10}$$

If $X = A$, then we write $\int^{\odot} f \, d\mu$.

Remark 4. For a discussion on the motivation and explanation of the formula of our fuzzy integral see [2], Sect. 3.2, where similar type of fuzzy integral (albeit on a different fuzzy measure space defined on algebra of fuzzy subsets of a crisp universal set M) is explained and picture showing how to compute its value for some individual f is provided.

[4] The denotation $]\bot, a[$ means the interval of all values b from L for which $\bot < b < a$.

Theorem 4. *Let $(A, \mathcal{F}, \mu)$ be a fuzzy measure space with $M = \mathrm{Dom}(A)$. If $B \in \mathcal{F}(M)$ such that $A \subseteq B$, $\mathcal{G} \in \mathbf{Alg}(B)$ and A is crisp, then $\mu' : \mathcal{G} \to L$ defined by*

$$\mu'(X) = \int^{\odot} X \; d\mu \tag{11}$$

is a fuzzy measure on $(B, \mathcal{G})$.

Let us show some of properties of the proposed $\odot$-fuzzy integral.

Theorem 5. *Let $(A, \mathcal{F}, \mu)$ be a fuzzy measure space. Then*

(i) $\int_X^{\odot}(f \cap g) \; d\mu \leq \int_X^{\odot} f \; d\mu \wedge \int_X^{\odot} g \; d\mu$,
(ii) $\int_X^{\odot}(f \cup g) \; d\mu \geq \int_X^{\odot} f \; d\mu \vee \int_X^{\odot} g \; d\mu$,
(iii) $\int_X^{\odot}(c \otimes f) \; d\mu \geq c \otimes \int_X^{\odot} f \; d\mu$,
(iv) $\int_X^{\odot}(c \cap f) \; d\mu = c \wedge \int_X^{\odot} f \; d\mu$,
(v) $\int_X^{\odot}(c \to f) \; d\mu \leq c \to \int_X^{\odot} f \; d\mu$,

hold for any $X \in \mathcal{F}$, $f, g : M \to L$ and $c \in L$.

Theorem 6. *Let $(A, \mathcal{F}, \mu)$ be a fuzzy measure space, $c \in L$ and X be a set. Then*

(i) $\int^{\otimes}(c \odot 1_X) \; d\mu = c \otimes \mu^*(1_X)$, *if* $X \subseteq \mathrm{Dom}(A)$,
(ii) $\int^{\wedge}(c \odot 1_X) \; d\mu \geq c \wedge \mu^*(1_X)$, *if* $X \subseteq \mathrm{Dom}(A)$,
(iii) $\int^{\otimes}(c \odot 1_X) \; d\mu = c \otimes \mu(1_X)$, *if* $1_X \in \mathcal{F}$,
(iv) $\int^{\wedge}(c \odot 1_X) \; d\mu \geq c \wedge \mu(1_X)$, *if* $1_X \in \mathcal{F}$,
(v) $\int^{\otimes} 1_X \; d\mu = \mu(1_X)$, *if* $1_X \in \mathcal{F}$,
(vi) $\int^{\wedge} 1_X \; d\mu \geq \mu(1_X)$, *if* $1_X \in \mathcal{F}$,
(vii) $\int^{\odot} c \; d\mu = c$.

If $\mathbf{L}$ is an MV-algebra, then all inequalities may be replaced by equalities.

The next theorem is important for stating a condition when fuzzy quantifiers defined using our integrals are isomorphism invariant (ISOM).

Theorem 7. *If $g^{\rightarrow}$ is an isomorphism between $(A, \mathcal{F}, \mu)$ and $(B, \mathcal{F}', \mu')$, $f : \mathrm{Dom}(A) \to L$ be a mapping and X be an $\mathcal{F}$-measurable fuzzy set, then*

$$\int_X^{\odot} f \; d\mu = \int_{g^{\rightarrow}(X)}^{\odot} f \circ g^{-1} \; d\mu'.$$

5 Conclusion and Future Work

In this paper we introduced a definition of fuzzy measures on algebras of fuzzy sets generated by some fuzzy set A. Further we defined fuzzy integral based on these fuzzy measures. We introduced several related notions (inner fuzzy measure, cardinal fuzzy measure space, etc.), investigated isomorphisms of fuzzy

measurable and measure spaces and show basic properties of fuzzy integrals based on these fuzzy measures.

As we mentioned in Sect. 1, our main motivation was the investigation of fuzzy quantifiers of type $\langle 1, 1\rangle$ and higher. We are currently working on semantical properties (permutation invariance, isomorphism invariance, extension, conservativity, monotonicity, etc. [4,1]) of these fuzzy quantifiers. These properties are connected to various notions introduced here, e.g. isomorphism invariance is based on isomorphisms of fuzzy measure spaces, permutation invariance on the notion of cardinal fuzzy measure space, etc. Further, we plan to use these quantifiers as models of so-called *intermediate fuzzy quantifiers* of Novák [13]. We also plan to apply these quantifiers in data mining, time series analysis and possibly also in other applications. For example, in time series analysis and modeling, we can use quantifiers like "many" for filtering off of outliers, i.e. values that are numerically distant from (the course of) the data.

References

1. Holčapek, M.: Monadic **L**-fuzzy quantifiers of the type $\langle 1^n, 1\rangle$. Fuzzy Sets and Systems 159(14), 1811–1835 (2008)
2. Dvořák, A., Holčapek, M.: **L**-fuzzy quantifiers of type $\langle 1\rangle$ determined by fuzzy measures. Fuzzy Sets and Systems 160(23), 3425–3452 (2009)
3. Barwise, J., Cooper, R.: Generalized quantifiers and natural language. Linguistics and Philosophy 4(2), 159–219 (1981)
4. Peters, S., Westerståhl, D.: Quantifiers in Language and Logic. Oxford University Press, New York (2006)
5. Glöckner, I.: Fuzzy Quantifiers: A Computational Theory. Springer, Berlin (2006)
6. Wang, Z., Klir, G.: Fuzzy Measure Theory. Plenum Press, New York (1992)
7. Grabisch, M., Murofushi, T., Sugeno, M. (eds.): Fuzzy Measures and Integrals. Theory and Applications. Studies in Fuzziness and Soft Computing. Physica Verlag, Heidelberg (2000)
8. Klement, E., Mesiar, R., Pap, E.: A universal integral based on measures of level sets. IEEE Transactions on Fuzzy Systems (to appear)
9. Ying, M.: Linguistic quantifiers modeled by Sugeno integrals. Artificial Intelligence 170(6), 581–606 (2006)
10. Bělohlávek, R.: Fuzzy Relational Systems: Foundations and Principles. Kluwer Academic Publisher, New York (2002)
11. Novák, V., Perfilieva, I., Močkoř, J.: Mathematical Principles of Fuzzy Logic. Kluwer Academic Publisher, Boston (1999)
12. Hájek, P.: Metamathematics of Fuzzy Logic. Kluwer Academic Publishers, Dordrecht (1998)
13. Novák, V.: A formal theory of intermediate quantifiers. Fuzzy Sets and Systems 159(10), 1229–1246 (2008)

Absolute Continuity of Monotone Measure and Convergence in Measure

Jun Li[1,⋆], Radko Mesiar[2,3], and Qiang Zhang[4]

[1] School of Science, Communication University of China, Beijing 100024, China
lijun@cuc.edu.cn
[2] Department of Mathematics and Descriptive Geometry,
Faculty of Civil Engineering, Slovak University of Technology,
SK-81368 Bratislava, Slovakia
mesiar@math.sk
[3] Institute for Research and Applications of Fuzzy Modeling,
University of Ostrava, 30. dubna 22, 701 03 Ostrava 1, Czech Republic
[4] School of Management and Economics, Beijing Institute of Technology,
Beijing 100081, People's Republic of China
qiangzhang@bit.edu.cn

Abstract. In this note, the concept of strong absolute continuity of set function is introduced in two different ways. By using the two types of strong absolute continuity of monotone measure, the inheriting of convergence a.e. and convergence in measure for sequence of measurable function under the common addition operation is shown, respectively.

Keywords: Monotone measure, strong absolute continuity, convergence in measure, pseudo-convergence in measure.

1 Introduction

In non-additive measure theory, there are several different kinds of convergence of sequence of measurable functions, such as almost everywhere convergence, pseudo-almost everywhere convergence, convergence in measure, and convergence pseudo-in measure, and theorems that describe implication relationship between such convergence concepts are fundamental and important. Generally, theorems in the classical measure theory no longer hold in non-additive measure theory, so that to find necessary and/or sufficient conditions for such theorems to hold is very important for the construction of non-additive measure theory. In this direction there are a lot of results [4,5,8,9,10,12].

In this paper, we shall introduce the concept of strong absolute continuity of set functions in two different ways. They are called strong absolute continuity of type I and strong absolute continuity of type VI, respectively. By using the two types of strong absolute continuity of monotone measure, the inheriting of convergence a.e and convergence in measure for sequence of measurable function under the common addition operation are investigated, respectively.

⋆ Corresponding author.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 500–504, 2010.

2 Preliminaries

Let X be a non-empty set, $\mathcal{F}$ a σ-algebra of subsets of X, and $(X, \mathcal{F})$ denotes the measurable space.

Definition 1. ([8,12]) *Set function $\mu : \mathcal{F} \to [0, +\infty]$ is called a monotone measure on $(X, \mathcal{F})$ iff it satisfies the following requirements:*

(1) $\mu(\emptyset) = 0$; *(vanishing at $\emptyset$)*
(2) $A \subset B$ *and* $A, B \in \mathcal{F} \ \Rightarrow \ \mu(A) \leq \mu(B)$. *(monotonicity)*

When μ is a monotone measure, the triple $(X, \mathcal{F}, \mu)$ is called a monotone measure space ([8,12]).

In some literature, a set function μ satisfying the conditions (1) and (2) of Definition 1 is called a fuzzy measure or a non-additive measure.

In this paper, all the considered sets are supposed to belong to $\mathcal{F}$ and μ is supposed to be a finite monotone measure, i.e., $\mu(X) < \infty$. All concepts and symbols not defined may be found in [8,12].

We define the conjugate $\overline{\mu}$ of μ by

$$\overline{\mu}(A) = \mu(X) - \mu(X \setminus A), \quad A \in \mathcal{F}.$$

Obviously, the conjugate $\overline{\mu}$ of a monotone measure μ is also a monotone measure, and it holds that $\overline{\overline{\mu}} = \mu$.

Let $\mathbf{F}$ be the class of all finite real-valued measurable functions on $(X, \mathcal{F}, \mu)$, and let $f \in \mathbf{F}, f_n \in \mathbf{F}$ $(n = 1, 2, \ldots)$ and $\{f_n\}$ denote a sequence of measurable functions. We say that $\{f_n\}$ *converges almost everywhere to f on X* , and denote it by $f_n \xrightarrow{a.e.} f$ $[\mu]$, if there is subset $E \subset X$ such that $\mu(E) = 0$ and $f_n \to f$ on $X \setminus E$; $\{f_n\}$ *converges pseudo-almost everywhere to f on X*, and denote it by $f_n \xrightarrow{p.a.e.} f$ $[\mu]$, if there is a subset $F \subset X$ such that $\mu(X \setminus F) = \mu(X)$ and $f_n \to f$ on $X \setminus F$; $\{f_n\}$ *converges in measure to f on X*, and denote it by $f_n \xrightarrow{\mu} f$, if for any given $\sigma > 0$, $\lim\limits_{n \to +\infty} \mu(\{|f_n - f| \geq \sigma\}) = 0$; $\{f_n\}$ *converges pseudo-in measure to f on X*, and denote it by $f_n \xrightarrow{p.\mu} f$, if for any given $\sigma > 0$, $\lim\limits_{n \to +\infty} \mu(\{|f_n - f| < \sigma\}) = \mu(X)$.

From the definitions above, we know that the convergence in measure (or almost everywhere) on X and the convergence pseudo-in measure (or pseudo-almost everywhere) on X are dual to each other ([6,7,10]). We state them in the following.

Proposition 1. *Let μ be a finite monotone measure. Then*

(1) $f_n \xrightarrow{\mu} f$ *iff* $f_n \xrightarrow{p.\overline{\mu}} f$;
(2) $f_n \xrightarrow{a.e.} f$ $[\mu]$ *iff* $f_n \xrightarrow{p.a.e.} f$ $[\overline{\mu}]$.

3 Strong Absolute Continuity and Null-Additivity of Set Function

Absolute continuity of set function plays an important role in measure theory. In non-additive measure theory, the researches on this matter were made and a lot

of results were obtained [13,2,3]. Now we shall introduce the concept of strong absolute continuity of set functions in two different ways. They are called strong absolute continuity of type I and strong absolute continuity of type VI, respectively. We discuss the properties of strong absolute continuity of a monotone measure μ with respect to its conjugate $\overline{\mu}$.

Definition 2. ([13]) *Let μ and ν be two monotone measures. We say that*

(1) μ is absolutely continuous of Type I with respect to ν, denoted by $\mu \ll_I \nu$, iff $\mu(A) = 0$ whenever $\nu(A) = 0$;

(2) μ is absolutely continuous of Type VI with respect to ν, denoted by $\mu \ll_{VI} \nu$, iff $\mu(A_n) \to 0$ $(n \to \infty)$ whenever $\nu(A_n) \to 0$ $(n \to \infty)$.

Obviously, $\mu \ll_{VI} \nu$ imply $\mu \ll_I \nu$. The inverse statement may not be true.

Now we introduce the concept of strong absolute continuity of monotone measures.

Definition 3. *Let μ and ν be two monotone measures. We say that*

(1) μ is strongly absolutely continuous of Type I with respect to ν, denoted by $\mu \ll_I^{(s)} \nu$, iff $\mu(A \cup B) = 0$ whenever $\nu(A) = \nu(B) = 0$;

(2) μ is strongly absolute continuous of Type VI with respect to ν, denoted by $\mu \ll_{VI}^{(s)} \nu$, iff $\mu(A_n \cup B_n) \to 0$ $(n \to \infty)$ whenever $\nu(A_n) \vee \nu(B_n) \to 0$ $(n \to \infty)$.

Proposition 2. *For monotone measures μ and ν, if $\mu \ll_{VI}^{(s)} \nu$, then $\mu \ll_I^{(s)} \nu$.*

The several kinds of null-additivity of monotone measures play an important role in discussing the convergence of measurable functions on monotone measure spaces.

A set function $\mu : \mathcal{F} \to [0, +\infty)$ is said to be (i) *weakly null-additive* [12] , if $\mu(E) = \mu(F) = 0$ imply $\mu(E \cup F) = 0$; (ii) *converse-null-additive at X* [10], if $\mu(X) = \mu(X - N)$ implies $\mu(N) = 0$; (iii) *null-subtractive at X* [10], if $\mu(N) = 0$ implies $\mu(X - N) = \mu(X)$;

Definition 4. *A set function $\mu : \mathcal{F} \to [0, +\infty)$ is said to be weakly pseudo-null-additive, if $\mu(X - E) = \mu(X - F) = \mu(X)$ imply $\mu(X - E \cup F) = \mu(X)$.*

Weak null-additivity and weak pseudo-null-additivity of monotone measure are dual to each other, i.e., μ is weakly pseudo-null-additive iff $\overline{\mu}$ is weakly null-additive, i.e., $\overline{\mu}(E) = \overline{\mu}(F) = 0$ imply $\overline{\mu}(E \cup F) = 0$.

Definition 5. ([1]) *A set function $\mu : \mathcal{F} \to [0, +\infty)$ is said to have pseudometric generating property (for short p.g.p), if for each $\epsilon > 0$ there is $\delta > 0$ such that for any $E, F \in \mathcal{F}$, $\mu(E) \vee \mu(F) < \delta$ implies $\mu(E \cup F) < \epsilon$, or equivalently, if for any $\{E_n\} \subset \mathcal{F}$ and $\{F_n\} \subset \mathcal{F}$, $\mu(E_n) \vee \mu(F_n) \to 0$ $(n \to \infty)$ imply $\mu(E_n \cup F_n) \to 0$ $(n \to \infty)$.*

Note: The concept of pseudometric generated property goes back to Dobrakov and Drewnowski in seventies, and this was related to Frechet-Nikodym topology [1,8].

Obviously, if μ has pseudometric generated property, then it is weakly null-additive. The inverse statement may not be true ([3,10]).

We can easily obtain the following propositions.

Proposition 3. (i) *μ is weakly null-additive iff $\overline{\mu}$ is weakly pseudo-null-additive.*
(ii) *μ is null-subtractive at X iff $\overline{\mu}$ is converse-null-additive at X.*

Proposition 4. *Let μ be monotone measure. Then*
(i) *μ is null-subtractive at X iff $\overline{\mu} \ll_I \mu$.*
(ii) *μ is converse-null-additive at X iff $\mu \ll_I \overline{\mu}$.*
(iii) *μ is weakly null-additive iff $\mu \ll_I^{(s)} \mu$.*
(v) *μ has p.g.p iff $\mu \ll_{VI}^{(s)} \mu$.*

Proposition 5. (i) *Let μ be null-subtractive at X. If μ is weakly null-additive, then $\overline{\mu} \ll_I^{(s)} \mu$.* (ii) *Let μ be converse-null-additive at X. If μ is weakly pseudo-null-additive, then $\mu \ll_I^{(s)} \overline{\mu}$.*

4 Strong Absolute Continuity and Convergence of Sequence of Measurable Functions

In this section , by using the two types of strong absolute continuity of a monotone measure μ with respect to its conjugate $\overline{\mu}$, we discuss the inheriting of convergence of sequence of measurable functions under the addition operation on monotone measure space, respectively. In the following we show our main results and omit their proofs.

Theorem 1. *Let μ be monotone measure. Then,*

(1) *$\mu \ll_I^{(s)} \mu$, i.e., μ is weakly null-additive iff for any $f, g, f_n, g_n \in \mathbf{F}$,*

$$f_n \xrightarrow{a.e} f\ [\mu] \text{ and } g_n \xrightarrow{a.e} g\ [\mu] \implies f_n + g_n \xrightarrow{a.e} f + g\ [\mu].$$

(2) *$\overline{\mu} \ll_I^{(s)} \overline{\mu}$, i.e., μ is weakly pseudo-null-additive iff for any $f, g, f_n, g_n \in \mathbf{F}$,*

$$f_n \xrightarrow{p.a.e} f\ [\mu] \text{ and } g_n \xrightarrow{p.a.e} g\ [\mu] \implies f_n + g_n \xrightarrow{p.a.e} f + g\ [\mu].$$

(3) *$\overline{\mu} \ll_I^{(s)} \mu$, i.e., μ is null-subtractive at X iff for any $f, g, f_n, g_n \in \mathbf{F}$,*

$$f_n \xrightarrow{a.e} f\ [\mu] \text{ and } g_n \xrightarrow{a.e} g\ [\mu] \implies f_n + g_n \xrightarrow{p.a.e} f + g\ [\mu].$$

(4) *$\mu \ll_I^{(s)} \overline{\mu}$, i.e., μ is converse-null-additive at X iff for any $f, g, f_n, g_n \in \mathbf{F}$,*

$$f_n \xrightarrow{p.a.e} f\ [\mu] \text{ and } g_n \xrightarrow{p.a.e} g\ [\mu] \implies f_n + g_n \xrightarrow{a.e} f + g\ [\mu].$$

Theorem 2. *Let μ be monotone measure. Then,*

(1) *μ has p.g.p, i.e., $\mu \ll_{VI}^{(s)} \mu$ iff for any $f, g, f_n, g_n \in \mathbf{F}$,*

$$f_n \xrightarrow{\mu} f \text{ and } g_n \xrightarrow{\mu} g \implies f_n + g_n \xrightarrow{\mu} f + g.$$

(2) $\overline{\mu}$ *has p.g.p, i.e.,* $\overline{\mu} \ll_{VI}^{(s)} \overline{\mu}$ *iff for any* $f, g, f_n, g_n \in \mathbf{F}$,

$$f_n \xrightarrow{p.\mu} f \text{ and } g_n \xrightarrow{p.\mu} g \implies f_n + g_n \xrightarrow{p.\mu} f + g.$$

(3) $\overline{\mu} \ll_{VI}^{(s)} \mu$ *iff for any* $f, g, f_n, g_n \in \mathbf{F}$,

$$f_n \xrightarrow{\mu} f \text{ and } g_n \xrightarrow{\mu} g \implies f_n + g_n \xrightarrow{p.\mu} f + g.$$

(4) $\mu \ll_{VI}^{(s)} \overline{\mu}$ *iff for any* $f, g, f_n, g_n \in \mathbf{F}$,

$$f_n \xrightarrow{p.\mu} f \text{ and } g_n \xrightarrow{p.\mu} g \implies f_n + g_n \xrightarrow{\mu} f + g.$$

Acknowledgements. This work was partially supported by the NSFC Grant No.70771010, and by Grants APVV-0012-07, VEGA 1/4024/07 and VZ MSM 619 8898 701. The support of bilateral project SK-CN-0016-07 is also kindly acknowledged.

References

1. Dobrakov, I., Farkova, J.: On submeasures II. Math. Slovaca 30, 65–81 (1980)
2. Jiang, Q., Suzuki, H., Wang, Z., Klir, G.J.: Exhaustivity and absolute continuity of fuzzy measures. Fuzzy Sets and Systems 96(2), 231–238 (1998)
3. Jiang, Q., Wang, S., Ziou, D., Wang, Z., Klir, G.J.: Pseudometric generated preporty and autocontinuity of fuzzy measure. Fuzzy Sets and Systems 112(2), 207–216 (2000)
4. Li, J., Yasuda, M., Jiang, Q., Suzuki, H., Wang, Z., Klir, G.J.: Convergence of sequence of measurable functions on fuzzy measure space. Fuzzy Sets and Systems 87(3), 385–387 (1997)
5. Li, J.: Order continuous of monotone set function and convergence of measurable functions sequence. Applied Mathematics and Computation 135(2-3), 211–218 (2003)
6. Li, J., Yasuda, M.: On Egoroff's theorem on finite monotone non-additive measure space. Fuzzy Sets and Systems 153(1), 71–78 (2005)
7. Murofushi, T.: Duality and ordinality in fuzzy measure theory. Fuzzy Sets and Systems 138(3), 523–535 (2003)
8. Pap, E.: Null-additive Set Functions. Kluwer Academic Press, Dordrecht (1995)
9. Sun, Q.: Property(s) of fuzzy measure and Riesz's theorem. Fuzzy Sets and Systems 62(1), 117–119 (1994)
10. Takahashi, M., Murofushi, T.: Relationship between convergence concepts ain fuzzy measure theory. In: 11th IFSA World Congress, Beijing, China, vol. I, pp. 467–473 (2005)
11. Uchino, K., Murofushi, T.: Relations between mathematical properties of fuzzy measures. In: 10th IFSA World Congress, Istanbul, Turkey, pp. 27–30 (2003)
12. Wang, Z., Klir, G.J.: Generalized Measure Theory. Springer, Heidelberg (2009)
13. Wang, Z., Klir, G.J., Wang, W.: Fuzzy measures defined by fuzzy integral and their absolute continuity. J. Math. Anal. Appl. 203(1), 150–165 (1996)

An Axiomatic Approach to Fuzzy Measures Like Set Cardinality for Finite Fuzzy Sets*

Michal Holčapek

University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, Ostrava, Czech Republic
michal.holcapek@osu.cz
http://irafm.osu.cz

Abstract. The aim of the paper is to propose an axiomatic system for fuzzy measures (denoted by fuzzy c-measures) like cardinality of sets for finite fuzzy sets that are defined in a universe of sets. The proposed axiomatic system is a generalization of the Casasnovas-Torrens axiomatic system introduced in [1]. We show a representation of fuzzy c-measures by a pair of special lattice homomorphisms and investigate several selected properties of fuzzy c-measure.

1 Introduction

In classical set theory, the cardinality of a set is a measure of the "number of elements of the set". A formal definition without knowing anything about numbers is based on the notion of a one-to-one correspondence between sets. In particular, we say that two sets A and B are *equipollent* (or also are *equipotent, equinumerous* etc.) and write $|A| = |B|$, if there exists a one-to-one mapping of A onto B. The notion of cardinality of sets introduced by this way has only a functional role. In order to express the cardinality of sets as a specified object itself, the equivalence of being equipollent (called *equipollency* or *equipotency, equinumerosity* etc.) on the class of all sets is introduced. The cardinality of a set A is then defined as the equivalence class of all equipollent sets with A or as the least ordinal number that is the element of this equivalence class for A.

The familiar approaches to the definition of cardinality of fuzzy sets describing the size of fuzzy sets are more or less analogous to the classical approaches for sets. Nevertheless, for fuzzy sets, the situation is much more complicated by the graduation of membership of elements of fuzzy sets. We can imagine two general approaches to the cardinality of fuzzy sets that are based on:

(a) (graded) equipollence of fuzzy sets and fuzzy classes as fuzzy cardinals,
(b) fuzzy measures generalizing the cardinality of sets.

This paper is devoted to (b), for the approach (a), we refer to [2,3,4,5,6,7]. More precisely, we consider a direction, when objects representing the cardinality of

* This paper has been supported by the Grant IAA108270901 of the GA AV ČR.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 505–514, 2010.

fuzzy sets are described by generalized ordinal numbers. Let us mention that the first definition of fuzzy cardinality of finite fuzzy sets, by means of functions from N to $[0,1]$, was done by L.A. Zadeh in [8]. A substantial improvement of the original definition was proposed again by L.A. Zadeh in [9] and gave rise to three types of cardinality, namely, FGCount, FLCount and FECount (their definitions are mentioned in Proposition 1, Sect. 3.1). For example, $\mathrm{FGCount}(A)(k)$ (see $\mathcal{C}_1(k)$) expresses a degree to which a fuzzy set A has "at least" k elements. A generalization of these measures using t-norms and t-conorms was proposed by M. Wygralak in [10] and a further development can be found in [7]. Note that both mentioned publications are an attempt to develop the cardinal theory for vaguely defined sets including a relation between equipollence and equality of cardinal numbers. An axiomatic approach to fuzzy cardinality of finite fuzzy sets covering some of the well-known definitions, defined by means of the generalized natural numbers, was proposed by J. Casasnovas and J. Torrens in [1].

In this contribution, we propose a generalization of Casasnovas-Torrens axiomatic approach to fuzzy measures generalized cardinality of sets. Although, the proposed system of fuzzy measures generalizes the cardinality of sets, some properties like the valuation property or the fundamental correspondence between the equipollence of sets and the equality of their cardinals are failed here in general. Therefore, we call these fuzzy measures as *fuzzy c-measures*, where the letter "c" is an abbreviation of "like cardinality". We consider fuzzy sets from a universe of sets (see [4]) that is derived from the Grothendieck universe (see e.g. [11]). This universe enables us to deal with fuzzy sets over different universes of discourse and seems to be a suitable framework for a development of the fuzzy set theory.

The paper is organized as follows: Sect. 2 is devoted to preliminaries, where we introduce a lattice structure called an rdr-lattice in which the membership degrees of fuzzy sets are interpreted and a universe of sets. Sect. 3 is devoted to the axiomatic approach to fuzzy c-measures, their representation by lattice homomorphisms and selected properties. The last section is a conclusion.

2 Preliminaries

2.1 Algebraic Structures of Membership Degrees of Fuzzy Sets

In this paper, the membership degrees of fuzzy sets are interpreted in a complete residuated lattice which is, moreover, extended by an adjoint pair of operations. We shall say that an algebra $\mathbf{L} = \langle L, \wedge, \vee, \otimes, \rightarrow, \oplus, \ominus, \bot, \top \rangle$ with six binary operations and two constants is a *residuated-dually residuated lattice* (shortly, an rdr-lattice), if

(i) $\langle L, \wedge, \vee, \bot, \top \rangle$ is a bounded lattice, where $\bot$ is the least element and $\top$ is the greatest element of L, respectively,
(ii) $\langle L, \otimes, \top \rangle$ and $\langle L, \oplus, \bot \rangle$ are commutative monoids,

(iii) the pairs $\langle \otimes, \rightarrow \rangle$ and $\langle \oplus, \ominus \rangle$ form adjoint pairs, i.e.

$$a \leq b \rightarrow c \quad \text{if and only if} \quad a \otimes b \leq c, \tag{1}$$
$$a \leq b \oplus c \quad \text{if and only if} \quad a \ominus b \leq c \tag{2}$$

hold for each $a, b, c \in L$ ($\leq$ denotes the corresponding lattice ordering).

The operations $\otimes$, $\rightarrow$, $\oplus$ and $\ominus$ are called *multiplication*, *residuum*, *addition* and *difference*, respectively.[1] We shall say that an rdr-lattice is *complete* (*linearly ordered*), if $\langle L, \wedge, \vee, \bot, \top \rangle$ is a complete (linearly ordered) lattice, respectively. Further, an rdr-lattice is *divisible*, if $a \otimes (a \rightarrow b) = a \wedge b$ holds for all $a, b, c \in L$, and *dually divisible*, if $(a \ominus b) \oplus b = a \vee b$ holds for all $a, b, c \in L$. To integrate some of alternative constructions based on the operations of $\wedge$ and $\otimes$, we shall use the common symbol $\odot$ (or $\bigodot$) for them. Analogously, we shall use $\overline{\odot}$ (or $\overline{\bigodot}$) to denote one of the operations $\vee$ or $\oplus$.

Example 1. Let T and S be a left continuous t-norm and a right continuous t-conorm, respectively, and define $\rightarrow_T$ and $\ominus_S$ by

$$a \rightarrow_T b = \bigvee \{c \in [0,1] \mid T(a,c) \leq b\}, \; a \ominus_S b = \bigwedge \{c \in [0,1] \mid S(b,c) \geq a\}.$$

The algebra $\mathbf{L} = \langle [0,1], \min, \max, T, \rightarrow_T, S, \ominus_S, 0, 1 \rangle$ is a complete rdr-lattice.

2.2 Fuzzy Sets and Fuzzy Classes in a Universe of Sets

In this paper, we deal with fuzzy sets and fuzzy classes from the universe of sets over an rdr-lattice $\mathbf{L}$ defined as follows.

Definition 1. A universe of sets over $\mathbf{L}$ *is a non-empty class* $\mathcal{U}$ *of sets having the following properties:*

(U1) $x \in y$ *and* $y \in \mathcal{U}$, *then* $x \in \mathcal{U}$,
(U2) $x, y \in \mathcal{U}$, *then* $\{x, y\} \in \mathcal{U}$,
(U3) $x \in \mathcal{U}$, *then* $\mathcal{P}(x) \in \mathcal{U}$,
(U4) $x \in \mathcal{U}$ *and* $y_i \in \mathcal{U}$ *for any* $i \in x$, *then* $\bigcup_{i \in x} y_i \in \mathcal{U}$,
(U5) $x \in \mathcal{U}$ *and* $f : x \rightarrow L$, *then* $\mathrm{Ran}(f) \in \mathcal{U}$,

where L *is the support of the complete residuated lattice* $\mathbf{L}$.

Simple examples of such universes are the classes of all, finite, or countable sets. If we extend the ZFC by the axiom accepting the existence of strongly inaccessible cardinals, then there are Grothendieck universes which are universes of sets over $\mathbf{L}$. From (U1), each set in $\mathcal{U}$ is a subclass of $\mathcal{U}$.[2] Notice that the opposite implication is not true. This motivates us to say that a collection of elements is

[1] Notice that if the adjoint pair $\langle \oplus, \ominus \rangle$ (or $\langle \otimes, \rightarrow \rangle$) is forgotten in the rdr-lattice $\mathbf{L}$ we obtain the *residuated lattice* (or *dually residuated lattice*).
[2] Note that each set is a subset of a Grothendieck universe.

a set (a class) in $\mathcal{U}$, if this collection is an element (a subset) of $\mathcal{U}$. We shall say that a class in $\mathcal{U}$ is *proper*, if it is not a set in $\mathcal{U}$. From now, we suppose that a universe of sets is given and we restrict ourselves to sets and classes in this universe.

Let us put $0 = \emptyset$, $1 = 0 \cup \{0\}$, $2 = 1 \cup \{1\}$, etc. Obviously, the natural numbers $0, 1, 2, \ldots$ belong to $\mathcal{U}$ and, according to (U4), an arbitrary finite set of natural numbers is also contained in $\mathcal{U}$. Nevertheless, the set of all natural numbers (we shall use N to denote it) need not be an element of $\mathcal{U}$ (it is sufficient to consider the universe of all finite sets) and, generally, it is only a class in $\mathcal{U}$.

Definition 2. *Let $\mathcal{U}$ be a universe of sets over* **L**. *A mapping $A : x \to L$ is called a fuzzy set in $\mathcal{U}$, if x is a set in $\mathcal{U}$.*

A consequence of (U2), (U4) and (U5) is that each fuzzy set (perceived as a relation $A \subseteq x \times L$) belongs to $\mathcal{U}$. Hence, all fuzzy sets form a class in $\mathcal{U}$. It is easy to see that fuzzy sets of higher order, i.e., fuzzy sets over sets of fuzzy sets, may be established inside of $\mathcal{U}$.

Let A be a fuzzy set in $\mathcal{U}$. The set $\mathrm{Dom}(A)$ is called a *universe of discourse of A*. Notice that the empty mapping of the empty set to L is also a fuzzy set with the empty universe of discourse. We use $\mathcal{F}$ to denote the class of all fuzzy sets in $\mathcal{U}$. The set $\mathrm{Supp}(A) = \{x \in \mathrm{Dom}(A) \mid A(x) > \bot\}$ is called a *support of fuzzy set A* and $A_a = \{x \in \mathrm{Dom}(A) \mid A(x) \geq a\}$ is called an *a-cut of fuzzy set A*. We shall say that a fuzzy set A is *finite*, if its support is a finite set and *infinite*, if it is not finite. We shall use $\mathcal{FIN}$ to denote the class of all finite fuzzy sets. A fuzzy set A is *crisp*, if $A(x) \in \{\bot, \top\}$ for any $x \in \mathrm{Dom}(A)$, and is called a *singleton*, if $A(x) > \bot$ for some $x \in \mathrm{Dom}(A)$ and $A(y) = \bot$ for any $y \in \mathrm{Dom}(A)$ with $y \neq x$. We shall use $\{a/x\}$ to denote a singleton A with $A(x) = a > \bot$.

An essential predicate in our theory is a binary relation saying that two fuzzy sets are the same fuzzy set. In the classical set theory we say that two sets are the same sets, if they have the same elements. In our theory we define some analogical notion as follows.

Definition 3. *We shall say that fuzzy sets A and B are* the same fuzzy sets *(symbolically, $A = B$), if* $\mathrm{Supp}(A) = \mathrm{Supp}(B)$ *and $A(x) = B(x)$ for any $x \in \mathrm{Supp}(A)$.*

For example, $A = \{0.9/a, 0/b\}$ and $B = \{0.9, /a\}$ are the same fuzzy sets. Now, we can introduce some operations with fuzzy sets in $\mathcal{U}$. Here, we demonstrate only the operations of the union and intersection, nevertheless, one can simply define the other operations as the difference, product, disjoint union etc.

Definition 4. *Let $A, B \in \mathcal{F}$, $x = \mathrm{Dom}(A) \cup \mathrm{Dom}(B)$ and $A' = A$, $B' = B$ such that* $\mathrm{Dom}(A') = \mathrm{Dom}(B') = x$. *Then*

- the union *of A and B is a mapping $A \cup B : x \to L$ defined by*

$$(A \cup B)(a) = A'(a) \vee B'(a) \tag{3}$$

for any $a \in x$,

- the intersection *of A and B is a mapping $A \cap B : x \to L$ defined by*

$$(A \cap B)(a) = A'(a) \wedge B'(a) \tag{4}$$

for any $a \in x$,

One can check easily that our definition is correct, i.e. there is no dependence on the choice of fuzzy sets A' and B' in the previous definition.

Definition 5. *We shall say that a fuzzy set A* is less than or equal to *a fuzzy set B (it is denoted by $A \subseteq B$), if* $\mathrm{Supp}(A) \subseteq \mathrm{Supp}(B)$ *and $A(x) \leq B(x)$ for any $x \in \mathrm{Supp}(A)$.*

It is easy to see that the (class) relation $\subseteq$ defines a partial ordering of fuzzy sets on $\mathcal{F}$.

Although the notion of fuzzy set in $\mathcal{U}$ is the most important concept in our theory, it seems to be useful (analogously to the classical set theory) to introduce the concept of fuzzy class in $\mathcal{U}$. Practically, a fuzzy class describes a family of sets from $\mathcal{U}$ that possess some property, but we cannot precisely decide, if some object has or has not the given property in general. The formal definition is as follows.

Definition 6. *Let $\mathcal{U}$ be a universe of sets over* **L**. *A mapping $\mathcal{A} : \mathcal{X} \to L$ is called a* fuzzy class in $\mathcal{U}$, *if $\mathcal{X}$ is a class in $\mathcal{U}$.*

It is self evident that each fuzzy set is also a fuzzy class, but the inverse implication is not true. We shall say that a fuzzy class in $\mathcal{U}$ is *proper*, if its domain is a proper class in $\mathcal{U}$. For example, a mapping $A : N \to [0,1]$ defined by $A(n) = \frac{1}{n}$ is a proper fuzzy class in the universe of all finite sets.

To define fuzzy measures like cardinality for finite fuzzy sets, we need to introduce a concept of generalized cardinal over N. Further, we will omit "over N", for simplicity, and say only "generalized cardinal". A reasonable condition (see e.g. [7,10]) for this concept is a convexity of fuzzy sets. The convexity of fuzzy sets may be defined as follows. Recall that $\odot \in \{\wedge, \otimes\}$.

Definition 7. *Let $A : x \to L$ be a fuzzy set and $\leq$ be a linear ordering on x. We shall say that A is an* $\odot$-convex fuzzy set, *if $A(a) \odot A(c) \leq A(b)$ holds for any $a, b, c \in x$ with $a \leq b \leq c$.*

Definition 8. A generalized cardinal *A is an $\odot$-convex fuzzy set $A : N \to L$.*

One can see that generalized cardinals may be finite or infinite[3] and the finite generalized cardinals are fuzzy sets belonging to each universe of sets.[4] We shall use $\mathcal{N}$ to denote the set of all fuzzy natural numbers and suppose that each fuzzy

[3] This is a reason why we use "generalized cardinals over N" instead of e.g. "fuzzy natural numbers".

[4] Note that infinite generalized cardinals are used for some fuzzy c-measures which are used, analogously to FLCount, to express the fact that a fuzzy set has "at most" $0, 1, 2, \ldots$ elements (e.g. for the fuzzy set $\{1/a, 0.5/b\}$ we can obtain by a fuzzy c-measures the generalized cardinal $\{0/0, 0.5/1, 1/2, 1/3, \ldots\}$).

natural number is a fuzzy set in $\mathcal{U}$. Notice that $\mathcal{N}$ is a class or a set in some cases of $\mathcal{U}$. To be able to manipulate with fuzzy natural numbers, we establish the operation of addition on $\mathcal{N}$ by

$$(A+B)(i) = \bigvee_{\substack{k,l\in N\\ k+l=i}} (A(k)\odot B(l)). \tag{5}$$

Obviously, we use the well-known Zadeh's extension principle with a more general operation $\odot$. Define $E : \mathcal{N} \to L$ by $E(k) = \top$, if $k = 0$, and $E(k) = \bot$ otherwise. One can prove easily the following statement saying that our definition of addition is correct.

Theorem 1. *The triplet $(\mathcal{N}, +, E)$ is a commutative monoid, where the divisibility of* **L** *has to be supposed for $\odot = \wedge$.*

3 Fuzzy c-Measures of Finite Fuzzy Sets

3.1 Definition and Examples

In the following text, we suppose that **L** is a complete rdr-lattice which is, moreover, divisible, whenever $\odot$ is considered to be $\wedge$.

Definition 9. *A class mapping $\mathcal{C} : \mathcal{FIN} \to \mathcal{N}$ is a* fuzzy c-measure of finite fuzzy sets*, if, for arbitrary $A, B \in \mathcal{FIN}$, it satisfies the following axioms*

(C1) if $A \cap B = \emptyset$, then $\mathcal{C}(A \cup B) = \mathcal{C}(A) + \mathcal{C}(B)$,
(C2) if $i, j \in N$ and $i > |\mathrm{Supp}(A)|$, $j > |\mathrm{Supp}(B)|$, then $\mathcal{C}(A)(i) = \mathcal{C}(B)(j)$,
(C3) if A is a crisp set, then $\mathcal{C}(A)$ is a crisp set and $\mathcal{C}(A)(|A|) = \top$,
(C4) if $a \in L$, $x, y \in \mathcal{U}$ and $i \in N$, then $\mathcal{C}(\{a/x\})(i) = \mathcal{C}(\{a/y\})(i)$,
(C5) if $a, b \in L$ and $x \in \mathcal{U}$, then

$$\mathcal{C}(\{a \,\overline{\odot}\, b/x\})(0) = \mathcal{C}(\{a/x\})(0) \odot \mathcal{C}(\{b/x\})(0) \qquad \textit{(C5a)},$$
$$\mathcal{C}(\{a \odot b/x\})(1) = \mathcal{C}(\{a/x\})(1) \odot \mathcal{C}(\{b/x\})(1) \qquad \textit{(C5b)}.$$

The axioms C1-C5 are called the *additivity, variability, consistency, singleton independency, preservation of non-existence* and *existence*, respectively. For simplicity, we shall often speak only about "c-measure" instead of "fuzzy c-measure of finite fuzzy sets" in the following text.

The first three axioms are used from [1] and their motivation is as follows. The additivity of c-measure is the property of cardinality of sets and it is very natural. The idea of variability is that the c-measure of fuzzy sets is only influenced by the elements that belong to its support. A consequence of this axiom is correctness of our axiomatic system, it means, the same finite fuzzy sets[5] have the same c-measure. The axiom of consistency ensures the fact that the c-measures are extensions of cardinality measure. The singleton independency guarantees that

[5] See Definition 3.

an arbitrary c-measure is not under influence of a form of considered elements. Finally, a value $\mathcal{C}(\{A(x)/x\})(0)$ expresses a degree to which $\{A(x)/x\}$ may be considered as the empty set $\{0/x\}$, i.e. x does not belong to A. Analogously, a value $\mathcal{C}(\{A(x)/x\})(1)$ determines a degree to which $\{A(x)/x\}$ may be considered as a singleton $\{1/x\}$. Since $a \mathbin{\overline{\odot}} b \geq a \vee b$, then a degree characterizing $\{a \mathbin{\overline{\odot}} b/x\}$ as the empty set cannot be greater than the degrees obtained for $\{a/x\}$ and $\{b/x\}$. This relation is specified in the axiom (C5a). An analogous consideration can be done also for the axiom (C5b).

The following proposition shows the FGCount, FLCount and FECount by means of c-measures of finite fuzzy sets for $\odot = \wedge$ (cf. [1,6,7]). Define a complement of A by $\overline{A}(x) = A(x) \to \bot$ for any $x \in \mathrm{Supp}(A)$ and $\overline{A}(x) = \bot$ for any $x \in \mathrm{Dom}(A) \setminus \mathrm{Supp}(A)$.[6]

Proposition 1. *Let $\odot = \wedge$. Then mappings of $\mathcal{FIN}$ to $\mathcal{N}$ defined by*

$$\mathcal{C}_1(A)(i) = \bigvee\{a \mid a \in L \text{ and } |A_a| \geq i\}, \tag{6}$$

$$\mathcal{C}_2(A)(i) = \mathcal{C}_1(\overline{A})(\overline{i}), \tag{7}$$

$$\mathcal{C}_3(A)(i) = \mathcal{C}_1(A)(i) \wedge \mathcal{C}_2(A)(i) \tag{8}$$

for each $i \in N$ are c-measures of finite fuzzy sets, where $\overline{i} = \max(0, |\mathrm{Dom}(A)| - i)$.

Example 2. Let $\mathbf{L}$ be the Łukasiewicz algebra, i.e. $a \otimes b = \max(0, a + b - 1)$ and $A = \{0.5/a, 0.8/b, 0.1/c, 0.4/d, 0/e\}$. Then we may write

$$\mathcal{C}_1(A) = \{1/0, 0.8/1, 0.5/2, 0.4/3, 0.1/4, 0/5, 0/6 \ldots\},$$
$$\mathcal{C}_2(A) = \{0/0, 0.2/1, 0.5/2, 0.6/3, 0.9/4, 1/5, 1/6, \ldots\},$$
$$\mathcal{C}_3(A) = \{0/0, 0.2/1, 0.5/2, 0.4/3, 0.1/4, 0/0, 0/6 \ldots\}.$$

Note that $\mathcal{C}_1$ do not define a c-measure for $\odot = \otimes$ in general. Two examples of c-measures for $\otimes$ are given below.

Proposition 2. *Let $\mathcal{C}_1$ be the c-measure defined above and $\odot = \otimes$. Then mappings of $\mathcal{FIN}$ to $\mathcal{N}$ defined by*

$$\mathcal{C}_4(A)(i) = \mathcal{C}_1(A_\top)(i) \tag{9}$$

$$\mathcal{C}_5(A)(i) = \begin{cases} \top, & i = 0, \\ \mathcal{C}_5(A)(i-1) \otimes \mathcal{C}_1(A)(i), & \text{otherwise}, \end{cases} \tag{10}$$

for each $i \in N$ are c-measures of finite fuzzy sets.

Example 3. Let us consider the same presumptions as in Example 2. Then we may write

$$\mathcal{C}_4(A) = \{1/0, 0/1, 0/2, \ldots\},$$
$$\mathcal{C}_5(A) = \{1/0, 0.8/1, 0.3/2, 0/3, 0/4, 6 \ldots\},$$

where, for example, $0.3 = 0.8 \otimes 0.5 = \max(0, 0.8 + 0.5 - 1)$.

[6] One can verify that $A = B$ implies $\overline{A} = \overline{B}$ and thus $=$ is a congruence with respect to the complement in $\mathcal{U}$. Note that the same result cannot be obtained for a definition of $\overline{A}$ where $\overline{A}(x) = \top$ for any $x \in \mathrm{Dom}(A) \setminus \mathrm{Supp}(A)$.

A straightforward consequence of the additivity of c-measures is the following theorem.

Theorem 2. *Let $\mathcal{C} : \mathcal{FIN} \to \mathcal{N}$ be a c-measure and $A \in \mathcal{FIN}$ such that* $\mathrm{Supp}(A) = \{x_1, \ldots, x_m\}$. *Then*

$$\mathcal{C}(A)(i) = \bigvee_{\substack{i_1,\ldots,i_m \in \{0,1\} \\ i_1+\cdots+i_m=i}} \bigodot_{k=1}^{m} \mathcal{C}(\{A(x_k)/x_k\})(i_k) \tag{11}$$

for each $i \in N$. Moreover, $\mathcal{C}(A)(i) = \bot$ or $\mathcal{C}(A)(i) = \top$ hold for any $m < i$.

3.2 Representation of c-Measures

In [1], there is shown a representation of cardinalities of finite fuzzy sets using two monotonic mappings $f, g : [0,1] \to [0,1]$. In order to introduce an analogical representation for c-measures, we need to establish a generalization of monotonic mappings used in [1].

Let $\mathbf{L}_1, \mathbf{L}_2$ be complete rdr-lattices. We shall say that $h : L_1 \to L_2$ is an $\odot$-*homomorphism* of $\mathbf{L}_1$ to $\mathbf{L}_2$, if h is a homomorphism of the reduct $(L_1, \odot_1, \top_1)$ of $\mathbf{L}_1$ to the reduct $(L_2, \odot_2, \top_2)$ of $\mathbf{L}_2$, i.e. $h(a \odot_1 b) = h(a) \odot_2 h(b)$ and $h(\top_1) = \top_2$.[7] Further, we shall say that $h : L_1 \to L_2$ is an $\overline{\odot}_{\mathrm{d}}$-*homomorphism*, if h is a homomorphism from the reduct $(L_1, \overline{\odot}_1, \bot_1)$ of $\mathbf{L}_1$ to the reduct $(L_2, \odot_2, \top_2)$ of $\mathbf{L}_2$, i.e. $h(a\,\overline{\odot}_1\, b) = h(a) \odot h(b)$ and $h(\bot_1) = \top_2$.

The following lemma shows a characterization of c-measures using $\odot$- and $\overline{\odot}_{\mathrm{d}}$-homomorphisms.

Lemma 1. *Let $f, g : L \to L$ be $\odot$- and $\overline{\odot}_{\mathrm{d}}$-homomorphisms from $\mathbf{L}$ to $\mathbf{L}$ such that $f(\bot) \in \{\bot, \top\}$ and $g(\top) \in \{\bot, \top\}$. Let $\mathcal{C}_{f,g} : \mathcal{FIN} \to \mathcal{N}$ be a mapping defined by the induction:*

$$\mathcal{C}_{f,g}(\{a/x\})(0) = g(a),\ \mathcal{C}_{f,g}(\{a/x\})(1) = f(a),\ \mathcal{C}_{f,g}(\{a/x\})(k) = f(\bot),\ k > 1$$

hold for each singleton $\{a/x\} \in \mathcal{FIN}$ and

$$\mathcal{C}_{f,g}(A) = \mathcal{C}_{f,g}(\{A(x_1)/x_1\}) + \cdots + \mathcal{C}_{f,g}(\{A(x_m)/x_m\})$$

holds for each $A \in \mathcal{FIN}$ with $\mathrm{Supp}(A) = \{x_1, \ldots, x_m\}$. *Then the mapping $\mathcal{C}_{f,g}$ is a c-measure of finite fuzzy sets.*

Theorem 3 (Representation of $\odot$-cardinality). *Let $\mathcal{C} : \mathcal{FIN} \to \mathcal{N}$ be a mapping satisfying the additivity axiom. Then the following statements are equivalent:*

(i) $\mathcal{C}$ is a c-measure of finite fuzzy sets,

(ii) there exist an $\odot$-homomorphism $f : \mathbf{L} \to \mathbf{L}$ and an $\overline{\odot}_{\mathrm{d}}$-homomorphism $g : \mathbf{L} \to \mathbf{L}$, such that $f(\bot) \in \{\bot, \top\}$, $g(\top) \in \{\bot, \top\}$ and

$$\mathcal{C}(\{a/x\})(0) = g(a),\ \mathcal{C}(\{a/x\})(1) = f(a),\ \mathcal{C}(\{a/x\})(k) = f(\bot)$$

hold for arbitrary $a \in L$, $x \in \mathcal{U}$ and $k \in N$, $k > 1$.

[7] Note that each homomorphism between rdr-lattices (or residuated lattices which are the reducts of original rdr-lattices) is also an $\odot$-homomorphism.

3.3 Selected Properties of c-Measures

It is well known that cardinality is order preserving (class) mapping, i.e. $A \subseteq B$ implies $|A| \leq |B|$. To investigate a monotonicity of c-measures, let us introduce the order preserving (reversing) homomorphisms and c-measures. We shall say that an $\odot$-homomorphism (an $\overline{\odot}_{\mathrm{d}}$-homomorphism) $h : \mathbf{L}_1 \to \mathbf{L}_2$ is *order preserving*, if $h(a) \leq h(b)$ holds for arbitrary $a, b \in L_1$ such that $a \leq b$.[8] Further, we shall say that a c-measure $\mathcal{C}$ is *order preserving* (*order reversing*), if $\mathcal{C}(A) \leq \mathcal{C}(B)$ holds for arbitrary $A, B \in \mathcal{FIN}$ such that $A \leq B$ ($B \leq A$). Finally, we shall say that an $\odot$-homomorphism f (an $\overline{\odot}_{\mathrm{d}}$-homomorphism g) is *trivial*, if $f(a) = \top$ ($g(a) = \bot$) for any $a \in L$.

Thus, we may characterize the monotonicity for c-measures as follows. Recall that, according to the representation theorem, we may freely use c-measures generated by $\odot$- and $\overline{\odot}_{\mathrm{d}}$-homomorphisms.

Theorem 4. *Let $\mathcal{C}_{f,g} : \mathcal{FIN} \to \mathcal{N}$ be a c-measure generated by order preserving $\odot$- and $\overline{\odot}_{\mathrm{d}}$-homomorphisms f and g, respectively. Then*

(i) $\mathcal{C}_{f,g}$ is order preserving if and only if g is trivial, i.e. $\mathcal{C}_{f,g} = \mathcal{C}_f$.
(ii) $\mathcal{C}_{f,g}$ is order reversing if and only if f is trivial, i.e. $\mathcal{C}_{f,g} = \mathcal{C}_g$.

In the cardinal theory of sets, there is a very important property of cardinality called the valuation property and characterized by the following formula

$$|A \cap B| + |A \cup B| = |A| + |B|, \tag{12}$$

A, B are arbitrary sets. Unfortunately, this property is not satisfied for all c-measures, in general.[9] A partial result is given in the following statement. Note that an analogical result for FGCount and FLCount has been proved by M. Wygralak in [7] (see Theorems 4.18 and 4.56).

Theorem 5. *Let $\mathcal{C} : \mathcal{FIN} \to \mathcal{N}$ be a c-measure. Then*

$$\mathcal{C}(A \cap B) + \mathcal{C}(A \cup B) \geq \mathcal{C}(A) + \mathcal{C}(B) \tag{13}$$

holds for arbitrary $A, B \in \mathcal{FIN}$. If $\mathbf{L}$ is linearly ordered, then $\mathcal{C}$ satisfies the valuation property.

In the cardinal theory, there is a fundamental relation between the cardinality of sets and bijective mappings saying that two sets have the same cardinality if and only if there is a one-to-one correspondence between them. Note that we usually say that two sets are equipotent or equipollent, if there exists a one-to-one correspondence between them.

[8] One can see that an $\odot$-homomorphism (an $\overline{\odot}_{\mathrm{d}}$-homomorphism) is trivially order preserving for $\odot = \wedge$ ($\overline{\odot} = \vee$).

[9] Note that a counterexample can be constructed in a complete *rdr*-lattice determined by an Archimedean continuous t-norm T and t-conorm S, where the T-intersection of fuzzy sets is defined by $(A \cap_T B)(x) = T(A(x), B(x))$ and the S-union of fuzzy sets by $(A \cup_S B)(x) = S(A(x), B(x))$.

Let $A, B \in \mathcal{FIN}$ be arbitrary **L**-fuzzy sets. We shall say that A and B are *equipollent*, if there exists a bijective mapping $f : \mathrm{Supp}(A) \to \mathrm{Supp}(B)$ such that $A(x) = B(f(x))$ for each $x \in \mathrm{Supp}(A)$. The fact that A and B are equipollent will be denoted by $A \equiv B$. The following theorem shows a "weaker implication" between the equipollency of finite fuzzy sets and the equality of their generalized cardinals. Note that the inverse implication cannot be principally satisfied in general because of a wide scale of possible definitions of c-measures.

Theorem 6. *Let $\mathcal{C}$ be a c-measure. If $A \equiv B$, then $\mathcal{C}(A) = \mathcal{C}(B)$.*

4 Conclusion

We presented an axiomatic approach to fuzzy measures like cardinality of sets for fuzzy sets and showed some of their properties including a representation by two lattice homomorphisms. One could see (the last theorem) that, contrary to the Cantor-Bernstein theorem, we proved only a weaker implication. An interesting question is whether there exists a wider class of c-measures provided with a reasonable defined equipollency of fuzzy sets making the desired equivalence true. Another interesting question is whether there exists a graded equipollency for which a graded version of Cantor-Bernstein theorem is satisfied.

References

1. Casasnovas, J., Torrens, J.: An axiomatic approach to fuzzy cardinalities of finite fuzzy sets. Fuzzy Sets and Systems 133, 193–209 (2003)
2. Gottwald, S.: Fuzzy uniqueness of fuzzy mappings. Fuzzy Sets and Systems 3, 49–74 (1980)
3. Gottwald, S.: A note on fuzzy cardinals. Kybernetika 16, 156–158 (1980)
4. Holčapek, M., Turčan, M.: Graded equipollence of fuzzy sets. In: Proceedings of IFSA World Congress/EUSFLAT Conference, Lisboa, Universidade Técnica de Lisboa, pp. 1565–1570 (2009)
5. Wygralak, M.: An axiomatic approach to scalar cardinalities of fuzzy sets. Fuzzy Sets and Systems 110(2), 175–179 (2000)
6. Wygralak, M.: Fuzzy sets with triangular norms and their cardinality theory. Fuzzy Sets and Systems 124(1), 1–24 (2001)
7. Wygralak, M.: Cardinalities of Fuzzy Sets. Kluwer Academic Publisher, Berlin (2003) ISBN: 3-5400-0337-1
8. Zadeh, L.: Similarity relations and fuzzy orderings. Inf. Science 3, 177–200 (1971)
9. Zadeh, L.: A computational approach to fuzzy quantifiers in natural languages. Comp. Math. with Applications 9, 149–184 (1983)
10. Wygralak, M.: Vaguely defined objects. Representations, fuzzy sets and nonclassical cardinality theory. Theory and Decision Library. Series B: Mathematical and Statistical Methods, vol. 33. Kluwer Academic Publisher, Dordrecht (1996)
11. McLarty, C.: Elementary Categories, Elementary Toposes. Oxford University Press, Oxford (1992)

Choquet-Integral-Based Evaluations by Fuzzy Rules: Methods for Developing Fuzzy Rule Tables on the Basis of Weights and Interaction Degrees

Eiichiro Takahagi

Senshu University, Kawasaki 214-8580, Japan
takahagi@isc.senshu-u.ac.jp

Abstract. Choquet-integral-based evaluations by fuzzy rules are comprehensive evaluation methods involving the use of a fuzzy rule table and the Choquet integral. Fuzzy measures are identified from the fuzzy rule table. In this paper, we propose methods for developing fuzzy rule tables for Choquet integral models on the basis of a basic fuzzy rule table and weights of evaluation items.

1 Introduction

Fuzzy rule tables and fuzzy reasoning models are very useful tools for developing fuzzy control models. Choquet integral models[1] are also very useful models for comprehensive evaluations in multiattribute decision making [2] [3]. Choquet-integral-based evaluations by fuzzy rules[4] inherit the merits of both the Choquet integral model and the fuzzy rule table. Fuzzy rule tables and "If-then Rules" are very convenient and well-known tools and can be understood intuitively. Therefore, they are used in a large number of fuzzy applications use them. However, max-min calculations do not have good properties such as the monotonicity property. In our method, in order to use Choquet integral models, max-min calculations are not performed.

The fuzzy rule table is of the same form as ordinal simplified fuzzy reasoning models, but the calculations involve the extended Choquet integral, min-max calculations or product-sum calculations are not performed. The comprehensive evaluation values satisfy continuous and piecewise linear outputs. Moreover, if fuzzy rules are monotone with respect to the inputs, the output values also satisfy the monotonicity property. The model can represent the cumulative prospect theory[5] and the bi-capacity models[6] [7]. Some types of the model are special cases of k-ary bicapacities and the Choquet integral[8],[9].

In section 3, Choquet-integral-based evaluations by fuzzy rules are described. In section 4, the basic fuzzy rule table is defined. The basic fuzzy rule table is a hypothetical fuzzy rule table that describes the degrees of interaction. From the basic fuzzy rule table and weights of evaluation items, a fuzzy rule table is developed.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 515–524, 2010.

2 Definitions

2.1 Notations

X is the set of evaluation items (n: number of evaluation items), x_i is the input value of the i^{th} item, and y is the comprehensive evaluation value.

2.2 Fuzzy Space Division Constraint

Each input i is divided into m fuzzy sets in which the membership functions are p_i^j for $i = 1, \ldots, n, j = 1, \ldots, m$. For each input item, all the membership functions satisfy the following conditions:

1. All fuzzy sets are normal and convex. The kernel of a membership function is unique $\forall i, j$; that is, there is a unique point v_i^j, where $p_i^j(v_i^j) = 1\ \forall i$ and j.
2. The sum of the membership values is 1; that is $\sum_j p_i^j(x_i) = 1, \forall x_i, i$.
3. There are one or two active membership functions $p_i^j(x_i) > 0\ \forall x_i, \forall i$.

An example of the membership function is a triangle membership function such as

$$p_i^1(x_i) = \max(-\frac{1}{50}x_i + 1, 0) \tag{1}$$

$$p_i^2(x_i) = \begin{cases} \frac{1}{50}x_i & \text{if } x_i \leq 50 \\ \max(-\frac{1}{50}(x_i - 50), 0) & \text{if } x_i > 50 \end{cases} \tag{2}$$

$$p_i^3(x_i) = \max(\frac{1}{50}(x_i - 50), 0). \tag{3}$$

where the domains of x_i are $[0, 100]$ and $p_i^1(x_i)$, $p_i^2(x_i)$, and $p_i^3(x_i)$ are "Small", "Middle", and "Big" membership functions respectively.

2.3 Fuzzy Rule Table

Representative points $(k_1, \ldots, k_n)$ are defined as n-tuple of kernel numbers. The fuzzy rule table c is defined as a function from the representative points to the output values:

$$c : \{1, \ldots, m\}^n \to \mathbb{R}. \tag{4}$$

Table 1 is an example of a fuzzy rule table.

2.4 Choquet Integral and Extended Choquet Integral

A non-monotone fuzzy measure $\mu^\diamondsuit$ is defined as

$$\mu^\diamondsuit : 2^X \to \mathbb{R}\ ,\ \ \mu^\diamondsuit(\emptyset) = 0. \tag{5}$$

Table 1. Fuzzy rule table ($n = 2$ and $m = 3$)

Input 2 \ 1	Small (1)	Middle (2)	Big (3)
Small (1)	0 (= $c(1,1)$)	50 (= $c(2,1)$)	80 (= $c(3,1)$)
Middle (2)	30 (= $c(1,2)$)	60 (= $c(2,2)$)	90 (= $c(3,2)$)
Big (3)	70 (= $c(1,3)$)	80 (= $c(2,3)$)	100 (= $c(3,3)$)

The Choquet integral with respect to $\mu^{\diamondsuit}$ is defined as

$$y = (\mathrm{C}) \int h d\mu \equiv \int_0^{\infty} \mu^{\diamondsuit}(\{x \mid h(x) > r\}) dr. \tag{6}$$

The extended fuzzy measure and extended Choquet integral [10,11] are proposed for handling the cases in which $\mu(\emptyset) \neq 0$. An extended fuzzy measure μ is defined as

$$\mu : 2^X \to \mathbb{R}. \tag{7}$$

As the values in the integrand h are membership values, that is, $h(i) \in [0,1], \forall i$ and $\mu(\emptyset) = 0$ is not assumed, the integration interval of the Choquet integral calculation is limited to $[0,1]$. The extended Choquet integral is defined as

$$y = (\mathrm{EC}) \int h d\mu \equiv \int_0^{1} \mu(\{x \mid h(x) > r\}) dr. \tag{8}$$

The extended Choquet integral can be calculated by using an ordinal Choquet integral:

$$(\mathrm{EC}) \int h d\mu = (\mathrm{C}) \int h d\mu^{\diamondsuit} + \mu(\emptyset) \ \text{ where } \mu^{\diamondsuit}(A) = \mu(A) - \mu(\emptyset), \forall A \in 2^X. \tag{9}$$

3 Calculation Method Based on Segment Division

3.1 Segmentation and Segment Selection

Segmentation is performed at kernel points v_i^j for all inputs. Figure 1 shows an example of the segmentation. Segment $S_{(k_1,\ldots,k_n)}(k_i < m, \forall i)$ is the n-dimensional rectangle whose vertices are the representative points $(k_1 + l_1, \ldots, k_n + l_n), \forall l_i \in \{0,1\}, i = 1, \ldots, n$. First, the segment that includes the input values $(x_1, \ldots, x_n)$ is selected. If $x_1 = 90$ and $x_2 = 20$, $S_{(2,1)}$ is selected.

In this model, each segment has a different extended fuzzy measure $\mu^{(k_1,\ldots,k_n)}$ and integrand $h^{(k_1,\ldots,k_n)}$, but the Choquet integral is calculated only for the selected segment.

3.2 Base Point Selection

The base point is the origin of the fuzzy measure and integrand for the segment. First, the base point is selected from among the representative points in

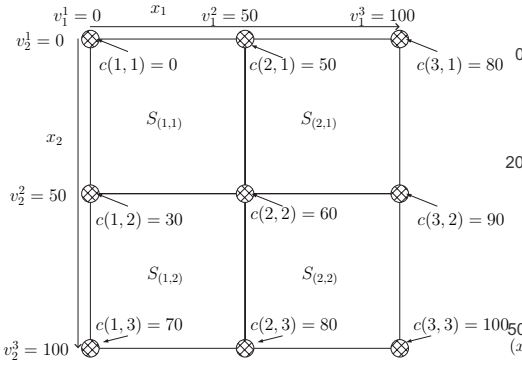


Fig. 1. Segmentation

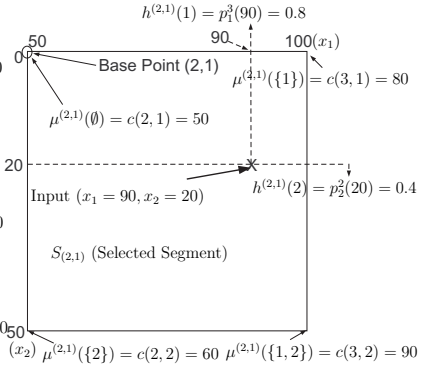


Fig. 2. Calculation Process

the selected segment. Alternative base points of segment $S_{(k_1,\ldots,k_n)}$ are $(k_1 + l_1, \ldots, k_n + l_n), \forall l_i \in \{0,1\}, i = 1,\ldots,n$.

The output values change with the methods for selecting the base points. TF-, BP-, and AV-type are proposed in [4]. In this paper, the TF-type is used, because this is the simplest type of method.

3.3 TF-Type

In TF-type models, the upper left representative points are selected as base points for all segments; that is, the base point of segment $S_{(k_1,\ldots,k_n)}$ is $(k_1,\ldots,k_n)$.

Extended fuzzy measure $\mu^{(k_1,\ldots,k_n)}$ assignment. Extended fuzzy measures $\mu^{(k_1,\ldots,k_n)}$ are assigned as follows:

$$\mu^{(k_1,\ldots,k_n)}(A) = c(l_1,\ldots,l_n), \forall A \in 2^X \tag{10}$$

$$\text{where } l_i = \begin{cases} k_i & \text{if } i \notin A \\ k_i + 1 & \text{if } i \in A. \end{cases}$$

Integrand $h^{(k_1,\ldots,k_n)}$ assignment. Integrands $h^{(k_1,\ldots,k_n)}$ are assigned as follows:

$$h^{(k_1,\ldots,k_n)}(i) = p_i^{k_i+1}(x_i), \quad i = 1,\ldots,n. \tag{11}$$

Extended Choquet integral. The output value y is calculated from the extended Choquet integral by using the fuzzy measure $\mu^{(k_1,\ldots,k_n)}$ and integrand $h^{(k_1,\ldots,k_n)}$ of the selected segment:

$$y = (\mathrm{EC}) \int h^{(k_1,\ldots,k_n)} d\mu^{(k_1,\ldots,k_n)} \tag{12}$$

3.4 Example

When $x_1 = 90$ and $x_2 = 20$, segment $S_{(2,1)}$ is selected, and $(2,1)$ is selected as the base point. From figure 2, $h^{(2,1)}(1) = 0.8$ and $h^{(2,1)}(2) = 0.4$; $\mu^{(2,1)}(\emptyset) = c(2,1) = 5$, $\mu^{(2,1)}(\{1\}) = c(3,1) = 80$, $\mu^{(2,1)}(\{2\}) = c(2,2) = 60$, and $\mu^{(2,1)}(\{1,2\}) = c(3,2) = 90$. The output is $y = (\mathrm{EC}) \int h^{(2,1)} d\mu^{(2,1)} = 78$.

3.5 K-ary Bicapacities and Choquet Integral Model

K-ary bicapacities and the Choquet integral model were proposed by Grabisch and Labreuche[8],[9]. F is a capacity on a lattice $L = m^n$. The fuzzy rule table c for equation (4) is F, that is $c(l_1, \ldots, l_n) = F(l_1 - 1, \ldots, l_n - 1)$. In the k-ary model, the domain of the integrand of the Choquet integral is the maximal chain of the join-irreducible elements of L. When $m = 3$ and $n = 2$, the value of $f(1,0)$, $f(2,0)$, $f(0,1)$, and $f(0,2)$ are given by DMs. The comprehensive value is calculated by the Choquet integral of f with respect to F.

The TF-type model is a special case of the k-ary capacities model. In the Choquet-integral-based evaluation models, there is only one k_i for which $f(0, \ldots, 0, k_i, 0 \ldots, 0) > 0, \forall i$. When $m = 3$ and $n = 2$ and the segment $S_{(i,j)}$ is selected, $f(i,0) = p_1^i(x_1)$, $f(0,j) = p_2^j(x_2)$, $f(k,0) = 0, \forall k \neq i$, and $f(0,k) = 0, \forall k \neq j$. The comprehensive value of the Choquet-integral-based evaluation is equal to the value of the Choquet integral of f with respect to F.

4 Method for Developing Fuzzy Rule Tables on the Basis of Weights and Interaction Degrees

It is not easy to identify all the output values of representative points if n is large. For example, if $n = 5$ and $m = 3$, then the number of representative points is $m^n = 243$. Therefore, we propose an identification method based on weights of evaluation items and a basic fuzzy rule table.

In this model, we assume that all evaluation items are homogeneous. This means that they have the same fuzzy space division; that is, each evaluation item has the same numbers of membership functions and the labels of the membership functions are identified. Interaction degrees have a constant value across all set of evaluation items such as in the case of λ fuzzy measure[12]. In this paper, we are concerned only with $m = 3$ case.

4.1 Basic Fuzzy Rule Table

Basic fuzzy rule tables (figure 3) show the interaction degrees among evaluation items and the output values of representative points with two homogeneous, equally weighted, hypothetical evaluation items. The table is developed by DMs. A basic fuzzy rule table b is a fuzzy rule table with $m = 3$ and $n = 2$:

$$b : \{1,2,3\} \times \{1,2,3\} \to \mathbb{R}. \tag{13}$$

Basic fuzzy rule tables are assumed to be based on monotone fuzzy rule; that is,

$$b(i_1, j_1) \geq b(i_2, j_2) \text{ if } i_1 \geq i_2, j_1 \geq j_2. \tag{14}$$

The output values are symmetrical, that is $b(i,j) = b(j,i), \forall i,j$. The output values are $b(k,k), k = 1,2,3$ when $x_i = v_i^k, i = 1,\ldots,n$. The value of $b(1,2)$ indicates the interaction degree when all inputs are in the low area. If $b(0,0) + \frac{1}{2}[b(1,1) - b(0,0)] < b(1,2) < b(2,2)$, then the interaction degree among low area inputs is substitute. If $b(0,0) < b(1,0) < b(0,0) + \frac{1}{2}[b(1,1) - b(0,0)]$, then the interaction degree is complementary. The value of $b(i+1,i)$ indicates the interaction degree when all inputs are in the diagonal segments $G_{(i,i)}, i = 1,2,3$.

DMs assign suitable output values that satisfy the monotony constraint of basic fuzzy rule table (equation (14)). Table 2 is an example of a basic fuzzy rule table. In the low area ($G_{(1,1)}$), the interaction degree is substitute. In the high area ($G_{(2,2)}$), the interaction degree is complementary.

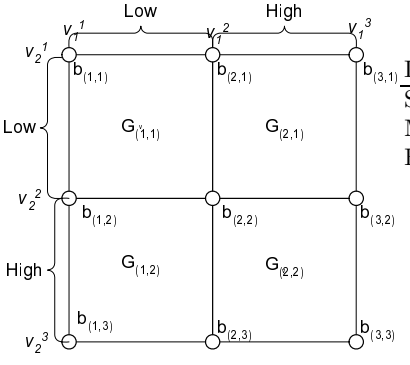


Fig. 3. Basic Fuzzy Rule Table

Table 2. Basic Fuzzy Rule Table

Input 2 \ 1	Small (l_1^1)	Middle (l_1^2)	Big (l_1^3)
Small (l_2^1)	0 ($b_{(1,1)}$)	30 ($b_{(2,1)}$)	35 ($b_{(3,1)}$)
Middle (l_2^2)	30 ($b_{(1,2)}$)	50 ($b_{(2,2)}$)	60 ($b_{(3,2)}$)
Big (l_2^3)	35 ($b_{(1,3)}$)	60 ($b_{(2,3)}$)	100 ($b_{(3,3)}$)

4.2 Measurement of Interaction Degrees

To measure the interaction degree, ξ of the ϕ_s transformation [14] is used. Unlike Murofusi and Soneda's interaction degree between pairs of items [16], ξ can show the interaction degree among two or more items such as λ.

***ϕ_s* Transformation.** $\phi_s : [0,1] \times [0,1] \to [0,1]$ is defined as

$$\phi_s(\xi, u) = \begin{cases} 1 & \text{if } (\xi = 1 \text{ and } u > 0) \text{ or } (\xi = 0 \text{ and } u = 1) \\ 0 & \text{if } (\xi = 1 \text{ and } u = 0) \text{ or } (\xi = 0 \text{ and } u < 1) \\ \frac{s^u - 1}{s - 1} & \text{otherwise,} \end{cases} \tag{15}$$

where $s = \frac{(1-\xi)^2}{\xi^2}$. The function ϕ_s is one of the scaling functions [13]. The fuzzy measure μ involving the use of the ϕ_s transformation is assigned as

$$\mu(A) = \phi_s(\xi, \sum_{i \in A} u_i), \tag{16}$$

where $\sum_{i=1}^{n} u_i = 1, u_i \in [0,1], i = 1, \ldots n$. The weights $u_i (i = 1, \ldots, n)$ are weights of evaluation items and ξ is an interaction degree to identify μ. If $0.5 < \xi \leq 1$, then μ is a sub-additive fuzzy measure. If $0 \leq \xi < 0.5$, then μ is a super-additive fuzzy measure. The inverse function of ϕ_s of u is

$$\phi_s^{-1}(\xi, v) = \frac{\log[v(s-1)+1]}{\log s}, \xi \in (0,1), s = \frac{(1-\xi)^2}{\xi^2}. \tag{17}$$

Measurement of Interaction Degree for Segment. To measure the interaction degree $\xi_{(i,j)}$ for segment $G_{(i,j)}$, the normal fuzzy measure $\mu_{(i,j)}$ is assigned on the basis of the values of the representative points in the segment,

$$\mu_{(i,j)}(\emptyset) = 0, \;\; \mu_{(i,j)}(\{1\}) = \frac{b(i+1,j) - b(i,j)}{b(i+1,j+1) - b(i,j)} \tag{18}$$

$$\mu_{(i,j)}(\{2\}) = \frac{b(i,j+1) - b(i,j)}{b(i+1,j+1) - b(i,j)}, \;\; \mu_{(i,j)}(\{1,2\}) = 1 \tag{19}$$

From the fuzzy measure $\mu_{(i,j)}$ and the ϕ_s transformation, $\alpha^P_{(i,j)}, \alpha^Q_{(i,j)} \in [0,1]$ and $\xi_{(i,j)} \in [0,1]$ are calculated by using

$$\phi_s(\xi_{(i,j)}, \alpha^P_{(i,j)}) = \mu_{(i,j)}(\{1\}), \;\; \phi_s(\xi_{(i,j)}, \alpha^Q_{(i,j)}) = \mu_{(i,j)}(\{2\}) \tag{20}$$

where $\alpha^P_{(i,j)} + \alpha^Q_{(i,j)} = 1$. As $\phi_s(\xi, u)$ is a strong increasing function of ξ and u, where $\xi \in (0,1)$ and $u \in (0,1)$, ξ and u can be calculated by repeated operations. The superscripts P and Q indicate the area name.

In the diagonal segments $G_{(1,1)}$ and $G_{(2,2)}$, as $\mu(\{1\}) = \mu(\{2\})$, $\alpha^L = \alpha^H = 0.5$. In the segments $G_{(1,2)}$ and $G_{(2,1)}$, the value of $\mu(\{1\})$ and $\mu(\{2\})$ are different. In segment $G_{(1,2)}$, as $\phi_s(0.387, 0.7565) = 0.6666 (= \mu_{(1,2)}(\{1\}))$ and $\phi_s(0.387, 0.2435) = 0.1667 (= \mu_{(1,2)}(\{2\}))$, $\xi_{(2,1)} = 0.387$, $\alpha^L_{(2,1)} = 0.7565$ and $\alpha^H_{(2,1)} = 0.2435$.

The weights α^L and α^H are called additional weights. Table 3 lists the identified interaction degrees and additional weights corresponding to tabele 2.

Table 3. Interaction Degrees and Additional Weights

Segment	Interaction Degree	Additional Weights
$G_{(1,1)}$	$\xi_{(1,1)} = 0.6$	$\alpha^L_{(1,1)} = 0.5$
$G_{(2,2)}$	$\xi_{(2,2)} = 0.2$	$\alpha^H_{(2,2)} = 0.5$
$G_{(2,1)}$	$\xi_{(2,1)} = 0.387$	$\alpha^H_{(2,1)} = 0.2435, \alpha^L_{(2,1)} = 0.7565$
$G_{(1,2)}$	$\xi_{(1,2)} = 0.387$	$\alpha^H_{(1,2)} = 0.2435, \alpha^L_{(1,2)} = 0.7565$

4.3 Identification of Fuzzy Rule Table

In this section, a fuzzy rule table identification method based on the use of the basic fuzzy rule table and weights of evaluation items is proposed. The example ($n = 3$) involvs the use of table 2, and weights are $w_1 = 0.5$, $w_2 = 0.3$, and $w_3 = 0.2$.

Identification of Fuzzy Measures. Normal fuzzy measures $\mu_{(i,\ldots,i)}$ of diagonal segments $S_{(i,\ldots,i)}$ are assigned as

$$\mu_{(i,\ldots,i)}(A) = \phi_s(\xi_{(i,i)}, \sum_{j \in A} w_j), \forall A, i = 1, 2. \tag{21}$$

As non diagonal segments have two or more values of ξ, the fuzzy measure values cannot be calculated by one ϕ_s transformation. Therefore, fuzzy measures are identified ϕ_s conversion methods[15].

Figure 4 shows the outline of the process for calculating $\mu_{(1,2,1)}$. $S_{(1,2,1)}$ has two inputs correspondin to the low area and one input corresponding to the high area. In the lower levels of the hierarchy diagram, the inputs in a single area are combined, and in the upper level, the different areas are combined. In the different area conbination, there are additional weights α^L and α^H. Therefore, in the connection, new weights ω are defined as

$$\omega_i = \frac{\alpha^{*_i}_{(1,2)} \times w_i}{\sum_j [\alpha^{*_j}_{(1,2)} \times w_j]} \tag{22}$$

where $*_i$ indicates the area to which the i-th input belongs.

As the interaction degrees in the lower and upper level of the hierarchy diagram are different, the weights that are calculated in the lower level are coverted to upper-level weights using the conversion ratio T, which is the ratio of the value of the lower label ϕ_s to upper level. The details are presented in [15].

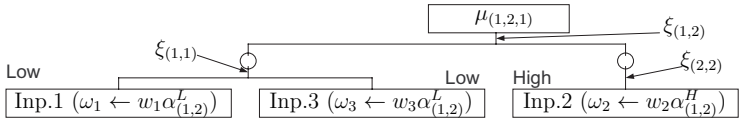


Fig. 4. Hierarchy Diagram of Segment $S_{(1,2,1)}$

Procedure for Assigning Values of Representative Points. The values of representative points are determined on the basis of fuzzy measure values of the segment. When $c(k_1, \ldots, k_n)$ and $c(k_1+1, \ldots, k_n+1)$ in a segment $S_{(k_1,\ldots,k_n)}$ have been previously assigned, the value of the representative point in the segment are as fellows

$$c(k_1 + l_1, \ldots, k_n + l_n) =$$
$$\mu_{(k_1,\ldots,k_n)}(\bigcup_{\{j;l_j=1\}} \{j\})[c(k_1 + 1, \ldots, k_n + 1) - c(k_1, \ldots, k_n)] + c(k_1, \ldots, k_n)$$
$$\forall l_i \in \{0, 1\}, i = 1, \ldots, n. \tag{23}$$

1. Assign diagonal representative points

$$c(i, \ldots, i) = b(i, i), \quad i = 1, 2, 3 \tag{24}$$

2. Assign the representative point values of the diagonal segments $S_{(i,\ldots,i)}, i = 1, 2$ using the equation (23).
3. Select a segment $S_{(k_1,\ldots,k_n)}$ and assign the unassigned representative point values using equation (23).
4. Repeat step 3 until all representative values are have been assigned.

Table 4 is the identified fuzzy rule table where $w_1 = 0.5$, $w_2 = 0.3$, and $w_3 = 0.2$, and table 2. Table 5 lists the Shapley values of the each segment.

Table 4. Identified Fuzzy Rule Table

x_1	x_3	x_2 Small (1)	Middle (2)	Big (3)
1	Big (3)	15.31	32.58	35.00
	Middle (2)	13.47	30.00	33.79
	Small (1)	0.00	19.44	22.17
2	Big (3)	41.67	52.47	60.00
	Middle (2)	38.98	50.00	54.32
	Small (1)	30.00	42.96	48.21
3	Big (3)	48.90	69.88	100.00
	Middle (2)	46.86	60.00	77.30
	Small (1)	35.00	51.79	58.41

Table 5. Shapley Value

Segments	1	2	3
$S_{(1,1,2)}$	0.59	0.35	0.06
$S_{(1,2,2)}$	0.72	0.16	0.11
$S_{(1,2,1)}$	0.64	0.11	0.25
$S_{(2,1,2)}$	0.37	0.47	0.16
$S_{(2,1,1)}$	0.26	0.44	0.30
$S_{(2,2,1)}$	0.41	0.27	0.32
$S_{(1,1,1)}$	0.50	0.30	0.20
$S_{(2,2,2)}$	0.47	0.31	0.22
Average	0.49	0.30	0.20

5 Conclusion

In this paper, we proposed a fuzzy rule table identification method based on fuzzy measure concepts. However, there are some problems that are yet to be solved. In this method, only the $m = 3$ case is defined. If $m > 3$, there are two or more ξ values among different areas.

In the fuzzy rule table identification procedure, some values of representative points can be identified in two or more segments; that is, some values are dependent on the segment selection order.

References

1. Choquet, G.: Theory of Capacities. Annales de l'Institut Fourier 5, 131–295 (1954)
2. Murofushi, T., Sugeno, M.: A theory of fuzzy measures: representations, the Choquet integral, and null sets. J. Math. Anal. Appl. 159(2), 532–549 (1991)
3. Grabisch, M., Nguyen, H.T., Walker, E.A.: Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference. Kluwer, Dordrecht (1995)
4. Takahagi, E.: Choquet-integral-based Evaluations by Fuzzy Rules. In: Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, Lisbon, pp. 843–848 (2009)
5. Tversky, A., Kahneman, D.: Advances in Prospect Theory, Cumulative Representation of Uncertainty. J. of Risk and Uncertainty 5, 297–323 (1992)

6. Grabisch, M., Labreuche, C.: Bi-capacities — Part I: definition, Möbius transform and interaction. Fuzzy Sets and Systems 151, 211–236 (2005)
7. Grabisch, M., Labreuche, C.: Bi-capacities — Part II: the Choquet integral. Fuzzy Sets and Systems 151, 237–259 (2005)
8. Grabisch, M., Labreuche, C.: Capacities on lattices and k-ary capacities. In: 3d Int, Conf. of the European Soc. for Fuzzy Logic and Technology (EUSFLAT 2003), Zittau (2003)
9. Grabisch, M., Labreuche, C.: Bipolarization of posets and natural interpolationstar. J. of Mathematical Analysis and Applications 343(2), 1080–1097 (2008)
10. Takahagi, E.: Fuzzy Integral Based Fuzzy Switching Functions. In: Peters, J.F., Skowron, A., Dubois, D., Grzymała-Busse, J.W., Inuiguchi, M., Polkowski, L. (eds.) Transactions on Rough Sets II. LNCS, vol. 3135, pp. 129–150. Springer, Heidelberg (2004)
11. Takahagi, E.: Fuzzy three-valued switching functions using Choquet integral. Journal of Advanced Computational Intelligence and Intelligent Informatics 7(1), 47–52 (2003)
12. Sugeno, M.: Theory of Fuzzy Integrals and its Applications, doctoral thesis, Tokyo Institute of Technology (1974)
13. Narukawa, Y., Torra, V.: Fuzzy Measure and Probability Distributions: Distorted Probabilities. IEEE Transactions on Fuzzy Systems 13(5), 617–629 (2005)
14. Tukamoto, Y.: A Measure Theoretic Approach to Evaluation of Fuzzy Set Defined on Probability Space. Journal of Fuzzy Math 2(3), 89–98 (1982)
15. Takahagi, E.: A Fuzzy Measure Identification Method by Diamond Pairwise Comparisons and ϕ_s Transformation. Fuzzy Optimization and Decision Making 7(3), 219–232 (2008)
16. Murofushi, T., Soneda, S.: Techniques for reading fuzzy measures (III): interaction index. In: 9th Fuzzy System Symposium, Sapporo, Japan, pp. 693–696 (1993) (in Japanese)

On a New Class of Implications in Fuzzy Logic

Yun Shi[1], Bart Van Gasse[1], Da Ruan[1,2], and Etienne Kerre[1]

[1] Fuzziness and Uncertainty Modelling Research Unit,
Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9 WE02), 9000 Gent, Belgium
[2] Belgian Nuclear Research Centre (SCK•CEN),
2400 Mol Belgium
{Yun.Shi,Bart.VanGasse,Etienne.Kerre}@UGent.be, DRuan@SCKCEN.be

Abstract. We define and study a new class of implications determined only by a negation. We examine under which conditions the most popular eight axioms for implications are satisfied. We obtain the intersection of the new class of implications with the S- and R- implications.

Keywords: fuzzy implication, negation, Sugeno negation, fuzzy implication axiom, S-implication, R-implication, Łukasiewicz implication.

1 Introduction

An implication in fuzzy logic is an extension of the implication in binary logic. It plays important roles in both mathematical and applied sides of fuzzy set theory ([1], [6], [9], [10], [11]).

Definition 1. An implication I is a $[0,1]^2 \to [0,1]$ mapping that satisfies, for all $x, y, z \in [0,1]$:

FI1. first place antitonicity FA: $x < y \Rightarrow I(x,z) \geq I(y,z)$;
FI2. second place isotonicity SI: $y < z \Rightarrow I(x,y) \leq I(x,z)$;
FI3. dominance of falsity of antecedent DF: $I(0,x) = 1$;
FI4. dominance of truth of consequent DT: $I(x,1) = 1$;
FI5. boundary condition BC: $I(1,0) = 0$.

Besides the basic axioms there are many potential axioms for implications, among which the following eight ones are widely used in the literature ([3], [4], [5], [13]): For all $x, y, z \in [0,1]$,

FI6. neutrality of truth NT: $I(1,x) = x$;
FI7. exchange principle EP: $I(x, I(y,z)) = I(y, I(x,z))$;
FI8. ordering principle OP: $I(x,y) = 1 \Leftrightarrow x \leq y$;
FI9. strong negation principle SN: the mapping N_I defined by $N_I(x) = I(x,0)$, is a strong negation;
FI10. consequent boundary CB: $I(x,y) \geq y$;
FI11. identity ID: $I(x,x) = 1$;

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 525–534, 2010.

FI12. contrapositive principle CP: there exists a strong negation N such that $I(x,y) = I(N(y), N(x))$;
FI13. continuity CO: I is a continuous mapping.

In our recent work [12], we have studied dependencies and independencies of these eight axioms, and found different implications satisfying different subgroups of these eight axioms. The following two implications were introduced:

$$I_6(x,y) = \begin{cases} 1, & x \le y \\ \frac{y}{1+\sqrt{1-x}} + \sqrt{1-x}, & x > y \end{cases}, \quad x, y \in [0,1]. \tag{1}$$

$$I_{10}(x,y) = \begin{cases} 1, & x \le y \\ \frac{(1-\sqrt{1-x^2})y}{x} + \sqrt{1-x^2}, & x > y \end{cases}, \quad x, y \in [0,1]. \tag{2}$$

I_6 satisfies NT, OP, CB, ID, CO but not EP, SN and CP. I_{10} satisfies NT, OP, SN, CB, ID, CO but not EP and CP. These implications actually have the same form. Indeed they can be represented by I^N where N is a negation:

$$I^N(x,y) = \begin{cases} 1, & x \le y \\ \frac{(1-N(x))y}{x} + N(x), & x > y \end{cases}, \quad x, y \in [0,1], \tag{3}$$

If I^N is always an implication, then (3) is an interesting new class of implications because it is only determined by a negation. In this paper we check that I^N is always an implication, and then study this new class of implications. In Section 2 we give some necessary basic notions about negations and some existing classes of implications generated by negations, conjunctions and disjunctions in fuzzy logic. In Section 3 we work out the axioms of the new class of implications. In Section 4 we obtain the intersection of the new class of implications with the S-implications and R-implications. In Section 5 we conclude our results and propose the possible generalization of the new class of implications.

2 Preliminaries

The concepts of order automorphism and conjugate are useful in the paper.

Definition 2. ([3], Definition 0) A mapping $\varphi : [a,b] \to [a,b]$ ($[a,b] \subset \mathbb{R}$) is *an order automorphism of the interval* $[a,b]$ if it is continuous, strictly increasing and satisfies the boundary conditions: $\varphi(a) = a$ and $\varphi(b) = b$.

Definition 3. ([2], Definition 2) Two mappings F, G: $[0,1]^2 \to [0,1]$ are *conjugate*, if there exists an order automorphism φ of the unit interval such that $G = F_\varphi$, where $F_\varphi(x,y) = \varphi^{-1}(F(\varphi(x), \varphi(y)))$, for all $x, y \in [0,1]$.

2.1 Negations in Fuzzy Logic

A negation in fuzzy logic is an extension of the negation in binary logic.

Definition 4. A mapping N: $[0,1] \to [0,1]$ is a *negation* if it satisfies:

N1. boundary conditions: $N(0) = 1$ and $N(1) = 0$,
N2. monotonicity: $(\forall (x,y) \in [0,1]^2)(x \leq y \Rightarrow N(x) \geq N(y))$.

A negation N is a *strong* negation if $N(N(x)) = x$, for all $x \in [0,1]$.

Strong negations are always continuous. But the converse is not true. One of the famous classes of strong negations are the Sugeno negations N_a: there exists an $a \in]-1, +\infty[$ such that for all $x \in [0,1]$, $N_a(x) = \frac{1-x}{1+ax}$. Notice that if $a = 0$, then N_a is the standard negation N_0, $N_0(x) = 1 - x$. An example of a class of non-continuous negations is:

$$N_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A \end{cases}, \quad x \in [0,1], \tag{4}$$

where $A = [0, \alpha[$ $(\alpha \in]0,1])$ or $A = [0, \alpha]$, $\alpha \in [0,1[$. Notice that N_A is the class of negations that take values only in $\{0,1\}$. Another class of negations that will be useful later is:

$$N_{A,\beta}(x) = \begin{cases} 1, & x \in A, \\ \frac{1-x}{1+\beta x}, & x \notin A \end{cases}, \quad x \in [0,1], \tag{5}$$

where $A = [0, \alpha[$ $(\alpha \in]0,1])$ or $A = [0, \alpha]$, $\alpha \in [0,1[$, and $\beta \in]-1, +\infty[$. Notice that $N_{\{0\},\beta}$ is the class of Sugeno negations.

2.2 Classes of Implications in Fuzzy Logic

A conjunction in fuzzy logic is an extension of the conjunction in binary logic. Widely used are *triangular norms* (*t-norms* for short).

Definition 5. A mapping T: $[0,1]^2 \to [0,1]$ is a t-norm if for all $x, y, z \in [0,1]$, $T(x,1) = x$ (boundary condition), $y \leq z \Rightarrow T(x,y) \leq T(x,z)$ (monotonicity), $T(x,y) = T(y,x)$ (commutativity), and $T(x, T(y,z)) = T(T(x,y), z)$ (associativity).

Four important t-norms ([7], Example 1.2, [5]) are commonly used:

1. $T_{\mathbf{M}}(x,y) = \min(x,y)$ (minimum),
2. $T_{\mathbf{P}}(x,y) = xy$ (product),
3. $T_{\mathbf{L}}(x,y) = \max(x + y - 1, 0)$ (Łukasiewicz t-norm),
4. $T_{\mathbf{D}}(x,y) = \begin{cases} \min(x,y), & \text{if } x = 1 \text{ or } y = 1 \\ 0, & \text{otherwise} \end{cases}$ (drastic product).

$T_{\mathbf{M}}$, $T_{\mathbf{P}}$ and $T_{\mathbf{L}}$ are continuous t-norms, $T_{\mathbf{D}}$ not (not even left-continuous).
A disjunction in fuzzy logic is an extension of the disjunction in binary logic. Widely used are *triangular conorms* (*t-conorms* for short).

Definition 6. A mapping S: $[0,1]^2 \to [0,1]$ is a t-conorm if for all $x, y, z \in [0,1]$, $S(x,0) = x$ (boundary condition), $y \leq z \Rightarrow S(x,y) \leq S(x,z)$ (monotonicity), $S(x,y) = S(y,x)$ (commutativity), and $S(x, S(y,z)) = S(S(x,y), z)$ (associativity).

Four important t-conorms ([7], Example 1.14) are commonly used:

1. $S_{\mathbf{M}}(x,y) = \max(x,y)$ (maximum),
2. $S_{\mathbf{P}}(x,y) = x + y - xy$ (probabilistic sum),
3. $S_{\mathbf{L}}(x,y) = \min(x+y,1)$ (Łukasiewicz t-conorm, bounded sum),
4. $S_{\mathbf{D}}(x,y) = \begin{cases} \max(x,y), & \text{if } x=0 \text{ or } y=0 \\ 1, & \text{otherwise} \end{cases}$ (drastic sum).

The most famous and important two existing classes of implications in fuzzy logic which are generated by negations, t-norms and t-conorms are strong implications (S-implications for short) and residuated implications (R-implications for short).

Definition 7. Let S be a t-conorm and N be a negation. An S-implication is defined by

$$I(x,y) = S(N(x),y), \quad \forall x,y \in [0,1]. \tag{6}$$

The four S-implications that are generated by the standard negation N_0 and the four aforementioned t-conorms are

(1) $I_{KD}(x,y) = \max(1-x,y)$ (Kleene-Dienes implication),
(2) $I_R(x,y) = 1 - x + xy$ (Reichenbach implication),
(3) $I_{\mathbf{L}}(x,y) = \min(1-x+y,1)$ (Łukasiewicz implication),
(4) $I_{LS}(x,y) = \begin{cases} y, & \text{if } x=1 \\ 1-x, & \text{if } y=0 \\ 1, & \text{otherwise} \end{cases}$.

Definition 8. Let T be a t-norm. An R-implication is defined by

$$I(x,y) = \sup\{t | T(x,t) \le y\}, \quad \forall x,y \in [0,1]. \tag{7}$$

The four R-implications that are generated by the four aforementioned t-norms are

(1) $I_{GD}(x,y) = \begin{cases} 1, & \text{if } x \le y \\ y, & \text{if } x > y \end{cases}$ (Gödel implication)
(2) $I_{GG}(x,y) \begin{cases} 1, & \text{if } x \le y \\ y/x, & \text{if } x > y \end{cases}$ (Goguen implication),
(3) $I_{\mathbf{L}}(x,y) = \min(1-x+y,1)$ (Łukasiewicz implication),
(4) $I_{LR}(x,y) = \begin{cases} y, & \text{if } x=1 \\ 1, & \text{if } x<1 \end{cases}$.

The Łukasiewicz implication $I_{\mathbf{L}}$ is both an S- and an R- implication. For $I_{\mathbf{L}}$ and implications that are conjugate with it we obtain:

Theorem 1. *([2]) A $[0,1]^2 \to [0,1]$ mapping satisfies the exchange principle (EP), the ordering principle (OP) and is continuous (CO) iff it is the Łukasiewicz implication $I_{\mathbf{L}}$, or it is conjugate with $I_{\mathbf{L}}$.*

3 A New Class of Implications Generated by a Negation

3.1 Is I^N Defined by (3) Always an Implication?

We examine whether the mapping I^N defined by (3) takes its values in $[0,1]$ and it satisfies axioms FI1-FI5. First we rewrite I^N as

$$I^N(x,y) = S_{\mathbf{P}}(N(x), I_{GG}(x,y)). \tag{8}$$

It is then straightforward that $I^N(x,y) \in [0,1]$, and I^N satisfies F1-F5. Therefore, I^N is an implication.

3.2 The Axioms of the New Class of Implications

Now we work out whether I^N defined by (3) satisfies the axioms FI6-FI13. If not always, then under which conditions I^N satisfies the axioms.
(1)NT: We see immediately from (8) that I^N always satisfies NT.
(2)EP: We obtain the following theorem:

Theorem 2. *The implication I^N defined by (3) satisfies EP iff N belongs to one of the following two classes of negations:*

(1) N_A defined by (4),
(2) $N_{A,\beta}$ defined by (5).

Proof. Necessity: Suppose I^N satisfies EP. We will show that if N is not of the form N_A, N must be of the form $N_{A,\beta}$. We will do this in three steps: first we will show that we can find a y_0 such that $0 < N(y_0) < y_0 < 1$. Second we prove that for $x \geq y_0$, $N(x) = \frac{1-x}{1+\beta x}$ for some fixed β. And finally we use this second step to prove that for $x < y_0$, $N(x) = 1$ or $N(x) = \frac{1-x}{1+\beta x}$.
Indeed, if I^N satisfies EP, then for all $x, y, z \in [0,1]$,
$I^N(x, I^N(y,z)) = I^N(y, I^N(x,z))$. Take $z = 0$, we obtain

$$(\forall (x,y) \in [0,1]^2)(I^N(x, N(y)) = I^N(y, N(x))). \tag{9}$$

Suppose $N \neq N_A$. Then in particular $N \neq N_{[0,1[}$. So there exists a $y_1 \in [0,1[$ such that $N(y_1) < 1$. Now take $y_0 \in \,]\max(y_1, N(y_1)), 1[$, then $N(y_0) \leq N(y) < y_0 < 1$. We first show that $N(y_0) > 0$. Indeed, if $N(y_0) = 0$, then for all $x \in [0,1]$, we obtain:

$$N(x) = I^N(x, N(y_0)) = I^N(y_0, N(x)) = \begin{cases} 1, & y_0 \leq N(x) \\ \frac{N(x)}{y_0}, & y_0 > N(x) \end{cases}$$

$$\Rightarrow N = N_A,$$

which we have already excluded. Therefore $N(y_0) > 0$. For all $x \in [y_0, 1[$, $x > N(y_0)$ and $N(x) < y_0$. We obtain:

$$(9) \Rightarrow \frac{1-N(x)}{x} N(y_0) + N(x) = \frac{1-N(y_0)}{y_0} N(x) + N(y_0)$$

$$\Rightarrow \frac{1-N(x)-x}{x} = \frac{1-N(y_0)-y_0}{y_0 N(y_0)} N(x)$$

If $N(x) = 0$, then $\frac{1-N(x)-x}{x} = 0 \Rightarrow x = 1$, which we have already excluded. Therefore we obtain:

$$\frac{1-N(x)-x}{xN(x)} = \frac{1-N(y_0)-y_0}{y_0N(y_0)}$$
$$\Rightarrow N(x) = \frac{1-x}{1+\beta x} \text{ (with } \beta = \frac{1-N(y_0)-y_0}{y_0N(y_0)},\ \beta \in]-1, +\infty[).$$

Now we prove, for any $x \in]0, y_0[$,that if $N(x) \neq 1$, then $N(x) = \frac{1-x}{1+\beta x}$. In other words that, because N is decreasing, $N = N_{A,\beta}$ defined by (5). Indeed, if $N(x) \neq 1$, then we can take y in $]\max(N(x), y_0), 1[$ such that $N(y) =< x$ (this is possible because we have just proven that for $y \in [y_0, 1[$, $N(y) = \frac{1-y}{1+\beta y}$). We obtain:

$$(9) \Rightarrow \frac{1-N(x)-x}{xN(x)} = \frac{1-N(y)-y}{yN(y)} = \beta.$$

Thus $N(x) = \frac{1-x}{1+\beta x}$.

Sufficiency of N_A: We obtain: $I^{N_A}(x,y) = \begin{cases} 1, & x \in A \\ I_{GG}(x,y), & x \notin A \end{cases}$. Thus

$$I^{N_A}(x, I^{N_A}(y,z)) = \begin{cases} 1, & x \in A \text{ or } y \in A \\ I_{GG}(x, I_{GG}(y,z)), & x \notin A \text{ and } y \notin A \end{cases}.$$

According to [2], I_{GG} satisfies EP. Therefore I^{N_A} satisfies EP.

Sufficiency of $N_{A,\beta}$, $A = [0, \alpha[$ $(\alpha \in]0, 1])$ or $A = [0, \alpha]$ $(\alpha \in [0, 1[)$: We obtain: $I^{N_{A,\beta}}(x,y) = \begin{cases} 1, & x \leq y \text{ or } x \in A \\ \frac{1-x+(1+\beta)y}{1+\beta x}, & x > y \text{ and } x \notin A \end{cases}$. Thus

$$I^{N_{A,\beta}}(x, I^{N_{A,\beta}}(y,z)) = \begin{cases} 1, & \text{if } x \in A \text{ or } y \in A \\ & \text{or } x + y + \beta xy \leq 1 + z + \beta z \\ \frac{2+\beta-x-y-\beta xy+(1+\beta)^2 z}{(1+\beta x)(1+\beta y)}, & \text{else} \end{cases}$$
$$= I^{N_{A,\beta}}(y, I^{N_{A,\beta}}(x,z)).$$

□

(3) OP: We obtain the following theorem:

Theorem 3. *The implication I^N defined by (3) satisfies OP iff $x > 0 \Rightarrow N(x) < 1$.*

Proof. This follows from, for all $0 \leq y < x \leq 1$,

$$I^N(x,y) < 1 \Leftrightarrow \frac{1-N(x)}{x} y + N(x) < 1 \Leftrightarrow 1 - N(x) > 0 \Leftrightarrow N(x) < 1.$$

□

(4) SN: It is straightforward that $N_{I^N}(x) = I^N(x,0)$ is a strong negation iff N is a strong negation, because $N_{I^N} = N$.
(5) CB: Because I_{GG} satisfies CB, I^N satisfies CB according to (8).
(6) ID: We see immediately through definition that $I^N(x,x) = 1$ for all $x \in [0,1]$.
(7) CP: We obtain the following theorem:

Theorem 4. *The implication I^N defined by (3) satisfies CP w.r.t. a strong negation N' iff N is a Sugeno negation N_a, $a \in]-1, +\infty[$, and $N' = N_a$.*

Proof. Necessity: Recall that I^N always satisfies NT. If I^N satisfies CP w.r.t. N', then according to ([12], Proposition 6.1), I^N also satisfies SN, and for all $x \in [0,1]$, $N'(x) = I^N(x,0) = N(x)$. Therefore, N is strong and I^N satisfies CP w.r.t. N. We obtain

$$\begin{aligned}
&I^N(N(y), N(x)) = I^N(x,y)\\
\Rightarrow &(\forall x \in]0,1[)(\forall y \in]0,x[)(\frac{1-y-N(y)}{N(y)}N(x) = \frac{1-N(x)-x}{x}y)\\
\Rightarrow &(\forall x \in]0,1[)(\forall y \in]0,x[)(\frac{1-y-N(y)}{yN(y)} = \frac{1-x-N(x)}{xN(x)})\\
\Rightarrow &(\exists a \in [-1,+\infty])(\forall x \in]0,1[)(\frac{1-x-N(x)}{xN(x)} = a).
\end{aligned}$$

If $a = -1$ or $a = +\infty$, then $N = N_A$ defined in (4) with $A = [0,1[$ or $A = \{0\}$, which is not a strong negation. Thus $N = N_a$, which is a Sugeno implication.
Sufficiency: If $N = N_a$, then $I^N(x,y) = \begin{cases} 1, & x \le y \\ \frac{(1+a)y+1-x}{1+ax}, & x > y \end{cases}$, and

$$\begin{aligned}
I^N(N(y), N(x)) &= \begin{cases} 1, & x \le y \\ \frac{1-y}{N(y)}N(x) + y, & x > y \end{cases}\\
&= \begin{cases} 1, & x \le y \\ \frac{(1+a)y+1-x}{1+ax}, & x > y \end{cases}.
\end{aligned}$$

Hence $I^N(x,y) = I^N(N(y), N(x))$. □

(8) CO: We obtain the following theorem:

Theorem 5. *The implication I^N defined by (3) satisfies CO iff N is continuous.*

Proof. It is easily verified that if N is continuous, I^N is continuous in each variable. Therefore by Corollary 1.2.2 in [1], I^N is continuous. The converse follows immediately from $I^N(x,0) = N(x)$. □

Combining the four theorems in this section and Theorem 1 in Section 2, we obtain the following two corollaries:

Corollary 1. *For the implication I^N defined in (3), the following four conditions are equivalent:*

(1) N is a Sugeno negation N_a, $a \in]-1, +\infty[$,
(2) I^N satisfies EP and N is a continuous negation,
(3) I^N satisfies CP (w.r.t. N),
(4) I^N is conjugate with the Łukasiewicz implication $I_{\mathbf{L}}$.

Notice that if $a = 0$, then $N = N_0$. Then $I^N = I_{\mathbf{L}}$.

Corollary 2. *An implication I^N defined by (3) satisfying EP and CO also satisfies OP.*

The converse of Corollary 2 is not true. For example, the implication $I^{N_{\{0\}}} = I_{GG}$: I_{GG} satisfies OP but it is not continuous at the point $(0, 0)$.

4 Intersection of the New Class of Implications with the S- and R- Implications

4.1 Intersection of the New Class of Implications and S-Implications

In this section we find the intersection of the new class of implications defined in (3) and the class of all S-implications as well as the class of all S-implications generated by a t-conorm and a strong negation.

Theorem 6. *The implication I^N defined in (3) is an S-implication $S(N'(x), y)$ iff $N = N'$ and N belongs to one of the following two negations:*

(1) N_A defined by (4) with $A = [0, 1[$,
(2) $N_{A,\beta}$ defined by (5).

Proof. Necessity: Because for all $x \in [0, 1]$,
$N(x) = I^N(x, 0) = S(N'(x), 0) = N'(x)$, $N = N'$.
According to ([1], Proposition 2.4.6), any S-implication satisfies EP. Then according to Theorem 2, if I^N is an S-implication, then $N = N_A$, $A = [0, \alpha[$ ($\alpha \in]0, 1]$) or $A = [0, \alpha]$ ($\alpha \in [0, 1[$), or $N = N_{A,\beta}$. Nevertheless,

$$I^{N_A}(x, y) = \begin{cases} 1, & x \le y \text{ or } x \in A \\ \frac{y}{x}, & x > y \text{ and } x \notin A \end{cases}, \tag{10}$$

while

$$S(N_A(x), y) = \begin{cases} 1, & x \in A \\ y, & x \notin A \end{cases}. \tag{11}$$

If $A \neq [0, 1[$, then we can take x and y such that $0 < y < x < 1$ and $x \notin A$. Then $S(N_A(x), y) = y \neq \frac{y}{x} = I^{N_A}(x, y)$. Thus (10)$\neq$(11) provided $A \neq [0, 1[$. Therefore $I^{N_{[0,\alpha[}}$ ($\alpha < 1$) and $I^{N_{[0,\alpha]}}$ are not S-implications.
Sufficiency of $N = N_{[0,1[}$: $I^{N_{[0,1[}}(x, y) = S(N_{[0,1[}(x), y)$ for any t-conorm S.
Sufficiency of $N = N_{A,\beta}$: Take $S(x, y) = \min(1, x + y + \beta xy)$. We can verify that S is a t-conorm (for the associativity:

$$S(x, S(y, z)) = \min(1, x + y + z + \beta xy + \beta yz + \beta xz + \beta^2 xyz) = S(S(x, y), z)),$$

and that

$$S(N_{A,\beta}(x), y) = \begin{cases} 1, & x \in A \text{ or } x \le y \\ \frac{1-x+y+\beta y}{1+\beta x}, & x \notin A \text{ and } x > y \end{cases} = I^{N_{A,\beta}}(x, y).$$

Consequently, $I^{N_{A,\beta}}$ is an S-implication. □

Combining Corollary 1 and Theorem 6 we obtain the following corollary.

Corollary 3. *For the implication I^N defined by (3), the following three conditions are equivalent:*

(1) I^N is an S-implication generated by a t-conorm and a strong negation,
(2) N is a Sugeno negation N_a, $a \in]-1, +\infty[$,
(3) I^N is conjugate with the Łukasiewicz implication $I_{\mathbf{L}}$.

4.2 Intersection of the New Class of Implications and R-Implications

In this section we find the intersection of the new class of implications defined in (3) and the class of the R-implications generated by left-continuous t-norms.

Theorem 7. *The implication I^N defined in (3) is an R-implication generated by a left-continuous t-norm iff N belongs to one of the following two negations:*

(1) a Sugeno negation N_a, $a \in]-1, +\infty[$,
(2) N_A defined by (4) with $A = \{0\}$.

Proof. Necessity: If I^N is an R-implication generated by a left-continuous t-norm, then according to ([4], Theorem 1.14), I^N satisfies EP and OP. According to Theorem 2, $N = N_A$ defined by (4), or $N = N_{A,\beta}$ defined by (5). According to Theorem 3, $N(x) < 1$ provided $x > 0$. Therefore $N = N_a$, or $N = N_{\{0\}}$.
Sufficiency of $N = N_a$: According to Corollary 1, if $N = N_a$, then I^N is conjugate with $I_{\mathbf{L}}$. According to Theorem 1, I^N is an R-implication.
Sufficiency of $N = N_{\{0\}}$: $I^{N_{\{0\}}} = I_{GG}$, the R-implication generated by the continuous t-norm $T_{\mathbf{P}}$. □

Notice that although $I^{N_{[0,1[}}$ is not an R-implication generated by a left-continuous t-norm, it is the R-implication I_{LR} generated by the non-left-continuous t-norm $T_{\mathbf{D}}$.

5 Conclusions

In this paper we studied a new class of implications determined by a negation N, i.e.,

$$I^N(x, y) = \begin{cases} 1, & x \le y \\ \frac{(1-N(x))y}{x} + N(x), & x > y \end{cases}, \quad x, y \in [0, 1].$$

We first checked that I^N is always an implication. For each of the axioms FI6-FI13 we found sufficient and necessary conditions in terms of N. We also obtained the intersection of I^N with any S-implications and R-implications generated by a left-continuous t-norm. An example of I^N being an R-implication generated by a non-left-continuous t-norm was also given.

It is worth mentioning that if we take N as a Sugeno negation, then I^N is an implication that is conjugate with the Łukasiewicz implication $I_{\mathbf{L}}$.

At the end of this paper we mention a possible generalization of the new class of implications I^N. If we replace $S_{\mathbf{P}}$ and I_{GG} in (8) with any t-conorm S and any implication I, then we obtain a class of implications defined by N, S and I:

$$I^{N,S,I}(x,y) = S(N(x), I(x,y)) \quad x,y \in [0,1].$$

Using the same proof for I^N we obtain that $I^{N,S,I}$ is always an implication. This class of implications helps us to generate new implications from existing ones, which will be the topic of further research.

References

1. Baczyński, M., Jayaram, B.: Fuzzy Implications. Springer, Heidelberg (2008)
2. Baczyński, M.: Residual implications revisited. Notes on the Smets-Magrez Theorem. Fuzzy Sets and Systems 145, 267–277 (2004)
3. Bustince, H., Burillo, P., Soria, F.: Automorphisms, negations and implication operators. Fuzzy Sets and Systems 134, 209–229 (2003)
4. Fodor, J.C., Roubens, M.: Fuzzy Preference Modelling and Multicriteria Decision Support. Kluwer Academic Publishers, Dordrecht (1994)
5. Fodor, J.C.: Contrapositive symmetry of fuzzy implications. Fuzzy Sets and Systems 69, 141–156 (1995)
6. Kerre, E., Nachtegael, M.: Fuzzy Techniques in Image Processing. Physica-Verlag, New York (2000)
7. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms. Kluwer Academic Publishers, Netherlands (2000)
8. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic, Theory and Applications. Prentice Hall, USA (1995)
9. Ruan, D., Kerre, E.: Fuzzy implication operators and generalized fuzzy method of cases. Fuzzy Sets and Systems 54, 23–37 (1993)
10. Ruan, D., Kerre, E.: Fuzzy IF-THEN Rules in Computational Intelligence: Theory and Applications. Kluwer Academic Publishers, Boston (2000)
11. Mas, M., Monserrat, M., Torrens, J., Trillas, E.: A survey on fuzzy implication functions. IEEE Transactions on Fuzzy Systems 15(6), 1107–1121 (2007)
12. Shi, Y., Van Gasse, B., Ruan, D., Kerre, E.: On dependencies and independencies of fuzzy implication axioms. Fuzzy Sets and Systems, doi:10.1016/j.fss.2009.12.003
13. Yager, R.R.: On some new classes of implication operators and their role in approximate reasoning. Information Sciences 167, 193–216 (2004)

Diagrams of Fuzzy Orderings*

Branimir Šešelja and Andreja Tepavčević

Department of Mathematics and Informatics
University of Novi Sad
Trg D. Obradovića 4, 21000 Novi Sad, Serbia
{seselja,andreja}@dmi.uns.ac.rs

Abstract. In the framework of residuated lattice valued fuzzy relations, we investigate connections among fuzzy orderings and corresponding Hasse-like diagrams. We prove that there is a link - a fuzzy totally intransitive relation which is used for determining properties of fuzzy order diagrams.

AMS Mathematics Subject Classification (2000): 06D72, 03E72.

Keywords and Phrases: fuzzy covering, fuzzy order, fuzzy totally intransitive, fuzzy diagram, residuated lattice.

1 Introduction

In the crisp, classical theory of relations, there is a one-to-one correspondence between finite orders and the corresponding coverings on the same set. This correspondence is represented by a Hasse diagram which determines both: an order and the corresponding covering relation.

Since we deal with analogue problems in fuzzy settings, let us mention briefly some historical data. Fuzzy orderings have been extensively investigated due to their applicability in situations in which different kinds of comparison appear. At the beginning, fuzzy ordering was introduced by Zadeh [26], and later investigated by many authors (see e.g., [19]) In the last decade the co-domain of orderings was taken to be a residuated lattice (see [1]). These investigations were applied in e.g., introducing fuzzy lattices [8,9], (see also [25]) and complementedness, [10]. We refer to the cited papers for more information.

Here we investigate fuzzy orders, coverings and diagrams in the framework of residuated lattice valued fuzzy relations. Namely, for the given fuzzy order we define fuzzy covering on the same set, and we prove that its essential property is fuzzy total intransitivity. This property on the other hand generates a fuzzy order for which it is a fuzzy covering. Though, this correspondence is not one-to-one: we show that every fuzzy totally intransitive relation is a fuzzy covering

* The research of both authors was supported by Serbian Ministry of Science and Tech. Development, Grant No. 144011 and by the Provincial Secretariat for Science and Tech. Development, Autonomous Province of Vojvodina, Grant "Lattice methods and applications".

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 535–544, 2010.

for a class of fuzzy orderings, among which there is a minimum one. In a finite case, fuzzy total intransitivity enable construction of a Hasse-like diagram, a particular fuzzy graph. Properties of such diagrams are investigated in terms of fuzzy total intransitivity and fuzzy order. We mention that fuzzy coverings and fuzzy total intransitivity were introduced in [20] and [23], but in different, cutworthy framework: only the basic lattice operations were used. Our present work deals with different operations (multiplication in a residuated lattice combined with meet and join). We do not investigate cuts of such relations, but we concentrate on their link to diagrams. Finally, let us comment our equality framework. The one we use here is crisp equality, though in several papers mentioned in references fuzzy orderings are defined and investigated with respect to fuzzy equalities (i.e., fuzzy equivalences with values 1 on the diagonal only). This approach would lead to essentially different fuzzy diagrams then the ones introduced here. Analogy with crisp cases would be less present, though such an investigation would be interesting.

2 Preliminaries

2.1 Residuated Lattices

Recall that a poset $(L, \leqslant)$ in which for every two-element subset $\{x, y\}$ there exist the infimum $(x \wedge y)$ and the supremum $(x \vee y)$ is a *lattice*. It is *complete* if the infimum and the supremum exist for every subset of L. As an algebraic structure, a complete lattice is usually denoted by $(L, \wedge, \vee, 0, 1)$, where 0 and 1 are the smallest and the greatest element under the order $\leqslant$ in L. Further on, a *commutative monoid* $(L, \otimes, 1)$ is a nonempty set L together with a commutative and associative binary operation $\otimes$ and a unit (neutral) element $1 \in L$ (for every $x \in L$, $1 \otimes x = x \otimes 1 = x$).

A **complete residuated lattice** is here considered to be a structure $(L, \wedge, \vee, \otimes, \rightarrow, 0, 1)$ in which

$(L, \wedge, \vee, 0, 1)$ is a *complete lattice*,

$(L, \otimes, 1)$ is a *commutative monoid*, and

the binary operations $\otimes$ and $\rightarrow$ form an *adjoint pair*, meaning that for all $x, y, z \in L$, $x \otimes y \leqslant z$ if and only if $x \leqslant y \rightarrow z$.

Residuated lattices were introduced in 1939. by Dilworth and Ward ([12]). Recently, Blount and Tsinakis investigated these algebras in a more general way ([2]). A detailed study of residuated lattices can be found in a recent book by Galatos et al. ([13]).

From the point of view of fuzzy mathematics, an extensive presentation of residuated lattices is given in a book by Bělohlávek ([1]). Here we deal with fuzzy relations, hence we use terminology, notions and related properties from this book. Apart from general lattice properties (see eg., [6]), concerning additional operations we use frequently the fact that *multiplication $\otimes$ is isotone with respect to the lattice order $\leqslant$ and distributive with respect to arbitrary infima.* Observe also that 0 *is absorbing with respect to multiplication*: $0 \otimes x = 0$, for every x.

There are many examples of complete residuated lattices whose lattice reduct (i.e., its part which is a complete lattice) is the unit interval of the real line, together with min and max as the lattice operations. Depending on how the other two binary operations are defined ($\otimes$ and $\rightarrow$), there are *Łukasiewicz structure*, *Gödel structure*, *Product structure* and others. All these are linearly ordered, but there are others which are not, like Heyting and Boolean lattices, both under suitably defined adjoint operations (for these, see also [1]).

2.2 Ordering Relation and Hasse Diagram

An **ordered set**, **poset**, is a pair $(A, \leqslant)$ consisting of a nonempty set A and an ordering (reflexive, antisymmetric and transitive) relation on A.

What we also use is a notion of a **directed graph**, which is an ordered pair (V, A) consisting of a set V of **vertices** and a binary relation A on V, whose elements are called **edges**. Directed graphs can be represented by a diagram, so that vertices are points in a plain and edges are arrows connecting them. A **path** is a sequence of vertices connected by edges. If the first and the last vertex of a path coincide, then this path is a **cycle**. In particular, a **loop** is an edge that connects a vertex to itself. A graph is **simple** if it has no loops. A graph without cycles is said to be **acyclic**.

As it is known, Hasse diagram of a finite poset is a graph visually representing the corresponding covering relation. Without referring to a partial order, Hasse diagram can be viewed as a directed acyclic graph which coincides with its transitive reduction. By the definition, the **transitive reduction** of a binary relation R on a set X is a minimal relation on X whose transitive closure coincides with the transitive closure of R.

Lemma 1. *Every directed acyclic graph which coincides with its transitive reduction uniquely represents an ordering relation on the set of its vertices.*

2.3 Fuzzy Relations

Let $\mathcal{L}=(L, \wedge, \vee, \otimes, \rightarrow, 0, 1)$ be a complete residuated lattice. Let X be a nonempty set. A **fuzzy relation** ρ is a mapping from X^2 to L.

Observe that there are several ways to name mappings from a set to a residuated lattice: fuzzy, L-valued, L-fuzzy, lattice valued... Our choice is the adjective 'fuzzy' because we use only one co-domain - a complete residuated lattice, hence no confusion could arise.

As for a fuzzy set, we say that the **support** of a fuzzy relation ρ on X is the crisp relation $\operatorname{supp}\rho \subseteq X^2$ determined by the non-zero values of ρ:

$$\operatorname{supp}\rho := \{(x, y) \in X^2 \mid \rho(x, y) > 0\}.$$

A fuzzy relation ρ on X is
(fuzzy) reflexive: if for all $x \in L$, $\rho(x, x) = 1$;
(fuzzy) antisymmetric: if for all $x, y \in X$, if $x \neq y$, then $\rho(x, y) \otimes \rho(y, x) = 0$;
(fuzzy) transitive: for all $x, y, z \in X$, $\rho(x, y) \geq \rho(x, z) \otimes \rho(z, y)$.

In connection with our comment about names, we sometimes omit adjective 'fuzzy' but only for the above three well known relational properties.

A relation ρ is a **fuzzy ordering** relation if it is reflexive, antisymmetric and transitive.

3 Fuzzy Covering

Let $\rho : X^2 \to L$ be a fuzzy ordering relation on a set X.

Define a fuzzy relation $\alpha_\rho : X^2 \to L$, as follows:

$\alpha_\rho(x,x) = 0$, and if $x \neq y$, then

$$\alpha_\rho(x,y) := \begin{cases} 0, & \text{if } \rho(x,z) \otimes \rho(z,y) > 0 \text{ for some } z \notin \{x,y\}; \\ \rho(x,y), & \text{if } \rho(x,z) \otimes \rho(z,y) = 0 \text{ for all } z \notin \{x,y\}. \end{cases}$$

We call the relation α_ρ a **fuzzy covering relation** induced by the fuzzy ordering ρ.

Remarks. 1. In paper [20], an analogue definition of a fuzzy covering is given in terms of basic lattice operations; here we use multiplication, and two notions differ not only in formal definitions, but also in properties, as indicated later.

2. The term *fuzzy covering* follows the analogue crisp terminology. The same notion could equivalently be named as *fuzzy neighboring*; we consider the former as more adequate.

Next we present a property which turn out to be essential for fuzzy coverings.

We say that a relation ρ on X is **totally fuzzy intransitive** if for every $n \in \mathbb{N}$, and all $x_1, \ldots, x_n \in X$,

from $\rho(x_1,x_2) \otimes \rho(x_2,x_3) \otimes \ldots \otimes \rho(x_{n-1},x_n) > 0$, it follows that

$x_1 \neq x_i$, for all $i = 2, \ldots, n$ and

$\rho(x_1,x_i) = 0$, for all $i = 3, \ldots, n$.

Theorem 1. *Every fuzzy covering relation induced by a fuzzy order is totally fuzzy intransitive.*

Proof. Let α_ρ be the fuzzy covering induced by a fuzzy order ρ. From

$$\alpha_\rho(x_1,x_2) \otimes \alpha_\rho(x_2,x_3) \otimes \ldots \otimes \alpha_\rho(x_{n-1},x_n) > 0,$$

we have that $\alpha_\rho(x_i,x_{i+1}) > 0$ for every $i = 1, \ldots, n-1$, and hence $\alpha_\rho(x_i,x_{i+1}) = \rho(x_i,x_{i+1}) > 0$ for each i. Since $\alpha_\rho(x_i,x_{i+1}) > 0$, by the definition of α_ρ, we have that $x_i \neq x_{i+1}$, for all $i = 1, ..., n-1$.

In case $x_i = x_1$, and $\rho(x_1,x_2) \otimes \rho(x_2,x_3) \otimes \ldots \otimes \rho(x_{i-1},x_1) > 0$, we would have $0 < \rho(x_1,x_2) \otimes \rho(x_2,x_3) \otimes \ldots \otimes \rho(x_{i-1},x_1) \leq \rho(x_1,x_{i-1}) \otimes \rho(x_{i-1},x_1)$, contradicting the antisymmetry of ρ. Thus, $x_1 \neq x_i$, for all $i = 2, \ldots, n$.

From $\alpha_\rho(x_1,x_i) > 0$, for some $i \in \{3, \ldots, n\}$ it follows that $\rho(x_1,x_i) > 0$. By the transitivity of ρ, we have that $\rho(x_1,x_{i-1}) \otimes \rho(x_{i-1},x_i) \geqslant \ldots \geqslant \rho(x_1,x_2) \otimes \rho(x_2,x_3) \otimes \ldots \otimes \rho(x_{i-1},x_i) > 0$. But then, $\alpha_\rho(x_1,x_i) = 0$, by the definition of α_ρ. Therefore, the assumption $\alpha_\rho(x_1,x_i) > 0$ leads to a contradiction, which proves the theorem. □

Every fuzzy covering is included in the corresponding fuzzy order. This simple property, which is more precisely formulated in the sequel, is a direct consequence of its definition.

Proposition 1. *If ρ is a fuzzy order on X and α_ρ the corresponding fuzzy covering, then for all $x, y \in X$,*
$\alpha_\rho(x, y)$ is either 0, or it coincides with the value of ρ.

Next we examine a kind of a converse of Theorem 1. We start with a totally fuzzy intransitive relation and we construct the corresponding fuzzy order.

Theorem 2. *Let X be a nonempty set and α a totally fuzzy intransitive relation on X. Then there exists a fuzzy ordering relation ρ such that its fuzzy covering relation is α.*

Proof. Let $\alpha : X^2 \to L$ be a totally fuzzy intransitive relation on X. The required fuzzy relation ρ is reflexive and for all distinct $x, y \in X$ fulfils the following:

$$\rho(x, y) := \bigvee_{n\in\mathbb{N}, x_i\in X} (\alpha(x, x_1) \otimes \alpha(x_1, x_2) \otimes \ldots \otimes \alpha(x_{n-1}, x_n) \otimes \alpha(x_n, y)).$$

To prove the antisymmetry, take $x \neq y$. Then,

$$\begin{aligned}\rho(x, y) \otimes \rho(y, x) = \\ &= \bigvee(\alpha(x, x_1) \otimes \ldots \otimes \alpha(x_n, y)) \otimes \bigvee(\alpha(y, y_1) \otimes \ldots \otimes \alpha(y_m, x)) \\ &= \bigvee(\alpha(x, z_1) \otimes \ldots \otimes \alpha(z_p, y) \otimes \alpha(y, z_{p+1}) \otimes \ldots \otimes \alpha(z_q, x)) = 0,\end{aligned}$$

by distributivity of $\otimes$ with respect to join and by the definition of total intransitivity (supremum is taken over all corresponding sequences).

For transitivity, we have
$\rho(x, z) \otimes \rho(z, y) = \bigvee(\alpha(x, z_1) \otimes \ldots \otimes \alpha(z_p, z) \otimes \alpha(z, z_{p+1}) \otimes \ldots \otimes \alpha(z_q, y)) \leqslant \bigvee(\alpha(x, x_1) \otimes \ldots \otimes \alpha(x_n, y)) = \rho(x, y)$, using distributivity of multiplication with respect to join.

Next we should prove that the fuzzy covering relation induced by ρ is α. By the definition, $\alpha(x, x) = 0$. Let $x \neq y$. If $\rho(x, z) \otimes \rho(z, y) > 0$ for some $z \notin \{x, y\}$, then by $\rho(x, z) \otimes \rho(z, y) = \bigvee(\alpha(x, z_1) \otimes \ldots \otimes \alpha(z_p, z) \otimes \alpha(z, z_{p+1}) \otimes \ldots \otimes \alpha(z_q, y))$ it follows that there exists a chain of elements $x, a_1 \ldots, a_n, z, b_1 \ldots, b_m, y$ such that $\alpha(x, a_1) \otimes \ldots \otimes \alpha(a_n, z) \otimes \alpha(z, b_1) \otimes \ldots \otimes \alpha(b_m, y) > 0$. This chain may be as short as x, z, y, but not shorter. Therefore, by total fuzzy intransitivity, it follows that $\alpha(x, y) = 0$.

Further, $\rho(x, z) \otimes \rho(z, y) = 0$ for all $z \notin \{x, y\}$ means that the join $\bigvee(\alpha(x, z_1) \otimes \ldots \otimes \alpha(z_p, z) \otimes \alpha(z, z_{p+1}) \otimes \ldots \otimes \alpha(z_q, y))$ equals 0 whenever there is a corresponding chain (with more than 2 elements) from x to y. Therefore, $\rho(x, y) = \alpha(x, y)$, finally proving that fuzzy covering relation of ρ is α. □

A following observation is a simple consequence of the fact that a fuzzy totally intransitive relation is contained the corresponding fuzzy order.

Corollary 1. *A fuzzy totally intransitive relation is antisymmetric.*

As we mentioned in Introduction, in the fuzzy settings there is no one-to-one correspondence among fuzzy orderings and the corresponding fuzzy coverings. One totally fuzzy intransitive relation can correspond to many fuzzy orderings on the same set. Among these, the order obtained in the proof of Theorem 2 is minimal, as shown in the sequel.

Let X be a nonempty set and α a totally fuzzy intransitive relation on X. Denote by $\overline{\rho_\alpha}$ the fuzzy order defined in the proof of Theorem 2:

$$\overline{\rho_\alpha}(x,y) := \begin{cases} 1, & \text{if } x = y \\ \bigvee_{n\in\mathbb{N}, x_i\in X}(\alpha(x,x_1)\otimes\ldots\otimes\alpha(x_n,y)), & \text{if } x \neq y. \end{cases}$$

Corollary 2. *Let X be a nonempty set and α a totally fuzzy intransitive relation on X. Then the fuzzy relation $\overline{\rho_\alpha}$ is the smallest fuzzy ordering on X whose fuzzy covering is α.*

Proof. $\overline{\rho_\alpha}$ is a fuzzy ordering on X by Theorem 2. What we have to prove is that every fuzzy ordering whose fuzzy covering is α is greater than $\overline{\rho_\alpha}$, in the sense of fuzzy inclusion. Indeed, let θ be a fuzzy ordering on X, such that the fuzzy covering of θ is α. By Proposition 1, for any distinct $x, y \in X$, if $\alpha(x,y) \neq 0$ then $\alpha(x,y) = \theta(x,y)$. Hence, by the definition of $\overline{\rho_\alpha}$, for $x \neq y$ we have

$\overline{\rho_\alpha}(x,y) = \bigvee_{n\in\mathbb{N}, x_i\in X}(\alpha(x,x_1)\otimes\alpha(x_1,x_2)\otimes\ldots\otimes\alpha(x_{n-1},x_n)\otimes\alpha(x_n,y)) \leqslant$
$\bigvee_{n\in\mathbb{N}, x_i\in X}(\theta(x,x_1)\otimes\theta(x_1,x_2)\otimes\ldots\otimes\theta(x_{n-1},x_n)\otimes\theta(x_n,y)) \leqslant \theta(x,y).$

In the last line of the above formula, we use associativity of $\otimes$, its isotonicity with respect to $\leqslant$, and transitivity of θ; obviously, supremum is idempotent, hence we obtain the upper bound $\theta(x,y)$. Therefore, for all $x, y \in X$, we have $\overline{\rho_\alpha}(x,y) \leqslant \theta(x,y)$, and thus $\overline{\rho_\alpha} \subseteq \theta$. □

Observe that the relation $\overline{\rho_\alpha}$, defined by a precise formula, could be considered as a main representative of the fuzzy ordering whose covering is α.

4 Fuzzy Diagram of a Fuzzy Order

Following the known terminology, we consider a **fuzzy directed graph** to be an ordered pair (X,ρ), where X is a nonempty set and ρ is a fuzzy relation on X. Observe that the definition adopted here is a special case of a fuzzy graph being an ordered triple (X,μ,ρ), where in addition to above, μ is a fuzzy set on X (see e.g., [18]). Hence, we consider a crisp set of vertices and a fuzzy relation, which give rise to values or weights associated to edges. As usual, we associate a diagram to a fuzzy directed graph (X,ρ): elements of X, **vertices**, are represented by points in a plain, and for $x, y \in X$, $\rho(x,y)$ is a **weighted edge**, represented by an arrow from x to y, carrying the *non-zero value (weight)* $\rho(x,y)$. Since we do not deal here with undirected graphs, we use a term fuzzy graph, meaning that the graph is directed. Finally, we also say that the graph, or its diagram, **corresponds** to a fuzzy relation ρ. The notions of **path, loop,**

acyclic and **simple** fuzzy graph are defined analogously as in the crisp case; namely, the only difference with the crisp notions is that the arrows representing edges carry non-zero weights.

As in the crisp case, we say that the **transitive reduction** of a fuzzy relation ρ on a set X is a minimal fuzzy relation on X whose transitive closure coincides with the transitive closure of ρ. As usual, the **transitive closure** of a fuzzy relation ρ on X is the smallest transitive relation on X, containing ρ.

We say that a residuated lattice L is **zero divisor free** if there are no zero divisors under $\otimes$ in L, i.e., if $x \otimes y = 0$ implies $x = 0$ or $y = 0$.

Proposition 2. *Let L be a zero divisor free residuated lattice. Then the support of a fuzzy ordering ρ on X is a crisp ordering on the same set.*

Proof. The relation $\operatorname{supp}\rho$ is obviously reflexive. It is antisymmetric: if $x \neq y$ and $(x, y) \in \operatorname{supp}\rho$, then $\rho(x, y) > 0$, hence by antisymmetry of ρ and by the assumption on L (no zero divisors under $\otimes$), we have $\rho(y, x) = 0$, hence $(y, x) \notin \operatorname{supp}\rho$. Transitivity of $\operatorname{supp}\rho$ follows similarly. □

Let $\rho : X^2 \to L$ be a fuzzy ordering relation on a finite set X. Then the fuzzy graph corresponding to the fuzzy covering relation induced by ρ is said to be the **fuzzy diagram** of ρ; we also say that the **diagram corresponds to this fuzzy order**.

Proposition 3. *Let $\alpha : X^2 \to L$ be a fuzzy totally intransitive relation on the finite set X. Then the fuzzy graph on X corresponding to α is simple and it fulfills the following:*

If there is an edge from x to y, then for any other path $x, y_1, \ldots, y_n, y$ from x to y,

$$\alpha(x, y_1) \otimes \alpha(y_1, y_2) \otimes \ldots \otimes \alpha(y_n, y) = 0. \tag{1}$$

Proof. The graph is obviously simple, since a fuzzy totally intransitive relation is irreflexive ($\alpha(x, x) = 0$ for every x). Formula (1) is obtained as a logical contraposition of the definition of fuzzy total intransitivity. □

Apart from reflexivity, the proof of the following lemma is identical to the proof of Theorem 2.

Lemma 2. *If $\alpha : X^2 \to L$ be a fuzzy totally intransitive relation on the finite set X, then the least transitive relation $\widehat{\alpha}$ containing α is*

$$\widehat{\alpha}(x, y) := \begin{cases} 0, & \text{if } x = y \\ \bigvee_{n \in \mathbb{N}, x_i \in X} (\alpha(x, x_1) \otimes \ldots \otimes \alpha(x_n, y)), & \text{if } x \neq y. \end{cases}$$

Next is a *characterization theorem* for fuzzy diagrams of fuzzy totally intransitive relations.

Theorem 3. *The fuzzy graph corresponding to a fuzzy totally intransitive relation coincides with its transitive reduction. If, in addition, the co-domain lattice is zero divisor free, then this fuzzy graph is acyclic.*

Proof. Let $\alpha : X^2 \to L$ be a fuzzy totally intransitive relation. Suppose contrary, that there is a fuzzy relation $\beta : X^2 \to L$ such that $\beta < \alpha$ and that $\widehat{\beta} = \widehat{\alpha}$. Hence there are $x, y \in X$, such that $\beta(x, y) < \alpha(x, y)$. Being less than α, fuzzy relation β is also fuzzy totally intransitive (which can be proved directly), and we can apply the previous lemma. Since $\alpha \leqslant \widehat{\alpha}$ and

$$\widehat{\beta}(x, y) := \begin{cases} 0, & \text{if } x = y \\ \bigvee_{n\in\mathbb{N}, x_i \in X} (\beta(x, x_1) \otimes \ldots \otimes \beta(x_n, y)), & \text{if } x \neq y, \end{cases}$$

it follows that there are elements x_i such that

$$\beta(x, x_1) \otimes \ldots \otimes \beta(x_n, y) > 0.$$

Therefore each factor is greater than 0, and the same holds for analogue values of the relation α (since $\beta < \alpha$). Since $\alpha(x, y) > 0$, we get the contradiction by Proposition 3.

If L is zero divisor free, then by the formula (1) in which y is replaced by x, there are no cycles. □

Conversely, we have the following claim, which is straightforward consequence of the definition of a fuzzy totally intransitive relation.

Proposition 4. *Let X be a finite set and (X, α) a fuzzy simple graph fulfilling the property (1). Then α is a fuzzy totally intransitive relation on X.*

As an obvious consequence of the unique correspondence among fuzzy graphs and fuzzy totally intransitive relations on X, we obtain a link between fuzzy orders and the corresponding fuzzy graphs.

Corollary 3. *Let $\rho : X^2 \to L$ be a fuzzy ordering relation on a finite set X. Then there is a unique fuzzy diagram corresponding to ρ.*

The converse of Corollary 3 does not hold as shown in the following proposition. It is easily proved by Theorem 2 and Corollary 2.

Corollary 4. *Let α be a totally fuzzy intransitive relation on a finite set X. Then the fuzzy graph (X, α) corresponds to every fuzzy order whose fuzzy covering coincides with α.*

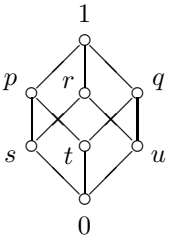
5 Examples

We provide two simple examples illustrating the foregoing properties of fuzzy orders, coverings and diagrams.

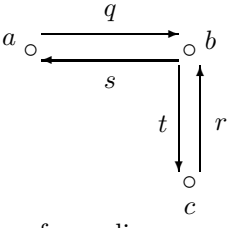
1. A fuzzy order ρ is given on a three-element set $\{a, b, c\}$, the lattice L is Boolean (Figure 1). On the same picture the fuzzy diagram is presented, and both, the fuzzy order ρ and the corresponding covering α_ρ are given by the tables.

ρ	a	b	c
a	1	q	u
b	s	1	r
c	0	t	1

α_ρ	a	b	c
a	0	q	0
b	s	0	r
c	0	t	0



L



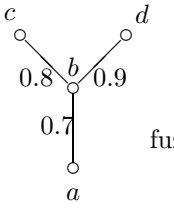
fuzzy diagram

Fig. 1.

2. Here we have the fuzzy order ρ on the set $\{a, b, c, d\}$, and the co-domain is the standard product algebra $([0, 1], \min, \max, \cdot, \rightarrow, 0, 1)$ (see [1]) which is zero division free. ρ is the minimum fuzzy order for which α_ρ is the fuzzy covering. The diagram is presented in Figure 2.

ρ	a	b	c	d
a	1	0.7	0.6	0.7
b	0	1	0.8	0.9
c	0	0	1	0
d	0	0	0	1

α_ρ	a	b	c	d
a	0	0.7	0	0
b	0	0	0.8	0.9
c	0	0	0	0
d	0	0	0	0



fuzzy diagram

Fig. 2.

6 Conclusion

The paper investigates connection of fuzzy orderings with Hasse-like diagrams and covering (totally intransitive) fuzzy relations, in the framework of residuated lattices. It turns out that this important field of fuzzy orderings is far from being simply analogue to the crisp one. Our next task would be to generalize this approach with respect to fuzzy equalities, in order to obtain fully fuzzified topic.

References

1. Bělohlávek, R.: Fuzzy relational systems. Kluwer Academic Publishers, Dordrecht (2002)
2. Blount, K., Tsinakis, C.: The structure of residuated lattices. Int. J. Algebr. Comput. 13, 437–461 (2003)
3. Bodenhofer, U.: Similarity-Based Generalization of Fuzzy Orderings preserving the classical axioms. Int. J. Uncertain. Fuzz. 8(3), 593–610 (2000)
4. Bodenhofer, U.: Representations and constructions of similarity-based fuzzy orderings. Fuzzy Set. Syst. 137, 113–136 (2003)
5. Bodenhofer, U., De Baets, B., Fodor, J.: A compendium of fuzzy weak orders: representations and characterizations. Fuzzy Set. Syst. 158, 811–829 (2007)
6. Davey, B.A., Pristley, H.A.: Introduction to lattices and order. Cambridge University Press, Cambridge (1992)
7. De Baets, B., De Meyer, H., Naessens, H.: A top-down algorithm for generating the Hasse tree of a fuzzy preorder closure. IEEE Trans. Fuzzy Systems 12, 838–848 (2004)
8. Demirci, M.: A Theory of vague lattices based on many-valued equivalence relations - I: General representation Results. Fuzzy Set. Syst. 151(3), 437–472 (2005)
9. Demirci, M.: A Theory of vague lattices based on many-valued equivalence relations - II: Complete lattices. Fuzzy Set. Syst. 151(3), 473–489 (2005)
10. Demirci, M., Eken, Z.: An Introduction to vague complemented ordered sets. Inform. Sci. 177(1), 150–160 (2007)
11. Diaz, S., De Baets, B., Montes, S.: Additive decomposition of fuzzy pre-orders. Fuzzy Set. Syst. 158, 830–842 (2007)
12. Dilworth, R.P., Ward, M.: Residuated lattices. Trans. Amer. Math. Soc. 45, 335–354 (1939)
13. Galatos, N., Jipsen, P., Kowalski, T., Ono, H.: Residuated lattices: An algebraic glimpse at substructural logics. Elsevier, Amsterdam (2007)
14. Goguen, J.A.: L-fuzzy Sets. J. Math. Anal. Appl. 18, 145–174 (1967)
15. Gottwald, S.: Fuzzy sets and fuzzy logic. Viewig, Wiesbaden (1993)
16. Harzheim, E.: Ordered sets. Springer, Heidelberg (2005)
17. Höhle, U., Blanchard, N.: Partial ordering in L-undeterminate sets. Inform. Sci. 35, 133–144 (1985)
18. Mordeson, J.N., Nair, P.S.: Fuzzy graphs and fuzzy hypergraphs. Physica-Verlag, Heidelberg (2000)
19. Ovchinnikov, S.V.: Similarity relations, fuzzy partitions, and fuzzy orderings. Fuzzy Set. Syst. 40(1), 107–126 (1991)
20. Šešelja, B.: L-fuzzy covering relation. Fuzzy Set. Syst. 158(22), 2456–2465 (2007)
21. Šešelja, B., Tepavčević, A.: Completion of ordered structures by cuts of fuzzy sets, An overview. Fuzzy Set. Syst. 136, 1–19 (2003)
22. Šešelja, B., Tepavčević, A.: Representing ordered structures by fuzzy sets, An overview. Fuzzy Set. Syst. 136, 21–39 (2003)
23. Šešelja, B.: Fuzzy covering relation and ordering - an abstract approach. In: Reusch, B. (ed.) Computational Intelligence, Theory and Applications. Advences in Soft Computing, pp. 295–300. Springer, Heidelberg (2006)
24. Skiena, S.: Implementing discrete mathematics: Combinatorics and graph theory with Mathematica. Addison-Wesley, Reading (1990)
25. Tepavčević, A., Trajkovski, G.: L-fuzzy lattices: an introduction. Fuzzy Set. Syst. 123, 209–216 (2001)
26. Zadeh, L.A.: Similarity relations and fuzzy orderings. Inform. Sci. 3, 177–200 (1971)

Fuzzy Relation Equations in Semilinear Spaces

Irina Perfilieva

Institute for Research and Applications of Fuzzy Modeling
University of Ostrava
30. dubna 22, 701 03 Ostrava 1, Czech Republic
Irina.Perfilieva@osu.cz

Abstract. We introduce a notion of an idempotent semilinear space and consider two systems of linear-like equations. These systems are equivalent to systems of fuzzy relation equations with sup-$*$ and inf-$\rightarrow$ compositions. We show that the theory of Galois connections can be successfully used in characterizing whether these systems are solvable and, if so, finding their solutions sets. Moreover, because the two types of systems of linear-like equations are dual according to this theory, it is sufficient to investigate only one system.

Keywords: Semilinear space, Residuated lattice, System of fuzzy relation equations, Fixed point.

1 Introduction

The aim of this paper is twofold: first, to aid in the formalization and unification of tools and methods used in the theory of fuzzy relation equations, and second, to propose a generalization of the theory of linear spaces. As is known from the extant literature, there are at least two types of systems of fuzzy relation equations that differ in types of composition [1,2,3,4]. However, results about the solvability and the structure of solution sets for both types of composition are, in some sense, dual. Additionally, there is a profound theory of linear spaces wherein the problem of determining the solvability of systems of linear equations is entirely solved. Thus, our motivation was to find a proper generalization of the theory of linear spaces that can also serve as a theoretical platform for the analysis of systems of fuzzy relation equations.

In this paper, we will show that the theory of Galois connections can be successfully used in characterizing the solvability and finding any solutions sets of systems of linear-like equations in semilinear spaces. If solvability is connected with a characterization of the vectors on the right-hand sides, then there exists a Galois connection between a set of admissible right-hand sides and a set of solutions. Moreover, on the basis of this theory, two types of systems of linear-like equations are dual, and thus it suffices to study only one of them.

2 Idempotent Semilinear Spaces

We recall that a linear (vector) space is a special case of a module over a ring, i.e., a linear space is a unitary module over a field [5]. In this paper, we will be dealing

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 545–552, 2010.

with a unitary semimodule over a commutative semiring [6,7], which will be called a *semilinear space*. Moreover, our semilinear space will be an idempotent structure with respect to its main operation.

Definition 1. *Let $\mathcal{R} = (R, +, \cdot, 0, 1)$ be a commutative semiring and $\mathcal{V} = (V, +, \bar{0})$ a commutative monoid. We say that $\mathcal{V}$ is a (left) semilinear space over $\mathcal{R}$ if an external (left) multiplication $\lambda : \bar{x} \mapsto \lambda\bar{x}$ is defined, where $\lambda \in R$ and $\bar{x} \in V$. Moreover, the following mutual properties are fulfilled for all $\bar{x}, \bar{y} \in V$ and $\lambda, \mu \in R$:*

(SLS1) $\lambda(\bar{x} + \bar{y}) = \lambda\bar{x} + \lambda\bar{y}$,
(SLS2) $(\lambda + \mu)\bar{x} = \lambda\bar{x} + \mu\bar{x}$,
(SLS3) $(\lambda \cdot \mu)\bar{x} = \lambda(\mu\bar{x})$,
(SLS4) $1\bar{x} = \bar{x}$,
(SLS5) $\lambda\bar{0} = \bar{0}$.

When $\mathcal{R}$ is clear from the context, we will shorten "left semilinear space over $\mathcal{R}$" to "semilinear space." Elements of a semilinear space will be distinguished by an overline.

Example 1. Let $\mathcal{R} = (R, +, \cdot, 0, 1)$ be a commutative semiring. Let R^n ($n \geq 1$) be the set of n-dimensional vectors whose components are elements of R, i.e. $R^n = \{\bar{x} = (x_1, \ldots, x_n) \mid x_1 \in R, \ldots, x_n \in R\}$. Let $\bar{0} = (0, \ldots, 0)$ and

$$\bar{x} + \bar{y} = (x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n).$$

Then, $\mathcal{R}^n = (R^n, +, \bar{0})$ is a commutative monoid. For any $\lambda \in R$, external multiplication $\lambda\bar{x}$ is defined by

$$\lambda\bar{x} = \lambda(x_1, \ldots, x_n) = (\lambda \cdot x_1, \ldots, \lambda \cdot x_n).$$

Then, $\mathcal{R}^n$ is a semilinear space over $\mathcal{R}$.

Semilinear space $\mathcal{R}^n$, $n \geq 1$, (see Example 1) will be called *vectorial semilinear space* over $\mathcal{R}$.

Definition 2. *Semilinear space $\mathcal{V}$ over $\mathcal{R}$ is called idempotent if the operations $+$ in both $\mathcal{V}$ and $\mathcal{R}$ are idempotent.*

Let $\mathcal{V} = (V, +, \bar{0})$ be an idempotent semilinear space. Then

$$\bar{x} \leq \bar{y} \iff \bar{x} + \bar{y} = \bar{y}, \tag{1}$$

is the *natural order* on $\mathcal{V}$. Therefore, $(V, \leq)$ is a bounded $\vee$-semilattice where $\bar{x} \vee \bar{y} = \bar{x} + \bar{y} = \sup\{\bar{x}, \bar{y}\}$, and $\bar{0}$ is a bottom element.

It may happen (see Example 2 below) that two idempotent semilinear spaces $\mathcal{V}_1 = (V, +_1, \bar{0}_1)$ and $\mathcal{V}_2 = (V, +_2, \bar{0}_2)$ with the same support V determine *dual* (or reverse) natural orders $\leq_1$ and $\leq_2$ on V, i.e.,

$$\bar{x} \leq_1 \bar{y} \iff \bar{y} \leq_2 \bar{x}.$$

In this case, $\leq_2$ is simply denoted $\geq_1$. With respect to $\leq_1$, $(\mathcal{V}_2, \geq_1)$ is a $\wedge$-semilattice with the top element $\bar{0}_2$ where $\bar{x} \wedge \bar{y} = \bar{x} +_2 \bar{y} = \inf\{\bar{x}, \bar{y}\}$. We will call $\mathcal{V}_1$ a $\vee$-*semilinear space*, and $\mathcal{V}_2$ a $\wedge$-*semilinear space*. Moreover, if $\mathcal{V}_1$ and $\mathcal{V}_2$ are idempotent semilinear spaces over the same semiring, then we will call them *dual*. It is easy to see that for dual semilinear spaces, the *Principle of Duality* for ordered sets holds true.

Example 2. Let $\mathcal{L} = (L, \vee, \wedge, *, \rightarrow, 0, 1)$ be an integral, residuated, commutative l-monoid and $\mathcal{L}_\vee = (L, \vee, *, 0, 1)$ a commutative $\vee$-semiring. L^n $(n \geq 1)$ is a set of n-dimensional vectors as in Example 1.

1. $\mathcal{L}_\vee^n = (L^n, \vee, \bar{0})$ is an idempotent commutative monoid, where $\bar{0} = (0, \ldots, 0) \in L^n$, and for any $\bar{x}, \bar{y} \in L^n$,

$$\bar{x} \vee \bar{y} = (x_1, \ldots, x_n) \vee (y_1, \ldots, y_n) = (x_1 \vee y_1, \ldots, x_n \vee y_n).$$

   The order on $\mathcal{L}_\vee^n$ is determined by $\vee$ so that $\bar{x} \leq \bar{y}$ if and only if $x_1 \leq y_1, \ldots, x_n \leq y_n$. For any $\lambda \in L$, external multiplication $\lambda\bar{x}$ is defined by

$$\lambda\bar{x} = \lambda(x_1, \ldots, x_n) = (\lambda * x_1, \ldots, \lambda * x_n).$$

   $\mathcal{L}_\vee^n$ with external multiplication $\lambda : \bar{x} \mapsto \lambda\bar{x}$ is an (idempotent) $\vee$-semilinear space over $\mathcal{L}_\vee$.
2. $\mathcal{L}_\wedge^n = (L^n, \wedge, \bar{1})$ is an idempotent commutative monoid where $\bar{1} = (1, \ldots, 1) \in L^n$, and for any $\bar{x}, \bar{y} \in L^n$,

$$(x_1, \ldots, x_n) \wedge (y_1, \ldots, y_n) = (x_1 \wedge y_1, \ldots, x_n \wedge y_n).$$

   The natural order on $\mathcal{L}_\wedge^n$ is determined by $\wedge$, and this ordering is dual to $\leq$, which was introduced on L^n in case 1 above. We will denote the natural order on $\mathcal{L}_\wedge^n$ by $\leq^d$, so that $\bar{x} \leq^d \bar{y}$ if and only if $\bar{x} \geq \bar{y}$. Alternatively, $\bar{x} \leq^d \bar{y}$ if and only if $x_1 \geq y_1, \ldots, x_n \geq y_n$. For any $\lambda \in L$, let us define external multiplication $\lambda \backslash \bar{x}$ by

$$\lambda \backslash (x_1, \ldots, x_n) = (\lambda \rightarrow x_1, \ldots, \lambda \rightarrow x_n).$$

   $\mathcal{L}_\wedge^n$ with the external multiplication $\lambda : \bar{x} \mapsto \lambda \backslash \bar{x}$ is an (idempotent) $\wedge$-semilinear space over $\mathcal{L}_\vee$.

$\vee$-semilinear space $\mathcal{L}_\vee^n$ and $\wedge$-semilinear space $\mathcal{L}_\wedge^n$ are duals.

2.1 Galois Connections in Semilinear Spaces

Let us recall that a Galois connection between two ordered sets $(A, \leq)$ and $(B, \leq)$ is a pair (h, g) of antitone mappings $h : A \rightarrow B$ and $g : B \rightarrow A$ such that $h \circ g \geq \mathrm{id}_A$ and $g \circ h \geq \mathrm{id}_B$ ($\circ$ denotes the composition of two mappings so that, e.g., for all $x \in A$, $h \circ g)(x) = g(h(x))$).

In this section, we will show that two dual idempotent semilinear spaces can be connected by Galois connections.

Theorem 1. *(i) Let $\mathcal{L}_\vee^n$ be a $\vee$-semilinear space and $\mathcal{L}_\wedge^n$ be a $\wedge$-semilinear space, both over $\mathcal{L}_\vee$. For each $\lambda \in L$, mappings $\bar{x} \mapsto \lambda\bar{x}$ and $\bar{y} \mapsto \lambda \backslash \bar{y}$ establish a Galois connection between $(\mathcal{L}_\vee^n, \leq)$ and $(\mathcal{L}_\wedge^n, \leq^d)$.*

(ii) Let $\mathcal{L}_\vee^m$, $m \geq 1$, be a $\vee$-semilinear space and $\mathcal{L}_\wedge^n$, $n \geq 1$, a $\wedge$-semilinear space, both over $\mathcal{L}_\vee$. Then for each $n \times m$ matrix $A \in L^{n \times m}$ with transpose A^*, mappings $h_A : L^m \rightarrow L^n$ and $g_{A^*} : L^n \rightarrow L^m$ given by

$$h_A(\bar{x})_i = a_{i1} * x_1 \vee \cdots \vee a_{im} * x_m, \quad i = 1, \ldots, n, \tag{2}$$

and

$$g_{A^*}(\bar{y})_j = (a_{1j} \to y_1) \wedge \cdots \wedge (a_{nj} \to y_n), \quad j = 1, \ldots, m, \tag{3}$$

establish a Galois connection between $(\mathcal{L}^n_\vee, \leq)$ *and* $(\mathcal{L}^n_\wedge, \leq^d)$.

Throughout this paper, let $\mathcal{L} = \langle L, \vee, \wedge, *, \to, 0, 1 \rangle$ be an integral, residuated, commutative l-monoid (*a residuated lattice*), U a non-empty set and L^U a set of L-valued functions on U. Fuzzy subsets of U are identified with L-valued functions on U (membership functions).

3 Systems of Fuzzy Relation Equations and Their Semilinear Analogs

Let U and V be two universes (not necessary different), $A_i \in L^U$, $B_i \in L^V$ arbitrarily chosen fuzzy subsets of respective universes, and $R \in L^{U \times V}$ a fuzzy subset of $U \times V$. This last item is called a fuzzy relation. Lattice operations $\vee$ and $\wedge$ are considered the union and intersection of fuzzy sets, respectively. Two other binary operations $*, \to$ of $\mathcal{L}$ are used for compositions—binary operations on $L^{U \times V}$. We will consider two of them: sup-$*$ -composition, usually denoted $\circ$, and inf-$\to$composition usually denoted $\triangleleft$. The first was introduced by L. Zadeh [8] and the second by W. Bandler and L. Kohout [9]. We will demonstrate definitions of both compositions on particular examples of set-relation compositions $A \circ R$ and $A \triangleleft R$, where $A \in L^U$ and $R \in L^{U \times V}$:

$$(A \circ R)(v) = \bigvee_{u \in U} (A(u) * R(u, v)),$$
$$(A \triangleleft R)(v) = \bigwedge_{u \in U} (A(u) \to R(u, v)).$$

Remark 1. Let us remark that both compositions can be considered as set-set compositions where R is assumed to be replaced by a fuzzy set. They are used in this reduced form later in instances of systems of fuzzy relation equations.

By a *system of fuzzy relation equations with sup-*-composition* (SFRE$*$), we mean the following system of equations

$$A_i \circ R = B_i, \text{ or } \bigvee_{u \in U} (A_i(u) * R(u, v)) = B_i(v), \quad 1 \leq i \leq n, \tag{4}$$

considered with respect to unknown fuzzy relation $R \in L^{U \times V}$. Its counterpart is a *system of fuzzy relation equations with* inf-$\to$*composition* (SFRE$\to$)

$$A_j \triangleleft R = D_j, \text{ or } \bigwedge_{u \in U} (A_j(u) \to R(u, v)) = D_j(v), \quad 1 \leq j \leq m, \tag{5}$$

also considered with respect to unknown $R \in L^{U \times V}$. System (4) and its potential solutions are well investigated in the literature (see e.g. [1,2,10,11,12,13,14,15]). On the other hand, investigations of the solvability of (5) are not so intensive (see [2,4,16]).

4 Systems of Equations in Semilinear Spaces $\mathcal{L}_{\vee}^{m}$ and $\mathcal{L}_{\wedge}^{n}$

4.1 System of Equations in Semilinear Space $\mathcal{L}_{\vee}^{m}$

Let $\mathcal{L}_{\vee}^{m}$, with $m \geq 1$, be a $\vee$-semilinear space over $\mathcal{L}_{\vee}$, and $\mathcal{L}_{\wedge}^{n}$, with $n \geq 1$, be a $\wedge$-semilinear space over $\mathcal{L}_{\vee}$. Let $n \times m$ matrix $A = (a_{ij})$, vector $\bar{b} = (b_1, \ldots, b_n) \in L^n$, and vector $\bar{d} = (b_1, \ldots, d_m) \in L^m$ have components from L. The systems of equations

$$\begin{aligned} a_{11} * x_1 \vee \cdots \vee a_{1m} * x_m &= b_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1} * x_1 \vee \cdots \vee a_{nm} * x_m &= b_n, \end{aligned} \tag{6}$$

and

$$\begin{aligned} (a_{11} \rightarrow y_1) \wedge \cdots \wedge (a_{n1} \rightarrow y_n) &= d_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ (a_{1m} \rightarrow y_1) \wedge \cdots \wedge (a_{nm} \rightarrow y_n) &= d_m, \end{aligned} \tag{7}$$

are considered with respect to unknown vectors $\bar{x} = (x_1 \ldots, x_m) \in L^m$ and $\bar{y} = (y_1 \ldots, y_n) \in L^n$. By (2) and (3), systems (6) and (7) can be respectively represented as follows:

$$h_A(\bar{x}) = \bar{b}.$$

and

$$g_{A^*}(\bar{y}) = \bar{d}.$$

It is easily seen that system (6) is an instance of (4) specified by $U = \{u_1, \ldots, u_m\}$, $V = \{v_1, \ldots, v_n\}$, $v \in V$ is fixed, $A_i(u_j) = a_{ij}$, $B_i(v) = b_i$, and $R(u_j, v) = x_j$. Similarly, system (7) is an instance of (5) specified by $U = \{u_1, \ldots, u_n\}$, $V = \{v_1, \ldots, v_m\}$, $v \in V$ is fixed, $A_i(u_j) = a_{ji}$, $D_i(v) = d_i$, and $R(u_j, v) = y_j$.

4.2 Solvability in Terms of Galois Connection

Below, we will give results regarding solvability and the solutions of systems (6) and (7) represented as:

$$\begin{aligned} A\bar{x} = \bar{b}, \quad (h_A(\bar{x}) = \bar{b}), \\ A^* \backslash \bar{y} = \bar{d}, \quad (g_{A^*}(\bar{y}) = \bar{d}), \end{aligned}$$

where mappings h_A, g_{A^*} establish a Galois connection between dually ordered spaces $(\mathcal{L}_{\vee}^{m}, \leq)$ and $(\mathcal{L}_{\wedge}^{n}, \leq^d)$. Therefore, any result about the solvability of one system has its dual counterpart, which can be obtained by

- replacing h_A by g_{A^*}, and vice versa,
- replacing $\leq$ by $\geq$, and vice versa,
- replacing $\vee$ by $\wedge$, and vice versa.

For the reader's convenience, we will formulate both dual results about the solvability and the solutions of systems (6) and (7).

Theorem 2. *Let A be a given matrix, and h_A and g_{A^*} establish a Galois connection between semilinear spaces $\mathcal{L}^m_\vee$ and $\mathcal{L}^n_\wedge$. Then,*

(i) System (6) is solvable if and only if $\bar{b}$ is a closed element of $\mathcal{L}^n_\wedge$ with respect to the closure operator $g_{A^*} \circ h_A$, or if and only if

$$\bar{b} = h_A(g_{A^*}(\bar{b})) = A(A^* \backslash \bar{b}). \tag{8}$$

(ii) System (7) is solvable if and only if $\bar{d}$ is a closed element of $\mathcal{L}^m_\vee$ with respect to the closure operator $h_A \circ g_{A^*}$, or if and only if

$$\bar{d} = g_{A^*}(h_A(\bar{d})) = A^* \backslash (A\bar{d}). \tag{9}$$

Remark 2. By (8), the right-hand side vector $\bar{b} \in L^n$ of a solvable system (6) is a *fixed point* of the closure operator $g_{A^*} \circ h_A$ determined by the matrix of coefficients A. Similarly by (9), the right-hand side vector $\bar{d} \in L^m$ of a solvable system (7) is a fixed point of the closure operator $h_A \circ g_{A^*}$.

Remark 3. By Theorem 1, $A(A^* \backslash \bar{y}) \leq \bar{y}$ so that the operator $g_{A^*} \circ h_A$ is a closure in $\mathcal{L}^n_\wedge$ ordered by the dual ordering $\leq^d$. In general, a closure operator in a dually ordered space is called an opening operator with respect to the reverse, i.e., genuine, ordering $\leq$. We will not, however, use this term.

Corollary 1. *Let the conditions of Theorem 2 be fulfilled. Then,*

(i) $\bar{b}$ is a fixed point of $g_{A^*} \circ h_A$ if and only if there exists $\bar{x} \in L^m$ such that $h_A(\bar{x}) = \bar{b}$, or $A\bar{x} = \bar{b}$.
(ii) $\bar{d}$ is a fixed point of $h_A \circ g_{A^*}$ if and only if there exists $\bar{y} \in L^n$ such that $g_{A^*}(\bar{y}) = \bar{d}$, or $A^* \backslash \bar{y} = \bar{d}$.

Corollary 2. *Let the conditions of Theorem 2 be fulfilled. Then,*

(i) for each $\bar{x} \in L^m$, $A(A^* \backslash A\bar{x}) = A\bar{x}$,
(ii) for each $\bar{y} \in L^n$, $A^* \backslash A(A^* \backslash \bar{y}) = A^* \backslash \bar{y}$.

Theorem 3. *Let A be a given matrix, $g_{A^*} \circ h_A$ a closure operator on $\mathcal{L}^n_\wedge$, $h_A \circ g_{A^*}$ a closure operator on $\mathcal{L}^m_\vee$. Then,*

(i) the set $cl^*_A(L^n)$ of fixed points of $g_{A^*} \circ h_A$ is a semilinear subspace of $\mathcal{L}^n_\vee$.
(ii) the set $cl_A(L^m)$ of fixed points of $h_A \circ g_{A^*}$ is a semilinear subspace of $\mathcal{L}^m_\wedge$.

Theorem 4. *Let systems (6) and (7) be specified by $n \times m$ matrix A and vectors $\bar{b} \in L^n$, $\bar{d} \in L^m$, respectively. Then,*

(i) if $\bar{b}$ is a fixed point of $g_{A^*} \circ h_A$, then $g_{A^*}(\bar{b}) = A^* \backslash \bar{b}$ is a solution of system (6),
(ii) if $\bar{d}$ is a fixed point of $h_A \circ g_{A^*}$, then $h_A(\bar{d}) = A\bar{d}$ is a solution of system (7).

Theorem 5. *Let systems (6) and (7) be specified by $n \times m$ matrix A and vectors $\bar{b} \in L^n$, $\bar{d} \in L^m$, respectively. Then,*

(i) h_A *restricted to the set of fixed points* $cl_A(L^m)$ *is a bijection between* $cl_A(L^m)$ *and* $cl_A^*(L^n)$,
(ii) g_{A^*} *restricted to the set of fixed points* $cl_A^*(L^n)$ *is a bijection between* $cl_A^*(L^n)$ *and* $cl_A(L^m)$,
(iii) *restriction* $g_{A^*}|_{cl_A^*(L^n)}$ *is inverse of the restriction* $h_A|_{cl_A(L^m)}$.

Let $\equiv_{h_A}$ be an equivalence relation on $\mathcal{L}_\vee^m$ such that

$$\bar{x}_1 \equiv_{h_A} \bar{x}_2 \iff A(\bar{x}_1) = A(\bar{x}_2).$$

Similarly, let $\equiv_{g_{A^*}}$ be an equivalence relation on $\mathcal{L}_\wedge^n$ such that

$$\bar{y}_1 \equiv_{g_{A^*}} \bar{y}_2 \iff A^* \backslash \bar{y}_1 = A^* \backslash \bar{y}_2.$$

Define $[\bar{x}]_{\equiv_{h_A}}$ as an equivalence class of $\bar{x}$ with respect to $\equiv_{h_A}$, and $[\bar{y}]_{\equiv_{g_{A^*}}}$ an equivalence class of $\bar{y}$ with respect to $\equiv_{g_{A^*}}$.

Lemma 1. *Let A be a* $n \times m$ *matrix. Then,*

(i) *For all* $\bar{x} \in L^m$, $[\bar{x}]_{\equiv_{h_A}} = [A^* \backslash A\bar{x}]_{\equiv_{h_A}}$, *where* $A^* \backslash A\bar{x} \in cl_A(L^m)$ *is a fixed point of* $h_A \circ g_{A^*}$.
(ii) *For all* $\bar{y} \in L^n$, $[\bar{y}]_{\equiv_{g_{A^*}}} = [A(A^* \backslash \bar{y})]_{\equiv_{g_{A^*}}}$, *where* $A(A^* \backslash \bar{y}) \in cl_A^*(L^n)$ *is a fixed point of* $g_{A^*} \circ h_A$.

Theorem 6. *Let systems (6) and (7) be specified by* $n \times m$ *matrix A and vectors* $\bar{b} \in L^n$, $\bar{d} \in L^m$, *respectively. Then,*

(i) $[A^* \backslash \bar{b}]_{\equiv_{h_A}}$ *is a set of solutions of (6) with the righthand side given by* $\bar{b}$, *i.e.,*

$$\bar{x} \in [A^* \backslash \bar{b}]_{\equiv_{h_A}} \Leftrightarrow A\bar{x} = \bar{b}.$$

Moreover, $A^* \backslash \bar{b} \in cl_A(L^m)$, *and* $A^* \backslash \bar{b}$ *is the greatest element in* $[A^* \backslash \bar{b}]$.
(ii) $[A\bar{d}]_{\equiv_{g_{A^*}}}$ *is a set of solutions of (7) with the righthand side given by* $\bar{d}$, *i.e.*

$$\bar{y} \in [A\bar{d}]_{\equiv_{g_{A^*}}} \Leftrightarrow A^* \backslash \bar{y} = \bar{d}.$$

Moreover, $A\bar{d} \in cl_A^*(L^n)$, *and* $A\bar{d}$ *is the least element in* $[A\bar{d}]$.

5 Conclusion

In this paper, we showed that the theory of Galois connections can be successfully used in characterizing the solvability and solutions sets of systems of linear-like equations in semilinear spaces. The solvability is characterized by the relationship between vectors of right-hand sides and solutions. Moreover, because two types of systems of linear-like equations are shown to be dual on the basis of this theory, only one of them was investigated.

Acknowledgement. The research has been supported by the grant IAA108270902 of GA AVČR, and by the project MSM 6198898701 of the MŠMT ČR.

References

1. Di Nola, A., Sessa, S., Pedrycz, W., Sanchez, E.: Fuzzy Relation Equations and Their Applications to Knowledge Engineering. Kluwer, Dordrecht (1989)
2. De Baets, B.: Analytical solution methods for fuzzy relation equations. In: Dubois, D., Prade, H. (eds.) The Handbooks of Fuzzy Sets Series, vol. 1, pp. 291–340. Kluwer, Dordrecht (2000)
3. Nosková, L., Perfilieva, I.: System of fuzzy relation equations with sup $-*$ composition in semi-linear spaces: minimal solutions. In: Proc. FUZZ-IEEE Conf. on Fuzzy Systems, London, pp. 1520–1525 (2007)
4. Perfilieva, I., Nosková, L.: System of fuzzy relation equations with inf $\rightarrow$ composition: complete set of solutions. Fuzzy Sets and Systems 159, 2256–2271 (2008)
5. Raikov, D.A.: Vector spaces. FizMatLit, Moscow (1962)
6. Kurosh, A.G.: Lectures on General Algebra. FizMatLit, Moscow (1962)
7. Golan, J.S.: Semirings and their Applications. Kluwer Academic Publishers, Dordrecht (1999)
8. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning I, II, III. Information Sciences 8-9, 199–257, 301–357, 43–80 (1975)
9. Bandler, W., Kohout, L.: Semantics of implication operators and fuzzy relational products. Int. J. Man-Machine Studies 12, 89–116 (1980)
10. Gottwald, S.: On the existence of solutions of systems of fuzzy equations. Fuzzy Sets and Systems 12, 301–302 (1984)
11. Klawonn, F.: Fuzzy points, fuzzy relations and fuzzy functions. In: Novák, V., Perfilieva, I. (eds.) Discovering the World with Fuzzy Logic, pp. 431–453. Springer, Berlin (2000)
12. Perfilieva, I., Tonis, A.: Compatibility of systems of fuzzy relation equations. Int. J. of General Systems 29, 511–528 (2000)
13. Perfilieva, I., Gottwald, S.: Solvability and approximate solvability of fuzzy relation equations. Int. J. of General Systems 32, 361–372 (2003)
14. Sanchez, E.: Resolution of composite fuzzy relation equations. Information and Control 30, 38–48 (1976)
15. Wang, X.: Infinite fuzzy relational equations on a complete Brouwerian lattice. Fuzzy Sets and Systems 138, 657–666 (2003)
16. Štěpnička, M., Jayaram, B.: On the suitability of the bandler-kohout subproduct as an inference mechanism. IEEE Trans. on Fuzzy Syst. (2010) (to appear)

Adaptive Rule Based-Reasoning by Qualitative Analysis

Marius Mircea Balas and Valentina Emilia Balas

"Aurel Vlaicu" University of Arad, 75 Bulevardul Revolutiei,
Arad, Romania
balas@inext.ro

Abstract. In some previous papers we presented a fuzzy-interpolative self-adaptive control architecture that is able to identify the current operating regime of the system by means of a qualitative analysis of the phase trajectory of the error. Here we are extending this technique in order to build a self-adjusting PID controller that can automatically set the input scaling factors. This way we can obtain simple interpolative controllers embedding conventional knowledge on the PID controller adjusting, adapted from the Ziegler-Nichols method.

Keywords: fuzzy-interpolative controller, knowledge embedding by control rules, phase trajectory of the error, iterative feedback tuning.

1 A Brief Introduction of the Fuzzy-Interpolative Methodology

Humans are reasoning in a *qualitative* and *symbolic* way. When we want to extend this kind of reasoning over the artificial intelligent systems we have to deal with the fact that digital computers are reasoning in a totally opposite manner, based on quantitative analysis and numeric algorithms. This is an opposition between the flexible and comprehensive heuristic solutions and the precise and specialized numeric algorithms. Thanks to the fuzzy sets theory and to the expert system technique, linguistic represented knowledge can now be embedded into digital systems. In other words, computers begin to operate with words and to accommodate with *world knowledge* WK [1]. We ask the question if low level computing devices, say microcontrollers, DSPs, etc. can benefit of WK, though sophisticated AI software, such as Matlab, are encountering difficulties?

Our answer to this question is the *Fuzzy-Interpolative Methodology* FIM [2] that is used for designing *the fuzzy-interpolative controllers* FIC. A FIC is a fuzzy controller, defined by a McVicar-Whelan table (a decision table), that can be automatically equaled with a corresponding look-up table with linear interpolations [3].

The main FICs' advantage is their extremely convenient implementation by means of interpolative networks, in any possible software technology. This is due to the fact that the interpolative networks can be directly associated to addressable memories, which is substantially helping the digital computation. In high level programming languages the look-up tables bring effectiveness, resources saving and quick developments. Digital hardware circuits (microcontrollers, DSPs) can also implement FICs thanks to their memory based architecture. Even analog circuitry such as the translinear analog CMOS can support FICs [4].

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 553–560, 2010.

FIM is taking advantage of both linguistic and interpolative nature of the FICs, merging their advantages: a) the use of linguistically represented expert knowledge in the conception and the development stages and b) the simplicity of the interpolative networks for the implementations. In close loop control FICs are perfectly matching a fundamental time analysis tool: *the phase trajectory of the error* PTE and their specific analysis method, the qualitative analysis.

This approach was inspired by the fundamental papers of William Clocksin and A.J. Morgan [5] and of Laurent Foulloy [6].

2 The Fuzzy Self-Adaptive Interpolative Controllers

The first controllers developed according to FIM are forming the family of the *fuzzy self-adaptive interpolative controllers* FSAIC [3]. There are three such control configurations, enclosing different amounts of symbolic knowledge.

The first level of symbolism is materialized by *the plane surface adaptive interpolative controller* PSAIC. During transient regimes PSAIC is PD (a 2D look-up-table). Its control surface is almost plane, in order to avoid the distortion of the phase trajectory of the error. During the steady regime an integrative effect is gradually introduced, the structure becoming a PID one (a 3D look-up-table). The inputs are the control error ε, its derivate ε' and its integrative $\int\varepsilon$. The different PD tables that are creating the $\int\varepsilon$ dimension differ only at their central rule, that is activated when ε=*zero* and ε'=*zero* (the steady regime). Thus the integrative effect is gradually activated only when steady regimes occur.

The PSAIC linguistic knowledge may be stated as follows: *the PD control is used for transient regimes while the integrative effect initiates progressively only for steady regimes*.

The next level of symbolism is belonging to proper FSAICs, which have a superior adaptive feature by using a specific fuzzy-interpolative adaptive corrector. The corrector, a Proportional Derivative FIC, is modulating the PSAIC output with a multiplicative factor *Gain*, which is resulting out of an on-line quantitative analysis of PTE. This analysis is performed with the help of the *Linguistic Phase Trajectory* LPT, the succession of the activated adaptive rules. The rules that are surely pointing the operating regimes, the so called *regime's signatures*, are identifying the respective regime simply by their activation.

We have considered four significant operating regimes: *transitory*, *steady*, *oscillating* and *unstable* [7]. Their signatures in terms of PTE are shown in Fig. 1. Each time the adaptive corrector recognize a change of the operating regime it will adjust *Gain*, according to our generic knowledge on the PID controllers [8].

For instance, the steady regime is identified by the activation of the central R13: "ε=*zero* and ε'=*zero*" (see Fig. 2) and has as result an increased *Gain*. This will increase the precision and will produce faster reactions of the control systems against perturbations.

The rules that are characterizing the oscillatory regime, R12 and R14: "ε=*zero* and ε'≠*zero*" or R8 and R18: "ε≠zero and ε'=*zero*" (see Fig. 2) must decrease *Gain*, as well as the rules that are typical for the unstable regime, R5 and R21: "ε=*big* and ε'=*big* and sign(ε x ε')=*negative*". The theory standing behind these actions is the Nyquist stability criterion.

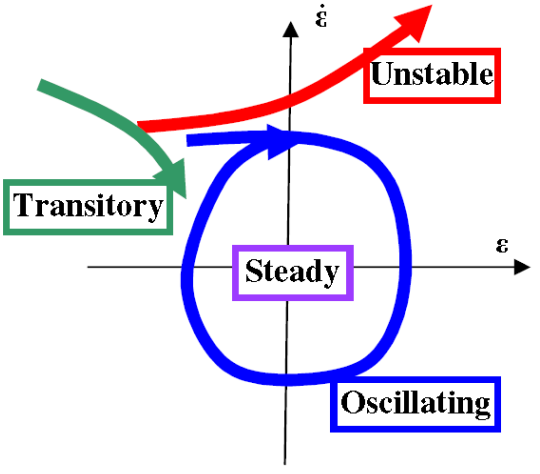


Fig. 1. PTE signatures of the main regimes

Such way, the general knowledge of the system theory helps us to define four rule clusters: *Steady* (R13), *Oscillating* (R8, R12, R14 and R18), *Unstable* (R5 and R21) and *Transitory* (the other remaining rules).

ε̇ \ ε	E_1	E_2	E_3	E_4	E_5
DE_1	R1	R6	R11	R16	R21
DE_2	R2	R7	R12	R17	R22
DE_3	R3	R8	R13	R18	R23
DE_4	R4	R9	R14	R19	R24
DE_5	R5	R10	R15	R20	R25

Fig. 2. A 5 x 5 McVicar-Whelan table with an associated linguistic phase trajectory

The previously presented principle can be further refined, with different goals. For example, the death time plants demand specific correctors, as shown in ref. [9]. This highest symbolic level is achieved by FFSAIC (*the fusioned fuzzy self-adaptive interpolative controller*) which may include different, even contradictory control strategies, into the same adaptive corrector [3].

Another version of our approach is the fuzzy-interpolative self-adaptive PID controller [12].

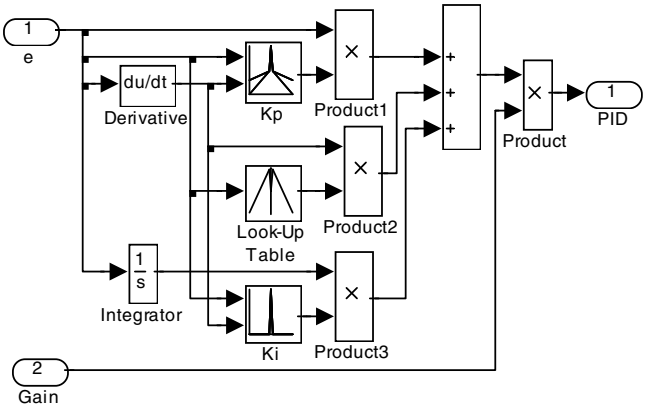


Fig. 3. The fuzzy-interpolative self-adaptive PID controller

As shown in Fig. 3, this controller has a PD fuzzy-interpolative adaptive corrector for each of the PID tuning parameters. They identify the operating regimes by their LPT signatures and, according to the activated rules, they modulate the values of the corresponding tuning parameter. The knowledge that supports the tuning is a heuristic PID tuning methodology, inspired by the Ziegler-Nichols method. The stability issues are considered in the light of the Nyquist criterion.

We illustrate this controller by the adaptive rules of the proportional effect K_P:

IF ε = *zero* and ε'= *zero* THEN K_P = *big*
IF ε = *big* or ε'= *big* THEN K_P= *small*
IF ε = *big* and ε'= *big* AND sign(ε x ε')=*negative* THEN K_P = *very small*
IF ε = *big* and ε'= *zero* OR ε = *zero* and ε'= *big* THEN K_P = *very small*
ELSE K_P = *medium*

The explanation of the rules is the following:

- during the steady regime K_P is big, for a good rejection of the disturbances;
- when the system is far from the steady regime K_P must decrease in order to avoid oscillations or overshoots; the error is anyway big so there is no need to amplify it excessively;
- when we have oscillatory or unstable regime K_P must be very small;

The following numerical implementation illustrates the procedure. Assume E as the limit of the steady regime. If *very small* = 0.2, *small* = 0.45, *medium* = 0.6 and *big* = 1, then the implementation look-up table is the following:

row (ε): [-1 –E 0 E 1]; (1)
column (ε'): [-1 –E' 0 E' 1];
K_P: [0.2 0.45 0.2 0.45 0.45; 0.6 0.6 0.6 0.6 0.6; 1 1 1 1 1; 0.6 0.6 0.6 0.6 0.6; 0.45 0.45 0.2 0.45 0.2].

The other parameters are designed in the same manner.

The primary reason of this adaptations by qualitative analysis is to extend as much as possible the flinging of our controllers, without a model of the controlled plant, on behalf of our generic knowledge of the automate control science. The ultimate objective of this effort is the well-known universal controller imagined by J.J. Buckley. FICs are giving us reasons to hope that comprehensive general purpose controllers may be achieved even at low costs, using low level digital controllers [10, 11].

3 The Self-Adaptive Scaling

The only important mechanism of a self-adaptive controller that was not described so far in our FIM dedicated works is the self-tuning of the scaling factors. The scaling is necessary since FICs are normalized, in order to match the input and the output values of the controller to the feedback transducer respectively to the actuator.

There are several options for a self-scaling architecture. A first option would be to directly address each of the controller's terminals: the three inputs *Proportional*, *Integrative* and *Derivative* and the *Output*, corresponding to the non-interacting version of the PID controller. However, because of the complex nonlinear interdependences between these variables, for the time being we could not reach an acceptable solution. That is why we preferred an architecture corresponding to the interacting PID version, where the main role will be played by the input scaling of the control error. The integrative and derivative effects will be tuned in a Ziegler-Nichols inspired manner, having in mind the PSAIC way of introducing the integrative.

We will consider the Fig. 4 test control system composed by a PID and the four scaling devices, each one controlled by its own corrector.

The simple fuzzy-interpolative correction mechanism previously described needs now to be upgraded because a fast and continuous shifting of the scaling factors is not recommended. The following architecture is a first step in this issue.

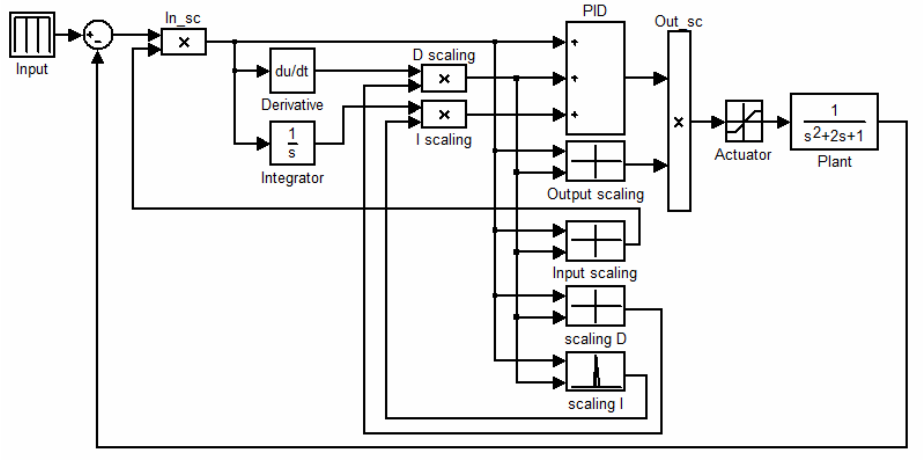


Fig. 4. The test control system

If the plant or the actuators are not changing the output scaling factor must remain fixe, in order to avoid interferences with the input scaling factors' variations. The correlation between the actuator's highest value *Act* and the output scaling factor *out_sc* is immediate: *out_sc* = *Act*, in order to avoid the actuator's windup. In a real application this can be done if the actuator is able to intercommunicate with the controller.

The integrative and the derivative components, as well as the proportional factor are usually tuned according to Ziegler-Nichols or to other similar method, as Tyreus-Luyblen for instance.

We will consider the interacting PID version:

$$c_{PID}(t) = K_P \cdot \left(e(t) + \frac{1}{T_I} \cdot \int_0^t e(\tau) \cdot d\tau + T_D \cdot \frac{de(t)}{dt} \right) \tag{2}$$

The manual tuning methods are first looking for K_C, the value of K_P that is beginning to produce oscillations with I and D effects turned off. This first off-line test must also identify T_C the period of the oscillations.

In the case of the PID control, The Ziegler-Nichols tuning is proposing the following values: $K_P = K_C / 1.7$, $T_I = T_C / 2$ and $T_D = T_C / 8$. The Tyreus-Luyblen version is $K_P = K_C / 2.2$, $T_I = 2.2 \cdot T_C$ and $T_D = T_C / 6.3$. Presenting a lower integrative effect, this tuning is meant to be more robust that the first one. We prefer the Ziegler-Nichols tuning because the PSAIC way of introducing the integrative effect is significantly increasing the control's robustness.

Since we want to develop an on-line automate tuning, we have to modify this procedure, in order to obtain an algorithm that is implementable with usual electronic circuitry. The first objectives are the identification of the oscillations and the measurement of their period T_C. These operations can be done with the help of the PTE, using the signature of the oscillating regimes, the succession of Fig. 2 rules R8 R12 R18 R14. The inversed time delay between the activation of the same rule is T_C. However, this on-line measurement is not performed in the same conditions as the conventional tuning, so we will accept the issued value only as a first guess.

Once fixed $T_I = T_C / 2$ and $T_D = T_C / 8$, the most difficult part remains the tuning of the input scaling factor. The main difficulty appears because the oscillations that are guiding the manual algorithm need special conditions to appear. Even if we identify oscillations on-line, we cannot be sure about their causes. That is why we have to try again a first guess, which is the half of K_C, where K_C is the proportional gain in the moment when we detect the oscillations.

Thus the first steps of our algorithm will be the followings:

1)The detection of an oscillation by a R8-R12-R18-R14 cycle.
2)The first guess of K_C, T_I and T_D.
3)The algorithm will continue by a tenure self-tuning stage that will end only when significant oscillations will occur, and a new tuning will become necessary.

A lot of self-tuning algorithms were already tested and applied so far [13]. In our case we do not know the mathematical model of the plant and the gradient techniques are not feasible. In the same time we want a minimalist configuration, easy to implement on cheap devices, which does not need open loop tests.

That is why we will use a fuzzy-interpolative version of the iterative feedback tuning [14]. Our approach follows a particular constraint: the algorithm must be representable in VHDL or in other hardware description language. Such way the FPGA or the ASIC applications will be encouraged.

4 The Fuzzy-Interpolative Iterative Feedback Tuning

The fuzzy-interpolative iterative feedback tuning is simply multiplying or dividing K_P by an iteration rate R, each time the Fig. 4 corrector *Input scaling* is identifying the signatures of the following situations: K_P *too small* and K_P *too big*. These signatures are defined for the Fig. 2 minimal rule base.

a) The K_P *too small* signature:
When K_P is too small, the control system is not able to force the plant to approach the small control errors zone. A typical step response of this kind would stabilize PTE at rules R3 or R23. The same case is appearing when the system is hanging too long, say longer that T_C, at rules R8 or R18. This signature can be linguistically described:

IF *R3 or R23* OR *R8 or R18 longer that* T_C THEN K_P *is too small*

This operation needs basic electronic circuits: comparators, a timer and logic gates. The result of K_P *is too small* is the multiplying of K_P by R.

b) The K_P *too big* signature:
When K_P is too big the control system has oscillatory tendencies during transient regimes and also chattering during the steady ones. The signature of the oscillations is already known: the succession R8-R12-R18-R14. This operation can be performed with a dedicated digital decoder.

The identification of the chattering can rely on the membership value of R13, denoted μ_{13}. It is now to underline that when we were talking about an activated rule in the definition of PTE we were meaning in fact the rule with the highest membership value, but this value is not necessarily 1. Several rules can be activated in the same time with different or equal membership values. As a result of this behavior, when the control system is firmly positioned on R13, $\mu_{13} = 1$, but when we have chattering, although μ_{13} is the highest comparing to other rules, $\mu_{13} < 1$.

The result of K_P *is too big* is the division of K_P by R.

It is to note that this paper is only a first step in the study of the fuzzy-interpolative iterative feedback tuned PID controllers. Further researches on the optimization of this method and a portfolio of applications are necessary. Some pending problems are:

- Which is the optimal iteration rate? For the time being we are using R = 2, proposed by the majority of manual tuning methods, but the choice of this value is heuristic;
- Which is the precise significance of all the rules of Fig. 2 rule base? Only statistic data issued from different applications could confirm our initial assumptions;
- Can we further detail the definition of the operating regimes signatures?

The considerations presented so far are identifying a research line into the adaptive control field. The previously described mechanisms were tested by simulations and our future work aims to integrate them into a coherent self-tuning PID structure.

5 Conclusions

The on-line qualitative analysis of the phase trajectory of the error, achieved by rule based reasoning and implemented by fuzzy-interpolative correctors, is able to support the self-adaptive control. A new application of this technique is presented: a self-tuning PID algorithm that is representable in hardware description languages.

This research is only at an early stage but the preliminary results are ensuring us to be on a good direction towards a cheap yet powerful general purpose self-tuning PID controller.

References

1. Zadeh, L.A.: From Search Engines to Question-Answering Systems - The Problems of World Knowledge, Relevance, Deduction and Precisiation. In: Invited Conference, IEEE-SOFA 2005, Arad, Romania (August 2005)
2. Balas, M.M.: The Fuzzy-Interpolative Methodology. In: Soft Computing Based Modeling in Intelligent Systems, pp. 145–167. Springer, Heidelberg (2009)
3. Balas, M.M., Balas, V.E.: The Family of Self Adaptive Interpolative Controllers. In: Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2004, Perugia, July 2004, pp. 2119–2124 (2004)
4. Balas, M.M., Balas, V.E.: Another Possible Way towards the Intelligent Nano-Systems: the Fuzzy-Interpolative. In: Proc. of 11th IPMU International Conference, Paris, July 2006, pp. 2135–2141 (2006)
5. Clocksin, W.F., Morgan, A.J.: Qualitative control. In: Proceedings of the European Conf. on Artificial Intelligence, Brighton (1986)
6. Foulloy, L.: Qualitative Control and Fuzzy Control: Towards a Writing Methodology. AICOM 6(3/4), 147–154 (1993)
7. Balas, V.E., Balas, M.M.: Observing Control Systems by Phase Trajectory of the Error. WSEAS Transactions on Systems 5(7), 1717–1722 (2006)
8. Van Doren, V.J.: Loop Tuning Fundamentals. Control Engineering (Red Business Information), July 1 (2003), http://www.controleng.com/article/CA307745.html
9. Balas, V.E., Balas, M.M.: Switching Adaptive Correctors for Fuzzy Fusion Self-Adapted Interpolative Controllers. In: Proc. of 3rd Int. Symposium on Computational Intelligence and Intelligent Informatics, ISCIII 2007, Agadir, Morocco, March 2007, pp. 67–72 (2007)
10. Balas, M.M., Balas, V.E.: World Knowledge for Applications by Fuzzy-Interpolative Systems. International Journal of Computers, Communications and Control III, 28–32 (2008)
11. Balas, M.M., Balas, V.E.: World Knowledge for Control Applications. In: Proc. of 11th International Conference on Intelligent Engineering Systems, INES 2007, Budapest, June 29 - July 1, pp. 225–228 (2007)
12. Balas, M.M.: The Fuzzy-Interpolative Adaptive Controllers. In: Proc. of the 10th WSEAS International Conference on Fuzzy Systems, Prague, March 2009, pp. 18–21 (2009)
13. Zaknich, A.: Principles of adaptive filters and self-learning systems. Springer, Heidelberg (2005)
14. Hjalmarsson, H., Gevers, M., Gunnarsson, S., Lequin, O.: Iterative feedback tuning: theory and applications. IEEE Control Systems Magazine 18(4), 26–41 (1998)

Fuzzy Regions: Adding Subregions and the Impact on Surface and Distance Calculation

Jörg Verstraete

Instytut Badań Systemowych, Polska Akademia Nauk (Systems Research Institute, Polish Academy of Sciences), Ul. Newelska 6, 01-447 Warszawa, Poland
DDCM, Dept. Telecommunications and Information Processing, Ghent University, Sint Pietersnieuwstraat 41, 9000 Ghent, Belgium
jorg.verstraete@{ibspan.waw.pl,telin.ugent.be}
http://www.ibspan.waw.pl, http://telin.ugent.be/ddcm

Abstract. In the concept of fuzzy regions we introduced before, a region was considered to be a fuzzy set of points, each having its own membership grade. While this allows the modelling of regions in which points only partly belong to the region, it has the downside that all the points are considered independently, which is too loose a restriction for some situations. The model is not able to support the fact that some points may be linked together. In this contribution, we propose an extension to the model, so that points can be made related to one another. It will permit the user to, for instance, specify points or even (sub)regions within the fuzzy region that are linked together: they all belong to the region to the same extent at the same time. By letting the user specify such subregions, the accuracy of the model can be increased: the model can match the real situation better; while at the same time decreasing the fuzziness: if points are known to be related, there is no need to consider them independently. As an example, the use of such a fuzzy region to represent a lake with a variable water level can be considered: as the water level rises, a set of points will become flooded; it is interesting to represent this set of points as a subset of the region, as these points are somewhat related (the same can be done for different water levels). The impact of this extension to the model on both surface area calculation an distance measurement are considered, and new appropriate definitions are introduced.

1 Introduction

The concept of the fuzzy regions was sparked by a lack of models capable of handling imperfect information, built on a solid, theoretical foundation, applicable for a wide range of representations and open to an implementation. In the presented concept, a fuzzy region was defined as a fuzzy set of points over the two dimensional domain. Each point therefore is given a membership grade indicating the extent to which it belongs to the region (i.e. a veristic interpretation [1]). While this broad definition is useful for many basic applications and allows for unlimited distributions of membership grades over the region, it treats all the

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 561–570, 2010.

elements independently of one another. In some situations however, it is possible a user has more information about a relation between elements of the region. Consider the example of a lake, where there is a variation possible in the level of water. At a random point in time we may not know the water level, but we do know that if a particular point would be under the water, so would other points (i.e. those at the same altitude just around the lake). If the likeliness of that water level is 0.4, then it makes sense to assign all these points that membership grade; however, as these points would be be at the same altitude, it is more natural not to consider them independently but rather as a group the points.

In this paper, a brief overview of the current model for fuzzy regions is given in section 2. Both the concept and the operations that are will be extended are considered here. Section 3 then deals with the proposed extension to fuzzy regions and the impact on the considered operations. The results are summarized in the conclusion.

2 Current Fuzzy Regions

2.1 Concept

A crisp region is often defined by means of its outline: a polygon or other closed line is defined as the boundary of the region, and the region is said to consist of the points inside this outline ([3]). To define fuzzy regions, a different point of view was adopted: rather than consider a crisp region to be defined by means of an outline, it was considered to be be a set of points. From this point of view, it is a small step to augment the definition to a fuzzy set ([8], [9]) of points. In [5], the fuzzy region was defined over $\mathbb{R}^2$, thus with each element (point) a membership grade was associated.

Definition 1 (Fuzzy region)

$$\tilde{R} = \{(p, \mu_{\tilde{R}}(p)) | p \in \mathbb{R}^2\} \tag{1}$$

A fuzzy region essentially is a fuzzy set defined over a two dimensional domain. As a result, the traditional fuzzy operations for intersection and union (by means of t-norms and t-conorms) are immiately applicable. Functionaly has been added to deal with spatial aspects, some examples include the distance between regions and the (fuzzy) surface area of a region. To determine topological relations however, appropriate definitions for the boundary, interior and exterior had to be derived from the initial given fuzzy set.

2.2 Operations

Surface area: cardinality. For the surface area of fuzzy regions, two definitions have been introduced ([6]). The first is an application of fuzzy cardinality, where each point is said to contribute its own membership grade.

$$\tilde{S}^c(\tilde{A}) = \int_{p(x,y) \in U} \mu_{\tilde{A}}(p(x,y)) d(x,y) \tag{2}$$

Where U is the considered universe (usually $\mathbb{R}^2$).

This definition has the advantage of yielding a crisp number that takes the fuzziness into account. The crisp number can easily be used in systems that cannot deal with fuzzy data.

Surface area: fuzzy number. The second definition for surface area yields a fuzzy number, such that all possible surface areas are possible values. This approach has the advantage of yielding a result that holds more information than the cardinality approach above, but it requires working with fuzzy numbers. For each possible surface area, the membership grade matches the possibility of the region being a region of this size. This is achieved by considering the surface area of all α-levels.

$$\tilde{S}^f(\tilde{A}) = \{(x, \mu_{\tilde{S}^f(\tilde{A})}(x)), x \in U\} \tag{3}$$

where U is the considered universe, most likely $\mathbb{R}^2$. For fuzzy regions defined as fuzzy sets over $\mathbb{R}^2$, the membership function was defined as:

$$\mu_{\tilde{S}^f(\tilde{A})}(x) : \mathbb{R} \to [0,1]$$
$$x \mapsto \begin{cases} 1 & \text{if } x = S(\tilde{A}_1) \\ \sup\{\alpha \mid S(\tilde{A}_{\overline{\alpha}}) \leq x \leq S(\tilde{A}_\alpha)\} & \text{if } S(\tilde{A}_{\overline{\alpha}}) \leq x \leq S(\tilde{A}_\alpha) \\ 0 & \text{elsewhere} \end{cases}$$

Distance. There are various options to define the distance between crisp regions; the most common one is the euclidean distance which represents the shortest distance between both regions. For fuzzy regions, the fuzzy distance $\tilde{d}(\tilde{R}_1, \tilde{R}_2)$ is a fuzzy number representing the possible distances between the α-cuts of R_1 and R_2.

$$\tilde{d}(\tilde{R}_1, \tilde{R}_2) = \{(x, \mu_{\tilde{d}(\tilde{R}_1, \tilde{R}_2)}(x)) \mid x \in \mathbb{R}\} \tag{4}$$

where

$$\mu_{\tilde{d}(\tilde{R}_1, \tilde{R}_2)} : \mathbb{R} \to [0,1]$$
$$x \mapsto \sup\{\alpha \mid d(\tilde{R}_{1\alpha}, \tilde{R}_{2\alpha}) \leq x \leq d(\tilde{R}_{1\overline{\alpha}}, \tilde{R}_{2\overline{\alpha}})\}$$

3 Extended Fuzzy Regions

3.1 Concept

The main problem with the above model is that all points are considered individually. In many applications, it would be beneficial to group elements of the fuzzy set together. In order to do this, the fuzzy set defining the region will be defined over the domain $\wp(\mathbb{R}^2)$; i.e. the powerset of $\mathbb{R}^2$. The powerset of a set is a new set containing all the possible subsets of that particular set. To illustrate this, consider the following example:

$$\wp(\{0,1,2\}) = \{\{\}, \{0\}, \{1\}, \{2\}, \{0,1\}, \{0,2\}, \{1,2\}, \{0,1,2\}\}$$

For $\mathbb{R}^2$ this becomes:

$$\wp(\mathbb{R}^2) = \left\{X | X \subseteq \mathbb{R}^2\right\}$$

At present, we are only considering the powerset over $\mathbb{R}^2$ as a means to further model the distribution of membership grades in fuzzy regions. The use of the fuzzy powerset over $\mathbb{R}^2$ (i.e. $\tilde{\wp}(\mathbb{R}^2)$, the set of all fuzzy sets over $\mathbb{R}^2$) is also under consideration, as it may allow the model to cope with more complex distributions of membership grades, but at an increased complexity. Definition 2 shows the new definitions for fuzzy regions.

Definition 2 (Fuzzy region).

$$\tilde{R} = \{(P, \mu_{\tilde{R}}(P)) | P \in \wp(\mathbb{R}^2) \wedge \forall P_1, P_2 \in \tilde{R} : P_1 \cap P_2 = \emptyset\} \tag{5}$$

Note that it is required that no two elements of the fuzzy region share points: the intersection between any two elements should be empty. A point can only be considered to belong to the region once, even if it is to a membership grade less than 1. An extension of this is under investigation, but is not elaborated upon for now. With each element of the fuzzy set, a membership grade is associated; it still has a veristic interpretation to indicate the degree to which this element belongs to the fuzzy region. Before considering the impact of this new definition on operations, it is interesting to verify that the model is a generalization of the previous model, and that it still is capable of representing fuzzy regions as defined before. A fuzzy region $\tilde{R}_1$, defined as:

$$\tilde{R}_1 = \{(p, \mu_{\tilde{R}_1}(p)) | p \in \mathbb{R}^2\}$$

can be defined using this new definition as a fuzzy region $\tilde{R}$, by using an appropriate membership function:

$$\tilde{R}_2 = \{(P, \mu_{\tilde{R}_2}(P)) | P \in \wp(\mathbb{R}^2)\}$$

where

$$\mu_{\tilde{R}_2}(P) = \begin{cases} \mu_{\tilde{R}_1}(p) & \text{if } P = \{p\} \\ 0 & \text{elsewhere} \end{cases}$$

The requirement that two elements need to be disjoint is automatically fulfilled by the given membership function.

3.2 Operators

The change in definition will of course impact the operators that have been defined. In [7], we presented two definitions for the surface area of fuzzy regions: the first is an extension of fuzzy cardinality (each point basically contributes its membership grade to the total area), which yielded a crisp number; whereas the second results in a fuzzy number representing the possible surface areas for the region. Both definitions (2.2) will now be adapted to suit the new definition. For the distance between fuzzy regions, two approaches have been considered: one using the α levels, and one based on topological aspects. In this contribution, only the first approach is considered, as topological aspects are currently under development.

Surface area: cardinality. While the meaning of the original definition (each element contributing its membership grade) remains the same, the present formal definition as shown in (2) needs to be changed slightly to deal with the fact that the basic elements of the fuzzy region are now sets rather than points. Each element will once again contribute to the extent of its membership grade, but the elements themselves are now possibly larger than a single point. As a result, the surface areas of each basic element also needs to be taken into consideration:

Definition 3 (surface area (cardinality)).

$$\tilde{S}^c(\tilde{A}) = \int_{X \in \wp(U)} \mu_{\tilde{A}}(X) \int_{p(x,y) \in X} d(x,y) dX \tag{6}$$

This definition yields the same result as the previous definition if all basic elements are singleton sets.

Surface area: fuzzy number. The surface area as a fuzzy number is impacted in a far greater way: a given set that is an element of the fuzzy region will contribute its area, or won't contribute at all, with nothing in between. Consequently, there is a smaller number of possible surface areas; as illustrated on figure 1. Figure 1a shows a fuzzy region in the classical definition and its surface area: each point may or may not belong to it, so it may or may not play a part in determining the surface area. The surface area is also shown, and every area between 0 and $2x$ is a possibility. Figure 1b shows a similar fuzzy region, but now the regions consists out of 2 sets: one set which is given membership grade 1, and one given membership grade 0.5. The expected surface area is also shown: both regions either contribute as a whole, or don't contribute at all. As a result, the only possible surface areas are x and $2x$. Figure 1c finally combines both aspects: the points with membership grade 0.5 each are counted individually, whereas the set with membership grade 1 is counted as one whole; this is also reflected in the surface area.

Note that currently, a sub-region is not considered to have 0 as a possible surface area, resulting in the fact that the total surface area for the region cannot be 0 (e.g. for the region on Figure 1b). This choice is open for discussion when the regions represent imprecision: some argue that there should always be one valid value, which implies that a region must have some elements and thus that 0 is not a possible surface area; whereas others can argue that an empty region is still a region. However if the regions are used to represent uncertainty, each sub-region must also include 0 as a possible surface area. This discussion only has a minor impact on the definition though.

In the new situation, where fuzzy regions are defined as fuzzy sets over $\wp(\mathbb{R}^2)$, the main problem with the original definition (definition 3) is that it makes all surface areas between every strong and the weak α level possible, which - as illustrated on fig 1b - may not be the case. As before, the surface area of the region should be a (fuzzy) number representing the possible surface areas and their membership grade. Furthermore each subregion is considered either as a

$$\tilde{A} = \tilde{A}_1 \cup \tilde{A}_2 = \{(p,1)|p \in A_1\} \cup \{(p,0.5)|p \in A_2\}$$

$$\tilde{B} = \{(B_1,1),(B_2,0.5)\}$$

$$\tilde{C} = \{(C_1,1)\} \cup \tilde{C}_2 = \tilde{C}_1 \cup \{(p,0.5)|p \in C_2\}$$

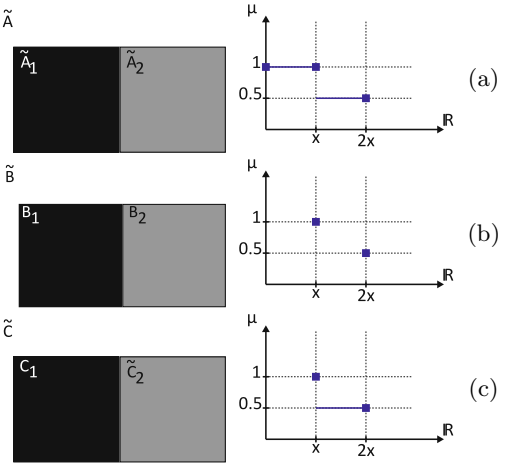


Fig. 1. Three different examples of fuzzy regions, with their surface areas: (a) the classical model, where each points is treated independently, (b) the new model, showing a region consisting of two subregions each counted as a whole, (c) a region where there is both a subregion where points are treated independently, and a subregion that is counted as a whole. For each region, a mathematical explanation of its elements, a graphical illustration, where the shade of grey is representative of the membership grade, and its surface area are shown.

whole, or not at all, so its only possible contribution is its surface area; the amount it contributes to the fuzzy number is the same is a the degree to which it belongs to the region.

Consider the example in fig 2a; the surface area is the sum of all the surface areas that make up the region. As each of these sub-regions are possible, their surface areas are considered to be (non-normalized) fuzzy numbers, with just one possible value which has a membership grade equal to the possibility of the sub-region in R (fig 2c).

$$\tilde{S}^f(\tilde{R}) = (S(R_1), \mu_{\tilde{R}}(R_1)) \tilde{+} (S(R_2), \mu_{\tilde{R}}(R_2)) \tilde{+} (S(R_3), \mu_{\tilde{R}}(R_3))$$

To add up all these regions, fuzzy arithmetic is used, which implies considering the different possible combinations:

- a surface area of x is only possible by considering either region 1, or region 2 or region 3; the membership grade for x therefore is: $\max(0.5, 1, 0.5) = 1$
- a surface area of $2x$ is possible by considering either region 1 and 2, 1 and 3 or 2 and 3; which yields: $\max(\min(0.5, 1), \min(0.5, 0.5), \min(1, 0.5)) = 0.5$
- a surface area of $3x$ is only possible by considering all three regions; yielding for $3x$ the membership grade: $\max(\min(0.5, 1, 0.5)) = 0.5$

The end result is

$$\tilde{S}^f(\tilde{R}) = \{(x,1),(2x,0.5),(3x,0.5\}$$

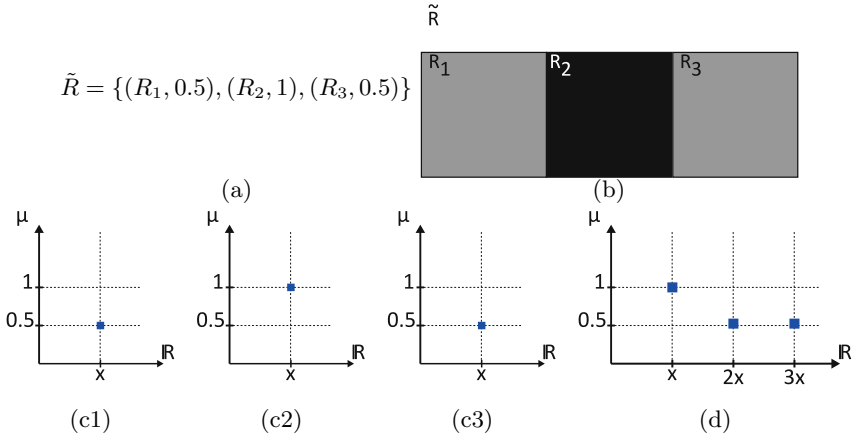


Fig. 2. Example to illustrate the definition: (a) definition of the region, (b) graphical representation, (ci) area of R_i, (d) total surface area of R

as also illustrated on fig 2d. In general, for any number of closed sets, the surface area can be defined as:

$$\tilde{S}^f(\tilde{R}) = \tilde{\sum}_{X|\mu_{\tilde{R}}(X)>0} \big(S(X), \mu_{\tilde{R}}(X)\big) \tag{7}$$

Generalizing results in the definition:

Definition 4 (surface area (fuzzy number))

$$\tilde{S}^f(\tilde{R}) = \tilde{\sum}_{X|\mu_{\tilde{R}}(X)>0} \left(\int_{p(x,y)\in X} d(x,y), \mu_{\tilde{R}}(X) \right) \tag{8}$$

In [2], a fuzzy number X is defined as a fuzzy set over $\mathbb{R}$ that satisfies the properties:

- $\tilde{X}$ is normalized, i.e. there is at least one element x for which $\mu_{\tilde{X}}(x) = 1$
- $\forall \alpha \in]0, 1] : \tilde{X}_{\overline{\alpha}}$ is a closed interval
- the support $\tilde{X}_0$ of $\tilde{X}$ must be bounded

For future computations, it is useful to have a fuzzy number as the result of a fuzzy surface area: calculations with different surface areas can be performed using fuzzy arithmetic. The new definition satisfies the first property if the region has at least one element with membership grade 1; and always satisfies the second and third properties if the fuzzy region is bound. While these assumptions are technically not required in the definition of fuzzy regions, they are reasonable assumptions which most likely will be made anyway.

Distance: α level approach. For the distance between fuzzy regions, two approaches have been considered [7]: one based on α-levels and one based on topological aspects. In this contribution, only the first approach is considered, in which the distance to a fuzzy region yields a fuzzy number, that models all the possible distances between the corresponding α-levels, not dissimilar to the concept of the surface area as a fuzzy number. To define the distance in the new concept, consider two fuzzy regions $\tilde{A}$ and $\tilde{B}$.

The original definition (definition 4) poses a problem similar to the calculation of the surface area: fewer distances than before need to be taken into consideration, as elements can now be sets that contribute only as a whole. In a similar approach to the surface calculation, the distances between the different elements (subregions) are first considered, and then combined to form the overall distance. As the elements of the newly defined fuzzy regions are basically crisp sets in $\mathbb{R}^2$, there are different options for the distance between any two elements of both fuzzy regions; here the shortest possible Euclidean distance is chosen.

$$d(A,B) = \min_{p_A \in A, p_B \in B} \{d(p_A, p_B)\}$$

where the distance between points is given by the Euclidean distance:

$$d(p_A, p_B) = \sqrt{(x_{p_A} - x_{p_B})^2 + (y_{p_A} - y_{p_B})^2}$$

with $p_i = (x_{p_i}, y_{p_i})$

The likelihood of each distance is given by the intersection of the membership grades of both elements considered; we will use the minimum as the intersection operator. The membership grade associated with $d(A,B)$ therefore is given by: $\min(\mu_{\tilde{R}_1}(A), \mu_{\tilde{R}_2}(B))$. The fuzzy number representing the distances is then obtained as the (fuzzy) union of all possible distances between any element of $\tilde{R_A}$ and any element of $\tilde{R_B}$. This leads us to define the distance between fuzzy regions as follows:

Definition 5 (distance)

$$\tilde{\bigcup}_{A|\mu_{\tilde{R}_1}(A)>0, B|\mu_{\tilde{R}_2}(B)>0} \left\{(d(A,B), \min(\mu_{\tilde{R}_1}(A), \mu_{\tilde{R}_2}(B)))\right\}$$

In the case where all element of both $\mu_{\tilde{R}_1}(A)$ and $\mu_{\tilde{R}_2}(B)$ are singleton sets, i.e. $A = \{p_A\}, B = \{p_B\}$, this definition yields the same result at the previous one:

$$\begin{aligned}
&\tilde{\bigcup}_{A|\mu_{\tilde{R}_1}(A)>0, B|\mu_{\tilde{R}_2}(B)>0} \left\{(d(A,B), \min(\mu_{\tilde{R}_1}(A), \mu_{\tilde{R}_2}(B)))\right\} \\
\Leftrightarrow\ & \mu_{\tilde{d}(\tilde{R}_1,\tilde{R}_2)}(x) = \sup\{\alpha \mid \alpha = \min(\mu_{\tilde{R}_1}(A), \mu_{\tilde{R}_2}(B)) \wedge d(A,B) = x\} \\
\Leftrightarrow\ & \mu_{\tilde{d}(\tilde{R}_1,\tilde{R}_2)}(x) = \sup\{\alpha \mid \alpha = \min(\mu_{\tilde{R}_1}(\{p_A\}), \mu_{\tilde{R}_2}(\{p_B\})) \wedge d(p_A,p_B) = x\} \\
\Leftrightarrow\ & \mu_{\tilde{d}(\tilde{R}_1,\tilde{R}_2)}(x) = \sup\{\alpha \mid d(\tilde{R}_{1\alpha}, \tilde{R}_{2\alpha}) \leq x \leq d(\tilde{R}_{1\overline{\alpha}}, \tilde{R}_{2\overline{\alpha}})\}
\end{aligned}$$

In the previous definition, the membership was defined on a per-element basis: for each distance, its membership was the highest value so that the distance

between the weak α-cuts of both regions at this level is less or equal, and the distance of the strong alpha cuts is greater or equal to it. The membership grade for each distance in the new definition is defined as the greatest value (due to the use of max as union operator) of the smallest membership grades of points in both regions that are this distance apart. This smallest membership grade is equivalent to the α-level in the previous definition: at a higher α level, the distance between these two points will no longer be taken into consideration as one point will not be present in the α-cut. As a result, these two definitions are equivalent if all elements of the newly defined regions are singleton sets.

4 Conclusion

In this contribution, the concept of fuzzy regions was extended with the ability to model groups of points belonging to some extent to the fuzzy region. This enriches the theoretical model of fuzzy regions, but at the same time opens the door for improvements to the derived practical models (both the bitmap based an the triangular network model that have been developed before [4],[6]). The new model can be used to incorporate possible subregions, and by carefully chosing them also possible boundaries, that need to be considered as a whole. Along with the changes to the model, some operators (namely surface calculation and distance) have been considered. The adaptation of the model is such that regions from the previous model can easily be represented with it; as such it is a generalization of the previous model (which in turn was a generalization of the crisp model). The next step is to consider the change that will occur in topological relations, and to optimize three derived models (contour lines, bitmaps, triangular networks) to handle the change in the theoretical model in an efficient manner.

References

1. Dubois, D., Prade, H.: Fundamentals of Fuzzy Sets. Kluwer Academic Publishers, Dordrecht (2000)
2. Klir, G.J., Yuan, B.: Fuzzy sets and fuzzy logic: Theory and applications. Prentice Hall, New Jersey (1995)
3. Rigaux, P., Scholl, M., Voisard, A.: Spatial Databases with Applications to GIS. Morgan Kaufman Publishers, San Francisco (2002)
4. Verstraete, J., De Tré, G., Hallez, A.: Adapting TIN-layers to Represent Fuzzy Geographic Information. In: The Seventh Meeting of the EURO Working Group on Fuzzy Sets, pp. 57–62 (2002)
5. Verstraete, J., De Tré, G., De Caluwe, R., Hallez, A.: Field Based Methods for the modelling of Fuzzy Spatial Data. In: Petry, F., Robinson, V., Cobb, M. (eds.) Fuzzy modeling with Spatial Information for Geographic Problems, pp. 41–69. Springer, Heidelberg (2005)

6. Verstraete, J., Hallez, A., De Tré, G.: Bitmap Based Structures for the modelling of Fuzzy Entities. Special issue of Control & Cybernetics 35(1), 147–164 (2006)
7. Verstraete, J.: Fuzzy regions: interpretations of surface area and distance. Control and Cybernetics 38, 509–528 (2009)
8. Zadeh, L.A.: Fuzzy Sets. Information and Control 1(3), 338–353 (1965)
9. Zimmerman, H.-J.: Practical Applications of Fuzzy Technologies. Kluwer Academic Publishers, Dordrecht (1999)

On Liu's Inference Rules for Fuzzy Inference Systems

Xin Gao[1], Dan A. Ralescu[2], and Yuan Gao[3]

[1] School of Mathematics and Physics, North China Electric Power University
Beijing, 102206, P.R. China
gao-xin00@mails.tsinghua.edu.cn
[2] Department of Mathematical Sciences, University of Cincinnati
Cincinnati, OH 45221-0025, USA
ralescd@uc.edu
[3] Department of Mathematical Sciences, Tsinghua University
Beijing, 100084, P.R. China
g-y08@mails.tsinghua.edu.cn

Abstract. Liu's inference is a process of deriving consequences from fuzzy knowledge or evidence via the tool of conditional credibility. Using membership functions, this paper derives some expressions of Liu's inference rule for fuzzy systems. This paper also gives some new inference rules with multiple antecedents and with multiple if-then rules.

Keywords: Fuzzy inference, inference rule, credibility theory.

1 Introduction

Fuzzy inference is a process of deriving consequences from fuzzy knowledge or evidence. Fuzzy inference has been studied by many scholars and a lot of inference rules have been suggested. Zadeh [11] outlined the compositional rule of fuzzy inference in 1973. Based on Zadeh's paper, Mamdani and Assilian [10] proposed an inference rule into a working control system. Besides, Lukasiewicz's inference rule is also widely used in fuzzy inference systems. Fuzzy inference systems have been successfully applied in fields such as automatic control, data classification, decision analysis, and expert systems.

Different from the above inference rules, a new type of inference rule was proposed by Liu [8] and revised by Liu [9] in 2009. The new inference rule derives the consequences by the tool of conditional fuzzy set. Such an inference rule was renamed Liu's inference rule by the community in order to differentiate it from Zadeh's rule, Mamdani's rule, etc.

In this paper, we discuss the expressions of Liu's inference rule for fuzzy systems when membership functions are given. Section 2 introduces some basic concepts and results on fuzzy sets as preliminary. By embedding the conditional credibility into the framework of Liu's inference rule, we obtain some expressions of Liu's inference rule by membership functions in Section 3. We also discuss Liu's inference rule with multiple antecedents and with multiple if-then rules in

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 571–580, 2010.

Section 4 and Section 5, respectively. Section 6 concludes this paper with a brief summary.

2 Preliminaries

Credibility theory, founded by Liu [5] in 2004 and refined by Liu [7] in 2007, is a branch of mathematics for studying the behavior of fuzzy phenomena. In this section, we will introduce some useful definitions about conditional credibility and fuzzy sets.

Let Θ be a nonempty set, and let $\mathcal{P}$ be the power set of Θ (i.e., all subsets of Θ). Each element in $\mathcal{P}$ is called an event. In order to measure the event, credibility measure Cr was introduced as a set function satisfying the normality, monotonicity, self-duality, and maximality axioms. The triplet $(\Theta, \mathcal{P}, \mathrm{Cr})$ is called a credibility space. In order to define the credibility measure of an event A after it has been learned that some other event B has occurred, Liu [7] defined a conditional credibility measure $\mathrm{Cr}\{A|B\}$ as follows:

Definition 1. *(Liu [7]) Let $(\Theta, \mathcal{P}, \mathrm{Cr})$ be a credibility space, and $A, B \in \mathcal{P}$. Then the conditional credibility measure of A given B is defined by*

$$\mathrm{Cr}\{A|B\} = \begin{cases} \dfrac{\mathrm{Cr}\{A \cap B\}}{\mathrm{Cr}\{B\}}, & \text{if } \dfrac{\mathrm{Cr}\{A \cap B\}}{\mathrm{Cr}\{B\}} < 0.5 \\ 1 - \dfrac{\mathrm{Cr}\{A^c \cap B\}}{\mathrm{Cr}\{B\}}, & \text{if } \dfrac{\mathrm{Cr}\{A^c \cap B\}}{\mathrm{Cr}\{B\}} < 0.5 \\ 0.5, & \text{otherwise} \end{cases}$$

provided that $\mathrm{Cr}\{B\} > 0$.

The concept of fuzzy set was initiated by Zadeh via membership function in 1965. In 2007, Liu [7] extended Zadeh's fuzzy set by redefining a fuzzy set as a function from a credibility space to a collection of sets.

Definition 2. *(Liu [7]) A fuzzy set is a function ξ from a credibility space $(\Theta, \mathcal{P}, \mathrm{Cr})$ to a collection of sets of real numbers.*

Liu's fuzzy set has the same fashion of random set that is a function from a probability space to a collection of sets.

A fuzzy set ξ on $(\Theta, \mathcal{P}, \mathrm{Cr})$ is said to have a membership function $\mu(x)$ with $0 \leq \mu(x) \leq 1$ if

$$\xi(\theta) = \{x \in \Re | \mu(x) \geq (2\mathrm{Cr}\{\theta\}) \wedge 1\}$$

for each $\theta \in \Theta$.

Definition 3. *(Liu [9]) Let ξ and η be two fuzzy sets on the credibility space $(\Theta, \mathcal{P}, \mathrm{Cr})$. Then the complement ξ^c of the fuzzy set ξ is*

$$\xi^c(\theta) = (\xi(\theta))^c, \ \forall \theta \in \Theta.$$

The union $\xi \cup \eta$ *of fuzzy sets* ξ *and* η *is*

$$(\xi \cup \eta)(\theta) = \xi(\theta) \cup \eta(\theta), \ \forall \theta \in \Theta.$$

The intersection $\xi \cap \eta$ *of fuzzy sets* ξ *and* η *is*

$$(\xi \cap \eta)(\theta) = \xi(\theta) \cap \eta(\theta), \ \forall \theta \in \Theta.$$

It is clear that

$$\xi \cup \xi^c \equiv \Re, \ \xi \cap \xi^c \equiv \emptyset$$

for any fuzzy set ξ*. This property is very different from Zadeh's fuzzy set.*

Definition 4. *(Liu [9]) Let* ξ *be a fuzzy set on* $(\Theta, \mathcal{P}, \mathrm{Cr})$*. A conditional fuzzy set of* ξ *given* B *is a function* $\xi|_B$ *from the conditional credibility space* $(\Theta, \mathcal{P}, \mathrm{Cr}\{\cdot|B\})$ *to a collection of sets of real numbers such that*

$$\xi|_B(\theta) \equiv \xi(\theta), \ \forall \theta \in \Theta$$

in the sense of classical set theory.

Let ξ and ξ^* be two fuzzy sets in the sense of Definition 2. In order to define the matching degree that ξ^* matches ξ, we first introduce the following symbols:

$$\{\xi^* \subset \xi\} = \{\theta \in \Theta \mid \xi^*(\theta) \subset \xi(\theta)\}$$

$$\{\xi^* \subset \xi^c\} = \{\theta \in \Theta \mid \xi^*(\theta) \subset \xi(\theta)^c\}$$

$$\{\xi^* \not\subset \xi^c\} = \{\theta \in \Theta \mid \xi^*(\theta) \cap \xi(\theta) \neq \emptyset\}.$$

Definition 5. *(Liu [9]) Let* ξ *and* ξ^* *be two fuzzy sets. Then the event* $\xi^* \rhd \xi$ *(i.e.,* ξ^* *matches* ξ*) is defined as*

$$\{\xi^* \rhd \xi\} = \begin{cases} \{\xi^* \subset \xi\}, \ \textit{if } \mathrm{Cr}\{\xi^* \subset \xi\} > \mathrm{Cr}\{\xi^* \subset \xi^c\} \\ \{\xi^* \not\subset \xi^c\}, \ \textit{if } \mathrm{Cr}\{\xi^* \subset \xi\} \leq \mathrm{Cr}\{\xi^* \subset \xi^c\}. \end{cases}$$

Definition 6. *(Liu [9]) Let* ξ *and* ξ^* *be two fuzzy sets. Then the matching degree of* $\xi^* \rhd \xi$ *is defined as*

$$\mathrm{Cr}\{\xi^* \rhd \xi\} = \begin{cases} \mathrm{Cr}\{\xi^* \subset \xi\}, \ \textit{if } \mathrm{Cr}\{\xi^* \subset \xi\} > \mathrm{Cr}\{\xi^* \subset \xi^c\} \\ \mathrm{Cr}\{\xi^* \not\subset \xi^c\}, \ \textit{if } \mathrm{Cr}\{\xi^* \subset \xi\} \leq \mathrm{Cr}\{\xi^* \subset \xi^c\}. \end{cases}$$

3 Inference Rule

Liu's Inference Rule 1. *(Liu [8,9]) Let* $\mathbb{X}$ *and* $\mathbb{Y}$ *be two concepts. Assume a rule "if* $\mathbb{X}$ *is a fuzzy set* ξ *then* $\mathbb{Y}$ *is a fuzzy set* η*". From* $\mathbb{X}$ *is a fuzzy set* ξ^* *we infer that* $\mathbb{Y}$ *is a fuzzy set*

$$\eta^* = \eta|_{\xi^* \rhd \xi}$$

which is the conditional fuzzy set of η *given* $\xi^* \rhd \xi$*. Liu's inference rule is represented by*

Rule: If $\mathbb{X}$ is ξ then $\mathbb{Y}$ is η
From: $\mathbb{X}$ is ξ^*
Infer: $\mathbb{Y}$ is $\eta^* = \eta|_{\xi^* \rhd \xi}$

Theorem 1. *Let ξ, ξ^* and η be independent fuzzy sets with membership functions μ, μ^* and ν, respectively. Then Liu's Inference Rule 1 yields that η^* has a membership function*

$$\nu^*(y) = \begin{cases} \dfrac{2\nu(y)}{\sup\limits_{x\in\Re} \mu(x) \wedge \mu^*(x)} \wedge 1, & \text{if } \sup\limits_{x\in\Re} \mu(x) \wedge \mu^*(x) < 1 \\ \dfrac{2\nu(y)}{2 - \sup\limits_{x\in\Re} \{\mu(x) | \mu(x) < \mu^*(x)\}} \wedge 1, & \text{if } \sup\limits_{x\in\Re} \mu(x) \wedge \mu^*(x) = 1. \end{cases}$$

□

Proof. From the definition of membership function, we have

$$\nu^*(y) = (2\mathrm{Cr}\{y \in \eta^*\}) \wedge 1 = (2\mathrm{Cr}\{y \in \eta | \xi^* \rhd \xi\}) \wedge 1. \tag{1}$$

By using Definition 1, we obtain

$$\mathrm{Cr}\{y \in \eta | \xi^* \rhd \xi\} = \begin{cases} \dfrac{\mathrm{Cr}\{(y \in \eta) \cap (\xi^* \rhd \xi)\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}}, \\ \qquad \text{if } \dfrac{\mathrm{Cr}\{(y \in \eta) \cap (\xi^* \rhd \xi)\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}} < 0.5 \\ 1 - \dfrac{\mathrm{Cr}\{(y \notin \eta) \cap (\xi^* \rhd \xi)\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}}, \\ \qquad \text{if } \dfrac{\mathrm{Cr}\{(y \notin \eta) \cap (\xi^* \rhd \xi)\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}} < 0.5 \\ 0.5, \text{otherwise} \end{cases}$$

$$= \begin{cases} \dfrac{\mathrm{Cr}\{y \in \eta\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}}, & \text{if } \dfrac{\mathrm{Cr}\{y \in \eta\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}} < 0.5 \\ 1 - \dfrac{\mathrm{Cr}\{y \notin \eta\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}}, & \text{if } \dfrac{\mathrm{Cr}\{y \notin \eta\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}} < 0.5 \\ 0.5, & \text{otherwise.} \end{cases} \tag{2}$$

If $\mathrm{Cr}\{y \in \eta\}/\mathrm{Cr}\{\xi^* \rhd \xi\} \geq 0.5$, then $\mathrm{Cr}\{y \in \eta | \xi^* \rhd \xi\} \geq 0.5$ by Equation (2). It follows from Equation (1) that

$$\nu^*(y) = 1 = \frac{2\mathrm{Cr}\{y \in \eta\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}} \wedge 1.$$

Therefore,

$$\nu^*(y) = \frac{2\mathrm{Cr}\{y \in \eta\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}} \wedge 1$$

for all $y \in \Re$. It follows from the definition of membership function that

$$\mathrm{Cr}\{y \in \eta\} = \begin{cases} \dfrac{1}{2}\nu(y), & \text{if } \nu(y) < 1 \\ 1, & \text{if } \nu(y) = 1. \end{cases}$$

If $\nu(y) = 1$, then

$$\frac{2\mathrm{Cr}\{y \in \eta\}}{\mathrm{Cr}\{\xi^* \rhd \xi\}} \wedge 1 = 1 = \frac{\nu(y)}{\mathrm{Cr}\{\xi^* \rhd \xi\}} \wedge 1.$$

Thus for all $y \in \Re$, we have

$$\nu^*(y) = \frac{\nu(y)}{\mathrm{Cr}\{\xi^* \rhd \xi\}} \wedge 1. \tag{3}$$

Now we calculate $\mathrm{Cr}\{\xi^* \rhd \xi\}$. It breaks down into two cases.

Case 1: $\sup_{x\in\Re} \mu(x) \wedge \mu^*(x) < 1$. We have $\mathrm{Cr}\{\xi^* \not\subset \xi^c\} = \frac{1}{2} \sup_{x\in\Re} \mu(x) \wedge \mu^*(x) < 0.5$ immediately. Since

$\{\xi^* \subset \xi\} \subset \{\xi^* \not\subset \xi^c\}$ and $\{\xi^* \not\subset \xi^c\} = \{\xi^* \subset \xi^c\}^c$, we have

$$\mathrm{Cr}\{\xi^* \subset \xi\} \leq \mathrm{Cr}\{\xi^* \not\subset \xi^c\} < 0.5,\ \mathrm{Cr}\{\xi^* \subset \xi^c\} = 1 - \mathrm{Cr}\{\xi^* \not\subset \xi^c\} > 0.5.$$

Hence $\mathrm{Cr}\{\xi^* \subset \xi\} < \mathrm{Cr}\{\xi^* \subset \xi^c\}$. It follows from Definition 6 that

$$\mathrm{Cr}\{\xi^* \rhd \xi\} = \mathrm{Cr}\{\xi^* \not\subset \xi^c\} = \frac{1}{2} \sup_{x\in\Re} \mu(x) \wedge \mu^*(x).$$

Case 2: $\sup_{x\in\Re} \mu(x) \wedge \mu^*(x) = 1$. For this case, $\mathrm{Cr}\{\xi^* \subset \xi^c\} = \mathrm{Cr}\{\emptyset\} = 0$. Thus $\mathrm{Cr}\{\xi^* \subset \xi\} \geq \mathrm{Cr}\{\xi^* \subset \xi^c\}$.

From Definition 6 we have

$$\mathrm{Cr}\{\xi^* \rhd \xi\} = \mathrm{Cr}\{\xi^* \subset \xi\} = 1 - \mathrm{Cr}\{\xi^* \not\subset \xi\} = 1 - \frac{1}{2} \sup_{x\in\Re} \{\mu(x) | \mu(x) < \mu^*(x)\}.$$

Hence

$$\mathrm{Cr}\{\xi^* \rhd \xi\} = \begin{cases} \frac{1}{2} \sup_{x\in\Re} \mu(x) \wedge \mu^*(x), & \text{if } \sup_{x\in\Re} \mu(x) \wedge \mu^*(x) < 1 \\ 1 - \frac{1}{2} \sup_{x\in\Re} \{\mu(x) | \mu(x) < \mu^*(x)\}, & \text{if } \sup_{x\in\Re} \mu(x) \wedge \mu^*(x) = 1, \end{cases} \tag{4}$$

It follows from Equation (3) and Equation (4) that

$$\nu^*(y) = \begin{cases} \frac{2\nu(y)}{\sup_{x\in\Re} \mu(x) \wedge \mu^*(x)} \wedge 1, & \text{if } \sup_{x\in\Re} \mu(x) \wedge \mu^*(x) < 1 \\ \frac{2\nu(y)}{2 - \sup_{x\in\Re} \{\mu(x) | \mu(x) < \mu^*(x)\}} \wedge 1, & \text{if } \sup_{x\in\Re} \mu(x) \wedge \mu^*(x) = 1. \end{cases}$$

□

Theorem 2. *Let ξ and η be independent fuzzy sets with membership functions μ and ν, respectively. If ξ^* is a constant a, then Liu's Inference Rule 7 yields that η^* has a membership function*

$$\nu^*(y) = \begin{cases} \dfrac{2\nu(y)}{\mu(a)} \wedge 1, & \text{if } \mu(a) < 1 \\ \nu(y), & \text{if } \mu(a) = 1. \end{cases}$$

□

Proof. From Equation (3), we have

$$\nu^*(y) = \frac{\nu(y)}{\mathrm{Cr}\{a \triangleright \xi\}} \wedge 1. \tag{5}$$

From Definition 6, we obtain

$$\begin{aligned} \mathrm{Cr}\{a \triangleright \xi\} &= \begin{cases} \mathrm{Cr}\{\{a\} \subset \xi\}, & \text{if } \mathrm{Cr}\{\{a\} \subset \xi\} > \mathrm{Cr}\{\{a\} \subset \xi^c\} \\ \mathrm{Cr}\{\{a\} \not\subset \xi^c\}, & \text{if } \mathrm{Cr}\{\{a\} \subset \xi\} \leq \mathrm{Cr}\{\{a\} \subset \xi^c\} \end{cases} \\ &= \begin{cases} \mathrm{Cr}\{a \in \xi\}, & \text{if } \mathrm{Cr}\{a \in \xi\} > \mathrm{Cr}\{a \in \xi^c\} \\ \mathrm{Cr}\{a \notin \xi^c\}, & \text{if } \mathrm{Cr}\{a \in \xi\} \leq \mathrm{Cr}\{a \in \xi^c\} \end{cases} \\ &= \mathrm{Cr}\{a \in \xi\} \\ &= \begin{cases} \mu(a)/2, & \text{if } \mu(a) < 1 \\ 1, & \text{if } \mu(a) = 1. \end{cases} \end{aligned} \tag{6}$$

It follows from Equation (5) and Equation (6) that

$$\nu^*(y) = \begin{cases} \dfrac{2\nu(y)}{\mu(a)} \wedge 1, & \text{if } \mu(a) < 1 \\ \nu(y), & \text{if } \mu(a) = 1. \end{cases}$$

□

4 Inference Rule with Multiple Antecedents

Liu's Inference Rule 2. *Let $\mathbb{X}$, $\mathbb{Y}$ and $\mathbb{Z}$ be three concepts. Assume a rule "if $\mathbb{X}$ is a fuzzy set ξ and $\mathbb{Y}$ is a fuzzy set η then $\mathbb{Z}$ is a fuzzy set τ". From $\mathbb{X}$ is a fuzzy set ξ^* and $\mathbb{Y}$ is a fuzzy set η^* we infer that $\mathbb{Z}$ is a fuzzy set*

$$\tau^* = \tau|_{(\xi^* \triangleright \xi) \cap (\eta^* \triangleright \eta)}$$

which is the conditional fuzzy set of τ given $\xi^* \triangleright \xi$ and $\eta^* \triangleright \eta$. Liu's inference rule is represented by

$$\frac{\begin{array}{l}\text{Rule: If } \mathbb{X} \text{ is } \xi \text{ and } \mathbb{Y} \text{ is } \eta \text{ then } \mathbb{Z} \text{ is } \tau \\ \text{From: } \mathbb{X} \text{ is } \xi^* \text{ and } \mathbb{Y} \text{ is } \eta^*\end{array}}{\text{Infer: } \mathbb{Z} \text{ is } \tau^* = \tau|_{(\xi^* \triangleright \xi) \cap (\eta^* \triangleright \eta)}}$$

Theorem 3. *Let ξ, ξ^*, η, η^* and τ be independent fuzzy sets with membership functions μ, μ^*, ν, ν^* and ψ respectively. Then Liu's Inference Rule 2 yields that τ^* has a membership function*

$$\psi^*(z) = \begin{cases} \dfrac{2\psi(z)}{\sup\limits_{x\in\Re} \mu(x) \wedge \mu^*(x) \wedge \sup\limits_{y\in\Re} \nu(y) \wedge \nu^*(y)} \wedge 1, \\ \qquad \textit{if } \sup\limits_{x\in\Re} \mu(x) \wedge \mu^*(x) \wedge \sup\limits_{y\in\Re} \nu(y) \wedge \nu^*(y) < 1 \\ \dfrac{2\psi(z)}{2 - \sup\limits_{x\in\Re} \{\mu(x) | \mu(x) < \mu^*(x)\} \vee \sup\limits_{y\in\Re} \{\nu(y) | \nu(y) < \nu^*(y)\}} \wedge 1, \\ \qquad \textit{if } \sup\limits_{x\in\Re} \mu(x) \wedge \mu^*(x) \wedge \sup\limits_{y\in\Re} \nu(y) \wedge \nu^*(y) = 1. \end{cases}$$

□

Proof. From the definition of membership function, we have

$$\psi^*(z) = (2\mathrm{Cr}\{z \in \tau^*\}) \wedge 1 = (2\mathrm{Cr}\{z \in \tau | (\xi^* \rhd \xi) \cap (\eta^* \rhd \eta)\}) \wedge 1. \tag{7}$$

By using Definition 1, we have

$$\mathrm{Cr}\{z \in \tau | (\xi^* \rhd \xi) \cap (\eta^* \rhd \eta)\} = \begin{cases} \dfrac{\mathrm{Cr}\{z \in \tau\}}{\mathrm{Cr}\{\xi^* \rhd \xi\} \wedge \mathrm{Cr}\{\eta^* \rhd \eta\}}, & \text{if } \dfrac{\mathrm{Cr}\{z \in \tau\}}{\mathrm{Cr}\{\xi^* \rhd \xi\} \wedge \mathrm{Cr}\{\eta^* \rhd \eta\}} < 0.5 \\ 1 - \dfrac{\mathrm{Cr}\{z \notin \tau\}}{\mathrm{Cr}\{\xi^* \rhd \xi\} \wedge \mathrm{Cr}\{\eta^* \rhd \eta\}}, & \text{if } \dfrac{\mathrm{Cr}\{z \notin \tau\}}{\mathrm{Cr}\{\xi^* \rhd \xi\} \wedge \mathrm{Cr}\{\eta^* \rhd \eta\}} < 0.5 \\ 0.5, & \text{otherwise.} \end{cases} \tag{8}$$

If $\mathrm{Cr}\{z \in \tau\} / (\mathrm{Cr}\{\xi^* \rhd \xi\} \wedge \mathrm{Cr}\{\eta^* \rhd \eta\}) \geq 0.5$, then $\mathrm{Cr}\{z \in \tau | (\xi^* \rhd \xi) \cap (\eta^* \rhd \eta)\} \geq 0.5$ by Equation (8). It follows from Equation (7) that

$$\psi^*(z) = (2\mathrm{Cr}\{z \in \tau | (\xi^* \rhd \xi) \cap (\eta^* \rhd \eta)\}) \wedge 1 = 1 = \frac{2\mathrm{Cr}\{z \in \tau\}}{\mathrm{Cr}\{\xi^* \rhd \xi\} \wedge \mathrm{Cr}\{\eta^* \rhd \eta\}} \wedge 1.$$

Therefore,

$$\psi^*(z) = \frac{2\mathrm{Cr}\{z \in \tau\}}{\mathrm{Cr}\{\xi^* \rhd \xi\} \wedge \mathrm{Cr}\{\eta^* \rhd \eta\}} \wedge 1$$

for all $z \in \Re$. It follows from the definition of membership function that

$$\mathrm{Cr}\{z \in \tau\} = \begin{cases} \dfrac{1}{2}\psi(z), & \text{if } \psi(z) < 1 \\ 1, & \text{if } \psi(z) = 1. \end{cases}$$

If $\psi(z) = 1$, then

$$\frac{2\mathrm{Cr}\{z \in \tau\}}{\mathrm{Cr}\{\xi^* \rhd \xi\} \wedge \mathrm{Cr}\{\eta^* \rhd \eta\}} \wedge 1 = 1 = \frac{\psi(z)}{\mathrm{Cr}\{\xi^* \rhd \xi\} \wedge \mathrm{Cr}\{\eta^* \rhd \eta\}} \wedge 1.$$

Thus for all $z \in \Re$, we have

$$\psi^*(z) = \frac{\psi(z)}{\mathrm{Cr}\{\xi^* \rhd \xi\} \wedge \mathrm{Cr}\{\eta^* \rhd \eta\}} \wedge 1. \tag{9}$$

It follows from Equation (4) that

$$\mathrm{Cr}\{\xi^* \rhd \xi\} = \begin{cases} \dfrac{1}{2} \sup\limits_{x \in \Re} \mu(x) \wedge \mu^*(x), & \text{if } \sup\limits_{x \in \Re} \mu(x) \wedge \mu^*(x) < 1 \\ 1 - \dfrac{1}{2} \sup\limits_{x \in \Re} \{\mu(x) | \mu(x) < \mu^*(x)\}, & \text{if } \sup\limits_{x \in \Re} \mu(x) \wedge \mu^*(x) = 1, \end{cases}$$

and

$$\mathrm{Cr}\{\eta^* \rhd \eta\} = \begin{cases} \dfrac{1}{2} \sup\limits_{y \in \Re} \nu(y) \wedge \nu^*(y), & \text{if } \sup\limits_{y \in \Re} \nu(y) \wedge \nu^*(y) < 1 \\ 1 - \dfrac{1}{2} \sup\limits_{y \in \Re} \{\nu(y) | \nu(y) < \nu^*(y)\}, & \text{if } \sup\limits_{y \in \Re} \nu(y) \wedge \nu^*(y) = 1. \end{cases}$$

Thus

$$\psi^*(z) = \begin{cases} \dfrac{2\psi(z)}{\sup\limits_{x \in \Re} \mu(x) \wedge \mu^*(x) \wedge \sup\limits_{y \in \Re} \nu(y) \wedge \nu^*(y)} \wedge 1, \\ \qquad \text{if } \sup\limits_{x \in \Re} \mu(x) \wedge \mu^*(x) \wedge \sup\limits_{y \in \Re} \nu(y) \wedge \nu^*(y) < 1 \\ \dfrac{2\psi(z)}{2 - \sup\limits_{x \in \Re} \{\mu(x) | \mu(x) < \mu^*(x)\} \vee \sup\limits_{y \in \Re} \{\nu(y) | \nu(y) < \nu^*(y)\}} \wedge 1, \\ \qquad \text{if } \sup\limits_{x \in \Re} \mu(x) \wedge \mu^*(x) \wedge \sup\limits_{y \in \Re} \nu(y) \wedge \nu^*(y) = 1. \end{cases}$$

□

Theorem 4. *Let ξ, η, and τ be independent fuzzy sets with membership functions μ, ν, and ψ, respectively. If ξ^* is a constant a and η^* is a constant b, then Liu's Inference Rule 2 yields that τ^* has a membership function*

$$\psi^*(z) = \begin{cases} \dfrac{2\psi(z)}{\mu(a) \wedge \nu(b)} \wedge 1, \text{ if } \mu(a) \wedge \nu(b) < 1 \\ \psi(z), \qquad \text{if } \mu(a) \wedge \nu(b) = 1. \end{cases}$$

□

Proof. From Equation (9), we have

$$\psi^*(z) = \frac{\psi(z)}{\mathrm{Cr}\{a \rhd \xi\} \wedge \mathrm{Cr}\{b \rhd \eta\}} \wedge 1. \tag{10}$$

It follows from Equation (6) that

$$\mathrm{Cr}\{a \rhd \xi\} = \begin{cases} \dfrac{1}{2}\mu(a), \text{ if } \mu(a) < 1 \\ 1, \qquad \text{if } \mu(a) = 1, \end{cases}$$

and

$$\mathrm{Cr}\{b \rhd \eta\} = \begin{cases} \frac{1}{2}\nu(b), & \text{if } \nu(b) < 1 \\ 1, & \text{if } \nu(b) = 1. \end{cases}$$

Therefore,

$$\mathrm{Cr}\{a \rhd \xi\} \wedge \mathrm{Cr}\{b \rhd \eta\} = \begin{cases} \frac{1}{2}\left(\mu(a) \wedge \nu(b)\right), & \text{if } \mu(a) \wedge \nu(b) < 1 \\ 1, & \text{if } \mu(a) \wedge \nu(b) = 1. \end{cases} \tag{11}$$

It follows from Equation (10) and Equation (11) that

$$\Box\psi^*(z) = \begin{cases} \dfrac{2\psi(z)}{\mu(a) \wedge \nu(b)} \wedge 1, & \text{if } \mu(a) \wedge \nu(b) < 1 \\ \psi(z), & \text{if } \mu(a) \wedge \nu(b) = 1. \end{cases}$$

□

5 Inference Rule with Multiple If-Then Rules

Liu's Inference Rule 3. *Let $\mathbb{X}$ and $\mathbb{Y}$ be two concepts. Assume two rules "if $\mathbb{X}$ is a fuzzy set ξ_1 then $\mathbb{Y}$ is a fuzzy set η_1" and "if $\mathbb{X}$ is a fuzzy set ξ_2 then $\mathbb{Y}$ is a fuzzy set η_2". From $\mathbb{X}$ is a fuzzy set ξ^* we infer that $\mathbb{Y}$ is a fuzzy set*

$$\eta^* = \frac{\mathrm{Cr}\{\xi^* \rhd \xi_1\} \cdot \eta_1|_{\xi^* \rhd \xi_1}}{\mathrm{Cr}\{\xi^* \rhd \xi_1\} + \mathrm{Cr}\{\xi^* \rhd \xi_2\}} + \frac{\mathrm{Cr}\{\xi^* \rhd \xi_2\} \cdot \eta_2|_{\xi^* \rhd \xi_2}}{\mathrm{Cr}\{\xi^* \rhd \xi_1\} + \mathrm{Cr}\{\xi^* \rhd \xi_2\}}. \tag{12}$$

The inference rule is represented by

Rule 1: If $\mathbb{X}$ is ξ_1 then $\mathbb{Y}$ is η_1
Rule 2: If $\mathbb{X}$ is ξ_2 then $\mathbb{Y}$ is η_2
From: $\mathbb{X}$ is ξ^*
Infer: $\mathbb{Y}$ is η^* determined by (12)

Theorem 5. *Let ξ_1, ξ_2, η_1, η_2 and ξ^*, be independent fuzzy sets, and let ξ_1, ξ_2, ξ^* have membership functions μ_1, μ_2, μ^*, respectively. Then Liu's Inference Rule 3 yields that*

$$\eta^* = \frac{c_1}{c_1 + c_2}\eta_1^* + \frac{c_2}{c_1 + c_2}\eta_2^*, \tag{13}$$

where $\eta_1^* = \eta_1|_{\xi^* \rhd \xi_1}$, $\eta_1^* = \eta_2|_{\xi^* \rhd \xi_2}$, and

$$c_1 = \begin{cases} \frac{1}{2}\sup\limits_{x \in \Re} \mu_1(x) \wedge \mu^*(x), & \text{if } \sup\limits_{x \in \Re} \mu_1(x) \wedge \mu^*(x) < 1 \\ 1 - \frac{1}{2}\sup\limits_{x \in \Re}\left\{\mu_1(x) | \mu_1(x) < \mu^*(x)\right\}, & \text{if } \sup\limits_{x \in \Re} \mu_1(x) \wedge \mu^*(x) = 1, \end{cases} \tag{14}$$

$$c_2 = \begin{cases} \frac{1}{2}\sup\limits_{x \in \Re} \mu_2(x) \wedge \mu^*(x), & \text{if } \sup\limits_{x \in \Re} \mu_2(x) \wedge \mu^*(x) < 1 \\ 1 - \frac{1}{2}\sup\limits_{x \in \Re}\left\{\mu_2(x) | \mu_2(x) < \mu^*(x)\right\}, & \text{if } \sup\limits_{x \in \Re} \mu_2(x) \wedge \mu^*(x) = 1. \end{cases} \tag{15}$$

□

Proof. Let $c_1 = \mathrm{Cr}\{\xi^* \rhd \xi_1\}$ and $c_2 = \mathrm{Cr}\{\xi^* \rhd \xi_2\}$. Then we get Equation (14) and Equation (15) immediately. According to Liu's Inference Rule 3, we obtain Equation (13). □

6 Conclusions

We obtained some expressions of Liu's inference rule by membership functions. Furthermore, we extended Liu's inference rule to fuzzy inference systems with multiple antecedents and with multiple if-then rules.

Acknowledgments

This work was supported by "the Fundamental Research Funds for the Central Universities" (No.09QL47). Dan A. Ralescu's work was partly supported by a Taft Travel for Research Grant.

References

1. Gao, X., Gao, Y., Ralescu, D.A.: On Liu's inference rule for uncertain systems. Int. J. Uncertainty Fuzziness Knowledge Based Syst. 18, 1–11 (2010)
2. Li, X., Liu, B.: A sufficient and necessary condition for credibility measures. Int. J. Uncertainty Fuzziness Knowledge Based Syst. 14, 527–535 (2006)
3. Li, X., Liu, B.: Foundation of credibilistic logic. Fuzzy Optim. Decis. Making 8, 91–102 (2009)
4. Liu, B., Liu, Y.K.: Expected value of fuzzy variable and fuzzy expected value models. IEEE Trans. Fuzzy Syst. 10, 445–450 (2002)
5. Liu, B.: Uncertainty Theory. Springer, Berlin (2004)
6. Liu, B.: A survey of credibility theory. Fuzzy Optim. Decis. Making 5, 387–408 (2006)
7. Liu, B.: Uncertainty Theory, 2nd edn. Springer, Berlin (2007)
8. Liu, B.: Some research problems in uncertainty theory. J. Uncertain Syst. 3, 3–10 (2009)
9. Liu, B.: Uncertainty Theory, 3rd edn., http://orsc.edu.cn/liu/ut.pdf
10. Mamdani, E.H., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man Mach. Stud. 7, 1–13 (1975)
11. Zadeh, L.A.: Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst. Man Cybern. SMC-3, 28–44 (1973)

A New Approach to the Distances between Intuitionistic Fuzzy Sets

Krassimir Atanassov

Centre of Biomedical Engineering,
Bulgarian Academy of Sciences
105 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
krat@bas.bg

1 Introduction and Preliminary Results

In a series of papers the author introduced a set of new negations and implications over Intuitionistic Fuzzy Sets (IFSs, for all notations about IFSs see [2]). Two of these negations were described in [6]. They generalize the classical negation over IFSs, but on the other hand, they have some non-classical properties. The set has the form

$$\mathcal{N} = \{\neg^{\varepsilon,\eta} \mid 0 \leq \varepsilon < 1 \ \& \ 0 \leq \eta < 1\},$$

where for each IFS $A = \{\langle x, \mu_A(x), \nu_A(x)\rangle | x \in E\}$:

$$\neg^{\varepsilon,\eta} A = \{\langle x, \min(1, \nu_A(x) + \varepsilon), \max(0, \mu_A(x) - \eta)\rangle | x \in E\}.$$

For numbers ε and η there are two cases.

- $\eta < \varepsilon$. As it is shown in [5], this case is impossible.
- $\eta \geq \varepsilon$. Let everywhere below $0 \leq \varepsilon \leq \eta < 1$ be fixed.

First, in [6] we showed that

$$\neg^{0,0} A = \{\langle x, \min(1, \nu_A(x)), \max(0, \mu_A(x))\rangle | x \in E\}$$

$$= \{\langle x, \nu_A(x), \mu_A(x)\rangle | x \in E\} = \neg_1 A,$$

where $\neg_1$ is the classical intuitionistic fuzzy negation.

Second, we checked that set $\neg^{\varepsilon,\eta} A$ is an IFS, because

$$\min(1, b + \varepsilon) + \max(0, a - \eta) = \min(1, b + \varepsilon) \leq 1.$$

Third, we constructed a new implication, generated by the new negation:

$$A \rightarrow^{\varepsilon,\eta} B = \{\langle x, \max(c, \min(1, b + \varepsilon)), \min(d, \max(0, a - \eta))\rangle | x \in E\}$$

$$= \{\langle x, \min(1, \max(c, b + \varepsilon)), \max(0, \min(d, a - \eta))\rangle | x \in E\}.$$

Fourth, in [2] the concepts of *Tautological Set (TS) and Intuitionistic Fuzzy Tautological Set* (IFTS) were introduced as follows: the IFS A is a TS iff for

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 581–590, 2010.

every $x \in E : \mu_A(x) = 1, \nu_A(x) = 0$; the IFS A is an IFTS iff for every $x \in E$: $\mu_A(x) \geq \nu_A(x)$. Obviously, each TS is an IFTS.

In [6], we checked the axioms of intuitionistic logic (see, e.g., [18]) in the case, when A, B and C are IFSs.

Fifth, in [7], a series of new versions of operation "subtraction" was introduced. As a basis of the new versions of operation "subtraction" from [7], the well-known formula from set theory:

$$A - B = A \cap \neg B \tag{1}$$

was used, A and B being given sets.

On the other hand, as we discussed in [3], the Law for Excluded Middle is not always valid in IFS theory. By this reason, we can introduce a new series of "subtraction" operations, that will have the form:

$$A -'' B = \neg\neg A \cap \neg B. \tag{2}$$

In some papers, e.g., [7,19], the properties of some IF-subtractions were studied.

2 Main Results

2.1 In [8], using (1), we obtained the following two operations of subtraction:

$$A -'^{\varepsilon,\eta} B = A \cap \neg^{\varepsilon,\eta} B$$

$$= \{\langle x, \min(\mu_A(x), 1, \nu_B(x) + \varepsilon), \max(\nu_A(x), 0, \mu_B(x) - \eta)\rangle | x \in E\}$$

$$= \{\langle x, \min(\mu_A(x), \nu_B(x) + \varepsilon), \max(\nu_A(x), \mu_B(x) - \eta)\rangle | x \in E\}.$$

Also, using (2) and having in mind that

$$\neg^{\varepsilon,\eta}\neg^{\varepsilon,\eta} A = \neg^{\varepsilon,\eta}\{\langle x, \min(1, \nu_A(x) + \varepsilon), \max(0, \mu_A(x) - \eta)\rangle x, | x \in E\}$$

$$= \{\langle x, \min(1, \max(0, \mu_A(x) - \eta) + \varepsilon), \max(0, \min(1, \nu_A(x) + \varepsilon) - \eta)\rangle x, | x \in E\}$$

$$= \{\langle x, \min(1, \max(\varepsilon, \mu_A(x) - \eta + \varepsilon)), \max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta))\rangle x, | x \in E\}$$

$$= \{\langle x, \max(\varepsilon, \mu_A(x) - \eta + \varepsilon), \max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta))\rangle x, | x \in E\}$$

we obtain the following form of the operation $-''^{\varepsilon,\eta}$:

$$A -''^{\varepsilon,\eta} B = \neg^{\varepsilon,\eta}\neg^{\varepsilon,\eta} A \cap \neg^{\varepsilon,\eta} B$$

$$= \{\langle x, \max(\varepsilon, \mu_A(x) - \eta + \varepsilon), \max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta))\rangle x, | x \in E\}$$

$$\cap \{\langle x, \min(1, \nu_B(x) + \varepsilon), \max(0, \mu_B(x) - \eta)\rangle x, | x \in E\}$$

$$= \{\langle x, \min(\max(\varepsilon, \mu_A(x) - \eta + \varepsilon), 1, \nu_B(x) + \varepsilon),$$

$$\max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta), \mu_B(x) - \eta)\rangle x, | x \in E\}$$

$$= \{\langle x, \min(\max(\varepsilon, \mu_A(x) - \eta + \varepsilon), \nu_B(x) + \varepsilon),$$

$$\max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta), \mu_B(x) - \eta)\rangle x, | x \in E\}.$$

Here, using the subtraction of sets A and $\neg^{\varepsilon,\eta}A$, we will introduce the following two (ε,η)-norms for element $x \in E$:

$$||x||'_{\varepsilon,\eta} = \langle \min(\mu_A(x), \nu_A(x)+\varepsilon), \max(\nu_A(x), \mu_A(x)-\eta)\rangle, \tag{3}$$

$$||x||''_{\varepsilon,\eta} = \langle \min(\max(\varepsilon, \mu_A(x)-\eta+\varepsilon), \nu_A(x)+\varepsilon), \max(0, \min(1-\eta, \nu_A(x)+\varepsilon-\eta), \mu_A(x)-\eta)\rangle. \tag{4}$$

Theorem 1. The two (ε,η)-norms are intuitionistic fuzzy pairs.

Proof. Let

$$X \equiv \min(\mu_A(x), \nu_A(x)+\varepsilon) + \max(\nu_A(x), \mu_A(x)-\eta).$$

If $\nu_A(x) \geq \mu_A(x)-\eta$, then

$$X = \min(\mu_A(x), \nu_A(x)+\varepsilon) + \nu_A(x) \leq \mu_A(x) + \nu_A(x) \leq 1.$$

If $\nu_A(x) > \mu_A(x)-\eta$, then

$$X = \min(\mu_A(x), \nu_A(x)+\varepsilon) + \mu_A(x) - \eta$$

$$\leq \nu_A(x) + \varepsilon + \mu_A(x) - \eta \leq 1 + \varepsilon - \eta < 1.$$

Therefore, the first norm is an intuitionistic fuzzy pair.

Let

$$Y \equiv \min(\max(\varepsilon, \mu_A(x)-\eta+\varepsilon), \nu_A(x)+\varepsilon) + \max(0, \min(1-\eta, \nu_A(x)+\varepsilon-\eta), \mu_A(x)-\eta).$$

If $\max(0, \min(1-\eta, \nu_A(x)+\varepsilon-\eta), \mu_A(x)-\eta) = 0$, then

$$Y = \min(\max(\varepsilon, \mu_A(x)-\eta+\varepsilon), \nu_A(x)+\varepsilon) + 0 \leq \max(\varepsilon, \mu_A(x)-\eta+\varepsilon) \leq 1.$$

If $\max(0, \min(1-\eta, \nu_A(x)+\varepsilon-\eta), \mu_A(x)-\eta) = \min(1-\eta, \nu_A(x)+\varepsilon-\eta)$, then

$$\min(1-\eta, \nu_A(x)+\varepsilon-\eta) \geq \mu_A(x)-\eta,$$

i.e., $\nu_A(x)+\varepsilon \geq \mu_A(x)$. Hence,

$$\mu_A(x)-\eta) \leq \nu_A(x)+\varepsilon-\eta) \leq \nu_A(x)$$

and therefore

$$Y = \min(\max(\varepsilon, \mu_A(x)-\eta+\varepsilon), \nu_A(x)+\varepsilon) + \min(1-\eta, \nu_A(x)+\varepsilon-\eta)$$

$$= \min(\varepsilon + \max(0, \mu_A(x)-\eta), \nu_A(x)+\varepsilon) + \min(1-\eta, \nu_A(x)+\varepsilon-\eta)$$

$$= \varepsilon + \min(\max(0, \mu_A(x)-\eta), \nu_A(x)) + \min(1-\eta, \nu_A(x)+\varepsilon-\eta)$$

$$= \varepsilon + \max(0, \mu_A(x)-\eta) + \min(1, \nu_A(x)+\varepsilon) - \eta$$

$$\leq \max(0, \mu_A(x) - \eta) + \min(1, \nu_A(x) + \varepsilon).$$

If $\mu_A(x) \geq \eta$, then

$$Y \leq \mu_A(x) - \eta + \min(1, \nu_A(x) + \varepsilon) \leq \mu_A(x) - \eta + \nu_A(x) + \varepsilon \leq 1 - \eta + \varepsilon \leq 1;$$

if $\mu_A(x) < \eta$, then $Y \leq \min(1, \nu_A(x) + \varepsilon) \leq 1$.

If $\max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta), \mu_A(x) - \eta) = \mu_A(x) - \eta$, then

$$Y = \min(\max(\varepsilon, \mu_A(x) - \eta + \varepsilon), \nu_A(x) + \varepsilon) + \mu_A(x) - \eta$$

$$\leq \nu_A(x) + \varepsilon + \mu_A(x) - \eta \leq 1 + \varepsilon - \eta \leq 1.$$

Therefore, the second norm is also an intuitionistic fuzzy pair.

All norms and distances, defined over IFSs up to now (see, e.g. [10,11,13,14, 15,16,20,21,22,23,24,25,26,27,28,29]), have been real numbers, that may in some cases be normalized to the $[0, 1]$ interval. As we see, the two norms (3) and (4) have the form of intuitionistic fuzzy pairs and they are the first of this form.

Let $e^*, o^*, u^* \in E$ so that $\mu_A(e^*) = 1,\ \nu_A(e^*) = 0,\ \mu_A(o^*) = 0,\ \nu_A(o^*) = 1,$ $\mu_A(u^*) = 0,\ \nu_A(u^*) = 0$. Then the norms of these three elements are:

x	$\|\|x\|\|'_{\varepsilon,\eta}$	$\|\|x\|\|''_{\varepsilon,\eta}$
e^*	$\langle \varepsilon, 1 - \eta \rangle$	$\langle \varepsilon, 1 - \eta \rangle\rangle$
o^*	$\langle 0, 1 \rangle$	$\langle \varepsilon, 1 - \eta \rangle$
u^*	$\langle 0, 0 \rangle$	$\langle \varepsilon, 0 \rangle$

2.2 There are two ways for introducing distances between elements of a fixed universe E, as it is shown on Fig. 1.

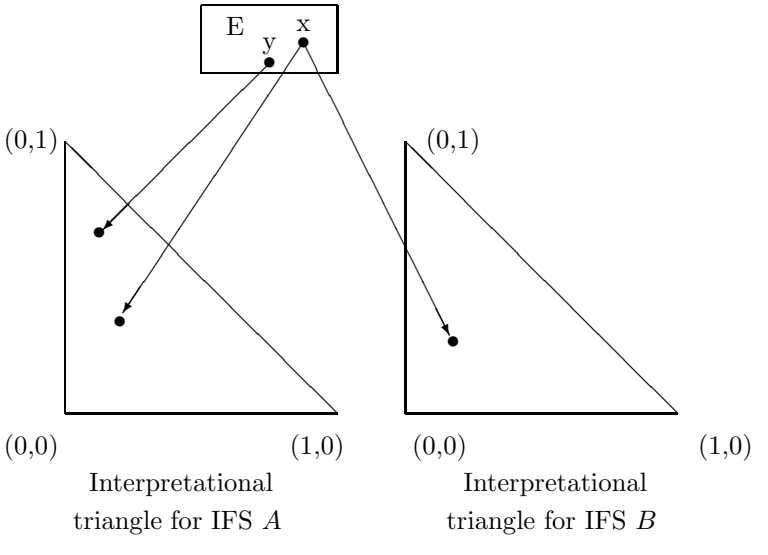


Fig. 1.

Up to now, all distances, similarly to all norms, have been real numbers. Here, and below, we will introduce distances, having forms of intuitionistic fuzzy pairs.

First, we will introduce the following five (ε,η)-distances between the values of one element $x \in E$ about two IFSs A and B, by analogy with the first (ε,η)-norm:

$$d'_{\varepsilon,\eta;str\_opt}(A,B)(x) = \langle \min(\mu_A(x), \nu_B(x)+\varepsilon) + \min(\mu_B(x), \nu_A(x)+\varepsilon) - \min(\mu_A(x), \nu_B(x)+\varepsilon).\min(\mu_B(x), \nu_A(x)+\varepsilon), \max(\nu_A(x), \mu_B(x)-\eta).\max(\nu_B(x), \mu_A(x)-\eta)\rangle,$$

$$d'_{\varepsilon,\eta;opt}(A,B)(x) = \langle \max(\min(\mu_A(x), \nu_B(x)+\varepsilon), \min(\mu_B(x), \nu_A(x)+\varepsilon)), \min(\max(\nu_A(x), \mu_B(x)-\eta), \max(\nu_B(x), \mu_A(x)-\eta))\rangle,$$

$$d'_{\varepsilon,\eta;aver}(A,B)(x) = \langle \frac{\min(\mu_A(x), \nu_B(x)+\varepsilon) + \min(\mu_B(x), \nu_A(x)+\varepsilon)}{2}, \frac{\max(\nu_A(x), \mu_B(x)-\eta) + \max(\nu_B(x), \mu_A(x)-\eta)}{2}\rangle,$$

$$d'_{\varepsilon,\eta;pes}(A,B)(x) = \langle \min(\min(\mu_A(x), \nu_B(x)+\varepsilon), \min(\mu_B(x), \nu_A(x)+\varepsilon)), \max(\max(\nu_A(x), \mu_B(x)-\eta), \max(\nu_B(x), \mu_A(x)-\eta))\rangle,$$

$$d'_{\varepsilon,\eta;str\_pes}(A,B)(x) = \langle \min(\mu_A(x), \nu_B(x)+\varepsilon).\min(\mu_B(x), \nu_A(x)+\varepsilon), \max(\nu_A(x), \mu_B(x)-\eta) + \max(\nu_B(x), \mu_A(x)-\eta) - \max(\nu_A(x), \mu_B(x)-\eta).\max(\nu_B(x), \mu_A(x)-\eta)\rangle.$$

Second, we will introduce the following five (ε,η)-distances between the values of two elements $x, y \in E$ about the IFS A, by analogy with the first (ε,η)-norm:

$$d'_{\varepsilon,\eta;str\_opt}(A)(x,y) = \langle \min(\mu_A(x), \nu_A(y)+\varepsilon) + \min(\mu_A(y), \nu_A(x)+\varepsilon) - \min(\mu_A(x), \nu_A(y)+\varepsilon).\min(\mu_A(y), \nu_A(x)+\varepsilon), \max(\nu_A(x), \mu_A(y)-\eta).\max(\nu_A(y), \mu_A(x)-\eta)\rangle,$$

$$d'_{\varepsilon,\eta;opt}(A)(x,y) = \langle \max(\min(\mu_A(x), \nu_A(y)+\varepsilon), \min(\mu_A(y), \nu_A(x)+\varepsilon)), \min(\max(\nu_A(x), \mu_A(y)-\eta), \max(\nu_A(y), \mu_A(x)-\eta))\rangle,$$

$$d'_{\varepsilon,\eta;aver}(A)(x,y) = \langle \frac{\min(\mu_A(x), \nu_A(y)+\varepsilon) + \min(\mu_A(y), \nu_A(x)+\varepsilon)}{2}, \frac{\max(\nu_A(x), \mu_A(y)-\eta) + \max(\nu_A(y), \mu_A(x)-\eta)}{2}\rangle,$$

$$d'_{\varepsilon,\eta;pes}(A)(x,y) = \langle \min(\min(\mu_A(x), \nu_A(y)+\varepsilon), \min(\mu_A(y), \nu_A(x)+\varepsilon)), \max(\max(\nu_A(x), \mu_A(y)-\eta), \max(\nu_A(y), \mu_A(x)-\eta))\rangle,$$

$$d'_{\varepsilon,\eta;str\_pes}(A)(x,y) = \langle \min(\mu_A(x), \nu_A(y)+\varepsilon).\min(\mu_A(y), \nu_A(x)+\varepsilon), \max(\nu_A(x), \mu_A(y)-\eta) + \max(\nu_A(y), \mu_A(x)-\eta) - \max(\nu_A(x), \mu_A(y)-\eta).\max(\nu_A(y), \mu_A(x)-\eta)\rangle.$$

Theorem 2. The ten (ε,η)-distances are intuitionistic fuzzy pairs.

The proof is similar to the above one.

As illustration, we will calculate the distances between the pairs (e^*,o^*), (e^*,u^*) and (u^*,o^*).

	(e^*,o^*)	(e^*,u^*)	(o^*,u^*)
$d'_{\varepsilon,\eta;str\_opt}(A)$	$\langle 1,0\rangle$	$\langle 0,1\rangle$	$\langle 0,0\rangle$
$d'_{\varepsilon,\eta;aver}(A)$	$\langle 1,0\rangle$	$\langle 0,1\rangle$	$\langle 0,0\rangle$
$d'_{\varepsilon,\eta;opt}(A)$	$\langle \frac{1}{2},\frac{1}{2}\rangle$	$\langle 0,1\rangle$	$\langle 0,\frac{1}{2}\rangle$
$d'_{\varepsilon,\eta;pes}(A)$	$\langle 0,1\rangle$	$\langle 0,1\rangle$	$\langle 0,1\rangle$
$d'_{\varepsilon,\eta;str\_pes}(A)$	$\langle 0,1\rangle$	$\langle 0,1\rangle$	$\langle 0,1\rangle$

Third, we will introduce the following five (ε,η)-distances between the values of two elements $x,y\in E$ about the IFS A, by analogy with the second (ε,η)-norm:

$$d''_{\varepsilon,\eta;str\_opt}(A,B)(x)=\langle \min(\max(\varepsilon,\mu_A(x)-\eta+\varepsilon),\nu_B(x)+\varepsilon)$$
$$+\min(\max(\varepsilon,\mu_B(x)-\eta+\varepsilon),\nu_A(x)+\varepsilon)-\min(\max(\varepsilon,\mu_A(x)-\eta+\varepsilon),\nu_B(x)+\varepsilon)$$
$$.\min(\max(\varepsilon,\mu_B(x)-\eta+\varepsilon),\nu_A(x)+\varepsilon),\max(0,\min(1-\eta,\nu_A(x)+\varepsilon-\eta),\mu_B(x)-\eta)$$
$$.\max(0,\min(1-\eta,\nu_B(x)+\varepsilon-\eta),\mu_A(x)-\eta)\rangle,$$
$$d''_{\varepsilon,\eta;opt}(A,B)(x)=\langle \max(\min(\max(\varepsilon,\mu_A(x)-\eta+\varepsilon),\nu_B(x)+\varepsilon),$$
$$\min(\max(\varepsilon,\mu_B(x)-\eta+\varepsilon),\nu_A(x)+\varepsilon)),$$
$$\min(\max(0,\min(1-\eta,\nu_A(x)+\varepsilon-\eta),\mu_B(x)-\eta),$$
$$\max(0,\min(1-\eta,\nu_B(x)+\varepsilon-\eta),\mu_A(x)-\eta))\rangle,$$
$$d''_{\varepsilon,\eta;aver}(A,B)(x)=\langle \frac{1}{2}.(\min(\max(\varepsilon,\mu_A(x)-\eta+\varepsilon),\nu_B(x)+\varepsilon)$$
$$+\min(\max(\varepsilon,\mu_B(x)-\eta+\varepsilon),\nu_A(x)+\varepsilon)),$$
$$\frac{1}{2}.(\max(0,\min(1-\eta,\nu_A(x)+\varepsilon-\eta),\mu_B(x)-\eta)$$
$$+\max(0,\min(1-\eta,\nu_B(x)+\varepsilon-\eta),\mu_A(x)-\eta))\rangle,$$
$$d''_{\varepsilon,\eta;pes}(A,B)(x)=\langle \min(\max(\varepsilon,\mu_A(x)-\eta+\varepsilon,\nu_B(x)+\varepsilon),$$
$$\max(\varepsilon,\mu_B(x)-\eta+\varepsilon),\nu_A(x)+\varepsilon),$$
$$\max(0,\min(1-\eta,\nu_A(x)+\varepsilon-\eta),\mu_B(x)-\eta,\min(1-\eta,\nu_B(x)+\varepsilon-\eta),\mu_A(x)-\eta)\rangle,$$
$$d''_{\varepsilon,\eta;str\_pes}(A,B)(x)=\langle \min(\max(\varepsilon,\mu_A(x)-\eta+\varepsilon),\nu_B(x)+\varepsilon).$$
$$\min(\max(\varepsilon,\mu_B(x)-\eta+\varepsilon),\nu_A(x)+\varepsilon),$$
$$\max(0,\min(1-\eta,\nu_A(x)+\varepsilon-\eta),\mu_B(x)-\eta)$$
$$+\max(0,\min(1-\eta,\nu_B(x)+\varepsilon-\eta),\mu_A(x)-\eta)$$
$$-\max(0,\min(1-\eta,\nu_A(x)+\varepsilon-\eta),\mu_B(x)-\eta)$$
$$.\max(0,\min(1-\eta,\nu_B(x)+\varepsilon-\eta),\mu_A(x)-\eta)\rangle.$$

Fourth, we will introduce the following five (ε, η)-distances between the values of two elements $x, y \in E$ about the IFS A, by analogy with the second (ε, η)-norm:

$$d''_{\varepsilon,\eta;str\_opt}(A)(x,y) = \langle \min(\max(\varepsilon, \mu_A(x) - \eta + \varepsilon), \nu_A(y) + \varepsilon)$$
$$+ \min(\max(\varepsilon, \mu_A(y) - \eta + \varepsilon), \nu_A(x) + \varepsilon)$$
$$- \min(\max(\varepsilon, \mu_A(x) - \eta + \varepsilon), \nu_A(y) + \varepsilon)$$
$$. \min(\max(\varepsilon, \mu_A(y) - \eta + \varepsilon), \nu_A(x) + \varepsilon),$$
$$\max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta), \mu_A(y) - \eta)$$
$$. \max(0, \min(1 - \eta, \nu_A(y) + \varepsilon - \eta), \mu_A(x) - \eta)\rangle,$$
$$d''_{\varepsilon,\eta;opt}(A)(x,y) = \langle \max(\min(\max(\varepsilon, \mu_A(x) - \eta + \varepsilon), \nu_A(y) + \varepsilon),$$
$$\min(\max(\varepsilon, \mu_A(y) - \eta + \varepsilon), \nu_A(x) + \varepsilon)),$$
$$\min(\max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta), \mu_A(y) - \eta),$$
$$\max(0, \min(1 - \eta, \nu_A(y) + \varepsilon - \eta), \mu_A(x) - \eta))\rangle,$$
$$d''_{\varepsilon,\eta;aver}(A)(x,y) = \langle \frac{1}{2}.(\min(\max(\varepsilon, \mu_A(x) - \eta + \varepsilon), \nu_A(y) + \varepsilon)$$
$$+ \min(\max(\varepsilon, \mu_A(y) - \eta + \varepsilon), \nu_A(x) + \varepsilon)),$$
$$\frac{1}{2}.(\max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta), \mu_A(y) - \eta)$$
$$+ \max(0, \min(1 - \eta, \nu_A(y) + \varepsilon - \eta), \mu_A(x) - \eta))\rangle,$$
$$d''_{\varepsilon,\eta;pes}(A)(x,y) = \langle \min(\max(\varepsilon, \mu_A(x) - \eta + \varepsilon, \nu_A(y) + \varepsilon),$$
$$\max(\varepsilon, \mu_A(y) - \eta + \varepsilon), \nu_A(x) + \varepsilon),$$
$$\max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta), \mu_A(y) - \eta,$$
$$\min(1 - \eta, \nu_A(y) + \varepsilon - \eta), \mu_A(x) - \eta)\rangle,$$
$$d''_{\varepsilon,\eta;str\_pes}(A)(x,y) = \langle \min(\max(\varepsilon, \mu_A(x) - \eta + \varepsilon), \nu_A(y) + \varepsilon)$$
$$. \min(\max(\varepsilon, \mu_A(y) - \eta + \varepsilon), \nu_A(x) + \varepsilon),$$
$$\max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta), \mu_A(y) - \eta)$$
$$+ \max(0, \min(1 - \eta, \nu_A(y) + \varepsilon - \eta), \mu_A(x) - \eta)$$
$$- \max(0, \min(1 - \eta, \nu_A(x) + \varepsilon - \eta), \mu_A(y) - \eta)$$
$$. \max(0, \min(1 - \eta, \nu_A(y) + \varepsilon - \eta), \mu_A(x) - \eta)\rangle.$$

Theorem 3. The last ten (ε, η)-distances are intuitionistic fuzzy pairs.

The proof is similar to the above one.

2.3 Now, we will discuss new distances between two given IFSs A and B. Up to now, they have also been real numbers. Here, for the first time, we will introduce a whole IFS, representing the distances between the origins of each element $x \in E$ with respect to the two sets. This set-form of distances can have different forms, but we will describe the following ten of them:

$$D'(A,B)_{\varepsilon,\eta;type} = \{\langle x, \mu_{d'(A,B)_{\varepsilon,\eta;type}(x)}, \nu_{d'(A,B)_{\varepsilon,\eta;type}(x)}\rangle | x \in E\},$$

$$D''(A,B)_{\varepsilon,\eta;type} = \{\langle x, \mu_{d''(A,B)_{\varepsilon,\eta;type}(x)}, \nu_{d''(A,B)_{\varepsilon,\eta;type}(x)}\rangle | x \in E\},$$

where

$$\langle \mu_{d'(A,B)_{\varepsilon,\eta;type}(x)}, \nu_{d'(A,B)_{\varepsilon,\eta;type}(x)}\rangle = d'_{\varepsilon,\eta;type}(A,B)(x),$$

$$\langle \mu_{d''(A,B)_{\varepsilon,\eta;type}(x)}, \nu_{d''(A,B)_{\varepsilon,\eta;type}(x)}\rangle = d''_{\varepsilon,\eta;type}(A,B)(x)$$

for "$type$" $\in$ {"str_opt", "opt", "$aver$", "pes", "str_pes"}.

The so constructed sets are IFSs and the proof of this fact is similar to the above proof. Now, we can introduce the numerical form of the distances between IFSs A and B by analogy with ordinary intuitionistic fuzzy distances (see, e.g., [2]).

3 Conclusion

The so constructed norms and distances generalize the norms and distances from [2]. The latter are obtained from the above ones in the case when $\varepsilon = \eta = 0$. On the other hand, the new objects generate different problems, e.g., for construction of their integral representations. The idea for set-forms of the distances can be transformed for all other forms of the intuitionistic fuzzy norms and distances, discussed in the literature.

On the other hand, the above idea was generated in respect of author's research on Conway's Game of Life (see, e.g., [12]). It is a popular zero-player game, devised by John Horton Conway in 1970, and it is the best-known example of a cellular automaton. Its "universe" is an infinite two-dimensional orthogonal grid of square cells, each of which is in one of two possible states, live or dead. Every cell interacts with its eight neighbours, which are the cells that are directly horizontally, vertically, or diagonally adjacent. In a stepwise manner, the state of each cell in the grid preserves or alternates with respect to a given list of rules.

In future we shall propose an intuitionistic fuzzy estimation of the cells' state in a modified Game of Life. For each cell we can define its IF estimation as a pair consisting of the degrees l_p and l_a, namely degrees of presence and absence of life, where $l_p + l_a \leq 1$. In the classical Conway's Game of Life, the live and dead states correspond to the elementary IF estimations $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$. In a future research, using the above formulas for norms and distances, we can calculate the IF state of liveliness of each cell, as functions of the current states of the cell's neighbours. Criteria of liveliness will also be determined in terms of IFS.

Acknowledgements

The author is very thankful to Prof. Janusz Kacprzyk for his advices for improving of present paper.

The author is grateful for the support provided by the projects DID-02-29 "Modelling processes with fixed development rules" and BIn-2/09 "Design and development of intuitionistic fuzzy logic tools in information technologies" funded by the National Science Fund, Bulgarian Ministry of Education, Youth, and Science.

References

1. Atanassov, K.: Two variants of intuitionistic fuzzy propositional calculus, Preprint IM-MFAIS-5-88, Sofia (1988)
2. Atanassov, K.: Intuitionistic Fuzzy Sets. Springer, Berlin (1999)
3. Atanassov, K.: On intuitionistic fuzzy negations and De Morgan Laws. In: Proc. of Eleventh International Conf., IPMU 2006, Paris, July 2-7, pp. 2399–2404 (2006)
4. Atanassov, K.: On some intuitionistic fuzzy implication. Comptes Rendus de l'Academie bulgare des Sciences 59(1), 19–24 (2006)
5. Atanassov, K.: Intuitionistic fuzzy implication $\rightarrow^{\varepsilon,\eta}$ and intuitionistic fuzzy negation $\neg^{\varepsilon,\eta}$. Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics 1, 1–10 (2008)
6. Atanassov, K.: 25 years of intuitionistic fuzzy sets or the most interesting results and the most important mistakes of mine. In: Advances in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics. I: Foundations, pp. 1–35. Academic Publishing House EXIT, Warszawa (2008)
7. Atanassov, K.: Remark on operations "subtraction" over intuitionistic fuzzy sets. Notes on Intuitionistic Fuzzy Sets 15(3), 24–29 (2009), http://ifigenia.org/wiki/issue:nifs/15/3/24-29
8. Atanassov, K.: Intuitionistic fuzzy subtractions $-'^{\varepsilon,\eta}$ and $-''^{\varepsilon,\eta}$. In: Eight Polish Conf. on Intuitionistic Fuzzy Sets and Generalized Nets, Warsaw, October 16 (in press, 2009)
9. Atanassov, K., Dimitrov, D.: On the negations over intuitionistic fuzzy sets. Part 1. Annual of "Informatics" Section Union of Scientists in Bulgaria 1, 49–58 (2008), http://ifigenia.org/wiki/issue:usb-2008-1-49-58
10. Ban, A.: Intuitionistic Fuzzy Measures. Theory and Applications. Nova Sci. Publishers, New York (2006)
11. Chen, T.-Y.: A note on distances between intuitionistic fuzzy sets and/or interval valued fuzzy sets based on the Hausdorff metric. Fuzzy Sets and Systems 158(22), 2523–2525 (2007)
12. Gardner, M.: Mathematical Games: The fantastic combinations of John Conway's new solitaire game "Life". Scientific American 223, 120–123 (1970)
13. Grzegorzewski, P.: Distances and orderings in a family of intuitionistic fuzzy numbers. In: Proc. of the Third Conf. of the European Society for Fuzzy Logic and Technology, EUSFLAT 2003, Zittau, September 10-12, pp. 223–227 (2003)
14. Grzegorzewski, P.: The Hamming distance between intuitionistic fuzzy sets. In: De Baets, B., Kaynak, O., Bilgiç, T. (eds.) IFSA 2003. LNCS (LNAI), vol. 2715, pp. 35–38. Springer, Heidelberg (2003)

15. Grzegorzewski, P.: On measuring distances between intuitionistic fuzzy sets. In: Atanassov, K., Kacprzyk, J., Krawczak, M. (eds.) Issues in Intuitionistic Fuzzy Sets and Generalized Nets, vol. 2, pp. 107–115. Wydawnictwo WSISiZ, Warszawa (2004)
16. Guo, X.J.: Li: Entropy, distance measure and similarity measure of intuitionistic fuzzy sets and their relations. Mathematics in Practice and Theory 37(4), 109–113 (2007)
17. Klir, G.: Bo Yuan: Fuzzy Sets and Fuzzy Logic. Prentice Hall, New Jersey (1995)
18. Rasiova, H., Sikorski, R.: The Mathematics of Metamathematics. Warszawa, Pol. Acad. of Sci. (1963)
19. Riečan, B., Boyadzhieva, D., Atanassov, K.: On intuitionistic fuzzy subtraction related to intuitionistic fuzzy negation $\neg_{11}$. Notes on Intuitionistic Fuzzy Sets 15(4), 1–6 (2009), http://ifigenia.org/wiki/issue:nifs/15/4/1-6
20. Song, Y., Zhang, Q., Zhou, X.: Supplier selection model based on distance measure between intuitionistic fuzzy sets. In: Proc. of IEEE International Conference on Systems, Man and Cybernetics, vol. 5, pp. 3795–3799
21. Szmidt, E., Kacprzyk, J.: Distances between intuitionistic fuzzy sets. Fuzzy Sets and Systems 114(4), 505–518 (2000)
22. Szmidt, E., Kacprzyk, J.: Distance from consensus under intuitionistic fuzzy preferences. In: Proceedings of EUROFUSE Workshop on Preference Modelling and Applications, EUROFUSE-PM 2001, Granada, April 25-27 (2001)
23. Szmidt, E., Kacprzyk, J.: Analysis of agreements in a group of experts via distances between intuitionistic fuzzy preferences. In: Proc. of the Ninth International Conf., IPMU 2002, Annecy, France, July 1-5, vol. III, pp. 1859–1865 (2002)
24. Szmidt, E., Kacprzyk, J.: Group agreement analysis via distances and entropy of intuitionistic fuzzy sets. In: Proceedings of Int. Conf. on Fuzzy Information Processing: Theories and Applications, Beijing, China, March 1-4, vol. I, pp. 631–635 (2003)
25. Szmidt, E., Kacprzyk, J.: Distances between intuitionistic fuzzy sets and their applications in reasoning. In: Halgamude, S., Wang, L. (eds.) Computational Intelligence for Modelling and Prediction, pp. 101–116. Springer, Berlin (2005)
26. Szmidt, E., Kacprzyk, J.: Distances between intuitionistic fuzzy sets: straightforward approaches may not work. In: Proc. of 3rd Int. IEEE Conf. "Intelligent Systems" IS 2006, London, September 4-6, pp. 716–721 (2006)
27. Tan, C., Zhang, Q.: Aggregation of opinion in group decision making based on intuitionistic fuzzy distances. Mathematics in Practice and Theory 36(2), 119–124 (2006) (in Chinese)
28. Wagn, W., Xin, X.: Distance measure between intuitionistic fuzzy sets. Pattern Recognition Letters 26(13), 2063–2069 (2005)
29. Yang, Y., Chiclana, F.: Intuitionistic fuzzy sets: Spherical representation and distances. Int. J. of Intelligent Systems 24, 399–420 (2009)

Atanassov's Intuitionistic Contractive Fuzzy Negations

Benjamin Bedregal[1], Humberto Bustince[2], Javier Fernandez[2], Glad Deschrijver[3], and Radko Mesiar[4]

[1] Departamento de Informática e Matemática Aplicada
Universidade Federal do Rio Grande do Norte
Brazil
bedregal@dimap.ufrn.br
[2] Departamento Automtica y Computacin
Universidad Publica de Navarra
{bustince,fcojavier.fernandez}@unavarra.es
[3] Department of Applied Mathematics and Computer Science
Ghent University
Glad.Deschrijver@ugent.be
[4] Department of Mathematics and Descriptive Geometry
Slovak University of Technology
mesiar@math.sk

Abstract. In this work we consider the concept of contractive Atanassov's intuitionistic mapping. In particular, we show that with our definitions, the only strict (strong) contractive negation is the standard one.

Keywords: Atanassov's intuitionistic fuzzy sets, fuzzy negation, Lipschitz property.

1 Introduction

Fuzzy sets theory, first introduced by Zadeh in [13], has shown itself a very valuable tool for problems that, by its very own nature, have to deal with imprecision, ignorance or vagueness. However, the definition of fuzzy set does not properly enough take into account this vagueness, since it imposes a single numerical value to measure the membership of an element to a given set. In this sense, Atanassov's intuitionistic fuzzy sets theory ([2,3]) makes far more richer the scope of the theory, since it allows to consider a second value, the so called non-membership value, which clarifies the information the expert is providing about the element and the set in consideration.

On the other hand, Lipschitzicianity is a very widely used concept in mathematical analysis. basically, it imposes a restriction in the way a function can increase or decrease, and it occurs in fields as different as topology (fixed point maps) or the study of ordinary differential equations ([9]).

In this work we intend to bring together this two concepts, by translating into the Atanassov's intuitionistic fuzzy sets setting the idea of 1-Lipschitzicianity.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 591–600, 2010.

In particular, this has led us to consider the concept of *contractive* negations and Atanassov's intuitionistic aggregation functions, as those functions such that, when applied to an Atanassov's intuitionistic value, give raise to another Atanassov's intuitionistic value such that its Atanassov's intuitionistic index is less than or to the Atanassov's intuitionistic index of the original pair. Since Atanassov's intuitionistic index can be used as a measure of how far a given Atanassov's intuitionistic set is from being an ordinary fuzzy set, by considering this kind of functions we are ensuring that "fuzziness" does not decrease. But this decreasing can also be understood as a reduction of the lack of knowledge on an expert when providing the membership and non-membership values, see [7].

Nevertheless, we want to stress that this work is only a first step in a very large field of possible research. For this reason, we have focused on negations, since they can be used as token for future developments of the theory, without involving too complicated issues.

The structure of this paper is the following. In the next section we present some preliminary results. In Section 3 we recall the concept of Lipschitz function. In Section 4 we deal with Atanassov's intuitionistic fuzzy negations. Sectio 5 is the core of this work, focusing on contractive Atanassov's intuitionistic fuzzy negations. We end with some conclusions and remarks.

2 Preliminaries

In this section we recall the main concepts about Atanassov's intuitionistic fuzzy sets. For us, an Atanassov's intuitionistic fuzzy set (AIFS) over a finite referential X is a set

$$A = \{(x, \mu_A(x), \nu_A(x))|\ x \in X\}$$

where $\mu_A, \nu_A : X \to [0,1]$ are called, respectively, the membership and non-membership functions and have to satisfy the inequality

$$\mu_A(x) + \nu_A(x) \leq 1$$

for all $x \in X$. In order to avoid notational problems, we will call Atanassov's intuitionistic pair any pair of numbers, denoted by $\langle \mu, \nu \rangle$, such that $\mu + \nu \leq 1$, with $\mu, \nu \in [0,1]$. We denote by $\mathcal{A}IFS$ the set of all Atanassov's intuitionistic pairs $P = \langle \mu_P, \nu_P \rangle$. Observe that any operation that we define over Atanassov's intuitionistic pairs can be naturally extended to act over Atanassov's intuitionistic fuzzy sets. In particular, given an Atanassov's intuitionistic pair $\langle \mu, \nu \rangle$, we can define its Atanassov's intuitionistic index as

$$\pi(\langle \mu, \nu \rangle) = 1 - \mu - \nu\ .$$

Over $\mathcal{A}IFS$ we can consider two different orderings. The first one is defined as $P \leq_L Q$ if and only if $\mu_P \leq \mu_Q$ and $\nu_P \geq \nu_Q$. The second order relation is $P \subseteq_L Q$ if and only if $\mu_P \geq \mu_Q$ and $\nu_P \geq \nu_Q$ (see [8] for more details).

Observe that none of the two orderings is complete.

3 Lipschitz Functions

We recall here the mathematical concepts of Lipschitz function, as well as some properties that will be of interest for us.

Definition 1. *A function $f : [0,1] \to [0,1]$ is called a Lipschitz mapping if there exists $K > 0$ such that, for all $x, y \in [0,1]$, the inequality*

$$|f(x) - f(y)| \leq K|x - y|$$

holds.

The smallest of such K is called the Lipschitz constant of the mapping f. If K is the Lipschitz constant of the mapping f, then f is also called a K-Lipschitz mapping. In particular, 1-Lipschitz mappings are also called short maps.

Observe that K-Lipschitzicianity does not allow f to increase or decrease faster than a given rate, which is defined by the constant K. In this sense, it is clear that any K-Lipschitz mapping f is also continuous. Moreover, it can be seen that a K-Lipschitz mapping is differentiable almost everywhere with respect to the Lebesgue measure. Also observe that if f is a K_1-Lipschitz function and g is a K_2-Lipschitz function, then $f \circ g$ (and $g \circ f$) are K_3-Lipschitz (K_4-Lipschitz) functions with $K_3(K_4) \leq K_1 K_2$.

For our following developments, we focus in bijective Lipschitz functions. In this sense, we start by introducing the concept of automorphism on the unit interval.

Definition 2. *[6] A mapping $\varphi : [0,1] \to [0,1]$ is an automorphism if it is strictly increasing and bijective. We denote by $Aut([0,1])$ the set of all automorphisms over $[0,1]$.*

Notice that for any automorphism φ the identities $\varphi(0) = 0$ and $\varphi(1) = 1$ hold. Observe also that any automorphism is in particular continuous, but not necessarily K-Lipschitz, as the family of automorphisms $\varphi(x) = x^p$ with $p < 1$ shows. On the other hand, we have the following result.

Proposition 1. *Let $\varphi \in Aut([0,1])$ be a K-Lipschitz automorphism. Then $K \geq 1$.*

Proof. By definition

$$|\varphi(1) - \varphi(0)| = 1 = 1 - 0$$

so the result is clear □

Observe that the only important point for the proof is that $\varphi(1) = 1$ and $\varphi(0) = 0$, regardless which the other values of φ are. In fact, not even monotonicity was necessary.

Example 1. Each of the automorphisms $\varphi_p(x) = x^p$ with $p \geq 1$ is p-Lipschitz. To see it, first of all notice that, from the mean value theorem, if $x > y$

$$\varphi_p(x) - \varphi_p(y) = \varphi_p'(c)(x - y) \leq p(x - y)$$

for some $c \in (x, y)$. So φ_p is Lipschitz with Lipschitz constant greater than or equal to p. On the other hand, $1 - x^p - p(1 - x) \to 0$ if $x \to 1$, so the p-Lipschitzicianity follows.

Moreover, 1-Lipschitzicianity completely determines an automorphism, as the next result shows.

Proposition 2. *An automorphism $\varphi \in Aut([0, 1])$ is 1-Lipschitz if and only if $\varphi(x) = x$ for all $x \in [0, 1]$.*

Proof. From the 1-Lipschitzicianity of φ, we have, on one hand, that

$$\varphi(x) = |\varphi(x) - \varphi(0)| \leq |x - 0| = x$$

whereas on the other hand

$$1 - \varphi(x) = |\varphi(1) - \varphi(x)| \leq |1 - x| = 1 - x$$

so we have that $\varphi(x) \geq x$, and the result follows from both inequalities. □

Remark. There is not a similar uniqueness result for K-Lipschitz automorphisms with $K > 1$. To see it, fix $K > 1$, and $s \in]0, 1/K[$. Then, the next automorphisms are K-Lipschitz:

$$\phi_{s,1}(x) = \min(Kx, x\frac{1 - Ks}{1 - s} + \frac{Ks - s}{1 - s})$$

and

$$\phi_{s,2}(x) = \max(Kx - K + 1, (1 - Ks)x) \ .$$

Note that $\sup\{\phi_{s,1}(x) | s \in]0, 1/K[\} = \min(Kx, 1)$ is the upper bound of all K-Lipschitz automorphisms *phi* from $Aut([0, 1])$. Nevertheless, it is not strictly monotone and thus not an automorphism. On the other hand, $\inf\{\phi_{s,2}(x) | s \in]0, 1/K[\} = \max(0, Kx - K + 1)$ is the lower bound of all K-Lipschitz automorphisms ϕ from $Aut([0, 1])$. As in the previous case, it is not strictly monotone and thus not an automorphism.

Of course, there is nothing specific from a mathematical point of view in the use of automorphisms. We can obtain a similar general result for any bijective continuous mapping, as the next result shows.

Proposition 3. *Let $f : [a, b] \to [c, d]$ be a bijective K-Lipschitz mapping. Then $K \geq \frac{d-c}{b-a}$ and the only $\frac{d-c}{b-a}$-Lipschitz monotone bijection is*

$$f(x) = \frac{d - c}{b - a}(x - a) + c$$

if f is increasing, or

$$f(x) = \frac{d - c}{b - a}(b - y) + c$$

if f is decreasing.

Proof. Suppose first that f is increasing. Define the mapping $g : [0,1] \to [0,1]$ as

$$g(x) = \frac{f((b-a)x + a) - c}{d - c}$$

We have that

$$\begin{aligned} |g(x) - g(y)| &\leq |\frac{f((b-a)x + a) - c}{d - c} - \frac{f((b-a)y + a) - c}{d - c}| \\ &= |\frac{f((b-a)x + a)}{d - c} - \frac{f((b-a)y + a)}{d - c}| \end{aligned}$$

and, since f is K-Lipschitz, this is smaller than or equal to

$$\frac{K}{d-c}|(b-a)x + a - (b-a)y - a| = K\frac{b-a}{d-c}|x - y|$$

so g is a Lipschitz automorphism. From Proposition 1, it follows that

$$K\frac{b-a}{d-c} \geq 1$$

or equivalently

$$K \geq \frac{d-c}{b-a}$$

as we intended to prove. If $K = \frac{d-c}{b-a}$, it follows from Proposition 2 that $g(x)$, and by clearing f in the definition of g, the result follows.

Finally, if f is decreasing, then the mapping $h(x) = f(b+a-x)$ is increasing, and the results follows from the calculations for the increasing case. □

3.1 Lipschitz Fuzzy Negations

In this section we analyze the relation between the Lipschitz property and the concept of fuzzy negation. Further considerations on the subject, as well as related developments, can be found in [10].

We start recalling the concept of (fuzzy) negation.

Definition 3. *A fuzzy negation is a nonincreasing mapping $N : [0,1] \to [0,1]$ such that $N(0) = 1$ and $N(1) = 0$. If N is continuous and strictly decreasing, i.e., if $x < y$ implies that $N(x) > N(y)$, then N is called a strict negation. A strict negation N which is involutive (i.e, such that $N(N(x)) = N(x)$ for all $x \in [0,1]$) is called a strong negation.*

The most representative example of negation is the so-called standard (or Zadeh's) negation $N_Z(x) = 1 - x$.

As a first result, we show that there are not purely contractive fuzzy negations, i.e., K-Lipschitz fuzzy negations with $K < 1$.

Proposition 4. *Let N be a K-Lipschitz fuzzy negation. Then $K \geq 1$.*

Proof. The proof runs similarly to that of Proposition 1, just recalling the remark after that result. □

Our aim is to see that the only 1-Lipschitz negation is Zadeh's negation. From Proposition 3 we have the following result.

Corollary 1. *Let N be a strict negation. Then N is 1-Lipschitz if and only if $N(x) = 1 - x$ for all $x \in [0, 1]$.*

Proof. Just observe that, if N is a strict negation, it is by definition a continuous, strictly decreasing bijection from $[0, 1]$ to $[0, 1]$. So, by Proposition 3, it follows that

$$N(x) = 1 - x$$

for all $x \in [0, 1]$ □

Now we want to drop out strictness. This can be done as follows.

Theorem 1. *Let N be a 1-Lipschitz negation. Then $N(x) = 1 - x$ for all $x \in [0, 1]$.*

Proof. Since N is 1-Lipschitz, we have that, for any $x \in [0, 1]$

$$1 - N(x) = N(0) - N(x) \leq x$$

so $N(x) \geq 1 - x$. Analogously

$$N(x) - 0 = N(x) - N(1) \leq 1 - x$$

so $N(x) \leq 1 - x$. The result follows. □

4 Atanassov's Intuitionistic Fuzzy Negations

Let $f : [0, 1] \to [0, 1]$ be a mapping. Then, the mapping $\hat{f} : \mathcal{AIFS} \to \mathcal{AIFS}$ given by

$$\hat{f}(P) = \{(x, \mu_{f(P)}, \nu_{f(P)})\}$$

with

$$\mu_{f(P)} = \inf\{f(t) |\ \mu_P \leq t \leq 1 - \nu_P\}$$

and

$$\nu_{f(P)} = 1 - \sup\{f(t) |\ \mu_P \leq t \leq 1 - \nu_P\}$$

is well defined.

In this paper we consider the following notions of continuity on $L([0, 1])$.

(i) Moore continuity [11]. It is defined by considering the metric given by the distance between two AIFSs $P, Q \in \mathcal{AIFS}$, which is defined by: $d_M(P, Q) = \max\{|\mu_P - \mu_Q|, |\nu_P - \nu_Q|\}$.

(ii) Scott continuity. It is defined considering the quasi-metric $q_S(P,Q) = \max\{\mu_P - \mu_Q, \nu_P - \nu_Q, 0\}$. It was introduced in [12] and [1].

The main result of [12] can be adapted to our setting as follows.

Theorem 2. *Let $f : [0,1] \to [0,1]$ be a mapping. Then, the following items are equivalent.*

(i) f is continuous;
(ii) $\hat{f}$ is Moore continuous;
(iii) $\hat{f}$ is Scott continuous.

By analogy to fuzzy negations, negations for AIFS can be defined as follows.

Definition 4. *A mapping $\mathbf{N} : \mathcal{AIFS} \to \mathcal{AIFS}$ is an AIFS negation if, for all $A, B \in \mathcal{AIFS}$, the following properties hold.*

(**N**1) $\mathbf{N}(0_L) = 1_L$ *and* $\mathbf{N}(1_L) = 0_L$, *where* $0_L = \langle 0,1\rangle$ *and* $1_L = \langle 1,0\rangle$;
(**N**2*a*) *If* $P \leq_L Q$, *then* $\mathbf{N}(Q) \leq_L \mathbf{N}(P)$;
(**N**2*b*) *If* $P \subseteq_L Q$, *then* $\mathbf{N}(P) \subseteq_L \mathbf{N}(Q)$;

If **N** *also satisfies the involutive property*

(**N**3) $\mathbf{N}(\mathbf{N}(P)) = P$ *for all* $P \in \mathcal{AIFS}$

then **N** *is said to be a strong Atanassov's intuitionistic fuzzy negation.*

A Moore (or Scott) continuous Atanassov's intuitionistic fuzzy negation is said to be strict if it also satisfies the following properties.

(**N**4*a*) *If* $P <_L Q$ *then* $\mathbf{N}(Q) <_L \mathbf{N}(P)$;
(**N**4*b*) *If* $P \subset_L Q$ *then* $\mathbf{N}(P) \subset_L \mathbf{N}(Q)$.

The proofs for the following propositions in this section are very similar to those in [4].

Proposition 5. *Let $\mathbf{N} : \mathcal{AIFS} \to \mathcal{AIFS}$ be a mapping. Given $P \in \mathcal{AIFS}$ we write $\mathbf{N}(P) = \langle \mu_{N(P)}, \nu_{N(P)}\rangle$. Then* **N** *is an (strict) Atanassov's intuitionistic fuzzy negation if and only if the mappings*

$$N_\mu(x) = \mu_{N(P)}(\langle x, 1-x\rangle)$$

and

$$N_\nu(x) = 1 - \nu_{N(P)}(\langle x, 1-x\rangle)$$

are (strict) fuzzy negations and

$$\mathbf{N}(\langle x,y\rangle) = \langle N_\mu(1-y), 1 - N_\nu(x)\rangle$$

for all $\langle x,y\rangle \in \mathcal{AIFS}$.

Proposition 6. *Let N_1, N_2 be (strict) fuzzy negations. If $N_1(x) \leq N_2(x)$ for all $x \in [0,1]$, then the mapping $I[N_1,N_2] : \mathcal{AIFS} \to \mathcal{AIFS}$ defined by*

$$I[N_1,N_2](\langle x,y\rangle) = \langle N_1(1-y), 1 - N_2(x)\rangle$$

is an Atanassov's intuitionistic fuzzy negation.

Notice that if $\mathbf{N} = I[N_1,N_2]$, then $N_\mu = N_1$ and $N_\nu = N_2$.

5 Contractive Fuzzy Negations

Indeterminacy index is a measure of how far an Atanassov's intuitionistic fuzzy number is from being a fuzzy set. In the following we are interested in identifying which AIFS negations do not increase the indeterminacy index. So we start formally defining what we are looking for.

Definition 5. *We say that an Atanassov's intuitionistic fuzzy negation* $\mathbf{N}$ *is contractive if for any Atanassov's intuitionistic pair* $\langle \mu(x), \nu(x) \rangle$ *the inequality*

$$\pi(\mathbf{N}(\langle \mu(x), \nu(x) \rangle)) \leq \pi(\langle \mu(x), \nu(x) \rangle)$$

holds.

Example 2. The Atanassov's intuitionistic fuzzy negation

$$\hat{N}_Z(\langle \mu(x), \nu(x) \rangle) = \langle \nu(x), \mu(x) \rangle$$

is clearly a contractive Atanassov's intuitionistic fuzzy negation.

In particular, observe that

$$\pi(\hat{N}_Z(\langle \mu(x), \nu(x) \rangle)) = \pi(\langle \mu(x), \nu(x) \rangle)$$

for any Atanassov's intuitionistic pair $\langle \mu(x), \nu(x) \rangle$. Moreover, we can assert the following.

Proposition 7. *Let* $\mathbf{N}$ *be an Atanassov's intuitionistic fuzzy negation such that*

$$\pi(\mathbf{N}(\langle x, y \rangle)) = \pi(\langle x, y \rangle)$$

for all $\langle x, y \rangle \in \mathcal{AIFS}$. *Then* $\mathbf{N} = \hat{N}_Z$.

Proof. Just observe that, by Proposition 5, there exist fuzzy negations N_μ and N_ν such that

$$\mathbf{N}(\langle x, y \rangle) = \langle N_\mu(1-y), 1 - N_\nu(x) \rangle$$

By hypothesis, we have that

$$N_\nu(x) - N_\mu(1-y) = 1 - x - y$$

for all $\langle x, y \rangle \in \mathcal{AIFS}$. In particular, if we take $y = 0$, we arrive at $N_\nu(x) = 1 - x$ for all $x \in [0, 1]$. Analogously, if $x = 0$ $1 - N_\mu(1-y) = 1 - y$. The result follows □

Our results on 1-Lipschitz negations together with Proposition 7 allow us to characterize contractive Atanassov's intuitionistic fuzzy negations. To start, we have the following result.

Now we have the following result.

Theorem 3. *Let* $\mathbf{N}$ *be an Atanassov's intuitionistic fuzzy negation . If* $\mathbf{N}$ *is contractive then there exists a fuzzy negation* N *such that* $\mathbf{N}(\langle x, y\rangle) = \langle N(1-y), 1-N(x)\rangle$ *for all* $\langle x, y\rangle \in \mathcal{AIFS}$

Proof. From Proposition 6, we know that there exist fuzzy negations N_μ and N_ν such that

$$\mathbf{N}(\langle x, y\rangle) = \langle N_\mu(1-y), 1-N_\nu(x)\rangle \ .$$

Since $\mathbf{N}$ is contractive, since $\pi(\langle x, 1-x\rangle = 0$ we arrive at

$$\pi(\mathbf{N}(\langle x, y\rangle)) = N_\nu(x) - N_\mu(x) \leq 0$$

for all $x \in [0,1]$, so the result follows. □

Notice that the converse of this theorem does not hold. In fact, if we consider the strict negation $N(x) = 1-x^2$, then $\mathbf{N}(\langle x, y\rangle) = \langle 1-(1-y)^2, x^2\rangle$. Since $\mathbf{N}(\langle 0.5, 0.4\rangle) = \langle 0.64, 0.25\rangle$, as $1-0.5-0.4 = 0.1 < 1-0.64-0.25 = 0.11$, it follows that $\mathbf{N}$ is not contractive.

The following corollary follows straightforward from the previous result.

Corollary 2. *An Atanassov's intuitionistic fuzzy negation* $\mathbf{N}$ *is contractive if and only if there exists a* 1*-Lipschitz fuzzy negation* N *such that* $\mathbf{N}(\langle x, y\rangle) = \langle N(1-y), 1-N(x)\rangle$ *for all* $\langle x, y\rangle \in \mathcal{AIFS}$*. In particular, The unique strict contractive Atanassov's intuitionistic fuzzy negation is*

$$\hat{N}_Z(\langle x, y\rangle) = \langle y, x\rangle \ .$$

Proof. From the previous theorem we have that there exists a fuzzy negation N such that

$$\mathbf{N}(\langle x, y\rangle) = \langle N(1-y), 1-N(x)\rangle \ .$$

Since $\mathbf{N}$ is contractive, it follows that

$$N(x) - N(1-y) \leq 1-x-y$$

for all $x, y \in [0,1]$ such that $x+y \leq 1$, or equivalently, $N(x)-N(t) \leq t-x$ for all $x, t \in [0,1]$ such that $x \leq t$. If $x > t$, then we have that $\pi(\mathbf{N}(\langle t, 1-x\rangle) \leq x-t$, or, equivalently, $N(t)-N(x) \leq x-t$. So we have that N is 1-Lipschitz. Since we already know that the only 1-Lipschitz strict negation is the standard one, the result follows. □

6 Conclusions

In this work we have related the concepts of Atanassov's intuitionistic fuzzy negation and the concept of Lipschitz function. We have in particular proved that the only 1-Lipschitz Atanassov's intuitionistic fuzzy negation is given by precisely by the extension of Zadeh's standard negation to the intuitionistic setting.

This work can be understood as a first step of a wider study on aggregation function over Atanassov's intuitionistic fuzzy sets. We hope that this study can be of interest, firstly, to relate analytical concepts with widely used fuzzy concepts; and secondly, in the applied field by providing a basis for the choice of some aggregation functions instead of others, depending on the considered application.

Acknowledgments. B. Bedregal was supported by a grant of the Brazilian funding agency CNPq (308256/2009-3). H. Bustince and J. Fernandez were supported by the project TIN2007-65981 of The Spanish Government. R. Mesiar was supported by grants VEGA 1/0080/10 and APVV-0012-07.

References

1. Acióly, B.M., Bedregal, B.C.: A quasi-metric topology compatible with inclusion monotonicity on interval space. Reliab. Comput. 3(3), 305–313 (1997)
2. Atanassov, K.T.: Intuitionistic fuzzy sets. In: Sgurev, V. (ed.) VII ITKR's Session, Sofia (June 1983) (deposed in Central Science and Technical Library, Bulgarian Academy of Sciences, 1697/84, in Bulgarian)
3. Atanassov, K.T.: Intuitionistic Fuzzy Sets. Fuzzy Sets Syst. 20(1), 87–96 (1986)
4. Bedregal, B.C.: On interval fuzzy negations. Technical report, DIMAP/UFRN (2009), http://www.dimap.ufrn.br/bedregal/main-publications.html
5. Bustince, H., Barrenechea, E., Pagola, M., Fernandez, J.: Interval-valued fuzzy sets constructed from matrices: Application to edge detection. Fuzzy Sets Syst. 160(13), 1819–1840 (2009)
6. Bustince, H., Burillo, P., Soria, F.: Automorphisms, negations and implication operators. Fuzzy Sets Syst. 134(2), 209–229 (2003)
7. Bustince, H., Pagola, M., Barrenechea, E., Fernandez, J., Melo-Pinto, P., Couto, P., Tizhoosh, H.R., Montero, J.: Ignorance functions. An application to the calculation of the threshold in prostate ultrasound images. Fuzzy Sets Syst. 161(1), 20–36 (2010)
8. Deschrijver, G., Kerre, E.: On the relationship between some extensions of fuzzy set theory. Fuzzy Sets Syst. 133(2), 227–235 (2003)
9. Evans, C., Gariepy, R.F.: Measure theory and fine properties of functions. Studies in advanced mathematics. CRC Press, Boca Raton (1992)
10. Kolesarova, A., Mesiar, R.: Lipschitzian De Morgan triplets of fuzzy connectives. Submitted to Inform. Sciences
11. Moore, R.E.: Methods and Applications of Interval Analysis. SIAM, Philadelphia (1979)
12. Santiago, R.H.N., Bedregal, B.C., Acióly, B.M.: Formal aspects of correctness and optimality in interval computations. Form. Asp. Comput. 18(2), 231–243 (2006)
13. Zadeh, L.A.: Fuzzy sets. Inform. Control 8, 338–353 (1965)

Trust Propagation Based on Group Opinion

Anna Stachowiak

Faculty of Mathematics and Computer Science
Adam Mickiewicz University
Umultowska 87, 61-614 Poznań, Poland
aniap@amu.edu.pl

Abstract. Modern technology, especially Internet, allows people to find the resources or knowledge they need by making use of the experiences and opinions of other people. It is easy to collect a vast amount of data, however, a problem of quality and reliability of these data is urgent. Trust networks seem to be the best solution but they need more research and attention. In this paper we join the discussion about trust representation and trust propagation. We follow the idea of modeling trust in gradual and dual form of trust and distrust degrees, using Atanassov's intuitionistic fuzzy sets (IFSs) theory as the basis. Moreover, we introduce a new trust propagation operator based on group opinion and on relative scalar cardinality of IFSs.

Keywords: Trust propagation, intuitionistic fuzzy sets, IFS, relative scalar cardinality of IFS.

1 Introduction

Establishing a trust network between users involved in some business or social activity through the Internet (e-business, recommender systems) contributes to its quality, effectiveness and reliability. A trust network is a special type of social network, where the relation between members is expressed by a trust degree. The notion of trust is hard to formalize, but we agree with the common opinion that trust should be modeled in a gradual way, to convey expressions from natural language like "to trust somebody very much" or "rather do not trust somebody", as the opposite to binary "trust" (fully) or "don't trust" (fully). Moreover, we believe that trust should be modeled in a dual form of trust and distrust. Trust would be defined as ones belief that the other user will behave in a dependable, ethical and honest way, and this belief comes from past experience (this user never/almost never let him down, usually gave useful information etc). On the other hand, distrust is one's belief of the other user lack of competence or even harmful intentions (a degree of distrust may arise if other member behaved in dishonest way, gave wrong information etc). What is important, distrust is not necessarily a simple negation of trust, what leaves a space for expressing a lack of knowledge or hesitation.

Besides the problem of trust modeling the crucial issue for trust network is a mechanism of trust propagation - a way of deriving trust information from a

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 601–610, 2010.

trusted third party. In a large system, it is unlikely for a single user to know all the other members well enough to express the level of trust in them. Instead, the user tries to compute trust information by consulting trusted third parties. This problem is difficult and still not well defined and researched.

This paper focuses mainly on the problem of trust propagation. In Section 2 we review an approach to trust representation based on IFS theory (Atanassov's intuitionistic fuzzy set theory) presented in [2]. In Section 3 we present some existing approaches to trust propagation and propose a new propagation operator, called $Prop_\sigma$, based on group opinion and on relative scalar cardinality of IFS, that can be used to calculate trust score in complex trust networks. Section 4 concludes the paper.

2 IFS in Trust and Distrust Modeling

In the literature we can find different approaches to trust modeling. A probabilistic approach deals with trust in a black or white fashion: a user can either be trusted or not (e.g. [7]). In a gradual approach (e.g. [2], [6], [9], [13]) trust is a matter of degree as opposed to being either right or wrong.

The approach adapted in this paper refers to the concept presented by De Cock et al. in [2]. The authors argued that representing trust as a combination of two values of trust and distrust degrees helps to preserve valuable information about the provenance of trust absence. It enables to differentiate between a situation of distrust (towards a malicious user) and a situation of lack of knowledge (towards an unknown user). This distinction is crucial for proper interpretation and propagation of trust.

The adequate method of modeling trust in such dual form is IFS theory (Atanassov's intuitionistic fuzzy sets theory, see [1]). An IFS $\mathcal{E}$ is a pair of fuzzy sets:

$$\mathcal{E} = (A^+, A^-),$$

where A^+ is a fuzzy set of elements that belong to $\mathcal{E}$, and A^- is a fuzzy set of elements that do not belong to $\mathcal{E}$. This theory, in contrast with fuzzy set theory, incorporates uncertainty about the membership of an element, as A^- is not necessarily a negation of A^+, but $A^- \subset (A^+)^c$ where the complement of fuzzy set A is defined as $A^c(x) = 1 - A(x)$ for each x (for a generalized definition of IFS with a use of a complement generated by an arbitrary strong negation see e.g. [10]). Therefore, the value $1 - A^+(x) - A^-(x)$ reflects uncertainty or hesitation about membership of an element x in IFS $\mathcal{E}$. Similarly, for example due to lack of knowledge, uncertainty is present when specifying trust and distrust degrees.

The following definition formalizes a notion of trust network based on IFSs.

Definition 1. A trust network is a couple (A, R) such that A is a set of sources and R is an $A \times A \to [0,1]^2$ mapping. For every a and b in A:

$$R(a,b) = (R^+(a,b), R^-(a,b))$$

where

- $R(a, b)$ is called the trust score of a in b;
- $R^+(a, b)$ is called the trust degree of a in b;
- $R^-(a, b)$ is called the distrust degree of a in b;
- $1 - R^+(a, b) - R^-(a, b)$ is called the uncertainty or hesitation margin.

Such representation let us express states like:

- source a doesn't know anything about b: $R(a, b) = (0, 0)$ (it neither trusts nor distrusts b); the hesitation is maximal and equal to 1,
- source a fully distrust b: $R(a, b) = (0, 1)$,
- source a has ambivalent attitude towards b: $R(a, b) = (0.5, 0.5)$,
- source a rather trusts b, but at the same time distrust it a bit: $R(a, b) = (0.6, 0.2)$.

3 Trust Propagation Operator

Let us assume that source a expressed its trust in source b: $R(a, b)$, and source b expressed its trust in source c: $R(b, c)$. Now we would like to estimate or predict the value of trust score of a in c, what is called a trust propagation (more precisely, it's a transitive model of trust propagation). We must be aware that propagating trust means in fact guessing a value of trust. This concept is still not clear enough, but we assume, that a propagation operator, used to estimate a new trust value, should depend on the value of the component trust in a monotonic way (the higher the component trust values, the higher the result trust value), but doesn't need to be commutative. Furthermore, this operator should depend on the distance between a and c (the longer the propagation chain, the smaller the final trust value). Finally, the operator should be attack resistant (it should not take into account opinions of false members).

In [13] four propagation operators of transitivity type were introduced. We use the following notation: T denotes an arbitrary t-norm, i.e. an increasing, commutative and associative $[0, 1]^2 \to [0, 1]$ mapping satisfying $T(1, x) = x$ for all $x \in [0, 1]$; S denotes an arbitrary t-conorm, i.e. an increasing, commutative and associative $[0, 1]^2 \to [0, 1]$ mapping satisfying $S(0, x) = x$ for all $x \in [0, 1]$; N denotes a negator, i.e. a decreasing $[0, 1] \to [0, 1]$ mapping satisfying $N(0) = 1$ and $N(1) = 0$. Moreover, we take $R(a, b) = (t_1, d_1)$ and $R(b, c) = (t_2, d_2)$.

Propagation operators proposed in [13] are defined as:

$$Prop_1((t_1, d_1), (t_2, d_2)) = (T(t_1, t_2), T(t_1, d_2))$$
$$Prop_2((t_1, d_1), (t_2, d_2)) = (T(t_1, t_2), T(N(d_1), d_2))$$
$$Prop_3((t_1, d_1), (t_2, d_2)) = (S(T(t_1, t_2), T(d_1, d_2)), S(T(t_1, d_2), T(d_1, t_2)))$$
$$Prop_4((t_1, d_1), (t_2, d_2)) = (T(t_1, t_2), S(T(t_1, d_2), T(d_1, t_2)))$$

3.1 Trust Propagation Operator Based on Group Opinion

In this paper we introduce a new propagation operator using a slightly different perspective. The approach is some extension of transitivity model, that allows

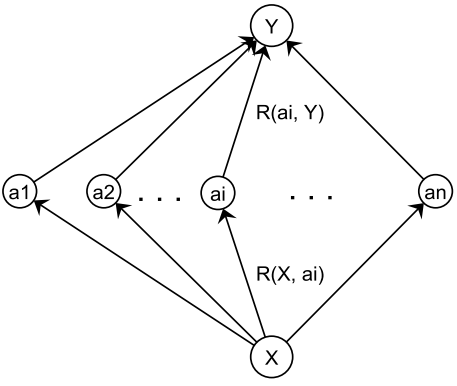


Fig. 1. A model of trust network

more than one intermediate source in trust network. Such extension allows to benefit from the big size of a trust network and variety of opinions. It is reasonable to take into account all available information to estimate a final trust score.

Intuitively, the method we propose resembles a behavior of a person that forms his opinion about somebody by asking "How many of my friends trust this person?". More formally, the simplified model of a trust network that we consider is presented in Fig. 1. A user X wants to estimate a trust score in user Y. X has a set of friends (direct neighbors that he trusts) $a_1, ..., a_m$ and some of them (n of them) have an opinion about Y. So, we may form a question:

"How many sources trusted by X trust Y?" (Q)

and take the answer to this question as a new trust score of X in Y.

We thus construct two IFSs: "sources trusted by X" and "sources that trust Y". Using notation from Fig.1. we define those sets as:

- $S_X : \{R(X, a_i)|i = 1, ..., n\}$ - a set $a_1, ..., a_n$ of sources trusted by X to a degree $R(X, a_1), ..., R(X, a_n)$;
- $S_Y : \{R(a_i, Y)|i = 1, ..., n\}$ - a set $a_1, ..., a_n$ of sources that trust Y to a degree $R(a_1, Y), ..., R(a_n, Y)$.

Let us notice that the question Q is in fact a question about relative scalar cardinality of IFS:

$$\sigma_I(S_Y|S_X) = \frac{\sigma_I(S_Y \cap_{T,S} S_X)}{\sigma_I(S_X)} \tag{1}$$

where:

$$\sigma_I(\mathcal{E}) = [\sigma(A^+), \sigma((A^-)^c)]$$

is a scalar cardinality of IFS $\mathcal{E} = (A^+, A^-)$, and

$$\sigma(A) = \sum_{x \in supp(A)} f(A(x)).$$

is a scalar cardinality of fuzzy set A with f being a non-decreasing function $f : [0,1] \to [0,1]$ such that $f(0) = 0$, $f(1) = 1$ called cardinality pattern. The intersection of two IFSs $\mathcal{E} = (A^+, A^-)$ and $\mathcal{F} = (B^+, B^-)$ is defined as $\mathcal{E} \cap_{T,S} \mathcal{F} = (A^+ \cap_T B^+, A^- \cup_S B^-)$.

Relative scalar cardinality of IFS (1) is thus a proportion of two intervals:

$$\sigma_I(A|B) = \frac{[\sigma(A^+ \cap_T B^+), \sigma((A^- \cup_S B^-)^c)]}{[\sigma(B^+), \sigma((B^-)^c)]}.$$

To compute this value we propose to approximate it by using the following formula:

$$\begin{aligned} \sigma_I(\mathcal{E}|\mathcal{F}) = \big[&min\left(\sigma(A^+|B^+), \sigma(A^+|(B^-)^c)\right), \\ &max\left(\sigma((A^-)^c|(B^-)^c), \sigma((A^-)^c|B^+)\right)\big]. \end{aligned} \tag{2}$$

Formula (2) returns an interval $[l, u]$ from $[0, 1]$ showing the proportion of number of sources that trust y and are trusted by x to the number of all sources trusted by x. This proportion is given by lower l and upper u bound, and the length of this interval expresses the hesitation or lack of knowledge about the trust. It was shown (e.g. in [3]) that such interval representation is equivalent to IFS defined as $(l, 1 - u)$. Now we can define the propagation operator.

Definition 2. The propagation operator based on relative scalar cardinality of IFS is defined as:

$$\begin{aligned} Prop_\sigma(S_X, S_Y) = \big(&min\left(\sigma(S_Y^+|S_X^+), \sigma(S_Y^+|(S_X^-)^c)\right), \\ &1 - max\left(\sigma((S_Y^-)^c|(S_X^-)^c), \sigma((S_Y^-)^c|S_X^+)\right)\big). \end{aligned}$$

Example
Let us consider a trust network from Fig.2. The values of trust score of X in Y, $R(X, Y)$, for different t-norms (minimum, algebraic, Łukasiewicz, Hamacher with $p = 3$ and Schweizer with $p = 0.2$) and cardinality pattern $f = id$ are:

- $T = \wedge : R(X, Y) = (0.63, 0.3)$
- $T = T_a : R(X, Y) = (0.52, 0.39)$
- $T = T_L : R(X, Y) = (0.4, 0.47)$
- $T = T_{H,3} : R(X, Y) = (0.46, 0.44)$
- $T = T_{S,0.2} : R(X, Y) = (0.5, 0.41)$

An atomic and general form of the operator $Prop_\sigma$ will be discussed in next subsections.

3.2 Atomic form of $Prop_\sigma$

The simplest case of a propagation is an atomic propagation, when number n of intermediate sources is equal to 1, so $S_X = \{R(X, a)\}, S_Y = \{R(a, Y)\}$.

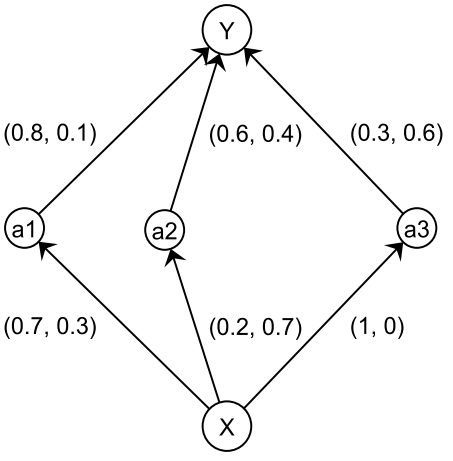


Fig. 2. Sample trust network

Then, the propagation operator $Prop_\sigma$ can be reduced to the form defined in Definition 3, which is in fact a proper form of propagation operator. To simplify the notation we will write (t_1, d_1) instead of $\{R(X, a)\}$ and (t_2, d_2) instead of $\{R(a, Y)\}$.

Definition 3. The atomic propagation operator $Prop_\sigma$ is defined by:

$$Prop_\sigma((t_1,d_1),(t_2,d_2)) = \left(min\left(\frac{T(t_2,t_1)}{t_1}, \frac{T(t_2,N(d_1))}{N(d_1)}\right), \right.$$
$$\left. 1 - max\left(\frac{T(N(d_2),N(d_1))}{N(d_1)}, \frac{T(N(d_2),t_1)}{t_1}\right)\right).$$

In the following we discuss some important properties connected with $Prop_\sigma$.

(P1) (*full trust*)

$$Prop_\sigma((1,0),(0.7,0.2)) = (0.7,0.2)$$

In fact, for all $(t, d) \in [0, 1]^2$ and for all t-norms T it holds that:

$$Prop_\sigma((1,0),(t,d)) = (t,d)$$

Therefore, the propagation operator $Prop_\sigma$ copies information from a fully trusted source. This basic property coincides with all operators $Prop_1$, $Prop_2$, $Prop_3$ and $Prop_4$ considered in [13].

(P2) (*non-commutative*) The propagation operator $Prop_\sigma$ is not commutative. For example:

$$Prop_\sigma((0.7,0.2),(0.4,0.5)) = (0,29,0,58) \text{ and}$$
$$Prop_\sigma((0.4,0.5),(0.7,0.2)) = (0,51,0,33)$$

for Hamacher t-norm with $p = 3$ and $f = id$.

(P3) (*trustful*) For each $t, d \in (0, 1)$:

$$Prop_\sigma((t, d), (1, 0)) = (1, 0),$$
$$Prop_\sigma((t, d), (0, 1)) = (0, 1),$$
$$Prop_\sigma((t, d), (0, 0)) = (0, 0).$$

This property may be a little bit controversial, however, it arises from $Prop_\sigma$ main idea of building a trust score on the basis of other's opinion. Being "trustful" means that opinions of other sources that know the source Y directly are of main significance for the final value of trust score of X in Y. This final trust score would be equal to some aggregation of direct trust scores of a group's member in Y, modified by a trust score of X to the group's members.

At the end we have to face a problem of division by 0, that appears in formula from Definition 3 when trust equals 0 (precisely, in two cases: $Prop_\sigma((0, 1), (t, d))$ and $Prop_\sigma((0, 0), (t, d))$). It means that there is no single source we can trust. In practise we probably will not take into account an opinion of sources that we don't trust to some acceptable degree, but from theoretical point of view there is a need to fill this gap. We propose to assume that $\frac{0}{0} = 0$ what leads to the following results:

$$Prop_\sigma((0, 1), (t, d)) = (0, 1),$$
$$Prop_\sigma((0, 0), (t, d)) = (0, d).$$

3.3 General form of $Prop_\sigma$

Now we will show how the propagation operator $Prop_\sigma$ can be used to calculate the trust score when (a) the propagation chain has an arbitrary length; (b) the information is delivered from many sources.

By the chain propagation we mean the situation when, in order to obtain a trust score in Y, a source X asks its direct friend, those sources in turn consult their friends and so on.

Let us consider the trust network from Fig.3. All the values $R(x_i, x_{i+1})$ are known. The trust score $R(X, Y)$ is defined by the following recursive procedure:

$$R(x_i, x_n) = Prop_\sigma(R(x_i, x_{i+1}), R(x_{i+1}, x_n)), \quad i = 1, 2, ..., n-1. \tag{3}$$

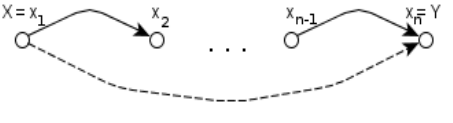


Fig. 3. Chain propagation

It means that X doesn't have to have knowledge about the whole network. If it knows a trust score $R(X, Y)$ then the procedure ends. Otherwise it asks its direct neighbor x_2 for the value $R(x_2, Y)$ and weights it with its own trust score $R(X, x_2)$. The source x_2 follows the same procedure.

In this context we can distinguish further properties of $Prop_\sigma$:

(P4) (*backward associativity*)
Propagation operator $Prop_\sigma$ is not associative. It is connected with the feature of being "trustful" - the most important is the opinion of a source that knows Y directly. That is why, as it was stated in (3), the propagation is backward (from Y to X).

(P5) (*Archimedean property*)
For Archimedean t-norms operator $Prop_\sigma$ meets the demand of "distance dependency" - the longer the propagation chain the smaller the trust degree and the bigger the hesitation degree. This property is intuitive, as the information from a far-distant party is less reliable then from a direct friend. For example, for Hamacher t-norm with $p = 3$ and $f = id$:

$$\begin{aligned} Prop_\sigma((0.7, 0.2), Prop_\sigma((0.7, 0.2), (0.7, 0.2))) &= Prop_\sigma((0.7, 0.2), (0.59, 0.26)) \\ &= (0.48, 0.33) \end{aligned}$$

with the final hesitation being equal to 0.2. That is why Archimedean t-norm seems to be the most adequate choice (in particular, more adequate than classical minimum t-norm).

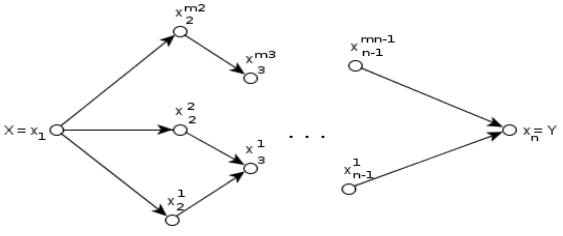


Fig. 4. Group propagation

Finally, let us consider the most general case when the source X receives information about trust score in Y from more than one indirect source. It can be combined with a chain propagation. We take the notation from Fig.4. Again, the procedure of calculating final trust score is recursive:

$$R(x_i, x_n) = Prop_\sigma(\{R(x_i, x_{i+1}^k)\}_{k=1,...,m_i}, \{R(x_{i+1}^k, x_n)\}_{k=1,...,m_i}), \quad i = 1, 2, ..., n-1. \tag{4}$$

The following remarks sum up our discussion.

(P6) (*unanimity*) In case of unanimous sources that are equally trusted by X the final trust score does not depend on the number n of intermediate sources. For example, for Hamacher t-norm with $p = 3$ and $f = id$:

$$Prop_\sigma(\{(0.2, 0.7)\}_1, \{(0.6, 0.4)\}_1) = Prop_\sigma(\{(0.2, 0.7)\}_n, \{(0.6, 0.4)\}_n) = (0.37, 0.62)$$

(P7) (*contradiction*) In case of contradictory opinions of fully trusted sources, the final trust score would always be $R(X, Y) = (s, s)$, $s \in (0, 0.5]$. For example, for Hamacher t-norm with $p = 3$ and $f = id$:

$$Prop_\sigma(\{(1.0, 0.0), (1.0, 0.0)\}, \{(0.8, 0.1), (0.1, 0.8)\}) = (0.45, 0.45)$$

(P8) (*monotonicity*) Adding a new source to the set of X neighbors with a higher trust score in Y, results in increasing the final trust score of X in Y. Similarly, adding a new source with a lower trust score in Y results in decreasing a final trust score of X in Y. For example, for Hamacher t-norm with $p = 3$ and $f = id$:

$$Prop_\sigma(\{(0.7, 0.2)\}, \{(0.8, 0.1)\}) = (0.71, 0.13)$$
$$Prop_\sigma(\{(0.7, 0.2), (0.7, 0.2)\}, \{(0.8, 0.1), (0.5, 0.5)\}) = (0.55, 0.36)$$

4 Conclusions and Further Work

The article introduced a new trust propagation operator $Prop_\sigma$ that is based on the opinion of a group of trusted sources, calculated with the use of relative scalar cardinality of IFS. This operator has a lot of desired features - it is attack resistant, it can deal with atomic propagation, chain propagation and group propagation. Some experiments shown that the best results are obtained for Archimedean t-norms. However, more researches need to be done, especially on big networks. Besides, cardinality function f and its influence on the final result needs more investigation - it seems that it can be useful for example for introducing individually acceptable trust threshold. Also distrust degree still need more closer look.

Acknowledgment. The research has been partially supported by the Ministry of Science and Higher Education Grant N N519 384936.

References

1. Atanassov, K.: Intuitionistic Fuzzy Sets. Theory and Applications. Physica-Verlag, Heidelberg (1999)
2. De Cock, M., Pinheiro da Silva, P.: A Many Valued Representation and Propagation of Trust and Distrust. In: Bloch, I., Petrosino, A., Tettamanzi, A.G.B. (eds.) WILF 2005. LNCS (LNAI), vol. 3849, pp. 108–113. Springer, Heidelberg (2006)
3. Cornelis, C., Atanassov, K., Kerre, E.: Intuitionistic fuzzy sets and interval-valued fuzzy sets: a critical comparison. In: Proc. Third European Conf. on Fuzzy Logic and Technol. (Eusflat 2003), Zittau, Germany, pp. 159–163 (2003)
4. Deschrijver, G., Kerre, E.E.: A generalization of operators on intuitionistic fuzzy sets using triangular norms and conorms. In: Notes on IFS 8, pp. 19–27 (2002)

5. Deschrijver, G., Kerre, E.E.: On the composition of intuitionistic fuzzy relations. Fuzzy Sets and Systems 136, 333–361 (2003)
6. Guha, R., Kumar, R., Raghavan, P., Tomkins, A.: Propagation of Trust and Distrust. In: Proc. WWW 2004, pp. 403–412 (2004)
7. Josang, A., Knapskog, S.J.: A Metric for Trusted Systems. In: Proc. NIST-NCSC, pp. 16–29 (1998)
8. Kral, P.: Scalar Cardinalities of IF Sets. Acta Univ. M. Belii Math. (12), 67–77 (2005)
9. Morzy, M., Wierzbicki, A.: Eksploracja aukcji internetowych w poszukiwaniu pozytywnej i negatywnej reputacji sprzedawców. In: II Krajowa Konferencja Naukowa Technologie Przetwarzania Danych TPD 2007, Pozna (2007)
10. Pankowska, A., Wygralak, M.: General IF-sets with triangular norms and their applications to group decision making. Information Sciences 176(18), 2713–2754 (2006)
11. Stachowiak, A.: Trust Propagation - Cardinality-Based Approach. In: Proceedings of IMCSIT 2009, pp. 125–129 (2009)
12. Wygralak, M.: Cardinalities of Fuzzy Sets. Springer, Heidelberg (2003)
13. Victor, P., Cornelis, C., De Cock, M., Pinheiro da Silva, P.: Gradual trust and distrust in recommender systems. Fuzzy Sets and Systems 160(10), 1367–1382 (2009)
14. Zadeh, L.A.: Fuzzy sets. Inform. and Control 8, 338–353 (1965)
15. Zhu, Y., Li, Y., Ren, Y.: Research on Propagation of Trust and Distrust by Means of Co-citation

Application of IF-Sets to Modeling of Lip Shapes Similarities*

Krzysztof Dyczkowski

Faculty of Mathematics and Computer Science,
Adam Mickiewicz University,
Umultowska 87, 61-614 Poznań, Poland
chris@amu.edu.pl

Abstract. This paper is an attempt to apply a similarity/disimilarity measure based on Atanassov IF-Sets to the comparison of lip shapes. A method of encoding lip shapes is presented and a comparison of numerical representations of lips based on the applied measure type is suggested.

Keywords: Atanassov IF-Sets, similarity/dissimilarity measure, lip shape encoding and comparison.

1 Introduction

Recognizing and comparing lip shapes is very important for applications in audio-visual speech recognition (cf. [11], [12]), lip reading (cf. [6], [9]) and phonetics research (cf. [8]). This article presents only one part of a larger project which deals with visual speech recognition based on the movements of the lips. Figure 1 shows a schematic diagram of the entire lip reading system (the elements of the system to which this work relates appear in the boldfaced frame).

We will present concepts relating to numerical modeling of lip shapes and their comparison, employing the idea of IF-Sets. Using such concepts makes it possible to compare lip shapes in a bipolar view. We can see how two lip shapes are similar to each other and simultaneously how they are dissimilar. We can also estimate our hesitation degree for lip similarity which expresses our ignorance in comparing them.

2 Preliminaries

The notion of intuitionistic fuzzy sets (IF-Sets) was introduced by Atanassov in 1986 ([1]). It can be viewed as a generalization of fuzzy sets as introduced by Zadeh in 1965 ([15]). They deliver tools for modeling simultaneously positive and negative information approaches to different phenomena and enable us to estimate our level of ignorance about them.

* The research project is supported by Ministry of Science and Higher Education grant NN519 384936.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 611–617, 2010.

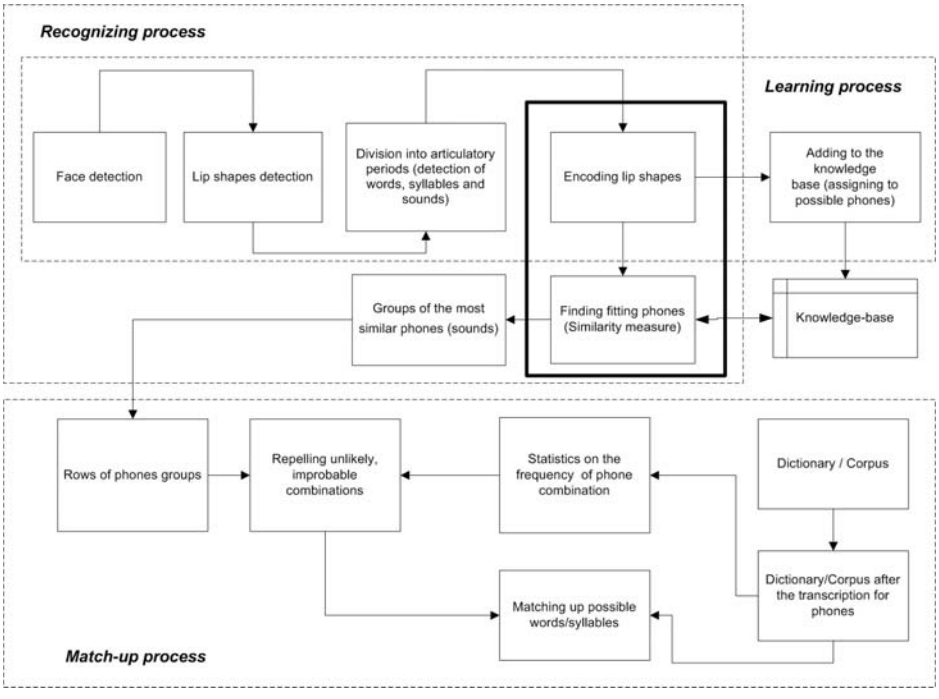


Fig. 1. Architecture of the lip reading system

Definition 1. *An intuitionistic fuzzy set (IF-Set) A in the universe X is an object of the form $A = \{\langle x; \mu(x); \nu(x)\rangle : x \in X\}$ where μ and ν are degrees of membership and non-membership of each $x \in X$, respectively, and $0 \leq \mu(x) + \nu(x) \leq 1$ for each $x \in X$.*

The number $\pi_A(x) = 1 - \mu_A(x) - \nu_A(x)$ is called the hesitation margin (or intuitionistic fuzzy index) of the IF-Set A.

More details on the theory of Atanassov intuitionistic fuzzy sets can be found in [1], [2], [7], [10] and [13,14].

Very interesting applications of IF-Sets for image processing and pattern recognition can be found in [3], [4] and [5].

3 The Lip Shape Encoding

The main task in the procedure of building a lip shape model is to construct an encoded pattern of a given lip shape that is representative and easy to compute and compare. We propose the following pattern encoding procedure.

In the first step, on the picture of the face we mark two points denoting the lip corners which are denoted by c_1 and c_5. The segment linking the two points

is divided into three parts denoted by c_2, c_3, c_4 through which segments parallel to the segment c_1, c_5 are drawn limited by edges of the lips, setting test points u_1 u_2 for the upper lips and l_1, l_2 l_3 for the lower lips. In this manner a pattern for the lip arrangement is constructed (see Fig. 2).

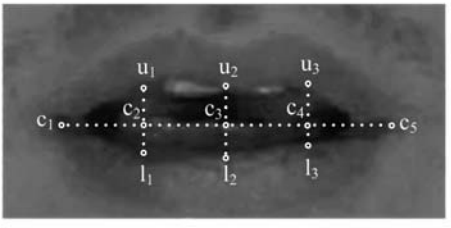


Fig. 2. Control points on the lip shape

In order to make this model independent of different sizes of lips or the picture the ratios of the lengths of vertical and horizontal segments are encoded in the pattern, rather than the distances between individual marked points are not (see Fig. 3).

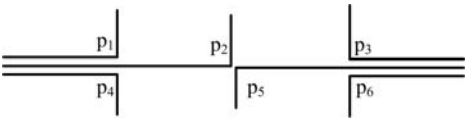


Fig. 3. Coding pattern of the lip shape

Definition 2. *Lip pattern* P *is defined as the vector:* $P = (p_1, p_2, p_3, p_4, p_5, p_6)$ *where* $p_i \in [0, 1]$ *and* $p_1 = \frac{|u1,c2|}{|c1,c2|}$, $p_2 = \frac{|u2,c3|}{|c1,c3|}$, $p_3 = \frac{|u3,c4|}{|c4,c5|}$, $p_4 = \frac{|l1,c2|}{|c1,c2|}$, $p_5 = \frac{|l2,c3|}{|c1,c2|}$, $p_6 = \frac{|l3,c4|}{|c4,c5|}$.

Here p_i *will be called the i-th pattern part of* P.

It should be emphasized that an individual element of the lip shape pattern p_i takes values from the period $[0, 1]$, and if p_i is closer to 1 then vertical and horizontal distances are more equal. If $p_i = 0$ then there is no vertical deviation from the line connecting the lip corners.

4 Similarity Measure

The construction of a lip shape similarity measure involves three steps:

1. calculation of a distance (it can be a simple module of difference) between particular pattern parts of two lip patterns;
2. interpretation of the distance by means of similarity and dissimilarity degree (construct the IF-Set of each distance between corresponding pattern parts);
3. aggregation to one similarity and one dissimilarity measure value.

Definition 3. *Let a_i, b_i be the $i-th$ pattern parts of the lip patterns A, B respectively. The similarity degree $sim_i(A,B)$ of the $i-th$ pattern parts is defined as $sim_i(A,B) = (1-|a_i - bi|)^2$ and the dissimilarity degree of the $i-th$ pattern parts is defined as $dis_i(A,B) = |a_i - b_i|^2$.*

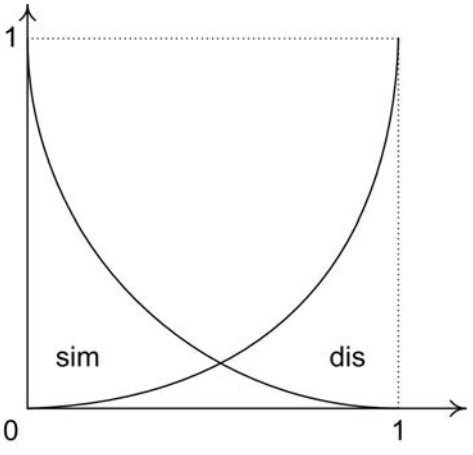


Fig. 4. Membership functions of similarity degree (sim) and dissimilarity degree (dis)

Definition 4. *Let A, B be lip patterns. The IFS-LIP measure $L(A,B)$ is a triple (SIM, DIS, HES) so such $L : A, B \to (SIM, DIS, HES)$ and*

$$SIM(A,B) = \frac{\sum_{i=1}^{6}(w_i * sim_i(A,B))}{\sum_{i=1}^{6} w_i},$$

$$DIS(A,B) = \frac{\sum_{i=1}^{6}(w_i * dis_i(A,B))}{\sum_{i=1}^{6} w_i},$$

$$HES(A,B) = 1 - SIM(A,B) - DIS(A,B),$$

$\forall_{A,B}\ \ 0 \leq SIM(A,B) + DIS(A,B) \leq 1$ *where w_i is the weight representing the importance of the $i-th$ lip pattern.*

As can be noticed, we have applied the weighted average of membership degrees as an aggregation operator. The selection of appropriate values of w_i is very important. If all w_i values are equal to 1, all fragments of the pattern will affect the measure in the same way. However experimental results have suggested that central elements of the lip pattern should have greater weights assigned. We therefore suggest using the vector of weights $w = (1, 2, 1, 1, 2, 1)$.

5 Example

Table 1 and Table 2 present eight pictures of lip arrangements with the corresponding patterns.

Table 3 presents values of the IFS-LIP measure between the lip arrangements in Table 1 and Table 2.

After verification of the results, it can be concluded that the constructed measure provides a more compressive form of information about the relations

Table 1. Lip shapes with encoded patterns P_1 - P_4

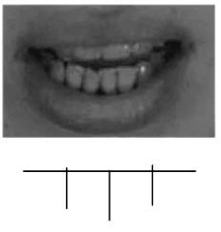
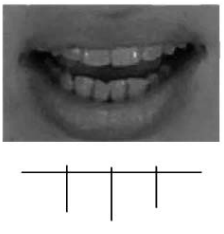
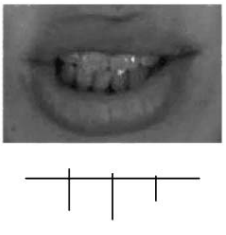
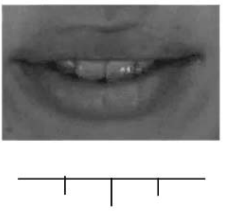
 $P_1 = (0.05, 0, 0.14, 0.89, 0.58, 0.78)$	 $P_2 = (0.15, 0.05, 0.12, 0.88, 0.54, 0.78)$
 $P_3 = (0.22, 0.05, 0.05, 0.73, 0.47, 0.51)$	 $P_4 = (0.13, 0, 0.03, 0.62, 0.44, 0.56)$

Table 2. Lip shapes with encoded patterns P_5 - P_8

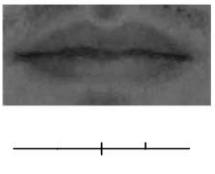
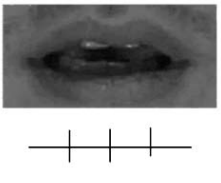
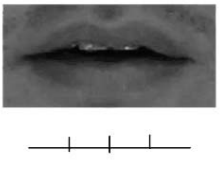
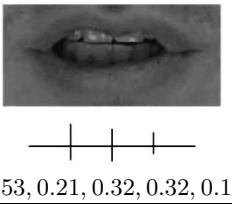
 $P_5 = (0, 0.07, 0.13, 0, 0.07, 0)$	 $P_6 = (0.41, 0.22, 0.47, 0.38, 0.18, 0.21)$
 $P_7 = (0.26, 0.14, 0.29, 0.09, 0.06, 0)$	 $P_8 = (0.53, 0.21, 0.32, 0.32, 0.18, 0.18)$

Table 3. Computed values of IFS-LIP measure

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
p_1	(1,0,0)	(0.93,0,0.07)	(0.75,0.02,0.23)	(0.75,0.03,0.22)	(0.51,0.28,0.21)	(0.37,0.17,0.45)	(0.4,0.27,0.33)	(0.38,0.19,0.43)
p_2	(0.93,0,0.07)	(1,0,0)	(0.81,0.02,0.17)	(0.78,0.02,0.2)	(0.5,0.27,0.23)	(0.42,0.15,0.43)	(0.45,0.25,0.3)	(0.42,0.17,0.42)
p_3	(0.75,0.02,0.23)	(0.81,0.02,0.17)	(1,0,0)	(0.88,0,0.11)	(0.52,0.17,0.31)	(0.51,0.09,0.4)	(0.51,0.15,0.34)	(0.5,0.09,0.4)
p_4	(0.75,0.03,0.22)	(0.78,0.02,0.2)	(0.88,0,0.11)	(1,0,0)	(0.53,0.14,0.33)	(0.5,0.1,0.41)	(0.47,0.14,0.39)	(0.48,0.1,0.42)
p_5	(0.51,0.28,0.21)	(0.5,0.27,0.23)	(0.52,0.17,0.31)	(0.53,0.14,0.33)	(1,0,0)	(0.55,0.08,0.37)	(0.82,0.02,0.16)	(0.59,0.08,0.33)
p_6	(0.37,0.17,0.45)	(0.42,0.15,0.43)	(0.51,0.09,0.4)	(0.5,0.1,0.41)	(0.55,0.08,0.37)	(1,0,0)	(0.69,0.03,0.28)	(0.88,0.01,0.11)
p_7	(0.4,0.27,0.33)	(0.45,0.25,0.3)	(0.51,0.15,0.34)	(0.47,0.14,0.39)	(0.82,0.02,0.16)	(0.69,0.03,0.28)	(1,0,0)	(0.73,0.03,0.24)
p_8	(0.38,0.19,0.43)	(0.42,0.17,0.42)	(0.5,0.09,0.4)	(0.48,0.1,0.42)	(0.59,0.08,0.33)	(0.88,0.01,0.11)	(0.73,0.03,0.24)	(1,0,0)

between lip arrangements. Apart from the information about similarities of lip patterns (e.g. the relatively high similarity between p_1 - p_2, p_6 - p_8) we additionally obtain the very essential information about hesitation in determining the power of similarity (e.g. p_1 - p_6 and p_1 - p_8 where hesitation margin is higher than the degree of similarity).

6 Summary

The idea of encoding lip shapes into patterns and comparing such patterns to the suggested measure IFS-LIP appears to be very promising. We have attempted to utilize the approach suggested by Atanassov which involves a the dual view of the data. Our work presents the initial stage of more complex research on modeling of lip shapes. Work is currently being undertaken on the implementation of algorithms of automatic lip shape pattern encoding on the basis of video processing.

References

1. Atanassov, K.: Intuitionistic fuzzy sets. Fuzzy Sets and Systems 20, 87–96 (1986)
2. Atanassov, K.: Intuitionistic fuzzy sets: Theory and Applications. Springer, Heidelberg (1999)
3. Bustincea, H., Pagola, M., Melo-Pinto, P., Barrenechea, E., Couto, P.: Use of Atanassov's intuitionistic fuzzy sets for modelling the uncertainty of the thresholds associated to an image. In: Fuzzy Sets and Their Extensions: Representation, Aggregation and Models. Intelligent Systems from Decision Making to DataMining, Web Intelligence and Computer Vision. Springer, Berlin (2008)
4. Couto, P., Pagola, M., Bustince, H., Barrenechea, E., Melo-Pinto, P.: Image segmentation using a-ifs. In: 12th Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems, Malaga, Spain, pp. 1620–1627 (2008)
5. Couto, P., Bustince, H., Filipe, V., Barrenechea, E., Pagola, M., Melo-Pinto, P.: Image threshold using A-IFSs based on bounded histograms. In: Melin, P., Castillo, O., Aguilar, L.T., Kacprzyk, J., Pedrycz, W. (eds.) IFSA 2007. LNCS (LNAI), vol. 4529, pp. 96–103. Springer, Heidelberg (2007)

6. Cox, S., Harvey, R., Lan, Y., Newman, J., Theobald, B.: The challenge of multi-speaker lip-reading. In: International Conference on Auditory-Visual Speech Processing (2008)
7. Dubois, D., Gottwald, S., Hajek, P., Kacprzyk, J., Prade, H.: Terminological difficulties in fuzzy set theory – the case of intuitionistic fuzzy sets. Fuzzy Sets and Systems 156, 485–491 (2005)
8. Dyczkowski, K., Nowakowski, P., Kordek, N., Stroński, K.: Computational tools in the analysis of phonetic grammar. Speech and Language Technology 11, 145–155 (2008)
9. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature 264, 746–748 (1976)
10. Pankowska, A., Wygralak, M.: On hesitation degrees in if-set theory. In: Rutkowski, L., Siekmann, J.H., Tadeusiewicz, R., Zadeh, L.A. (eds.) ICAISC 2004. LNCS (LNAI), vol. 3070, pp. 338–343. Springer, Heidelberg (2004)
11. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.: Recent advances in the automatic recognition of audio-visual speech. Proceedings of the IEEE 92, 1306–1326 (2003)
12. Potamianos, G., Neti, C., Luettin, J., Matthews, I.: Audio-Visual Automatic Speech Recognition: An Overview, ch. 10. MIT Press, Cambridge (2004)
13. Szmidt, E., Kacprzyk, J.: Distances between intuitionistic fuzzy sets. Fuzzy Sets and Systems 114, 505–518 (2000)
14. Szmidt, E., Kacprzyk, J.: A similarity measure for intuitionistic fuzzy sets and its application in supporting medical diagnostic reasoning. In: Rutkowski, L., Siekmann, J.H., Tadeusiewicz, R., Zadeh, L.A. (eds.) ICAISC 2004. LNCS (LNAI), vol. 3070, pp. 388–393. Springer, Heidelberg (2004)
15. Zadeh, L.: Fuzzy sets. Information and Control 8, 338–353 (1965)

A Random Set and Prototype Theory Interpretation of Intuitionistic Fuzzy Sets

Jonathan Lawry

Department of Engineering Mathematics,
University of Bristol,
Bristol, UK
j.lawry@bris.ac.uk

Abstract. An interpretation of intuitionistic fuzzy sets is proposed based on random set theory and prototype theory. The extension of fuzzy labels are modelled by lower and upper random set neighbourhoods, identifying those element of the universe within an uncertain distance threshold of a set of prototypical elements. These neighbourhoods are then generalised to compound fuzzy descriptions generated as logical combinations of basic fuzzy labels. The single point coverage functions of these lower and upper random sets are then shown to generate lower and upper membership functions satisfying the min-max combination rules of interval fuzzy set theory, the latter being isomophic to intuitionistic fuzzy set theory.

1 Introduction

Intuitionistic fuzzy sets (IFS) were first proposed by Atanassov [1] as a bipolar model of fuzzy sets where membership and non-membership are considered separately. The basis of IFS are two measures τ and ν where, for x an element of the underlying universe and θ a fuzzy description generated recursively from a set of basic fuzzy labels through application of logical connectives $\wedge, \vee$ and $\neg$, $\tau_\theta(x)$ corresponds to the membership degree of x in the extension of θ[1] and $\nu_\theta(x)$ is the non-membership degree of x in the extension of θ. A duality relationship is then defined between τ and ν such that, $\tau_{\neg\theta}(x) = \nu_\theta(x)$ and $\nu_{\neg\theta}(x) = \tau_\theta(x)$. It is also assumed that $\tau_\theta(x) + \nu_\theta(x) \leq 1$. Furthermore, τ and ν are fully truth-functional satisfying the following combination rules for $\wedge$ and $\vee$: For any fuzzy descriptions θ and φ, and element x,

- $\tau_{\theta\wedge\varphi}(x) = \min(\tau_\theta(x), \tau_\varphi(x))$, $\nu_{\theta\wedge\varphi}(x) = \max(\nu_\theta(x), \nu_\varphi(x))$
- $\tau_{\theta\vee\varphi}(x) = \max(\tau_\theta(x), \tau_\varphi(x))$, $\nu_{\theta\vee\varphi}(x) = \min(\nu_\theta(x), \nu_\varphi(x))$

As shown by Atanassov and Gargov [2] and discussed by Dubois etal. [5], there is an isomorphic relationship between IFS and an older notion of interval fuzzy sets independently introduced by Zadeh [21], Grattan-Guiness [7], Jahn [11] and

[1] The extension of θ is the set of elements to which the description θ can be appropriately applied.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 618–628, 2010.

Sambuc [19]. In this framework lower and upper membership degrees are defined, where $\underline{\mu}_\theta(x)$ is the lower membership degree of element x in the extension of θ, and $\overline{\mu}_\theta(x)$ is the upper membership degree of x in θ. These lower and upper memberships then satisfy the following properties: For any element x and fuzzy descriptions θ and φ

- $\underline{\mu}_\theta(x) \leq \overline{\mu}_\theta(x)$
- $\underline{\mu}_{\neg\theta}(x) = 1 - \overline{\mu}_\theta(x)$, and $\overline{\mu}_{\neg\theta}(x) = 1 - \underline{\mu}_\theta(x)$
- $\underline{\mu}_{\theta\wedge\varphi}(x) = \min(\underline{\mu}_\theta(x), \underline{\mu}_\varphi(x))$, $\overline{\mu}_{\theta\wedge\varphi}(x) = \min(\overline{\mu}_\theta(x), \overline{\mu}_\varphi(x))$
- $\underline{\mu}_{\theta\vee\varphi}(x) = \max(\underline{\mu}_\theta(x), \underline{\mu}_\varphi(x))$, $\overline{\mu}_{\theta\vee\varphi}(x) = \max(\overline{\mu}_\theta(x), \overline{\mu}_\varphi(x))$

The mapping between interval fuzzy sets and IFS is then obtained by taking $\underline{\mu}_\theta(x) = \tau_\theta(x)$ and $\overline{\mu}_\theta(x) = 1 - \nu_\theta(x)$. In fact it is for this interval valued fuzzy set theory for which we shall propose a direct interpretation based on random set theory and prototype theory.

Prototype theory has been proposed by Rosch [16] [17] as an alternative model of concepts in natural language. The fundamental idea is that concepts, instead of being defined by formal rules or mappings, are represented by a set of prototypical cases. These cases correspond to those elements of the underlying universe Ω, which it is certain satisfy the concept. Categorization of elements from Ω is then based on similarity to the prototypes as quantified by a distance metric defined on Ω (see [10] for an overview). By taking typicallity to be a decreasing function of distance from prototypes, this approach would naturally explain the fact that some instances are seen as being more typical exemplars of a concept than others. For instance, robins are viewed as being a more typical example of the concept *bird* than penguins, since the latter have certain atypical characteristics such as the inability to fly. This notion of typicality is also strongly related to concept vagueness where borderline cases have an intermediate range of typicality values. In other words, such cases are not sufficiently similar to the prototypes to be judged as having certain membership in the category but are also not sufficiently dissimilar to the prototypes to be ruled out as being certainly outside the category.

Random set theory has been proposed by Goodman and Nguyen as a framework for linguistic reasoning in rule based systems [14], [8], [9]. Stated simply, random sets are set-valued variables with an associated probability measure. In Goodman and Nguyen's work they provide a model of vague concepts from the perspective that the extension of such a concept is an uncertain set. This is an implicitly epistemic model of vagueness since by using a random set to model a concept an intelligent agent is assuming that there is a correct extension set about which they are uncertain. Notice that this does not require that there is actually some objectively correct definition of the concept (as suggested by Williamson [20]), but rather that the agent *assumes*, for the purposes of decision making and communication, that such a definition exists (see Lawry [13] for a discussion of this *epistemic stance*).

Dubois et al. [6] have identified both random sets and prototype theory as possible interpretations of fuzzy set membership functions. More specifically,

given a random set $\mathcal{R}$ modelling a concept, the fuzzy set membership value of an element x in $\mathcal{R}$ is then taken to be the probability that the value of $\mathcal{R}$ is a set which contains x. This is the single point coverage function of the random set $\mathcal{R}$. For the prototype theory model it is assumed that there exists a similarity measure between the elements of Ω, which takes values in $[0, 1]$. Given a set of prototypical elements, the membership of x in the associated fuzzy set is then taken as corresponding to the similarity between x and these prototypes [18]. In [12] we have proposed a natural combination of prototype theory and random set theory to model linguistic labels and descriptions. The idea behind this approach is that, in order to decide whether the assertion 'x is L_i' is appropriate for element x and label L_i with prototypes P_i, an agent would threshold the distance between x and P_i. In other words, L_i would be deemed an appropriate label for x provided that $d(x, P_i) \leq \epsilon$, for some distance function $d : \Omega^2 \to [0, \infty)$ and threshold $\epsilon \geq 0$. However, the inherent uncertainty about the extension of L_i would naturally result in uncertainty about the value of threshold ϵ. Consequently, the extension of L_i would correspond to a random set neighbourhood of the prototypes of L_i as defined by those elements which lie within the uncertain threshold ϵ of P_i. In the sequel we extend this idea, so as to generate lower and upper neighbourhoods as extensions of a concept by introducing lower and upper thresholds.

2 Lower and Upper Membership Functions

We envisage a population of communicating agents applying a finite set of labels to describe the elements of an underlying universe of discourse. Given an element x an agent must decide which labels and compound descriptions are *appropriate* to describe x, where appropriateness is governed by the linguistic conventions of the population. Agents adopt the *epistemic stance* [13] by assuming that there is an uncertain but crisp division between those labels which are, and those which are not appropriate to describe a given element. Now since an agent's knowledge of these linguistic conventions, obtained through their experience of communication with others, is partial and often conflicting they will have significant uncertainty about the appropriateness of labels. It is assumed, however, that there will be prototypical elements for which they will be certain that a given label can describe. These prototypes will then form the basis of the agent's representation of each label. More formally:

Let Ω denote the underlying universe of discourse and $LA = \{L_1, \ldots, L_n\}$ be a finite set of labels for describing elements of Ω. LE then corresponds to the set of compound express generated by recursive application of the connectives $\wedge$, $\vee$ and $\neg$ to the labels in LA. For example, if LA contains labels *red* and *blue*, then LE contains expressions including *red and blue*, *red or blue*, *not red*, *not blue*, *red and not blue* etc. For each label L_i there is a set of prototypical elements $P_i \subseteq \Omega$, such that L_i is *certainly appropriate* to describe any prototypical elements in P_i. Given a distance function $d : \Omega^2 \to [0, \infty)$ satisfying $d(x, x) = 0$ and $d(x, y) = d(y, x)$ for all $x, y \in \Omega$, lower and upper extensions of each label are defined to be those

elements of Ω with distance from P_i less than or equal to a lower and an upper threshold value respectively. In other words, the lower extension of L_i is taken to be $\{x \in \Omega : d(x, P_i) \leq \underline{\epsilon}\}$ and the upper extension $\{x \in \Omega : d(x, P_i) \leq \overline{\epsilon}\}$, where $\underline{\epsilon} \leq \overline{\epsilon}$ and $d(x, P_i) = \inf\{d(x, y) : y \in P_i\}$. Here, we further assume that both $\underline{\epsilon}$ and $\overline{\epsilon}$ are functions of a single parameter α taking values in $[0, 1]$. The underlying intuition is that α quantifies an agent's overall level of imprecision in their definition of labels, so that as α increases the difference between the upper extension of a label and its lower extension decreases. In effect this means that there exists an increasing function $\underline{f} : [0, 1] \to [0, \infty)$ and a decreasing function $\overline{f} : [0, 1] \to [0, \infty)$ such that $\underline{f} \leq \overline{f}$ and for which $\underline{\epsilon} = \underline{f}(\alpha)$ and $\overline{\epsilon} = \overline{f}(\alpha)$.

Definition 1. *Lower and Upper Threshold Functions*
$\underline{f} : [0, 1] \to [0, \infty)$ and $\overline{f} : [0, 1] \to [0, \infty)$ where $\underline{f}$ is an increasing function and $\overline{f}$ is a decreasing function satisfying $\forall \alpha \in [0, 1]$ $\underline{f}(\alpha) \leq \overline{f}(\alpha)$.

The lower and upper extension of the labels and the compound descriptions in LE are then defined recursively as follows:

Definition 2. *Lower and Upper Random Neighbourhoods*

- $\forall L_i \in LA$ $\underline{\mathcal{N}}^{\alpha}_{L_i} = \{x : d(x, P_i) \leq \underline{f}(\alpha)\}$, $\overline{\mathcal{N}}^{\alpha}_{L_i} = \{x : d(x, P_i) \leq \overline{f}(\alpha)\}$.
- $\forall \theta, \varphi \in LE$ $\underline{\mathcal{N}}^{\alpha}_{\theta \wedge \varphi} = \underline{\mathcal{N}}^{\alpha}_{\theta} \cap \underline{\mathcal{N}}^{\alpha}_{\varphi}$, $\overline{\mathcal{N}}^{\alpha}_{\theta \wedge \varphi} = \overline{\mathcal{N}}^{\alpha}_{\theta} \cap \overline{\mathcal{N}}^{\alpha}_{\varphi}$.
- $\forall \theta, \varphi \in LE$ $\underline{\mathcal{N}}^{\alpha}_{\theta \vee \varphi} = \underline{\mathcal{N}}^{\alpha}_{\theta} \cup \underline{\mathcal{N}}^{\alpha}_{\varphi}$, $\overline{\mathcal{N}}^{\alpha}_{\theta \vee \varphi} = \overline{\mathcal{N}}^{\alpha}_{\theta} \cup \overline{\mathcal{N}}^{\alpha}_{\varphi}$.
- $\forall \theta \in LE$ $\underline{\mathcal{N}}^{\alpha}_{\neg \theta} = (\overline{\mathcal{N}}^{\alpha}_{\theta})^c$, $\overline{\mathcal{N}}^{\alpha}_{\neg \theta} = (\underline{\mathcal{N}}^{\alpha}_{\theta})^c$

Now in view of the distributed manner in which language is learnt through the interaction and communications between a population of agents, it is likely that an individual agent will be uncertain as to which value of α should be adopted in a given context. Here, in keeping with the epistemic stance, we model this uncertainty by a probability density function δ on α. The lower and upper membership functions of expression $\theta \in LE$ for element $x \in \Omega$ are then given by the probability of a value of α such that $x \in \underline{\mathcal{N}}^{\alpha}_{\theta}$ and the probability of an α such that $x \in \overline{\mathcal{N}}^{\alpha}_{\theta}$ respectively.

Definition 3. *Lower and Upper Membership Functions*
Let δ be a density function on $[0, 1]$. Then $\forall \theta \in LE$, $\forall x \in \Omega$ we define $\underline{\mu}_{\theta}(x) = \delta(\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\theta}\})$ and $\overline{\mu}_{\theta}(x) = \delta(\{\alpha : x \in \overline{\mathcal{N}}^{\alpha}_{\theta}\})$

Here $\underline{\mu}_{\theta}(x)$ quantifies the agent's belief that expression θ is *definitely appropriate* to describe x, and $\overline{\mu}_{\theta}(x)$ is the belief that θ is *possibly appropriate* to describe x. These lower and upper measures attempt to capture the intuition that 'appropriateness' or 'assertability' of descriptions is inherently bipolar. This bipolarity manifests itself in the distinction between those descriptions which convention would deem clearly appropriate to describe an element x, and those which convention would not classify as incorrect, or perhaps even dishonest, descriptions. Parikh [15] observes that:

> Certain sentences are assertible in the sense that we might ourselves assert them and other cases of sentences which are non-assertible in the sense that we ourselves (and many others) would reproach someone who used them. But there will also be the intermediate kind of sentences, where we might allow their use.

For example, consider a witness in a court of law describing a suspect as being *tall*. Depending on the actual height of the suspect this statement may be deemed as clearly true or clearly false, in which latter case the witness could be accused of perjury. However, there will also be an intermediate height range for which, while there may be doubt and differing opinions concerning the use of the description *tall*, it would not be deemed as definitely inappropriate and hence the witness would not be viewed as committing perjury.

We now investigate some basic properties of lower and upper neighbourhoods and argue that $\underline{\mu}_\theta$ and $\overline{\mu}_\theta$ can indeed be viewed as lower and upper membership functions according to the random set interpretation of fuzzy sets. The following theorem shows that the lower neighbourhood is, as intended, a subset of the upper neighbourhood for any expression in LE.

Theorem 1. $\forall \Psi \in LE$, $\forall \alpha \in [0,1]$ $\underline{\mathcal{N}}^\alpha_\Psi \subseteq \overline{\mathcal{N}}^\alpha_\Psi$

Proof. Let $LE^{(1)} = LA$ and $LE^{(k)} = LE^{(k-1)} \cup \{\theta \wedge \varphi, \theta \vee \varphi, \neg\theta : \theta, \varphi \in LE^{(k-1)}\}$. We now proceed by induction on k. If $\Psi = L_i$ then $\underline{\mathcal{N}}^\alpha_{L_i} = \{x : d(x, P_i) \leq \underline{f}(\alpha)\} \subseteq \{x : d(x, P_i) \leq \overline{f}(\alpha)\} = \overline{\mathcal{N}}^\alpha_{L_i}$. Now assuming the result holds for $\Psi \in LE^{(k)}$ we show that it holds for $\Psi \in LE^{(k+1)}$. If $\Psi \in LE^{(k+1)}$ then either $\Psi \in LE^{(k)}$, in which case the result holds trivially, or $\exists \theta, \varphi \in LE^{(k)}$ for which one of the following holds:

- $\Psi = \theta \wedge \varphi$ in which case $\underline{\mathcal{N}}^\alpha_\Psi = \underline{\mathcal{N}}^\alpha_{\theta\wedge\varphi} = \underline{\mathcal{N}}^\alpha_\theta \cap \underline{\mathcal{N}}^\alpha_\varphi \subseteq \overline{\mathcal{N}}^\alpha_\theta \cap \overline{\mathcal{N}}^\alpha_\varphi$ (by the inductive step) $= \overline{\mathcal{N}}^\alpha_{\theta\wedge\varphi} = \overline{\mathcal{N}}^\alpha_\Psi$.
- $\Psi = \theta \vee \varphi$ in which case $\underline{\mathcal{N}}^\alpha_\Psi = \underline{\mathcal{N}}^\alpha_{\theta\vee\varphi} = \underline{\mathcal{N}}^\alpha_\theta \cup \underline{\mathcal{N}}^\alpha_\varphi \subseteq \overline{\mathcal{N}}^\alpha_\theta \cup \overline{\mathcal{N}}^\alpha_\varphi$ (by the inductive step) $= \overline{\mathcal{N}}^\alpha_{\theta\vee\varphi} = \overline{\mathcal{N}}^\alpha_\Psi$.
- $\Psi = \neg\theta$. Now by induction $\overline{\mathcal{N}}^\alpha_\theta \supseteq \underline{\mathcal{N}}^\alpha_\theta$ and therefore $(\overline{\mathcal{N}}^\alpha_\theta)^c \subseteq (\underline{\mathcal{N}}^\alpha_\theta)^c$. Hence, in this case $\underline{\mathcal{N}}^\alpha_\Psi = \underline{\mathcal{N}}^\alpha_{\neg\theta} = (\overline{\mathcal{N}}^\alpha_\theta)^c \subseteq (\underline{\mathcal{N}}^\alpha_\theta)^c = \overline{\mathcal{N}}^\alpha_{\neg\theta} = \overline{\mathcal{N}}^\alpha_\Psi$.

Corollary 1. $\forall \theta \in LE$, $\forall x \in \Omega$ $\underline{\mu}_\theta(x) \leq \overline{\mu}_\theta(x)$

For any expression θ, $\underline{\mathcal{N}}^\alpha_\theta$ and $\overline{\mathcal{N}}^\alpha_\theta$ are both random sets taking as values subsets of Ω. From this perspective $\underline{\mu}_\theta$ and $\overline{\mu}_\theta$ are the single point coverage functions of $\underline{\mathcal{N}}^\alpha_\theta$ and $\overline{\mathcal{N}}^\alpha_\theta$ respectively. Hence, according to the random set interpretation of fuzzy sets proposed in [8], [9] and [6], $\underline{\mu}_\theta(x)$ and $\overline{\mu}_\theta(x)$ can be viewed as membership values of x in the lower and upper extension of θ respectively.

Theorem 2. $\forall \alpha, \alpha' \in [0,1]$ *where* $\alpha \leq \alpha'$ *it holds that* $\forall \theta \in LE$ $\underline{\mathcal{N}}^\alpha_\theta \subseteq \underline{\mathcal{N}}^{\alpha'}_\theta$ *and* $\overline{\mathcal{N}}^\alpha_\theta \supseteq \overline{\mathcal{N}}^{\alpha'}_\theta$.

Proof. Let $LE^{(1)} = LA$ and $LE^{(k)} = LE^{(k-1)} \cup \{\theta \wedge \varphi, \theta \vee \varphi, \neg\theta : \theta, \varphi \in LE^{(k-1)}\}$. We now proceed by induction on k. If $\Psi = L_i$ then $\underline{\mathcal{N}}^{\alpha}_{L_i} = \{x : d(x, P_i) \leq \underline{f}(\alpha)\} \subseteq \{x : d(x, P_i) \leq \underline{f}(\alpha')\} = \underline{\mathcal{N}}^{\alpha'}_{L_i}$ since $\underline{f}$ is an increasing function. Also $\overline{\mathcal{N}}^{\alpha}_{L_i} = \{x : d(x, P_i) \leq \overline{f}(\alpha)\} \supseteq \{x : d(x, P_i) \leq \overline{f}(\alpha')\} = \overline{\mathcal{N}}^{\alpha'}_{L_i}$ since $\overline{f}$ is a decreasing function. Now assuming the result holds for $\Psi \in LE^{(k)}$ we show that it holds for $\Psi \in LE^{(k+1)}$. If $\Psi \in LE^{(k+1)}$ then either $\Psi \in LE^{(k)}$, in which case the result holds trivially, or $\exists \theta, \varphi \in LE^{(k)}$ for which one of the following holds:

- $\Psi = \theta \wedge \varphi$. In this case $\underline{\mathcal{N}}^{\alpha}_{\Psi} = \underline{\mathcal{N}}^{\alpha}_{\theta\wedge\varphi} = \underline{\mathcal{N}}^{\alpha}_{\theta} \cap \underline{\mathcal{N}}^{\alpha}_{\varphi} \subseteq \underline{\mathcal{N}}^{\alpha'}_{\theta} \cap \underline{\mathcal{N}}^{\alpha'}_{\varphi}$ (by induction) $= \underline{\mathcal{N}}^{\alpha'}_{\theta\wedge\varphi} = \underline{\mathcal{N}}^{\alpha'}_{\Psi}$. Also $\overline{\mathcal{N}}^{\alpha}_{\Psi} = \overline{\mathcal{N}}^{\alpha}_{\theta\wedge\varphi} = \overline{\mathcal{N}}^{\alpha}_{\theta} \cap \overline{\mathcal{N}}^{\alpha}_{\varphi} \supseteq \overline{\mathcal{N}}^{\alpha'}_{\theta} \cap \overline{\mathcal{N}}^{\alpha'}_{\varphi}$ (by induction) $= \overline{\mathcal{N}}^{\alpha'}_{\theta\wedge\varphi} = \overline{\mathcal{N}}^{\alpha'}_{\Psi}$.
- $\Psi = \theta \vee \varphi$. In this case $\underline{\mathcal{N}}^{\alpha}_{\Psi} = \underline{\mathcal{N}}^{\alpha}_{\theta\vee\varphi} = \underline{\mathcal{N}}^{\alpha}_{\theta} \cup \underline{\mathcal{N}}^{\alpha}_{\varphi} \subseteq \underline{\mathcal{N}}^{\alpha'}_{\theta} \cup \underline{\mathcal{N}}^{\alpha'}_{\varphi}$ (by induction) $= \underline{\mathcal{N}}^{\alpha'}_{\theta\vee\varphi} = \underline{\mathcal{N}}^{\alpha'}_{\Psi}$. Also $\overline{\mathcal{N}}^{\alpha}_{\Psi} = \overline{\mathcal{N}}^{\alpha}_{\theta\vee\varphi} = \overline{\mathcal{N}}^{\alpha}_{\theta} \cup \overline{\mathcal{N}}^{\alpha}_{\varphi} \supseteq \overline{\mathcal{N}}^{\alpha'}_{\theta} \cup \overline{\mathcal{N}}^{\alpha'}_{\varphi}$ (by induction) $= \overline{\mathcal{N}}^{\alpha'}_{\theta\vee\varphi} = \overline{\mathcal{N}}^{\alpha'}_{\Psi}$.
- $\Psi = \neg\theta$. In this case $\underline{\mathcal{N}}^{\alpha}_{\Psi} = \underline{\mathcal{N}}^{\alpha}_{\neg\theta} = (\overline{\mathcal{N}}^{\alpha}_{\theta})^c \subseteq (\overline{\mathcal{N}}^{\alpha'}_{\theta})^c$ (by induction) $= \underline{\mathcal{N}}^{\alpha'}_{\neg\theta} = \underline{\mathcal{N}}^{\alpha'}_{\Psi}$. Also $\overline{\mathcal{N}}^{\alpha}_{\Psi} = \overline{\mathcal{N}}^{\alpha}_{\neg\theta} = (\underline{\mathcal{N}}^{\alpha}_{\theta})^c \supseteq (\underline{\mathcal{N}}^{\alpha'}_{\theta})^c$ (by induction) $= \overline{\mathcal{N}}^{\alpha'}_{\neg\theta} = \overline{\mathcal{N}}^{\alpha'}_{\Psi}$.

Corollary 2. $\forall \theta, \varphi \in LE$, $\forall x \in \Omega$

- $\underline{\mu}_{\theta\wedge\varphi}(x) = \min(\underline{\mu}_{\theta}(x), \underline{\mu}_{\varphi}(x))$, $\overline{\mu}_{\theta\wedge\varphi}(x) = \min(\overline{\mu}_{\theta}(x), \overline{\mu}_{\varphi}(x))$
- $\underline{\mu}_{\theta\vee\varphi}(x) = \max(\underline{\mu}_{\theta}(x), \underline{\mu}_{\varphi}(x))$, $\overline{\mu}_{\theta\vee\varphi}(x) = \max(\overline{\mu}_{\theta}(x), \overline{\mu}_{\varphi}(x))$
- $\underline{\mu}_{\neg\theta}(x) = 1 - \overline{\mu}_{\theta}(x)$, $\overline{\mu}_{\neg\theta}(x) = 1 - \underline{\mu}_{\theta}(x)$

Proof. From theorem 2 we have that $\forall \theta, \varphi \in LE$ either $\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\theta}\} \subseteq \{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\varphi}\}$ or $\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\theta}\} \supseteq \{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\varphi}\}$ and either $\{\alpha : x \in \overline{\mathcal{N}}^{\alpha}_{\theta}\} \subseteq \{\alpha : x \in \overline{\mathcal{N}}^{\alpha}_{\varphi}\}$ or $\{\alpha : x \in \overline{\mathcal{N}}^{\alpha}_{\theta}\} \supseteq \{\alpha : x \in \overline{\mathcal{N}}^{\alpha}_{\varphi}\}$. Now assume w.l.o.g that $\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\theta}\} \subseteq \{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\varphi}\}$ then:

$$\underline{\mu}_{\theta\wedge\varphi}(x) = \delta(\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\theta\wedge\varphi}\}) = \delta(\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\theta}\} \cap \{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\varphi}\})$$
$$= \delta(\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\theta}\}) = \underline{\mu}_{\theta}(x) = \min(\underline{\mu}_{\theta}(x), \underline{\mu}_{\varphi}(x)) \text{ and}$$
$$\underline{\mu}_{\theta\vee\varphi}(x) = \delta(\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\theta\vee\varphi}\}) = \delta(\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\theta}\} \cup \{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\varphi}\})$$
$$= \delta(\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\varphi}\}) = \underline{\mu}_{\varphi}(x) = \max(\underline{\mu}_{\theta}(x), \underline{\mu}_{\varphi}(x))$$

The result also follows similarly for $\overline{\mu}_{\theta\wedge\varphi}(x)$ and $\overline{\mu}_{\theta\vee\varphi}(x)$. Furthermore,

$$\underline{\mu}_{\neg\theta}(x) = \delta(\{\alpha : x \in \underline{\mathcal{N}}^{\alpha}_{\neg\theta}\}) = \delta(\{\alpha : x \in (\overline{\mathcal{N}}^{\alpha}_{\theta})^c\}) = \delta(\{\alpha : x \in \overline{\mathcal{N}}^{\alpha}_{\theta}\}^c)$$
$$= 1 - \delta(\{\alpha : x \in \overline{\mathcal{N}}^{\alpha}_{\theta}\}) = 1 - \overline{\mu}_{\theta}(x)$$

Similarly $\overline{\mu}_{\neg\theta}(x) = 1 - \underline{\mu}_{\theta}(x)$

Also notice from theorem 2 that $\forall \alpha \leq \alpha'$ and $\forall \theta \in LE$, $\overline{\mathcal{N}}^{\alpha'}_{\theta} - \underline{\mathcal{N}}^{\alpha'}_{\theta} \subseteq \overline{\mathcal{N}}^{\alpha}_{\theta} - \underline{\mathcal{N}}^{\alpha}_{\theta}$, and hence, in accordance with the original intuition, the parameter α is a direct indicator of the imprecision associated with the definition of θ.

Theorem 3. *$\forall \theta, \varphi \in LE$, $\forall \alpha \in [0,1]$ the following hold:*

- $\underline{\mathcal{N}}^{\alpha}_{\neg(\neg\theta)} = \underline{\mathcal{N}}^{\alpha}_{\theta}$, *and* $\overline{\mathcal{N}}^{\alpha}_{\neg(\neg\theta)} = \overline{\mathcal{N}}^{\alpha}_{\theta}$
- $\underline{\mathcal{N}}^{\alpha}_{\neg(\theta\wedge\varphi)} = \underline{\mathcal{N}}^{\alpha}_{\neg\theta\vee\neg\varphi}$ *and* $\overline{\mathcal{N}}^{\alpha}_{\neg(\theta\wedge\varphi)} = \overline{\mathcal{N}}^{\alpha}_{\neg\theta\vee\neg\varphi}$
- $\underline{\mathcal{N}}^{\alpha}_{\neg(\theta\vee\varphi)} = \underline{\mathcal{N}}^{\alpha}_{\neg\theta\wedge\neg\varphi}$ *and* $\overline{\mathcal{N}}^{\alpha}_{\neg(\theta\vee\varphi)} = \overline{\mathcal{N}}^{\alpha}_{\neg\theta\wedge\neg\varphi}$
- $\underline{\mathcal{N}}^{\alpha}_{\theta\wedge\neg\theta} = \emptyset$ *and* $\overline{\mathcal{N}}^{\alpha}_{\theta\vee\neg\theta} = \Omega$

Proof.
- $\underline{N}^{\alpha}_{\neg(\neg\theta)} = (\overline{N}^{\alpha}_{\neg\theta})^c = ((\underline{\mathcal{N}}^{\alpha}_{\theta})^c)^c = \underline{\mathcal{N}}^{\alpha}_{\theta}$ and similarly $\overline{N}^{\alpha}_{\neg(\neg\theta)} = (\underline{N}^{\alpha}_{\neg\theta})^c = ((\overline{\mathcal{N}}^{\alpha}_{\theta})^c)^c = \overline{\mathcal{N}}^{\alpha}_{\theta}$
- $\underline{\mathcal{N}}^{\alpha}_{\neg(\theta\wedge\varphi)} = (\overline{\mathcal{N}}^{\alpha}_{\theta\wedge\varphi})^c = (\overline{\mathcal{N}}^{\alpha}_{\theta} \cap \overline{\mathcal{N}}^{\alpha}_{\varphi})^c = (\overline{\mathcal{N}}^{\alpha}_{\theta})^c \cup (\overline{\mathcal{N}}^{\alpha}_{\varphi})^c = \underline{\mathcal{N}}^{\alpha}_{\neg\theta} \cup \underline{\mathcal{N}}^{\alpha}_{\neg\varphi} = \underline{\mathcal{N}}^{\alpha}_{\neg\theta\vee\neg\varphi}$ and similarly $\overline{\mathcal{N}}^{\alpha}_{\neg(\theta\wedge\varphi)} = (\underline{\mathcal{N}}^{\alpha}_{\theta\wedge\varphi})^c = (\underline{\mathcal{N}}^{\alpha}_{\theta} \cap \underline{\mathcal{N}}^{\alpha}_{\varphi})^c = (\underline{\mathcal{N}}^{\alpha}_{\theta})^c \cup (\underline{\mathcal{N}}^{\alpha}_{\varphi})^c = \overline{\mathcal{N}}^{\alpha}_{\neg\theta} \cup \overline{\mathcal{N}}^{\alpha}_{\neg\varphi} = \overline{\mathcal{N}}^{\alpha}_{\neg\theta\vee\neg\varphi}$
- $\underline{\mathcal{N}}^{\alpha}_{\neg(\theta\vee\varphi)} = (\overline{\mathcal{N}}^{\alpha}_{\theta\vee\varphi})^c = (\overline{\mathcal{N}}^{\alpha}_{\theta} \cup \overline{\mathcal{N}}^{\alpha}_{\varphi})^c = (\overline{\mathcal{N}}^{\alpha}_{\theta})^c \cap (\overline{\mathcal{N}}^{\alpha}_{\varphi})^c = \underline{\mathcal{N}}^{\alpha}_{\neg\theta} \cap \underline{\mathcal{N}}^{\alpha}_{\neg\varphi} = \underline{\mathcal{N}}^{\alpha}_{\neg\theta\wedge\neg\varphi}$ and similarly $\overline{\mathcal{N}}^{\alpha}_{\neg(\theta\vee\varphi)} = (\underline{\mathcal{N}}^{\alpha}_{\theta\vee\varphi})^c = (\underline{\mathcal{N}}^{\alpha}_{\theta} \cup \underline{\mathcal{N}}^{\alpha}_{\varphi})^c = (\underline{\mathcal{N}}^{\alpha}_{\theta})^c \cap (\underline{\mathcal{N}}^{\alpha}_{\varphi})^c = \overline{\mathcal{N}}^{\alpha}_{\neg\theta} \cap \overline{\mathcal{N}}^{\alpha}_{\neg\varphi} = \overline{\mathcal{N}}^{\alpha}_{\neg\theta\wedge\neg\varphi}$
- $\underline{\mathcal{N}}^{\alpha}_{\theta\wedge\neg\theta} = \underline{\mathcal{N}}^{\alpha}_{\theta} \cap \underline{\mathcal{N}}^{\alpha}_{\neg\theta} = \underline{\mathcal{N}}^{\alpha}_{\theta} \cap (\overline{\mathcal{N}}^{\alpha}_{\theta})^c \subseteq \underline{\mathcal{N}}^{\alpha}_{\theta} \cap (\underline{\mathcal{N}}^{\alpha}_{\theta})^c = \emptyset$ by theorem 1.
- $\overline{\mathcal{N}}^{\alpha}_{\theta\vee\neg\theta} = \overline{\mathcal{N}}^{\alpha}_{\theta} \cup \overline{\mathcal{N}}^{\alpha}_{\neg\theta} \supseteq \underline{\mathcal{N}}^{\alpha}_{\theta} \cup \overline{\mathcal{N}}^{\alpha}_{\neg\theta} = \Omega$ by theorem 1.

Example 1. Let $\Omega = \mathbb{R}$ and L_i be a label with prototype $P_i = \{10\}$ (i.e. L_i denotes *about 10*). Let $\underline{f}(\alpha) = 2\alpha$ and $\overline{f}(\alpha) = 4 - 2\alpha$ (see figure 1) and also let δ be a gaussian distribution with mean 0.5 and standard deviation 0.15 normalised so as to have integral 1 on $[0,1]$ (see figure 2). From this we have the following lower and upper neighbourhoods:

$$\underline{\mathcal{N}}^{\alpha}_{L_i} = [10 - 2\alpha, 10 + 2\alpha] \text{ and } \overline{\mathcal{N}}^{\alpha}_{L_i} = [6 + 2\alpha, 14 - 2\alpha]$$

Hence, the lower and upper membership functions are given by (see figure 3):

$$\underline{\mu}_{L_i}(x) = \begin{cases} \int_{\frac{10-x}{2}}^{1} \delta(\epsilon)d\epsilon : 8 \leq x \leq 10 \\ \int_{\frac{x-10}{2}}^{1} \delta(\epsilon)d\epsilon : 10 < x \leq 12 \\ 0 : \text{otherwise} \end{cases} \qquad \overline{\mu}_{L_i}(x) = \begin{cases} \int_{0}^{\frac{x-6}{2}} \delta(\epsilon)d\epsilon : 6 \leq x \leq 10 \\ \int_{0}^{\frac{14-x}{2}} \delta(\epsilon)d\epsilon : 10 < x \leq 14 \\ 0 : \text{otherwise} \end{cases}$$

3 Discussion and Conclusions

We have introduced a random set and prototype theory interpretation of lower and upper fuzzy membership functions. In particular, we have proposed lower and upper random set extensions of fuzzy descriptions generated as recursive combinations of random set neighbourhoods of prototypes defined for a set of basic fuzzy labels. Each such extension then identifies those elements of Ω which can be *appropriately described* by the associated fuzzy description. Random sets are defined based on lower and upper threshold distances from prototypes which are taken to be functions of a single parameter α indicating the overall level of imprecision associated with concept definition. Uncertainty associated with

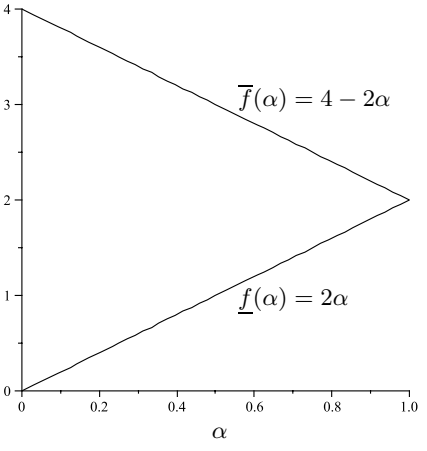


Fig. 1. Example of lower and upper threshold functions

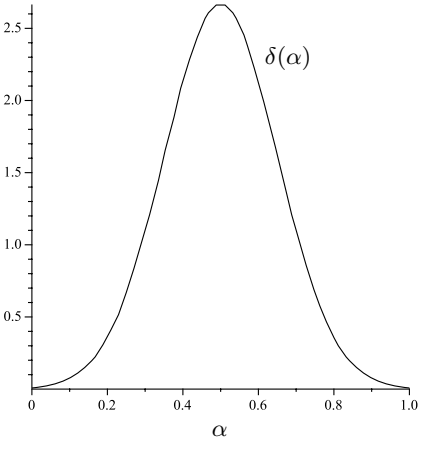


Fig. 2. Normalised gaussian density function δ with mean 0.5 and standard deviation $0,15$

the correct level of α is modelled by a probability density function δ, according to which we can calculate the lower and upper membership functions of x in θ, as the probabilities of those α values for which x is in the lower and upper extensions of θ respectively. In effect these two measures are the single point coverage functions of the lower and upper random sets, and hence according to the random set interpretation of membership functions, can be viewed as lower and upper memberships functions of the extension of the fuzzy concept.

Based on this definition we have then shown that the lower and upper membership functions are fully truth-functional satisfying the min and max rules for conjunction and disjunction as proposed for interval fuzzy sets. However, the

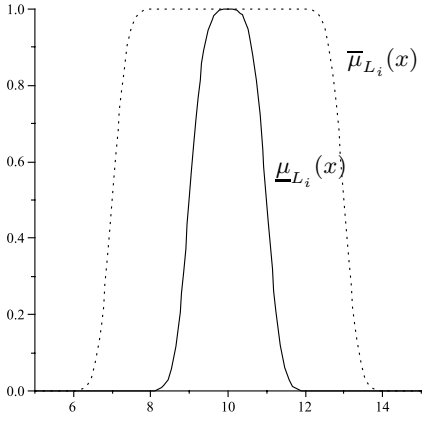


Fig. 3. Example of lower and upper membership functions for label L_i with prototype $P_i = \{10\}$

interpretation also imposes other additional constraint on the calculus for lower and upper membership functions. For instance, from theorem 3 it follows immediately that $\forall\theta \in LE$ and $\forall x \in \Omega$, $\underline{\mu}_{\theta\wedge\neg\theta}(x) = 0$ and $\overline{\mu}_{\theta\vee\neg\theta}(x) = 1$. Hence, by corollary 2, it holds that $\min(\underline{\mu}_\theta(x), 1 - \overline{\mu}_\theta(x)) = 0$. In other words, for any element x and expression θ, either $\underline{\mu}_\theta(x) = 0$ or $\overline{\mu}_\theta(x) = 1$. Clearly then by applying the mapping $\tau_\theta(x) = \underline{\mu}_\theta(x)$ and $\nu_\theta(x) = 1 - \overline{\mu}_\theta(x) = \underline{\mu}_{\neg\theta}(x)$ we obtain a calculus for membership and non-membership degree identical to that proposed for IFS by Atanassov [1], but with the additional constraint that either $\tau_\theta(x) = 0$ or $\nu_\theta(x) = 0$ for any x and θ.

Dubois etal. [5] question the interpretation of IFS as an intuitionistic theory. For example, unlike intuitionistic logic, IFS satisfies double negation while it does not satisfy the law of non-contradiction [2]. This criticism would seem to be borne out under the current interpretation, since the fundamental notion underlying the measures $\underline{\mu}$ and $\overline{\mu}$ is that of random set neighbourhoods, which do not seem to be at all intuitionistic in nature. For instance, from theorem 3 we see that the double negation law is strongly validated since it holds for both lower and upper neighbourhoods. Also theorem 3 shows that the behaviour of lower and upper neighbourhoods with regard to the laws of excluded middle and non-contradiction differs significantly from intuitionistic logic. Indeed, while the lower neighbourhood does not satisfy excluded middle, the upper neighbourhood does. Similarly, while the lower neighbourhood satisfies the law of non-contradiction, the upper neighbourhood does not. Indeed the behaviour of lower and upper random set neighbourhoods with regard to these laws is exactly what would be expected from two criteria, one weaker and one stronger, related in a bipolar

[2] We refer here to the standard min, max calculus for IFS where $\tau_{\neg\theta} = \nu_\theta$ and $\nu_{\neg\theta} = \tau_\theta$. Atanassov [3] shows that for other choices of negation operator, the law of non-contradiction may be satisfied.

manner as outlined in [4] , rather than being based on the notion of justifiability as is the case in intuitionistic logic. In particular, lower and upper membership functions would seem to be a special case of what Dubois and Prade [4] refer to as *symmetric bivariate unipolarity*, whereby judgments are made according to two distinct evaluations on unipolar scales. In the current context, we have a strong and a weak evaluation criterion where the former corresponds to definite appropriateness and the latter to possible appropriateness. As with many examples of this type of bipolarity there is a natural duality between the two evaluation criterion in that a description θ is definitely appropriate to describe element x if and only if $\neg\theta$ is not possibly appropriate to describe x.

References

1. Atanassov, K.: Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems 20, 87–96 (1986)
2. Atanassov, K., Gargov, G.: Interval Valued Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems 31(3), 343–349 (1989)
3. Atanassov, K.: On Intuitionistic Fuzzy Negations and De Morgan Laws. In: Proceedings of the Eleventh International Conference IPMU 2006, Paris (2006)
4. Dubois, D., Prade, H.: An Introduction to Bipolar Representations of Information and Preference. International Journal of Intelligent Systems 23, 866–877 (2008)
5. Dubois, D., Gottwald, S., Hajek, P., Kacprzyk, J., Prade, H.: Terminological Difficulties in Fuzzy Set Theory - The Case of 'Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems 156, 485–491 (2005)
6. Dubois, D., Prade, H.: The three semantics of fuzzy sets. Fuzzy Sets and Systems 90, 141–150 (1997)
7. Grattan-Guiness, I.: Fuzzy Membership Mapped onto Interval and Many-Valued Quantities. Z. Math. Logik Grundladen Math. 22, 149–160 (1975)
8. Goodman, I.R.: Fuzzy Sets as Equivalence Classes of Random Sets. In: Yager, R. (ed.) Fuzzy Set and Possibility Theory, pp. 327–342 (1982)
9. Goodman, I.R., Nguyen, H.T.: Uncertainty Models for Knowledge Based Systems. North Holland, Amsterdam (1985)
10. Hampton, J.A.: Concepts as Prototypes. In: Ross, B.H. (ed.) The Psychology of Learning and Motivation: Advances in Research and Theory, vol. 46, pp. 79–113 (2006)
11. Jahn, K.U.: Intervall-wertige Mengen. Math. Nach. 68, 115–132 (1975)
12. Lawry, J., Tang, Y.: Uncertainty modelling for vague concepts: A prototype theory approach. Artificial Intelligence 173, 1539–1558 (2009)
13. Lawry, J.: Appropriateness measures: an uncertainty model for vague concepts. Synthese 161, 255–269 (2008)
14. Nguyen, H.: On modeling of linguistic information using random sets. Information Sciences 34, 265–274 (1984)
15. Parikh, R.: Vague Predicates and Language Games. Theoria (Spain) XI(27), 97–107 (1996)
16. Rosch, E.H.: Natural Categories. Cognitive Psychology 4, 328–350 (1973)
17. Rosch, E.H.: Cognitive Representation of Semantic Categories. Journal of Experimental Psychology: General 104, 192–233 (1975)

18. Ruspini, E.H.: On the Semantics of Fuzzy Logic. International Journal of Approximate Reasoning 5, 45–88 (1991)
19. Sambuc, R.: Fonctions-floues. Application a l'aide au diagnostic en pathologie thyroidienne, PhD Thesis, Univ. Marseille, France (1975)
20. Williamson, T.: Vagueness, Routledge (1994)
21. Zadeh, L.A.: The Concept of a Linguistic Variable and its Application to Approximate Reasoning:I. Information Sciences 8, 199–249 (1975)

Hesitation Degrees as the Size of Ignorance Combined with Fuzziness

Maciej Wygralak

Faculty of Mathematics and Computer Science, Adam Mickiewicz University
Umultowska 87, 61-614 Poznań, Poland
wygralak@amu.edu.pl

Abstract. We present a general, triangular norm-based approach to hesitation degrees related to I-fuzzy sets, Atanassov's intuitionistic fuzzy sets. Our main aim will be a closer look at hesitation degrees generated by the three basic t-norms. We will give some illustrative examples showing a true practical sense of those degrees and emphasizing that the nature of triangular norm-based hesitation degrees is generally twofold. They combine the size of ignorance and a fuzziness factor whenever the minimum or a strict t-norm is involved.

Keywords: Atanassov's intuitionistic fuzzy set, hesitation degree, triangular norm, fuzziness measure.

1 Introduction

The recent decade has brought a rapid growth of interest and research efforts in the areas of theory and applications of I-fuzzy sets, Atanassov's intuitionistic fuzzy sets. This is hardly surprising as I-fuzzy sets turn out to be a convenient and effective tool for representing incompletely known fuzzy sets, i.e. a tool for modeling a mixture of *imprecision* and *incompleteness* of information.

As one knows, I-fuzzy sets are formally equivalent with interval-valued fuzzy sets (see e.g. [4]). However, from a practical viewpoint, those two concepts are different. In contrast to interval-valued fuzzy sets, I-fuzzy sets make it possible and even force us to think about and look, say, at given decision alternatives in the language of positive and negative information: in the language of advantages and disadvantages, positive and negative features, satisfaction and dissatisfaction, trust and distrust, etc. And just this bipolar optics offered by I-fuzzy sets seems to be difficult to overestimate as psychological investigations clearly suggest that decision makers have a tendency to focus on positive sides and, simultaneously, to forget about negative sides of decisions they consider.

Let us move on to a more formal discussion. Recollect that an *I-fuzzy set* is a pair $\mathcal{E} = (A^+, A^-)$ of fuzzy sets $A^+, A^-: \boldsymbol{U} \to [0, 1]$ with (see [1, 2])

$$\text{i.e.} \quad \begin{gathered} A^+ \subset (A^-)', \\ A^+(x) + A^-(x) \leq 1 \ \text{ for each } x \in \boldsymbol{U}. \end{gathered} \tag{1}$$

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 629–636, 2010.

A^+ is then understood as a *membership function*, whereas A^- is interpreted as a *nonmembership function*. Consequently, $A^+(x)$ forms a *membership degree* and $A^-(x)$ is a *nonmembership degree* of x. $\mathcal{E}$ models an incompletely known fuzzy set A. Speaking more precisely, we assume all one knows about A is that

$$A^+ \subset A \subset (A^-)' \text{ and, hence, } A^- \subset A' \subset (A^+)'. \tag{2}$$

The component A^+, *positive information* about A, is thus a lower bound on A, whereas A^-, *negative information*, forms a lower bound on A':

$$A(x) \in [A^+(x), 1 - A^-(x)], \quad A'(x) \in [A^-(x), 1 - A^+(x)]. \tag{3}$$

A fundamental characterization of an I-fuzzy set is its *hesitation area*, a fuzzy set $\chi_{\mathcal{E}}$ with

$$\chi_{\mathcal{E}}(x) = 1 - A^+(x) - A^-(x), \tag{4}$$

which is known as the *degree* (*index*, *margin*) *of hesitation* concerning x. It is the *size of ignorance* as to $A(x)$, the length of any of the intervals in (3). Trivially, one has

$$A^+(x) + \chi_{\mathcal{E}}(x) = (A^-)'(x). \tag{5}$$

The subject literature proposes at least two more general approaches to hesitation degrees. The first one offers a triangular-norm based formulation of those degrees ([7]; see also [8,12]). The second approach, presented in [3], can be characterized in terms of implication operators. We like to focus on the former. Section 2 will recollect its idea and general properties. However, our main aim will be a look at hesitation degrees based on the three basic t-norms (Section 3). The Łukasiewicz t-norm $\boldsymbol{t}_Ł$ with $a \boldsymbol{t}_Ł b = 0 \vee (a + b - 1)$ then leads to the standard hesitation index from (4). Hesitation degrees generated by the minimum $\wedge$ or the product t-norm $\boldsymbol{t}_P$ with $a \boldsymbol{t}_P b = ab$ are different. They combine the size of ignorance as to $A(x)$ and a fuzziness factor. Section 4 will present illustrative examples showing a true practical sense of those combined hesitation degrees. Throughout, we will assume that the reader is familiar with basic notions and facts from the area of triangular norms (see e.g. [6, 11]).

2 I-Fuzzy Sets and Hesitation Degrees – A General Approach

First, let us replace (1) by a more general relationship between the components A^+ and A^- of an I-fuzzy set $\mathcal{E} = (A^+, A^-)$, namely

$$\text{i.e.} \quad \begin{gathered} A^+ \subset (A^-)^{\nu}, \\ A^+(x) \leq \nu(A^-(x)) \text{ for each } x \in \boldsymbol{U}, \end{gathered} \tag{6}$$

where ν is a strong negation and B^{ν} denotes the complement of B generated by ν. This generalization implies obvious modifications in (2)-(3). Clearly, (6) collapses to (1) if the standard Łukasiewicz negation $\nu_Ł$ is used, $\nu_Ł(a) = 1 - a$.

The *hesitation area* $\chi_{\mathcal{E}}$ of $\mathcal{E}$ can then be defined as

$$\chi_{\mathcal{E}} = (A^+)^{\nu} \cap_{\boldsymbol{t}} (A^-)^{\nu},$$

i.e.

$$\chi_{\mathcal{E}}(x) = \nu(A^+(x)) \,\boldsymbol{t}\, \nu(A^-(x)) \tag{7}$$

with a t-norm $\boldsymbol{t}$. (7) reflects a simple intuition saying that hesitation consists in "not *yes* and not *no*", "not *pro* and not *contra*", etc. For instance, using a Schweizer-Sklar t-norm $\boldsymbol{t}_{S,p}$ with

$$a \,\boldsymbol{t}_{S,p}\, b = [0 \vee (a^p + b^p - 1)]^{1/p}, \quad p > 0,$$

one gets

$$\chi_{\mathcal{E}}(x) = [1 - (A^+(x))^p - (A^-(x))^p]^{1/p}, \tag{8}$$

whereas (6) collapses to

$$(A^+(x))^p + (A^-(x))^p \leq 1 \quad \text{for each } x \in \boldsymbol{U}. \tag{9}$$

Let us formulate a few general properties of (7). We easily notice that generally

$$\chi_{\mathcal{E}}(x) = 1 \;\Leftrightarrow\; A^+(x) = A^-(x) = 0. \tag{10}$$

It is convenient, and quite sufficient from the viewpoint of applications, to restrict our further discussion to Archimedean t-norms and $\wedge$. $\boldsymbol{t}^{\nu}$ will denote the t-conorm which is ν-dual to $\boldsymbol{t}$, $a \,\boldsymbol{t}^{\nu}\, b = \nu(\nu(a) \,\boldsymbol{t}\, \nu(b))$, whereas the negation induced by $\boldsymbol{t}$ will be denoted by $\nu_{\boldsymbol{t}}$, $\nu_{\boldsymbol{t}}(a) = \sup\{c: a \,\boldsymbol{t}\, c = 0\}$.

Theorem 1. *Let* $\mathcal{E} = (A^+, A^-)$. *If* $\boldsymbol{t}$ *is nilpotent,* $\nu = \nu_{\boldsymbol{t}}$, *and h is the normed generator of* $\boldsymbol{t}^{\nu}$, *then*

(a) $\chi_{\mathcal{E}}(x) = h^{-1}(1 - h(A^+(x)) - h(A^-(x)))$,

(b) $A^+(x) \,\boldsymbol{t}^{\nu}\, \chi_{\mathcal{E}}(x) = \nu(A^-(x))$

for each $x \in \boldsymbol{U}$.

If a nilpotent t-norm is used, hesitation degrees thus form (a sort of) a "pure" size of ignorance as to $A(x)$: $\chi_{\mathcal{E}}(x)$ is isomorphic to $1 - A^+(x) - A^-(x)$, the length of the interval $[A^+(x), \nu_{\mathcal{L}}(A^-(x))]$ (see e.g. (8)). Moreover, by (b), $A^+(x)$ and $\chi_{\mathcal{E}}(x)$ "sum up" to $\nu(A^-(x))$ (cf. (5)). By (a), one also has for nilpotent t-norms

$$\chi_{\mathcal{E}}(x) = 0 \;\Leftrightarrow\; A^-(x) = \nu(A^+(x)),$$

i.e.

$$\chi_{\mathcal{E}} = 1_{\varnothing} \;\Leftrightarrow\; \mathcal{E} \text{ collapses to a fuzzy set}. \tag{11}$$

The task of interpreting the $\chi_{\mathcal{E}}(x)$'s becomes more sophisticated if $\boldsymbol{t}$ is strict or $\boldsymbol{t} = \wedge$, i.e. $\boldsymbol{t}$ has no zero divisors. By (7), we then have

$$\chi_{\mathcal{E}}(x) = 0 \;\Leftrightarrow\; (A^+(x), A^-(x)) \in \{(1, 0), (0, 1)\}$$

and, hence,

$$\chi_{\mathcal{E}} = 1_{\varnothing} \;\Leftrightarrow\; \mathcal{E} \text{ is a set}. \tag{12}$$

If $\mathcal{E}$ forms an ordinary fuzzy set, $\mathcal{E} = (A, A^{\nu})$, one gets

$$\chi_{\mathcal{E}}(x) = A(x)\, \boldsymbol{t}\, \nu(A(x)), \tag{13}$$

i.e. $\chi_{\mathcal{E}}(x)$ collapses to a $\boldsymbol{t}$-based fuzziness index of $A(x)$ (see Section 3). So, generally, $\chi_{\mathcal{E}}(x)$ seems to be the size of ignorance as to $A(x)$ combined with a fuzziness index of $A(x)$. Details and examples of this combination are the subject of the next sections.

3 A Look through the Three Basic t-Norms

Recollect that if the standard negation $\nu_{\text{Ł}}$ is used, then

$$Fuzz(B) = \sum_{x \in \boldsymbol{U}} \varphi(B(x)) \tag{14}$$

is called the *fuzziness measure* (*entropy measure of fuzziness*, being more specific) of a fuzzy set B whenever φ: [0, 1] → [0, 1] satisfies the following (see [5, 9]):

(a) $\varphi(0) = \varphi(1) = 0$ and $\varphi(a) > 0$ for each $a \in (0, 1)$,
(b) $\varphi(a) < \varphi(0.5)$ for each $a \neq 0.5$,
(c) φ is nondecreasing on $[0, 0.5]$ and nonincreasing on $[0.5, 1]$,
(d) $\varphi(a) = \varphi(1-a)$ for each $a \in [0, 1]$.

$Fuzz(B)$ says how fuzzy B is. In other words, it is a measure of imprecision of information carried by B. $\varphi(a)$ can be viewed as a *fuzziness index* of a and, thus, $Fuzz(B)$ is just the sum of all the fuzziness indices of the membership degrees in B. If convenient, one can refer to $\varphi(B(x))$ as the fuzziness index of the very x.

We are now ready to move on to a discussion of the hesitation degrees from (7) for the three basic t-norms.

- $\boldsymbol{t} = \boldsymbol{t}_{\text{Ł}}$ and $\nu = \nu_t$, i.e. $\nu = \nu_{\text{Ł}}$. $\chi_{\mathcal{E}}(x)$ is then nothing else than the hesitation degree

$$\chi_{\mathcal{E}}(x) = 1 - A^{+}(x) - A^{-}(x)$$

from (4), i.e. it is just the size of ignorance as to $A(x)$.

- $\boldsymbol{t} = \wedge$ and $\nu = \nu_{\text{Ł}}$. Now

$$\chi_{\mathcal{E}}(x) = 1 - A^{+}(x) \vee A^{-}(x) = 1 - A^{+}(x) - A^{-}(x) + A^{+}(x) \wedge A^{-}(x). \tag{15}$$

Since $\varphi(a) = a \wedge (1-a)$ is a fuzziness index and

$$A^{+}(x) \wedge A^{-}(x) = \varphi(A^{+}(x)) \wedge \varphi(1 - A^{-}(x)), \tag{16}$$

we conclude that $A^{+}(x) \wedge A^{-}(x)$ is the minimum possible fuzziness index of $A(x)$ (see also Section 4). By (15), $\chi_{\mathcal{E}}(x)$ thus becomes the sum of the size of ignorance as to $A(x)$ and the minimum possible fuzziness index of $A(x)$.

- $t = t_P$ and $v = v_L$. Then

$$\chi_{\mathcal{E}}(x) = 1 - A^+(x) - A^-(x) + A^+(x) \cdot A^-(x). \tag{17}$$

We see that

$$A^+(x) \cdot A^-(x) \le \varphi(A^+(x)) \wedge \varphi(1 - A^-(x)) \tag{18}$$

with fuzziness index $\varphi(a) = a \cdot (1-a)$. Consequently, $\chi_{\mathcal{E}}(x)$ is the sum of the size of ignorance as to $A(x)$ and a lower evaluation of the minimum possible fuzziness index of $A(x)$.

4 Illustrative Examples

Consider the following collection of 8 bottles of water which are more or less full. The water level in some bottles is partially unknown as they are partially covered up and, thus, invisible. That level is somewhere between the horizontal sides of the black rectangles (curtains) covering up the bottles.

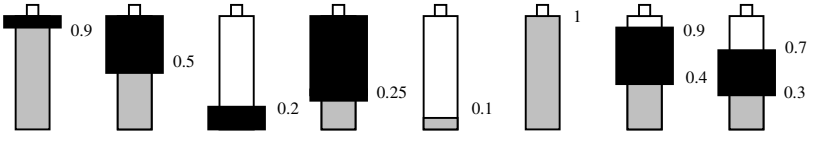


Fig. 1. Incompletely known fuzzy set of full bottles

In terms of fuzzy sets, we thus deal with a universe of 8 bottles, $\boldsymbol{U} = \{b_1, \ldots, b_8\}$, and a fuzzy set A of full bottles. However, A is incompletely known: our knowledge about $A(b_i)$, the water level in bottle b_i, is generally only partial. We see that Fig. 1 presents a model situation which can be easily transferred, say, to

- a collection of resources (financial, natural, …) whose levels are partially unknown,
- preferences with incompletely known degrees of intensity,
- modules of a system with incompletely known extents of faults,
- decisions with incompletely known consequences,
- partially unknown levels of trust, satisfaction, …

Trying to represent A as an I-fuzzy set, let us look at each bottle from two opposite perspectives: in terms of its advantages and disadvantages, of being full and being empty. We then get:

$$\begin{gathered} A^+(b_1) = 0.9,\ A^+(b_2) = 0.5,\ A^+(b_3) = 0,\ A^+(b_4) = 0.25, \\ A^+(b_5) = 0.1,\ A^+(b_6) = 1,\ A^+(b_7) = 0.4,\ A^+(b_8) = 0.3, \\ \text{and} \\ A^-(b_1) = 0,\ A^-(b_2) = 0,\ A^-(b_3) = 0.8,\ A^-(b_4) = 0, \\ A^-(b_5) = 0.9,\ A^-(b_6) = 0,\ A^-(b_7) = 0.1,\ A^-(b_8) = 0.3, \end{gathered} \tag{19}$$

Speaking practically, $A^+(b_i)$ is the visible degree to which b_i is full, whereas $A^-(b_i)$ forms the visible degree to which b_i is empty. We are ready to move on to the issue of related hesitation degrees involving the three basic t-norms.

- $t = t_Ł$ and $\nu = \nu_Ł$. So,

$$\chi_{\mathcal{E}}(b_i) = 1 - A^+(b_i) - A^-(b_i)$$

with $\chi_{\mathcal{E}}(b_i)$ collapsing to the size of ignorance as to $A(b_i)$, the height of the curtain covering up bottle b_i. Using (19), we get the following numerical results:

$$\begin{gathered}\chi_{\mathcal{E}}(b_1) = 0.1,\ \chi_{\mathcal{E}}(b_2) = 0.5,\ \chi_{\mathcal{E}}(b_3) = 0.2,\ \chi_{\mathcal{E}}(b_4) = 0.75,\\ \chi_{\mathcal{E}}(b_5) = 0,\ \chi_{\mathcal{E}}(b_6) = 0,\ \chi_{\mathcal{E}}(b_7) = 0.5,\ \chi_{\mathcal{E}}(b_8) = 0.4.\end{gathered} \tag{20}$$

Worth recollecting is here the notion of the *sigma count* $\sigma(B)$ of a fuzzy set B, the sum of all the membership degrees in B forming a frequently used type of cardinality of B ([14]; see also [10, 13]). By (20), we obtain

$$\sigma(\chi_{\mathcal{E}}) = \sum_i \chi_{\mathcal{E}}(b_i) = 2.45.$$

This is a total size of ignorance as to the content of the bottles. That lack of knowledge is equivalent to a complete lack of information about the content of 2.45 bottles. As we thus see, the cardinality $\sigma(\chi_{\mathcal{E}})$ offers a very interesting type of information.

- $t = \wedge$ and $\nu = \nu_Ł$. By (15),

$$\chi_{\mathcal{E}}(b_i) = 1 - A^+(b_i) - A^-(b_i) + A^+(b_i) \wedge A^-(b_i).$$

This hesitation degree is now the sum of the size of ignorance as to $A(b_i)$ and the minimum possible fuzziness index of $A(b_i)$. Using (19), we get

$$\begin{gathered}\chi_{\mathcal{E}}(b_1) = 0.1,\ \chi_{\mathcal{E}}(b_2) = 0.5,\ \chi_{\mathcal{E}}(b_3) = 0.2,\ \chi_{\mathcal{E}}(b_4) = 0.75,\\ \chi_{\mathcal{E}}(b_5) = 0.1,\ \chi_{\mathcal{E}}(b_6) = 0,\ \chi_{\mathcal{E}}(b_7) = 0.6,\ \chi_{\mathcal{E}}(b_8) = 0.7.\end{gathered} \tag{21}$$

In other words, (21) gives us considerably more information (2 in 1) in comparison with (20). By (21), we get

$$\sigma(\chi_{\mathcal{E}}) = \sum_i (1 - A^+(b_i) - A^-(b_i)) + \sum_i A^+(b_i) \wedge A^-(b_i) = 2.45 + 0.5 = 2.95,$$

which is a total size of ignorance plus the minimum possible entropy measure of fuzziness of A, i.e. the minimum possible amount of water which has to be poured out from or poured into the bottles if one likes to make all of them totally classifiable, totally full or totally empty. The second component, 0.5, is thus the minimum possible distance of the system of bottles from a totally classifiable system.

- $t = t_P$ and $\nu = \nu_Ł$. Then (see (17))

$$\chi_{\mathcal{E}}(b_i) = 1 - A^+(b_i) - A^-(b_i) + A^+(b_i) \cdot A^-(b_i)$$

and, applying (19), one gets

$$\begin{aligned} &\chi_{\mathcal{E}}(b_1) = 0.1,\ \chi_{\mathcal{E}}(b_2) = 0.5,\ \chi_{\mathcal{E}}(b_3) = 0.2,\ \chi_{\mathcal{E}}(b_4) = 0.75, \\ &\chi_{\mathcal{E}}(b_5) = 0.09,\ \chi_{\mathcal{E}}(b_6) = 0,\ \chi_{\mathcal{E}}(b_7) = 0.54,\ \chi_{\mathcal{E}}(b_8) = 0.49. \end{aligned} \tag{22}$$

$\chi_{\mathcal{E}}(b_i)$ becomes the sum of the size of ignorance as to $A(b_i)$ and a lower evaluation of the minimum possible fuzziness index $\varphi(a) = a \cdot (1-a)$ of $A(b_i)$. Finally,

$$\sigma(\chi_{\mathcal{E}}) = \sum_i (1 - A^+(b_i) - A^-(b_i)) + \sum_i A^+(b_i) \cdot A^-(b_i) = 2.45 + 0.22 = 2.67.$$

This is a total size of ignorance plus a lower evaluation of the minimum possible entropy measure of fuzziness of A (see (18)).

5 Conclusions

We presented a natural, triangular norm-based generalization of hesitation degrees related to I-fuzzy sets. It seems that this generalized look casts new light on the very essence of hesitation degrees by emphasizing their relationship with fuzziness measures whenever the minimum or a strict t-norm is involved. This was illustrated by simple examples involving the three basic t-norms. The results will be applied in group decision making algorithms.

Acknowledgments. This research was supported by the Ministry of Science and Higher Education Grant N N519 384936.

References

1. Atanassov, K.T.: Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems 20, 87–96 (1986)
2. Atanassov, K.T.: Intuitionistic Fuzzy Sets. Theory and Applications. Physica-Verlag, Heidelberg (1999)
3. Barrenechea, E., Bustince, H., Pagola, M., Fernandez, J., Sanz, J.: Generalized Atanassov's Intuitionistic Fuzzy Index. Construction Methods. In: 2009 IFSA World Congress and 2009 EUSFLAT Conference, Lisbon, Portugal, pp. 478–482 (2009)
4. Bustince, H., Montero, J., Pagola, M., Barrenechea, E., Gómez, D.: A Survey of Interval-Valued Fuzzy Sets. In: Pedrycz, W., Skowron, A., Kreinovich, V. (eds.) Handbook of Granular Computing, pp. 491–516. Wiley, New York (2008)
5. De Luca, A., Termini, S.: On Some Algebraic Aspects of the Measures of Fuzziness. In: Gupta, M.M., Sanchez, E. (eds.) Fuzzy Information and Decision Processes, pp. 17–24. North-Holland, Amsterdam (1982)
6. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms. Kluwer, Dordrecht (2000)

7. Pankowska, A., Wygralak, M.: On Hesitation Degrees in IF-Set Theory. In: Rutkowski, L., Siekmann, J.H., Tadeusiewicz, R., Zadeh, L.A. (eds.) ICAISC 2004. LNCS (LNAI), vol. 3070, pp. 338–343. Springer, Heidelberg (2004)
8. Pankowska, A., Wygralak, M.: General IF-sets with Triangular Norms and Their Applications to Group Decision Making. Information Sciences 176, 2713–2754 (2006)
9. Pedrycz, W., Gomide, F.: Fuzzy Systems Engineering: Towards Human Centric Computing. Wiley, Hoboken (2007)
10. Szmidt, E., Kacprzyk, J.: Entropy for Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems 118, 467–477 (2001)
11. Wygralak, M.: Cardinalities of Fuzzy Sets. Springer, Heidelberg (2003)
12. Wygralak, M.: Representing Incomplete Knowledge about Fuzzy Sets. In: 5th EUSFLAT Conference, Ostrava, Czech Rep., vol. II, pp. 287–292 (2007)
13. Wygralak, M.: Model Examples of Applications of Sigma f -Counts to Counting in Fuzzy and I-Fuzzy Sets. In: Atanassov, K.T., Hryniewicz, O., Kacprzyk, J., Krawczak, M., Nahorski, Z., Szmidt, E., Zadrożny, S. (eds.) Advances in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics. Applications, vol. II, pp. 187–196. EXIT Acad. Publ., Warsaw (2008)
14. Zadeh, L.A.: A Computational Approach to Fuzzy Quantifiers in Natural Languages. Comput. and Math. with Appl. 9, 149–184 (1983)

On the Distributivity of Implication Operations over t-Representable t-Norms Generated from Strict t-Norms in Interval-Valued Fuzzy Sets Theory

Michał Baczyński

Institute of Mathematics, University of Silesia,
40-007 Katowice, ul. Bankowa 14, Poland
michal.baczynski@us.edu.pl

Abstract. Recently, many papers have appeared dealing with the distributivity of fuzzy implications over t-norms, t-conorms and uninorms. These equations have a very important role to play in efficient inferencing in approximate reasoning, especially fuzzy control systems. In [2] we have discussed the distributivity of continuous functions, in particular implication operations, over t-representable t-norms, generated from the product t-norm, in intuitionistic fuzzy sets theory. In this work we continue investigations for interval-valued fuzzy sets theory, but without any regular assumptions on an unknown function $\mathcal{I}$.

Keywords: Interval-valued fuzzy sets, Intuitionistic fuzzy sets, Fuzzy implication, Triangular norm, Distributivity, Functional equations.

1 Introduction

Distributivity of fuzzy implications over different fuzzy logic connectives has been studied in the recent past by many authors (see [1], [19], [4], [17], [18], [3]). These equations have a very important role to play in efficient inferencing in approximate reasoning, especially fuzzy control systems (see [6]). Recently, in [2], we have discussed the continuous solutions $\mathcal{I}$ of the following distributive equation

$$\mathcal{I}(x,\mathcal{T}(y,z)) = \mathcal{T}(\mathcal{I}(x,y),\mathcal{I}(x,z)),$$

where $\mathcal{T}$ is a t-representable t-norm defined over the interval-valued fuzzy sets and generated from the product t-norm . In this paper we continue these investigations, but without any assumption on function $\mathcal{I}$. As a byproduct result we have obtained the solutions of the following functional equation

$$f(u_1+v_1,u_2+v_2) = f(u_1,u_2)+f(v_1,v_2), \qquad (u_1,u_2),(v_1,v_2)\in L^{\infty},$$

where $L^{\infty} = \{(u_1,u_2)\in[0,\infty]^2 \,:\, u_1 \geq u_2\}$ and $f\colon L^{\infty}\to[0,\infty]$ is an unknown function. Observe that this equation is correctly written, since if $u,v\in L^{\infty}$, then $u+v\in L^{\infty}$. Such theoretical developments connected with solutions of different

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 637–646, 2010.

functional equations can be also useful in other topics like fuzzy mathematical morphology (see [7]) or similarity measures (cf. [5]).

We assume that the reader is familiar with the notion of intuitionistic (by Atanassov) fuzzy sets theory and interval-valued fuzzy sets theory. It is important to notice that in [8] it is shown that both theories are equivalent from the mathematical point of view. Since we are limited in number of pages, in this article we discuss main results in the language of interval-valued fuzzy sets, but they can be easily transformed to the intuitionistic fuzzy case. Let us define

$$L^I = \{(x_1, x_2) \in [0,1]^2 \ : \ x_1 \leq x_2\},$$
$$(x_1, x_2) \leq_{L^I} (y_1, y_2) \iff x_1 \leq y_1 \wedge x_2 \leq y_2.$$

In the sequel, if $x \in L^I$, then we denote it by $x = [x_1, x_2]$. One can easily observe that $\mathcal{L}^I = (L^I, \leq_{L^I})$ is a complete lattice with units $0_{\mathcal{L}^I} = [0,0]$ and $1_{\mathcal{L}^I} = [1,1]$. An interval-valued fuzzy set on X is a mapping $A\colon X \to L^I$.

2 Basic Fuzzy Connectives

We also assume that the reader is familiar with the classical results concerning basic fuzzy logic connectives, but we briefly mention some of the results employed in the rest of the work.

Definition 2.1. *Let $\mathcal{L} = (L, \leq_L)$ be a complete lattice. An associative, commutative operation $\mathcal{T}\colon L^2 \to L$ is called a t-norm if it is increasing and $1_{\mathcal{L}}$ is the neutral element of $\mathcal{T}$.*

Definition 2.2. *We say that a t-norm T on $([0,1], \leq)$ is strict, if it is continuous and strictly monotone, i.e., $T(x,y) < T(x,z)$ whenever $x > 0$ and $y < z$.*

The following characterization of strict t-norms is very well-known.

Theorem 2.3 ([14]). *A function $T\colon [0,1]^2 \to [0,1]$ is a strict t-norm if and only if there exists a continuous, strictly decreasing function $t\colon [0,1] \to [0,\infty]$ with $t(1) = 0$ and $t(0) = \infty$, which is uniquely determined up to a positive multiplicative constant, such that*

$$T(x,y) = t^{-1}(t(x) + t(y)), \qquad x, y \in [0,1].$$

T-norms on $\mathcal{L}^I$ can be defined in many ways. In our article we shall consider the following special class of t-norms.

Definition 2.4 (see [9]). *A t-norm $\mathcal{T}$ on $\mathcal{L}^I$ is called t-representable if there exist t-norms T_1 and T_2 on $([0,1], \leq)$ such that $T_1 \leq T_2$ and*

$$\mathcal{T}([x_1, x_2], [y_1, y_2]) = [T_1(x_1, y_1), T_2(x_2, y_2)], \qquad [x_1, x_2], [y_1, y_2] \in L^I.$$

It should be noted that not all t-norms on $\mathcal{L}^I$ are t-representable (see [9]).

One possible definition of an implication on $\mathcal{L}^I$ is based on the well-accepted notation from fuzzy sets theory introduced by Fodor and Roubens [13].

Definition 2.5. *Let $\mathcal{L} = (L, \leq_L)$ be a complete lattice. A function $\mathcal{I}\colon L^2 \to L$ is called a fuzzy implication on $\mathcal{L}$ if it is decreasing with respect to the first variable, increasing with respect to the second variable and fulfills the following conditions: $\mathcal{I}(0_\mathcal{L}, 0_\mathcal{L}) = \mathcal{I}(1_\mathcal{L}, 1_\mathcal{L}) = 1_\mathcal{L}$ and $\mathcal{I}(1_\mathcal{L}, 0_\mathcal{L}) = 0_\mathcal{L}$.*

Directly from the above definition we can deduce that each implication $\mathcal{I}$ on $\mathcal{L}$ satisfies also the normality condition $\mathcal{I}(0_\mathcal{L}, 1_\mathcal{L}) = 1_\mathcal{L}$. Consequently, every implication restricted to the set $\{0_\mathcal{L}, 1_\mathcal{L}\}^2$ coincides with the classical implication.

When $\mathcal{L} = ([0,1], \leq)$, then $\mathcal{I}$ is called a fuzzy implication, while if $\mathcal{L} = \mathcal{L}^I$, then $\mathcal{I}$ is called an interval-valued fuzzy implication. Detailed investigations on different classes of implications on above lattices and their algebraic properties were presented in [10] and [16].

3 Some New Results Pertaining to Functional Equations

In this section we show one new result related to functional equations, which will be crucial in the proof of main results. It can be seen as a generalization of the classical facts from the theory of functional equations and as a full version of Proposition 8 from [2]. In the proof we will use the following fact.

Proposition 3.1 ([3], Proposition 2). *For a function $f\colon [0,\infty] \to [0,\infty]$ the following statements are equivalent:*

(i) f satisfies the additive Cauchy functional equation $f(x+y) = f(x) + f(y)$, for all $x, y \in [0,\infty]$.

(ii) Either $f = \infty$, or $f = 0$, or $f(x) = \begin{cases} 0, & \text{if } x = 0 \\ \infty, & \text{if } x \in (0,\infty] \end{cases}$, or $f(x) = \begin{cases} 0, & \text{if } x \in [0,\infty) \\ \infty, & \text{if } x = \infty \end{cases}$, or there exists a unique constant $c \in (0,\infty)$ such that $f(x) = cx$, for all $x \in [0,\infty]$.

Proposition 3.2. *Let $L^\infty = \{(u_1, u_2) \in [0,\infty]^2 \ : \ u_1 \geq u_2\}$. For a function $f\colon L^\infty \to [0,\infty]$ the following statements are equivalent:*

(i) f satisfies the functional equation

$$f(u_1 + v_1, u_2 + v_2) = f(u_1, u_2) + f(v_1, v_2), \quad (u_1, u_2), (v_1, v_2) \in L^\infty. \tag{B1}$$

(ii) Either $f = 0$, or $f = \infty$, or

$$f(u_1, u_2) = \begin{cases} 0, & \text{if } u_2 = 0, \\ \infty, & \text{if } u_2 > 0, \end{cases} \tag{1}$$

or

$$f(u_1, u_2) = \begin{cases} 0, & \text{if } u_2 < \infty, \\ \infty, & \text{if } u_2 = \infty, \end{cases} \tag{2}$$

or

$$f(u_1,u_2)=\begin{cases}0, & \text{if } u_1=0,\\ \infty, & \text{if } u_1>0,\end{cases} \tag{3}$$

or

$$f(u_1,u_2)=\begin{cases}0, & \text{if } u_1=u_2<\infty,\\ \infty, & \text{if } u_2=\infty \text{ or } u_1>u_2,\end{cases} \tag{4}$$

or

$$f(u_1,u_2)=\begin{cases}0, & \text{if } u_2=0 \text{ and } u_1<\infty,\\ \infty, & \text{if } u_2>0 \text{ or } u_1=\infty,\end{cases} \tag{5}$$

or

$$f(u_1,u_2)=\begin{cases}0, & \text{if } u_1<\infty,\\ \infty, & \text{if } u_1=\infty,\end{cases} \tag{6}$$

or there exists unique $c\in(0,\infty)$ *such that*

$$f(u_1,u_2)=cu_2, \tag{7}$$

or

$$f(u_1,u_2)=\begin{cases}cu_1, & \text{if } u_1=u_2,\\ \infty, & \text{if } u_1>u_2,\end{cases} \tag{8}$$

or

$$f(u_1,u_2)=\begin{cases}cu_2, & \text{if } u_1<\infty,\\ \infty, & \text{if } u_1=\infty,\end{cases} \tag{9}$$

or

$$f(u_1,u_2)=\begin{cases}cu_1, & \text{if } u_2=0,\\ \infty, & \text{if } u_2>0,\end{cases} \tag{10}$$

or

$$f(u_1,u_2)=\begin{cases}c(u_1-u_2), & \text{if } u_2<\infty,\\ \infty, & \text{if } u_2=\infty,\end{cases} \tag{11}$$

or

$$f(u_1,u_2)=cu_1, \tag{12}$$

or there exist unique $c_1,c_2\in(0,\infty)$, $c_1\neq c_2$ *such that*

$$f(u_1,u_2)=\begin{cases}c_1(u_1-u_2)+c_2u_2, & \text{if } u_2<\infty,\\ \infty, & \text{if } u_2=\infty,\end{cases} \tag{13}$$

for all $(u_1,u_2)\in L^\infty$.

Proof. $(ii)\Longrightarrow(i)$ It is a direct calculation that the above functions satisfy (B1).

$(i)\Longrightarrow(ii)$ Let a function $f\colon L^\infty\to[0,\infty]$ satisfy equation (B1) for all $(u_1,u_2),(v_1,v_2)\in L^\infty$. Setting $u_1=v_1=\infty$ in (B1) we get

$$f(\infty,u_2+v_2)=f(\infty,u_2)+f(\infty,v_2),\qquad u_2,v_2\in[0,\infty].$$

Let us denote $f_\infty(x) := f(\infty, x)$, for $x \in [0, \infty]$. Therefore, we get

$$f_\infty(u_2 + v_2) = f_\infty(u_2) + f_\infty(v_2), \qquad u_2, v_2 \in [0, \infty].$$

For this equation we can use solutions described in Proposition 3.1. We have 5 possible cases for the function f_∞.

1. If $f_\infty = 0$, then putting $u_1 = u_2 = \infty$ in (B1) we have

$$f(\infty, \infty) = f(\infty, \infty) + f(v_1, v_2), \qquad (v_1, v_2) \in L^\infty,$$

thus $0 = 0 + f(v_1, v_2)$, hence we get first possible solution $f = 0$.

2. If $f_\infty(x) = \begin{cases} 0, & \text{if } x = 0 \\ \infty, & \text{if } x \in (0, \infty] \end{cases}$, then putting $u_1 = \infty$ in (B1) we have

$$f(\infty, u_2 + v_2) = f(\infty, u_2) + f(v_1, v_2), \qquad (v_1, v_2) \in L^\infty.$$

If we take $u_2 = v_2 = 0$ above, then we get $0 = 0 + f(v_1, 0)$, thus $f(v_1, 0) = 0$ for all $v_1 \in [0, \infty]$. If we take $u_2 = 0$ and $v_2 > 0$ above, then we get $\infty = 0 + f(v_1, v_2)$, thus $f(v_1, v_2) = \infty$. In summary, we get the solution (1).

3. If $f_\infty(x) = \begin{cases} 0, & \text{if } x < \infty \\ \infty, & \text{if } x \in [0, \infty) \end{cases}$, then putting $u_1 = \infty$ and $u_2 = 0$ in (B1) we have $f(\infty, v_2) = f(v_1, v_2)$, so we get the solution (2) in this case.
4. If $f_\infty(x) = cx$ with some real $c > 0$, then putting $u_1 = \infty$ and $u_2 = 0$ in (B1) we have $f(\infty, v_2) = f(v_1, v_2)$, so we get the solution (7) in this case.

Therefore, we need to solve our equation with the assumption that $f_\infty = \infty$. Setting now $u_2 = v_2 = 0$ in (B1) we get

$$f(u_1 + v_1, 0) = f(u_1, 0) + f(v_1, 0), \qquad u_1, v_1 \in [0, \infty].$$

Let us denote $f^0(x) := f(x, 0)$, for $x \in [0, \infty]$. Hence, we obtain

$$f^0(u_1 + v_1) = f^0(u_1) + f^0(v_1), \qquad u_1, v_1 \in [0, \infty].$$

For this equation we again can use solutions described in Proposition 3.1. We have 5 possible cases for the function f^0.

1. If $f^0 = 0$, then $f(\infty, 0) = 0$, which contradicts our assumption $f_\infty = \infty$.
2. If $f^0 = \infty$, then putting $u_1 = u_2 = 0$ in (B1) we have

$$f(v_1, v_2) = f(0, 0) + f(v_1, v_2), \qquad (v_1, v_2) \in L^\infty,$$

thus $f(v_1, v_2) = \infty + f(v_1, v_2)$, hence we get next possible solution $f = \infty$.

3. If $f^0(x) = \begin{cases} 0, & \text{if } x = 0 \\ \infty, & \text{if } x \in (0, \infty] \end{cases}$, then putting $u_2 = 0$ in (B1) we have

$$f(u_1 + v_1, v_2) = f(u_1, 0) + f(v_1, v_2), \qquad u_1 \in [0, \infty], (v_1, v_2) \in L^\infty.$$

Let us assume that $u_1 > 0$ and $v_1 = v_2$ above. Then we get

$$f(u_1 + v_2, v_2) = f(u_1, 0) + f(v_2, v_2), \qquad u_1 \in (0, \infty], v_2 \in [0, \infty].$$

hence

$$f(u_1 + v_2, v_2) = \infty, \qquad u_1 \in (0, \infty], v_2 \in [0, \infty].$$

Since $u_1 + v_2 \in (v_2, \infty]$, we have obtained the result that $f(x_1, x_2) = \infty$ for any $(x_1, x_2) \in L^\infty$ such that $x_1 > x_2$.
Let us take now $u_2 = u_1$ and $v_2 = v_1$ in (B1). Then we have

$$f(u_1 + v_1, u_1 + v_1) = f(u_1, u_1) + f(v_1, v_1), \qquad u_1, v_1 \in [0, \infty].$$

Let us denote $g(x) := f(x, x)$, for $x \in [0, \infty]$. Therefore, we get

$$g(u_1 + v_1) = g(u_1) + g(v_1), \qquad u_1, v_1 \in [0, \infty].$$

For this equation we again can use solutions described in Proposition 3.1. We have 5 possible cases for the function g.

(a) If $g = 0$, then $f(\infty, \infty) = 0$, which contradicts our assumption $f_\infty = \infty$.
(b) If $g = \infty$, then $f(0, 0) = \infty$, which contradicts our assumption 3. on function f^0.
(c) If $g(x) = \begin{cases} 0, & \text{if } x = 0 \\ \infty, & \text{if } x \in (0, \infty] \end{cases}$, then we get the solution (3) in this case.
(d) If $g(x) = \begin{cases} 0, & \text{if } x < \infty \\ \infty, & \text{if } x \in [0, \infty) \end{cases}$, then we get the solution (4) in this case.
(e) If $g(x) = cx$ with some $c > 0$, then we get the solution (8) in this case.

4. If $f^0(x) = \begin{cases} 0, & \text{if } x < \infty \\ \infty, & \text{if } x \in [0, \infty) \end{cases}$, then putting $u_2 = u_1$ and $v_2 = v_1$ in (B1) we have

$$f(u_1 + v_1, u_1 + v_1) = f(u_1, u_1) + f(v_1, v_1), \qquad u_1, v_1 \in [0, \infty].$$

Let us denote $g(x) := f(x, x)$, for $x \in [0, \infty]$. Therefore, we get

$$g(u_1 + v_1) = g(u_1) + g(v_1), \qquad u_1, v_1 \in [0, \infty].$$

For this equation we again can use solutions described in Proposition 3.1. We have 5 possible cases for the function g.

(a) If $g = 0$, then $f(\infty, \infty) = 0$, which contradicts our assumption $f_\infty = \infty$.
(b) If $g = \infty$, then $f(0, 0) = \infty$, which contradicts our assumption 4. on function f^0.
(c) If $g(x) = \begin{cases} 0, & \text{if } x = 0 \\ \infty, & \text{if } x \in (0, \infty] \end{cases}$, then taking into account all assumptions we get the solution (5).

(d) If $g(x) = \begin{cases} 0, & \text{if } x < \infty \\ \infty, & \text{if } x \in [0, \infty) \end{cases}$, then taking into account all assumptions we get the solution (6).

(e) If $g(x) = cx$ with some $c > 0$, then we get the solution (9) in this case.

5. Let $f^0(x) = c_1 x$, with some real $c_1 > 0$ for $x \in [0, \infty]$. Putting $u_2 = u_1$ and $v_2 = v_1$ in (B1) we have

$$f(u_1 + v_1, u_1 + v_1) = f(u_1, u_1) + f(v_1, v_1), \qquad u_1, v_1 \in [0, \infty].$$

Let us denote $g(x) := f(x, x)$, for $x \in [0, \infty]$. Therefore, we get

$$g(u_1 + v_1) = g(u_1) + g(v_1), \qquad u_1, v_1 \in [0, \infty].$$

For this equation we again can use solutions described in Proposition 3.1. We have 5 possible cases for the function g.

(a) If $g = 0$, then $f(\infty, \infty) = 0$, which contradicts our assumption $f_\infty = \infty$.

(b) If $g = \infty$, then $f(0, 0) = \infty$, which contradicts our assumption 5. on function f^0.

(c) If $g(x) = \begin{cases} 0, & \text{if } x = 0 \\ \infty, & \text{if } x \in (0, \infty] \end{cases}$, then putting $u_2 = 0$ and $v_1 = v_2$ in (B1) we get

$$f(u_1 + v_2, v_2) = f(u_1, 0) + f(v_2, v_2), \qquad u_1, v_2 \in [0, \infty],$$

thus

$$f(u_1 + v_2, v_2) = c_1 u_1 + \begin{cases} 0, & \text{if } v_2 = 0 \\ \infty, & \text{if } v_2 > 0 \end{cases} = \begin{cases} c_1 u_1, & \text{if } v_2 = 0 \\ \infty, & \text{if } v_2 > 0 \end{cases},$$

for any $u_1, v_2 \in [0, \infty]$. Therefore, this solution can be written as (10).

(d) If $g(x) = \begin{cases} 0, & \text{if } x < \infty \\ \infty, & \text{if } x \in [0, \infty) \end{cases}$, then similarly as earlier putting $u_2 = 0$ and $v_1 = v_2$ in (B1) we get

$$f(u_1 + v_2, v_2) = f(u_1, 0) + f(v_2, v_2), \qquad u_1, v_2 \in [0, \infty],$$

thus

$$f(u_1 + v_2, v_2) = c_1 u_1 + \begin{cases} 0, & \text{if } v_2 < \infty \\ \infty, & \text{if } v_2 = \infty \end{cases} = \begin{cases} c_1 u_1, & \text{if } v_2 < \infty \\ \infty, & \text{if } v_2 = \infty \end{cases},$$

for any $u_1, v_2 \in [0, \infty]$. Therefore, this solution can be written as (11).

(e) If $g(x) = c_2 x$ with some real $c_2 > 0$, then similarly as earlier putting $u_2 = 0$ and $v_1 = v_2$ in (B1) we get

$$f(u_1 + v_2, v_2) = f(u_1, 0) + f(v_2, v_2), \qquad u_1, v_2 \in [0, \infty],$$

thus

$$f(u_1 + v_2, v_2) = c_1 u_1 + c_2 v_2, \qquad u_1, v_2 \in [0, \infty].$$

If $c_1 = c_2$, then $f(u_1 + v_2, v_2) = c_1(u_1 + v_2)$, therefore this solution can be written as (12). If $c_1 \neq c_2$, then we get the solution (13) in this case. □

4 Distributive Equation for t-Representable t-Norms

In this section we will show how we can use solutions presented in Proposition 3.2 to obtain all solutions, in particular fuzzy implications, of our main distributive equation

$$\mathcal{I}(x,\mathcal{T}(y,z))=\mathcal{T}(\mathcal{I}(x,y),\mathcal{I}(x,z)),\qquad x,y,z\in L^I, \tag{14}$$

when a t-norm $\mathcal{T}$ on $\mathcal{L}^I$ is t-representable and generated from strict t-norms.

Let us assume that a t-representable t-norm $\mathcal{T}$ and a function $\mathcal{I}$ are the solutions of the functional equation (14) satisfying the required properties. Assume that projection mappings on $\mathcal{L}^I$ are defined as the following:

$$pr_1([x_1,x_2])=x_1,\qquad pr_2([x_1,x_2])=x_2,\qquad \text{for } [x_1,x_2]\in L^I.$$

At this situation our distributive equation has the following form

$$\begin{aligned}&\mathcal{I}([x_1,x_2],[T_1(y_1,z_1),T_2(y_2,z_2)])\\ &\quad=[T_1(pr_1(\mathcal{I}([x_1,x_2],[y_1,y_2])),pr_1(\mathcal{I}([x_1,x_2],[z_1,z_2]))),\\ &\qquad T_2(pr_2(\mathcal{I}([x_1,x_2],[y_1,y_2])),pr_2(\mathcal{I}([x_1,x_2],[z_1,z_2])))],\end{aligned}$$

for all $[x_1,x_2],[y_1,y_2],[z_1,z_2]\in L^I$. As a consequence we obtain the following two equations

$$\begin{aligned}&pr_1(\mathcal{I}([x_1,x_2],[T_1(y_1,z_1),T_2(y_2,z_2)]))\\ &\qquad=T_1(pr_1(\mathcal{I}([x_1,x_2],[y_1,y_2])),pr_1(\mathcal{I}([x_1,x_2],[z_1,z_2]))),\\ &pr_2(\mathcal{I}([x_1,x_2],[T_1(y_1,z_1),T_2(y_2,z_2)]))\\ &\qquad=T_2(pr_2(\mathcal{I}([x_1,x_2],[y_1,y_2])),pr_2(\mathcal{I}([x_1,x_2],[z_1,z_2]))),\end{aligned}$$

which are satisfied for all $[x_1,x_2],[y_1,y_2],[z_1,z_2]\in L^I$. Now, let us fix arbitrarily $[x_1,x_2]\in L^I$ and define two functions $g^1_{[x_1,x_2]},g^2_{[x_1,x_2]}\colon L^I\to L^I$ by

$$g^1_{[x_1,x_2]}(\cdot):=pr_1\circ\mathcal{I}([x_1,x_2],\cdot),\qquad g^2_{[x_1,x_2]}(\cdot):=pr_2\circ\mathcal{I}([x_1,x_2],\cdot).$$

As a consequence we get the following two equations

$$\begin{aligned}g^1_{[x_1,x_2]}([T_1(y_1,z_1),T_2(y_2,z_2)])&=T_1(g^1_{[x_1,x_2]}([y_1,y_2]),g^1_{[x_1,x_2]}([z_1,z_2])),\\ g^2_{(x_1,x_2)}([T_1(y_1,z_1),T_2(y_2,z_2)])&=T_2(g^2_{[x_1,x_2]}([y_1,y_2]),g^2_{[x_1,x_2]}([z_1,z_2])).\end{aligned}$$

When we assume that $T_1=T_2=T$, then in both cases we have obtained the bisymmetry equation. The continuous and strictly increasing solutions are known even for domain L^I (see [15]). But in our investigation t-norms are not strictly increasing, since $T(0,y)=0$, so we cannot use results from the previous article. Let us assume that $T_1=T_2=T$ is a strict t-norm. Using the representation theorem of strict t-norms (Theorem 2.3) we can transform our problem to the following equation (for a simplicity we deal only with g^1 now):

$$\begin{aligned}&g^1_{[x_1,x_2]}([t^{-1}(t(y_1)+t(z_1)),t^{-1}(t(y_2)+t(z_2))])\\ &\qquad=t^{-1}(t(g^1_{[x_1,x_2]}([y_1,y_2]))+t(g^1_{[x_1,x_2]}([z_1,z_2]))).\end{aligned}$$

Hence

$$t \circ g^1_{[x_1,x_2]}([t^{-1}(t(y_1)+t(z_1)),t^{-1}(t(y_2)+t(z_2))]) = t \circ g^1_{[x_1,x_2]}([y_1,y_2]) + t \circ g^1_{[x_1,x_2]}([z_1,z_2]).$$

This equation can be written in the following form:

$$t \circ g^1_{[x_1,x_2]}([t^{-1}(t(y_1)+t(z_1)),t^{-1}(t(y_2)+t(z_2))]) = t \circ g^1_{[x_1,x_2]}([t^{-1}(t(y_1)),t^{-1}(t(y_2))]) + t \circ g^1_{[x_1,x_2]}([t^{-1}(t(z_1)),t^{-1}(t(z_2))]).$$

Let us put $t(y_1) = u_1$, $t(y_2) = u_2$, $t(z_1) = v_1$ and $t(z_2) = v_2$. Of course $u_1, u_2, v_1, v_2 \in [0,\infty]$. Moreover $[y_1,y_2],[z_1,z_2] \in L^I$, thus $y_1 \leq y_2$ and $z_1 \leq z_2$. The generator t is strictly decreasing, so $u_1 \geq u_2$ and $v_1 \geq v_2$. If we put

$$f_{[x_1,x_2]}(a,b) := t \circ pr_1 \circ \mathcal{I}([x_1,x_2],[t^{-1}(a),t^{-1}(b)]),$$

for $a, b \in [0,\infty]$ and $a \geq b$, then we get the following functional equation

$$f_{[x_1,x_2]}(u_1+v_1,u_2+v_2) = f_{[x_1,x_2]}(u_1,u_2) + f_{[x_1,x_2]}(v_1,v_2), \tag{15}$$

where $(u_1,u_2),(v_1,v_2) \in L^\infty$. In a same way we can repeat all the above calculations but for the function g^2, to obtain the following functional equation

$$f^{[x_1,x_2]}(u_1+v_1,u_2+v_2) = f^{[x_1,x_2]}(u_1,u_2) + f^{[x_1,x_2]}(v_1,v_2), \tag{16}$$

where

$$f^{[x_1,x_2]}(a,b) := t \circ pr_2 \circ \mathcal{I}([x_1,x_2],[t^{-1}(a),t^{-1}(b)]).$$

Observe that (15) and (16) are exactly our functional equation (B1). Therefore, using solutions of Proposition 3.2, we are able to obtain the description of the vertical section $\mathcal{I}([x_1,x_2],\cdot)$ for a fixed $[x_1,x_2] \in L^I$. Since in this proposition we have 15 possible solutions, we should have 225 different solutions of (14). Observe now that some of these solutions are not good, since the range of $\mathcal{I}$ is L^I. Since in this paper it is not possible to discuss all solutions we only present one positive example and one negative.

Let us firstly assume that $f_{[x_1,x_2]}(a,b) = 0$ and $f^{[x_1,x_2]}(a,b) = \infty$. Then $t \circ pr_1 \circ \mathcal{I}([x_1,x_2],[t^{-1}(a),t^{-1}(b)]) = 0$ and $t \circ pr_2 \circ \mathcal{I}([x_1,x_2],[t^{-1}(a),t^{-1}(b)]) = \infty$, thus $pr_1 \circ \mathcal{I}([x_1,x_2],[t^{-1}(a),t^{-1}(b)]) = 1$ and $pr_2 \circ \mathcal{I}([x_1,x_2],[t^{-1}(a),t^{-1}(b)]) = 0$, so

$$\mathcal{I}([x_1,x_2],[y_1,y_2]) = [1,0], \qquad [y_1,y_2] \in L^I,$$

for a fixed $[x_1,x_2] \in L^I$. This solution is not good, since $[1,0] \notin \mathcal{L}^I$. Now, let us assume that $f_{[x_1,x_2]}(a,b) = \infty$ and $f^{[x_1,x_2]}(a,b) = 0$. Then

$$\mathcal{I}([x_1,x_2],[y_1,y_2]) = [0,1], \qquad [y_1,y_2] \in L^I,$$

for a fixed $[x_1,x_2] \in L^I$. This solution is correct, since $[0,1] \in \mathcal{L}^I$.

We need to notice that not all obtained vertical solutions in $\mathcal{L}^I$ can be used for obtaining fuzzy implication on $\mathcal{L}^I$ in the sense of Definition 2.5. One can easily see that the above positive example is not a fuzzy implication. Also, for example, the following vertical sections are not possible in this case: $[0,0]$, $[0,1]$.

Acknowledgment

This work has been supported by the Polish Ministry of Science and Higher Education Grant Nr N N519 384936.

References

1. Baczyński, M.: On a class of distributive fuzzy implications. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 9, 229–238 (2001)
2. Baczyński, M.: On the distributivity of implication operations over t-representable t-norms generated from strict t-norms in Atanassov's intuitionistic fuzzy sets theory. In: Magdalena, L., Ojeda-Aciego, M., Verdegay, J.L. (eds.) Proc. 12th Int. Conf. IPMU 2008, Torremolinos (Málaga), Spain, pp. 1612–1619 (2001)
3. Baczyński, M., Jayaram, B.: On the distributivity of fuzzy implications over nilpotent or strict triangular conorms. IEEE Trans. Fuzzy Systems 17, 590–603 (2009)
4. Balasubramaniam, J., Rao, C.J.M.: On the distributivity of implication operators over T- and S-norms. IEEE Trans. Fuzzy Systems 12, 194–198 (2004)
5. Bustince, H., Pagola, M., Barrenechea, E.: Construction of fuzzy indices from fuzzy DI-subsethood measures: Application to the global comparison of images. Inform. Sci. 177, 906–929 (2007)
6. Combs, W.E., Andrews, J.E.: Combinatorial rule explosion eliminated by a fuzzy rule configuration. IEEE Trans. Fuzzy Systems 6, 1–11 (1998)
7. De Baets, B., Kerre, E., Gupta, M.: The fundamentals of fuzzy mathematical morphology, Part 1: Basic Concepts. Int. J. Gen. Syst. 23, 155–171 (1994)
8. Deschrijver, G., Kerre, E.E.: On the relationship between some extensions of fuzzy set theory. Fuzzy Sets and Systems 133, 227–235 (2003)
9. Deschrijver, G., Cornelis, C., Kerre, E.E.: On the representation of intuitionistic fuzzy t-norms and t-conorms. IEEE Trans. Fuzzy Systems 12, 45–61 (2004)
10. Deschrijver, G., Cornelis, C., Kerre, E.E.: Implication in intuitionistic and interval-valued fuzzy set theory: construction, classification and application. Internat. J. Approx. Reason. 35, 55–95 (2004)
11. Dubois, D., Gottwald, S., Hajek, P., Kacprzyk, J., Prade, H.: Terminological difficulties in fuzzy set theory - The case of "Intuitionistic Fuzzy Sets". Fuzzy Sets and Systems 156, 485–491 (2005)
12. Dubois, D., Prade, H.: An introduction to bipolar representations of information and preference. Int. J. Intell. Syst. 23, 866–877 (2008)
13. Fodor, J., Roubens, M.: Fuzzy preference modelling and multicriteria decision support. Kluwer, Dordrecht (1994)
14. Klement, E.P., Mesiar, R., Pap, E.: Triangular norms. Kluwer, Dordrecht (2000)
15. Kocsis, I.: A bisymmetry equation on restricted domain. Aequationes Math. 73, 280–284 (2007)
16. Mas, M., Monserrat, M., Torrens, J., Trillas, E.: A survey on fuzzy implication functions. IEEE Trans. Fuzzy Systems 15, 1107–1121 (2007)
17. Ruiz-Aguilera, D., Torrens, J.: Distributivity of strong implications over conjunctive and disjunctive uninorms. Kybernetika 42, 319–336 (2005)
18. Ruiz-Aguilera, D., Torrens, J.: Distributivity of residual implications over conjunctive and disjunctive uninorms. Fuzzy Sets and Systems 158, 23–37 (2007)
19. Trillas, E., Alsina, C.: On the law $[p \wedge q \to r] = [(p \to r) \vee (q \to r)]$ in fuzzy logic. IEEE Trans. Fuzzy Systems 10, 84–88 (2002)

Properties of Interval-Valued Fuzzy Relations, Atanassov's Operators and Decomposable Operations

Barbara Pękala

Institute of Mathematics, University of Rzeszów
Rejtana 16a, 35-310 Rzeszów, Poland
bpekala@univ.rzeszow.pl

Abstract. In this paper we study properties of interval-valued fuzzy relations which were introduced by L.A. Zadeh in 1975. Fuzzy set theory turned out to be a useful tool to describe situations in which the data are imprecise or vague. Interval-valued fuzzy set theory is a generalization of fuzzy set theory which was introduced also by Zadeh in 1965. We examine some properties of interval-valued fuzzy relations in the context of Atanassov's operators and decomposable operations in interval-valued fuzzy set theory.

Keywords: Fuzzy relations, interval-valued fuzzy relations, Atanassov's operators, decomposable operations.

1 Introduction

The idea of a fuzzy relation was defined in [25]. An extension of fuzzy set theory is interval-valued fuzzy set theory. Any interval-valued fuzzy set is defined by an interval-valued membership function: a mapping from the given universe to the set of all closed subintervals of [0,1]. In this work we study preservation of properties of interval-valued fuzzy relations by the powers of this relations, by lattice operations and Atanassov's operators. Consideration of diverse properties of the composition is interesting not only from a theoretical point of view but also for applications, since the composition of interval-valued fuzzy relations is proved to be useful in several fields, see for example [17] (performance evaluation), [22] (genetic algorithm), [16] (approximate reasoning) or [1], [14]. In [8] is analyzed some concept of interval-valued fuzzy set theory and their application to edge detection in grayscale images. Where authors chose an Atanassov's operator and analyzed the number of the elements belonging to the edge. Moreover, interval-valued fuzzy relations are applied in classification and in decision making. Furthermore, it is interesting to use Atanassov's operators in intuitionistic fuzzy systems, i.e. we can change the values of the intuitionistic fuzzy systems by the Atanassov's operators [3]. Interval-valued fuzzy relations (sets) are equivalent to some other extensions of fuzzy relations (sets) (see [11]). Among others, interval-valued fuzzy relations are isomorphic to Atanassov's intuitionistic fuzzy relations. This fact was noticed by several authors [2], [9], [11].

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 647–655, 2010.

An Atanassov's intuitionistic fuzzy relation is a pair of fuzzy relations, namely a membership and a nonmembership function, which represent positive and negative aspects of the given information. These objects introduced by Atanassov and originally called intuitionistic fuzzy relations were recently suggested to be called Atanassov's intuitionistic fuzzy relations or just bipolar fuzzy relations [13]. In this work we recall some concepts and results useful in our further considerations. Next, we study properties of decomposable operations and their relationship to t-norms in interval-valued fuzzy set theory. Finally, we consider some properties of interval-valued fuzzy relations and we study connections of its properties with lattice operations and some Atanassov's operators, so we consider preservation of some properties of interval-valued fuzzy relations by lattice operations and some Atanassov's operators.

2 Basic Definitions

First we recall the notion of the lattice operations and the order in the family of interval-valued fuzzy relations. Let X, Y, Z be non-empty sets and $Int([0,1]) = \{[x_1, x_2] : x_1, x_2 \in [0,1],\ x_1 \leq x_2\}$.

Definition 1 (cf. [23], [24]). *An interval-valued fuzzy relation R between universes X, Y is a mapping $R : X \times Y \to Int([0,1])$ such that*

$$R(x,y) = [\underline{R}(x,y), \overline{R}(x,y)] \in Int([0,1]), \tag{1}$$

for all pairs $(x,y) \in (X \times Y)$.
The class of all interval-valued fuzzy relations between universes X, Y will be denoted by $IVFR(X \times Y)$ or $IVFR(X)$ for $X = Y$.

Interval-valued fuzzy relations reflect the idea that membership grades are often not precise and the intervals represent such uncertainty.

The boundary elements in $IVFR(X \times Y)$ are $\mathbf{1} = [1,1]$ and $\mathbf{0} = [0,0]$.

Let $S, R \in IVFR(X \times Y)$. Then for every $(x,y) \in (X \times Y)$ we can define

$$S(x,y) \leq R(x,y) \Leftrightarrow \underline{S}(x,y) \leq \underline{R}(x,y), \overline{S}(x,y) \leq \overline{R}(x,y), \tag{2}$$

$$(S \vee R)(x,y) = [\max(\underline{S}(x,y), \underline{R}(x,y)), \max(\overline{S}(x,y), \overline{R}(x,y))], \tag{3}$$

$$(S \wedge R)(x,y) = [\min(\underline{S}(x,y), \underline{R}(x,y)), \min(\overline{S}(x,y), \overline{R}(x,y))], \tag{4}$$

where operations $\vee$ and $\wedge$ are the supremum and the infimum in $IVFR(X \times Y)$, respectively. Similarly for arbitrary set $T \neq \emptyset$

$$(\bigvee_{t \in T} R_t)(x,y) = [\bigvee_{t \in T} \underline{R_t}(x,y), \bigvee_{t \in T} \overline{R_t}(x,y)], \tag{5}$$

$$(\bigwedge_{t \in T} R_t)(x,y) = [\bigwedge_{t \in T} \underline{R_t}(x,y), \bigwedge_{t \in T} \overline{R_t}(x,y)]. \tag{6}$$

The pair $(IVFR(X \times Y), \leq)$ is a partially ordered set. As a result, the family $(IVFR(X \times Y), \vee, \wedge)$ is a lattice (for the notion of a lattice and other related concepts see [4]) which is a consequence of the fact that $([0,1], \max, \min)$ is a lattice. The lattice $IVFR(X \times Y)$ is complete. This fact follows from the notion of the supremum $\bigvee$ and the infimum $\bigwedge$ and from the fact that the values of fuzzy relations are from the interval $[0,1]$ which, with the operations maximum and minimum, forms a complete lattice. As a result $(IVFR(X \times Y), \vee, \wedge)$ is a complete, infinitely distributive lattice.

3 Properties of Decomposable Operations

In this paper we also study some properties of decomposable operations on an interval-valued fuzzy set. A characterization of such properties of operations is interesting not only from a theoretical point of view, but also for their applications, since they have proved to be useful in several fields. In [12] properties of decomposable operation were considered with a definition:

Definition 2. *An operation* $\mathcal{D} : (Int([0,1]))^2 \to Int([0,1])$ *is called decomposable if there exist operations* $D_1, D_2 : [0,1]^2 \to [0,1]$ *such that for all* $x, y \in Int([0,1])$

$$\mathcal{D}(x,y) = [D_1(x_1, y_1), D_2(x_2, y_2)],$$

where $x = [x_1, x_2]$, $y = [y_1, y_2]$.

Lemma 1 (cf. [12]). *Increasing operations* $D_1, D_2 : [0,1]^2 \to [0,1]$ *lead to the decomposable operation* $\mathcal{D}$ *if and only if* $D_1 \leq D_2$.

Special classes of operations in $Int([0,1])$ are interval-valued triangular norms (IV t-norms) and triangular conorms (IV t-conorms), which are useful in approximate reasoning, e.g. for medical diagnosis and information retrieval.

Definition 3 (cf. [10], [19]). *An IV t-norm* $\mathcal{T}$ *on* $Int([0,1])$ *is an increasing, commutative, associative operation* $\mathcal{T} : (Int([0,1]))^2 \to Int([0,1])$ *with a neutral element* $\mathbf{1}$.
An IV t-conorm $\mathcal{S}$ *on* $Int([0,1])$ *is an increasing, commutative, associative operation* $\mathcal{S} : (Int([0,1]))^2 \to Int([0,1])$ *with a neutral element* $\mathbf{0}$.

Theorem 1 (cf. [12]). *Let* $\mathcal{D} : (Int([0,1]))^2 \to Int([0,1])$ *be a decomposable binary operation such that* $\mathcal{D} = [D_1, D_2]$. *The decomposable operation* $\mathcal{D}$ *is a IV t-norm (IV t-conorm) if and only if* D_1 *and* D_2 *are t-norms (t-conorms) and* $D_1 \leq D_2$.

A decomposable operation $\mathcal{D}$ fulfilling conditions of Theorem 1 we call t-representable (s-representable).

We will give the notions of some Atanassov's operators for interval-valued fuzzy relations. These notions follow the ones introduced by Atanassov for intuitionistic fuzzy relations [3].

Definition 4. *Let $R \in IVFR(X \times Y)$, $\alpha,\ \beta \in [0,1]$, $\alpha + \beta \leq 1$. We define the operator $F_{\alpha,\beta} : Int([0,1]) \to Int[0,1]$ such that*

$$F_{\alpha,\beta}(R(x,y)) = [\underline{R}(x,y) + \alpha(\overline{R}(x,y) - \underline{R}(x,y)), \overline{R}(x,y) - \beta(\overline{R}(x,y) - \underline{R}(x,y))].$$

In particular, if $\beta = 1 - \alpha$, then we denote $F_{\alpha,1-\alpha}(R) = K_\alpha(R)$ and
$K_\alpha(R) = \underline{R}(x,y) + \alpha(\overline{R}(x,y) - \underline{R}(x,y))$ and
$F_{0,1}(R) = K_0(R) = \underline{R}$, $F_{1,0}(R) = K_1(R) = \overline{R}$.

Now, we consider more general definition of a decomposable operation:

Definition 5. *An operation $\mathcal{D} : (Int([0,1]))^2 \to Int([0,1])$ is called decomposable if there exist operations $D_1, D_2 : [0,1]^2 \to [0,1]$ such that for all $x, y \in Int([0,1])$*

$$\mathcal{D}(x,y) = [D_1(\underline{F}_{\alpha,\beta}(x), \underline{F}_{\alpha,\beta}(y)), D_2(\overline{F}_{\alpha,\beta}(x), \overline{F}_{\alpha,\beta}(y))], \tag{7}$$

where $x = [x_1, x_2]$, $y = [y_1, y_2]$, $\alpha, \beta \in [0,1]$, $\alpha + \beta \leq 1$ and
$\underline{F}_{\alpha,\beta}(x) = \underline{x} + \alpha(\overline{x} - \underline{x})$, $\overline{F}_{\alpha,\beta}(x) = \overline{x} - \beta(\overline{x} - \underline{x})$.

In a similar way to [7] and [12] we obtain

Theorem 2. *An operation $\mathcal{D} : (Int([0,1]))^2 \to Int([0,1])$ satisfying (7) is an t-representable (s-representable) if and only if operations $D_1, D_2 : [0,1]^2 \to [0,1]$, $D_1 \leq D_2$ are t-norms (t-conorms) and $F_{\alpha,\beta} = F_{0,0}$.*

4 Properties of Interval-Valued Fuzzy Relations

Let us recall the notion of the composition in $IVFR$.

Definition 6 (cf. [6], [15]). *Let $S \in IVFR(X \times Y)$, $R \in IVFR(Y \times Z)$. By the $\sup - \min$ composition of the relations S and R we call the relation $S \circ R \in IVFR(X \times Z)$,*

$$(S \circ R)(x,z) = [(\underline{S} \circ \underline{R})(x,z), (\overline{S} \circ \overline{R})(x,z)],$$

where

$$(\underline{S} \circ \underline{R})(x,z) = \bigvee_{y \in Y} (\underline{S}(x,y) \wedge \underline{R}(y,z)), \quad (\overline{S} \circ \overline{R})(x,z) = \bigvee_{y \in Y} (\overline{S}(x,y) \wedge \overline{R}(y,z))$$

and $(\underline{S} \circ \underline{R})(x,z) \leq (\overline{S} \circ \overline{R})(x,z)$.

$(IVFR(X), \circ)$ is a semigroup (see [20]), then we can consider the powers of its elements i.e. relations R^n for $R \in IVFR(X)$, $n \in \mathbb{N}$. Analogously to [18] we define

Definition 7. *By a power of relation $R \in IVFR(X)$ we mean*

$$R^1 = R,\ R^{n+1} = R^n \circ R,\ n \in \mathbb{N}.$$

The sequence (R^n) is convergent, if

$$\exists_{k \in \mathbb{N}} R^{k+1} = R^k.$$

In this section we continue the study from [21] and we examine transitivity property. Similarly to definitions of properties of fuzzy relations considered by Kaufmann we have

Definition 8. *Let $R \in IVFR(X)$. A relation R is called transitive if $R^2 \leq R$.*

The transitivity guarantees the convergence of (R^n). Hence, the transitivity property is one of the most interesting properties of crisp, fuzzy and bipolar fuzzy relations. The Atanassov's operator $F_{\alpha,\beta}$ may be applied, as indicated in the introduction, to edge detection (see [8]). We examine preservation of transitivity by operator $F_{\alpha,\beta}$ and lattice operations. In [5] the authors introduced the concept of a "partially included relation" for Atanassov's intuitionistic fuzzy relations. The justification for consideration of partially included relations is connected with the fact that Atanassov's operators $K_\alpha(R)$, where $\alpha \in [0,1]$, do not generally fulfil transitivity property of an Atanassov's intuitionistic fuzzy relation. However, for Atanassov's intuitionistic fuzzy relations which are partially included, the transitivity of an Atanassov's intuitionistic fuzzy relation ρ is equivalent to the transitivity of an operator $K_\alpha(\rho)$ for each $\alpha \in [0,1]$. We consider a more general form of the concept of a partially included Atanassov's intuitionistic fuzzy relation. Namely

Definition 9 (cf. [5]). *A relation $R \in IVFR(X)$ is called partially included if*

$$sgn(\underline{R}(x,z) - \underline{R}(z,y)) = sgn(\overline{R}(x,z) - \overline{R}(z,y)), \quad x,y,z \in X. \tag{8}$$

Proposition 1. *Let $R \in IVFR(X)$, $\alpha, \beta \in [0,1]$, $\alpha + \beta \leq 1$. If R is partially included and transitive, then $F_{\alpha,\beta}(R)$ is transitive.*

Proof. Let $R^2 \leq R$ and R be partially included, $x, y \in X$. From (8) we obtain
$((1-\alpha)\underline{R}(x,z) + \alpha\overline{R}(x,z)) \wedge ((1-\alpha)\underline{R}(z,y) + \alpha\overline{R}(z,y)) =$
$(1-\alpha)(\underline{R}(x,z) \wedge \underline{R}(z,y)) + \alpha(\overline{R}(x,z) \wedge \overline{R}(z,y))$.
Then $F^2_{\alpha,\beta}(R)(x,y) =$
$[(\underline{R}(x,y) + \alpha(\overline{R}(x,y) - \underline{R}(x,y)))^2, (\overline{R}(x,y) - \beta(\overline{R}(x,y) - \underline{R}(x,y)))^2] =$
$[((1-\alpha)\underline{R}(x,y) + \alpha\overline{R}(x,y))^2, ((1-\beta)\overline{R}(x,y) + \beta\underline{R}(x,y))^2] =$
$[\bigvee_{z\in X}((1-\alpha)\underline{R}(x,z) + \alpha\overline{R}(x,z)) \wedge ((1-\alpha)\underline{R}(z,y) + \alpha\overline{R}(z,y)),$
$\bigvee_{z\in X}((1-\beta)\overline{R}(x,z) + \beta\underline{R}(x,z)) \wedge ((1-\beta)\overline{R}(z,y) + \beta\underline{R}(z,y))]$.
From the above considerations we have
$[\bigvee_{z\in X}((1-\alpha)(\underline{R}(x,z) \wedge \underline{R}(z,y)) + \alpha(\overline{R}(x,z) \wedge \overline{R}(z,y))),$
$\bigvee_{z\in X}((1-\beta)(\overline{R}(x,z) \wedge \overline{R}(z,y)) + \beta(\underline{R}(x,z) \wedge \underline{R}(z,y)))] \leq$
$[\bigvee_{z\in X}(1-\alpha)(\underline{R}(x,z) \wedge \underline{R}(z,y)) + \bigvee_{z\in X}\alpha(\overline{R}(x,z) \wedge \overline{R}(z,y)),$
$\bigvee_{z\in X}(1-\beta)(\overline{R}(x,z) \wedge \overline{R}(z,y)) + \bigvee_{z\in X}\beta(\underline{R}(x,z) \wedge \underline{R}(z,y))] =$
$F_{\alpha,\beta}(R^2)(x,y)$, so by the isotonicity of $F_{\alpha,\beta}$ we obtain
$F^2_{\alpha,\beta}(R)(x,y) \leq F_{\alpha,\beta}(R)(x,y)$.

In [21] we considered preservation of reflexivity ($R(x,x) = \mathbf{1}$, $x \in X$) and irreflexivity of R ($R(x,x) = \mathbf{0}$, $x \in X$) by operator $F_{\alpha,\beta}$. Now, we consider more general properties i.e. local reflexivity and local irreflexivity. On the finite set X the local reflexivity is called dominating principle and similarly to transitivity guarantees convergence of the sequence (R^n), $R \in IVFR(X)$.

Definition 10. *An interval-valued fuzzy relation $R \in IVFR(X)$ is:*
locally reflexive, if

$$\forall_{x\in X}(R(x,x) = \bigvee_{y\in X} R(x,y) \ \text{ and } \ R(x,x) = \bigvee_{y\in X} R(y,x)), \tag{9}$$

locally irreflexive, if

$$\forall_{x\in X}(R(x,x) = \bigwedge_{y\in X} R(x,y) \ \text{ and } \ R(x,x) = \bigwedge_{y\in X} R(y,x)). \tag{10}$$

We study connections of the above properties with some Atanassov operator:

Lemma 2. *Let $\alpha, \beta \in [0,1],\ \alpha+\beta \le 1$.*
If $R \in IVFR(X)$ is locally reflexive, then for all $x \in X$

$$F_{\alpha,\beta}(\bigvee_{y\in X} R(x,y)) = \bigvee_{y\in X} F_{\alpha,\beta}(R(x,y)),\ F_{\alpha,\beta}(\bigvee_{y\in X} R(y,x)) = \bigvee_{y\in X} F_{\alpha,\beta}(R(y,x)). \tag{11}$$

If $R \in IVFR(X)$ is locally irreflexive, then for all $x \in X$

$$F_{\alpha,\beta}(\bigwedge_{y\in X} R(x,y)) = \bigwedge_{y\in X} F_{\alpha,\beta}(R(x,y)),\ F_{\alpha,\beta}(\bigwedge_{y\in X} R(y,x)) = \bigwedge_{y\in X} F_{\alpha,\beta}(R(y,x)). \tag{12}$$

Proof. Let $R \in IVFR(X)$ be locally reflexive, $x \in X$. For lattice operations we have the following inequalities:
$F_{\alpha,\beta}R(x,x) \le \bigvee_{y\in X} F_{\alpha,\beta}R(x,y) =$
$[\bigvee_{y\in X}((1-\alpha)\underline{R}(x,y) + \alpha\overline{R}(x,y)), \bigvee_{y\in X}((1-\beta)\overline{R}(x,y) + \beta\underline{R}(x,y))] \le$
$[\bigvee_{y\in X}(1-\alpha)\underline{R}(x,y) + \bigvee_{y\in X}\alpha\overline{R}(x,y), \bigvee_{y\in X}(1-\beta)\overline{R}(x,y) + \bigvee_{y\in X}\beta\underline{R}(x,y)] =$
$F_{\alpha,\beta}(\bigvee_{y\in X} R(x,y)) = F_{\alpha,\beta}R(x,x)$.
Thus $F_{\alpha,\beta}(\bigvee_{y\in X} R(x,y)) = \bigvee_{y\in X} F_{\alpha,\beta}(R(x,y))$.
Analogously, we may prove the formula
$F_{\alpha,\beta}(\bigvee_{y\in X} R(y,x)) = \bigvee_{y\in X} F_{\alpha,\beta}(R(y,x))$
and the condition (12).

Theorem 3. *Let $\alpha, \beta \in [0,1],\ \alpha+\beta \le 1$. If an interval-valued fuzzy relation $R \in IVFR(X)$ is locally reflexive (locally irreflexive), then $F_{\alpha,\beta}(R(x,y))$ is also locally reflexive (locally irreflexive).*

Proof. Let $x \in X$ and R be locally reflexive. We have
$\underline{R}(x,x)(1-\alpha) = \bigvee_{y\in X}\underline{R}(x,y)(1-\alpha), \quad \alpha\overline{R}(x,x) = \bigvee_{y\in X}\alpha\overline{R}(x,y),$
$\overline{R}(x,x)(1-\beta) = \bigvee_{y\in X}\overline{R}(x,y)(1-\beta), \quad \beta\underline{R}(x,x) = \bigvee_{y\in X}\beta\underline{R}(x,y).$
Hence,
$F_{\alpha,\beta}(R(x,x)) = [\underline{R}(x,x)(1-\alpha) + \alpha\overline{R}(x,x), \overline{R}(x,x)(1-\beta) + \beta\underline{R}(x,x)] =$
$[\bigvee_{y\in X}\underline{R}(x,y)(1-\alpha) + \bigvee_{y\in X}\alpha\overline{R}(x,y), \bigvee_{y\in X}\overline{R}(x,y)(1-\beta) + \bigvee_{y\in X}\beta\underline{R}(x,y)] =$
$F_{\alpha,\beta}(\bigvee_{y\in X} R(x,y))$
and by Lemma 2 we obtain

$F_{\alpha,\beta}(R(x,x)) = \bigvee_{y\in X} F_{\alpha,\beta}(R(x,y))$.
Similarly we prove the formula
$F_{\alpha,\beta}(R(x,x)) = \bigvee_{y\in X} F_{\alpha,\beta}(R(y,x))$
and the preservation of local irreflexivity property.

Moreover, we observe preservations of local reflexivity and local irreflexivity by lattice operations in $IVFR(X)$.

Theorem 4. *Let $P, R \in IVFR(X)$.*
If P, R are locally reflexive, then $P \vee R$ and $P \wedge R$ are locally reflexive.
If P, R are locally irreflexive, then $P \vee R$ and $P \wedge R$ are locally irreflexive.

Proof. Let $x \in X$. First we consider locally reflexive relations P, R with respect to second variable. Then for $Q = P \wedge R$ we have

$$Q(x,x) = P(x,x) \wedge R(x,x) = \bigvee_{y\in X} P(x,y) \wedge \bigvee_{z\in X} R(x,z).$$

So, for infinitely distributive lattice $(IVFR(X \times Y), \vee, \wedge)$ we observe
$Q(x,x) = \bigvee_{y\in X} \bigvee_{z\in X} (P(x,y) \wedge R(x,z)) \geq \bigvee_{y\in X} (P(x,y) \wedge R(x,y)) \geq$
$P(x,x) \wedge R(x,x) = Q(x,x)$.
As a consequence we have

$$Q(x,x) = \bigvee_{y\in X} (P(x,y) \wedge R(x,y)) = \bigvee_{y\in X} Q(x,y),$$

which proves local reflexivity of $P \wedge R$ with respect to the second variable. We may prove the other conditions in a similar way.

The above theorem is not true for weak local reflexivity and weak local irreflexivity, where the relation R is defined as weakly locally reflexive, if

$$\forall_{x\in X}(R(x,x) = \bigvee_{y\in X} R(x,y) \ \ or \ \ R(x,x) = \bigvee_{y\in X} R(y,x))$$

and the relation R is defined as weakly locally irreflexive, if

$$\forall_{x\in X}(R(x,x) = \bigwedge_{y\in X} R(x,y) \ \ or \ \ R(x,x) = \bigwedge_{y\in X} R(y,x)).$$

Example 1. Let $n = 3$. Relations

$$P = \begin{bmatrix} 0.6 & 0.6 & 0.6 \\ 0.4 & 0.4 & 0.4 \\ 0.7 & 0.7 & 0.7 \end{bmatrix}, \quad R = \begin{bmatrix} 0.5 & 0.3 & 0.7 \\ 0.5 & 0.3 & 0.7 \\ 0.5 & 0.3 & 0.7 \end{bmatrix}$$

are weakly locally reflexive, but the relation

$$P \vee R = \begin{bmatrix} 0.6 & 0.6 & 0.7 \\ 0.5 & 0.4 & 0.7 \\ 0.7 & 0.7 & 0.7 \end{bmatrix}$$

is not weakly locally reflexive. Relations P, R are weakly locally irreflexive, but

$$P \wedge R = \begin{bmatrix} 0.5 & 0.3 & 0.6 \\ 0.4 & 0.3 & 0.4 \\ 0.5 & 0.3 & 0.7 \end{bmatrix}$$

is not weakly locally irreflexive.

Moreover, since every increasing t-representable operation is distributive over t-conorm $S = [\max, \max]$ and t-norm $T = [\min, \min]$ (see [12]) we can obtain a more general theorem, i.e. an arbitrary t-representable t-norm (t-conorm) preserves property of local reflexivity and local irreflexivity.

5 Conclusion

In this work we consider preservation of some properties of interval-valued fuzzy relations by lattice operations and Atanassov's operators $F_{\alpha,\beta}$. These properties may be applied to solving the problem of a convergence of sequence of powers of an interval-valued fuzzy relation problem. We can also consider other Atanassov's operators and other properties interesting because of its applications.

Acknowledgments. This paper is partially supported by the Ministry of Science and Higher Education Grant Nr N N519 384936.

References

1. Agustench, E., Bustince, H., Mohedano, V.: Method of Least Squares Applied to the GMP with International-Valued Fuzzy Sets. Mathware Soft Comp. 6(2-3), 267–276 (1999)
2. Atanassov, K.T., Gargov, G.: Interval Valued Intuitionistic Fuzzy Sets. Fuzzy Sets Syst. 31, 343–349 (1989)
3. Atanassov, K.: Intuitionistic Fuzzy Sets: Theory and Applications. Springer, Heidelberg (1999)
4. Birkhoff, G.: Lattice Theory. AMS Coll. Publ. 25, Providence (1967)
5. Burillo, P., Bustince, H.: Intuitionistic Fuzzy Relations. Effect of Atanassov's Operators on the Properties of the Intuitionistic Fuzzy Relations. Math. Soft Comp. 2, 117–148 (1995)
6. Bustince, H., Burillo, P.: Mathematical Analysis of Interval-Valued Fuzzy Relations: Application to Approximate Reasoning. Fuzzy Sets Syst. 113, 205–219 (2000)
7. Bustince, H., Barrenechea, E., Pagola, M.: Generation of Interval-Valued Fuzzy and Atannasov's Intuitionistic Fuzzy Connectives from Fuzzy Connectives and from K_α Operators: Laws for Conjunction and Disjunction. Amplitude. Inter. J. Intel. Syst. 23, 680–714 (2008)
8. Bustince, H., Barrenechea, E., Pagola, M., Fernandez, J.: Interval-Valued Fuzzy Sets Constructed from Matrices: Application to Edge Detection. Fuzzy Sets Syst. 160, 1819–1840 (2009)
9. Deschrijver, G., Kerre, E.E.: On the Relationship between some Extensions of Fuzzy Set Theory. Fuzzy Sets Syst. 133(2), 227–235 (2003)

10. Deschrijver, G., Cornelis, C., Kerre, E.E.: On the Representation of Intuitonistic Fuzzy t-Norms and t-Conorms. IEEE Trans. Fuzzy Syst. 12, 45–61 (2004)
11. Deschrijver, G., Kerre, E.E.: On the Position of Intuitionistic Fuzzy Set Theory in the Framework of Theories Modelling Imprecision. Inform. Sci. 177, 1860–1866 (2007)
12. Drygaś, P., Pękala, B.: Properties of Decomposable Operations on some Extension of the Fuzzy Set Theory. In: Atanassov, K.T., Hryniewicz, O., Kacprzyk, J., et al. (eds.) Advances in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics, pp. 105–118. EXIT, Warsaw (2008)
13. Dubois, D., Gottwald, S., Hajek, P., Kacprzyk, J., Prade, H.: Terminological Difficulties in Fuzzy Set Theory - The Case of "Intuitionistic Fuzzy Sets". Fuzzy Sets Syst. 156, 485–491 (2005)
14. Dubois, D., Prade, H., Sessa, S.: Recent Literature. Fuzzy Sets Syst. 159, 857–865 (2008)
15. Goguen, A.: L-Fuzzy Sets. J. Math. Anal. Appl. 18, 145–174 (1967)
16. Gorzałczany, M.B.: A Method of Inference in Approximate Reasoning Based on Interval-Valued Fuzzy Sets. Fuzzy Sets Syst. 21(1), 1–17 (1987)
17. Guh, Y.-Y., Yang, M.-S., Po, R.-W., Lee, E.S.: Interval-Valued Fuzzy Relation-Based Clustering with its Application to Performance Evaluetion. Comp. Math. Appl. 57, 841–849 (2009)
18. Kaufmann, A.: Introduction to the Theory of Fuzzy Subsets. Acad. Press, New York (1975)
19. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms. Kluwer Acad. Publ., Dordrecht (2000)
20. Pękala, B.: Operations on Interval Matrices. In: Kryszkiewicz, M., Peters, J.F., Rybiński, H., Skowron, A. (eds.) RSEISP 2007. LNCS (LNAI), vol. 4585, pp. 613–620. Springer, Heidelberg (2007)
21. Pękala, B.: Preservation of Properties of Interval-Valued Fuzzy Relations. In: Carvalho, J.P., et al. (eds.) IFSA/EUSFLAT 2009, pp. 1206–1210 (2009)
22. Sánchez, L., Couso, I.: Advocating the Use of Imprecisely Observed Data in Genetic Fuzzy Systems. IEEE Trans. Fuzzy Syst. 15(4), 551–562 (2007)
23. Sambuc, R.: Fonctions ϕ-floues: Application á l'aide au diagnostic en pathologie thyroidienne. Ph.D. Thesis, Université de Marseille, France (1975) (in French)
24. Zadeh, L.A.: The Concept of a Linguistic Variable and its Application to Approximate Reasoning-I. Inform. Sci. 8, 199–249 (1975)
25. Zadeh, L.A.: Fuzzy Sets. Inform. Control 8, 338–353 (1965)

Cardinality and Entropy for Bifuzzy Sets

Vasile Pătraşcu

Department of Informatics Technology, Tarom Company
Bucharest, Romania
patrascu.v@gmail.com

Abstract. In this paper, one presents entropy and cardinality measure for bifuzzy sets. All these are constructed in the framework of a penta-valued representation. This representation uses the following five indexes: index of truth, index of falsity, index of incompleteness, index of inconsistency and index of indeterminacy. Also a new metric distance for bounded real interval is defined.

Keywords: cardinality, entropy, penta-valued representation, bifuzzy sets,intuitionistic fuzzy sets, fuzzy sets.

1 Introduction

Similarity measures play an important role in different research topics such as image analysis, pattern recognition, decision making and market prediction. In the same way, distance measure is an important tool which describes differences between two objects and considered as a dual concept of similarity measure [5]. The choice of a distance measure or a similarity measure for any fields of research is not trivial [13], [14], [17]. In this paper, one presents a new measure of distance for the interval $[a,b]$. Then, using the similarities or distances, entropy and cardinality measures are constructed for bifuzzy sets. All these measures are done for bifuzzy values or bifuzzy sets in the framework of penta-valued representation. The proposed measure of entropy reaches its maximum for elements for which membership function is equal to non-membership function. The structure of entropy offers the advantage we immediately are able to indicate possible reasons of the missing information: the incompleteness, the inconsistency or the lack of distinguish between membership/non-membership and their negations.

The paper has the following structure: section 2 presents the bifuzzy sets and their particular forms: intuitionistic fuzzy sets, paraconsistent fuzzy sets and fuzzy sets. Also, the main operators for bifuzzy sets are presented. Section 3 presents a representation of bifuzzy sets considering the *net truth* and *definedness*. These parameters can have positive or negative values. Supplementary, it is defined the *index of indeterminacy*. Next, it is constructed a penta-valued representation. Section 4 presents a new metric distance for the interval $[a,b]$ and its particular form for $[0,1]$. Section 5 presents measures for entropy and cardinality. Finally, the conclusions are presented in section 6.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 656–665, 2010.

2 The Bifuzzy Sets and Their Particular Forms

Let X be a crisp set (the space of points), with a generic element of X denoted by x. In the framework of Zadeh theory [17], a *fuzzy set A* in X is characterized by the membership function $\mu : X \to [0,1]$. The non-membership function $\nu : X \to [0,1]$ is obtained by negation and thus both functions define a partition of unity, namely:

$$\mu + \nu = 1 \ . \tag{2.1}$$

Atanassov has extended the fuzzy sets to the *intuitionistic fuzzy sets* [1]. Atanassov has relaxed the condition (2.1) to the following inequality:

$$\mu + \nu \le 1 \ .$$

He has used the third function, the Atanassov intuitionistic fuzzy index or the index of undefinedness π. It is defined by:

$$\pi = 1 - \mu - \nu \ .$$

In the same way, we can consider instead of (2.1) the following condition:

$$\mu + \nu \ge 1 \ .$$

Thus, we obtain the *paraconsistent fuzzy sets* [15], [12] and one can define the index of contradiction or index of inconsistency [12]:

$$\kappa = \mu + \nu - 1 \ .$$

More generally, we can consider a set A, defined by two functions totally independent $\mu : X \to [0,1]$ and $\nu : X \to [0,1]$. In this paper, we will use for this type of sets the name *bifuzzy sets* (BFS). For this kind of sets, one defines the union, the intersection, the complement, the negation and the dual operators [6]:
The union $A \cup B$ for two sets $A, B \in BFS$ is defined by:

$$(\mu_{A\cup B}, \nu_{A\cup B}) = (\mu_A \vee \mu_B, \nu_A \wedge \nu_B) \ . \tag{2.2}$$

The intersection $A \cap B$ between two sets $A, B \in BFS$ is defined by:

$$(\mu_{A\cap B}, \nu_{A\cap B}) = (\mu_A \wedge \mu_B, \nu_A \vee \nu_B) \tag{2.3}$$

In formulae (2.2) and (2.3), the symbols "$\vee$" and "$\wedge$" represent any couple of t-conorm, t-norm.

The complement A^c for the set $A \in BFS$ is defined by the formula:

$$(\mu_{A^c}, \nu_{A^c}) = (\nu_A, \mu_A) \ .$$

The negation A^n for the set $A \in BFS$ is defined by the formula:

$$(\mu_{A^n}, \nu_{A^n}) = (1 - \mu_A, 1 - \nu_A) \ .$$

The dual A^d for the set $A \in BFS$ is defined by the formula:

$$(\mu_{A^d}, \nu_{A^d}) = (1 - \nu_A, 1 - \mu_A) \ .$$

3 From Bifuzzy Sets Representation with Positive and Negative Values to a Penta-Valued One

Considering a bifuzzy value $(\mu,\nu)\in[0,1]^2$, we can define the following two parameters:

net truth: $\tau=\mu-\nu$. (3.1)

definedness $\delta=\mu+\nu-1$. (3.2)

The *net truth* τ and the *definedness* δ verify the following two equalities:

$$\begin{cases}\tau=\mu\circ\overline{\nu}-\overline{\mu}\circ\nu\\ \delta=\mu\circ\nu-\overline{\mu}\circ\overline{\nu}\end{cases}$$

where $\overline{x}=1-x$ and the symbol $\circ$ represents any Frank t-norm [7]. From (3.1) and (3.2) it results the following inequality:

$$|\tau|+|\delta|\leq 1\,. \tag{3.3}$$

and we can define the third parameter:

$$i=1-|\tau|-|\delta|\ . \tag{3.4}$$

Taking into account the Belnap logic [4] that uses four logical values: *true*, *false*, *contradictory* and *undefined*, we will consider the following four points: $T=(1,0)$, $F=(0,1)$, $C=(1,1)$, $U=(0,0)$. The value $|\tau|$ shows how close the point (μ,ν) to the set $\{T,F\}$ is and the value $|\delta|$ shows how close the point (μ,ν) to the set $\{U,F\}$ is. Also, the value i shows how close the point (μ,ν) to the center of square $(0.5,0.5)$ is. Because the point $(0.5,0.5)$ is equidistant to the four corners T,F,U,C, we can say that parameter i is *index of indeterminacy* and we will consider the fifth point $I=(0.5,0.5)$. Also, we define the following indexes:

index of truth: $t=\max(\tau,0)$. (3.5)

index of falsity: $f=\max(-\tau,0)$. (3.6)

index of inconsistency: $c=\max(\delta,0)$ (3.7)

index of incompleteness: $u=\max(-\delta,0)$. (3.8)

From (3.4)-(3.8) it results the following penta-valued fuzzy partition of unity:

$$t+f+u+c+i=1\ .$$

The penta-valued representation (t,f,u,c,i) is related to the five points $\{T,F,U,C,I\}$ and it is connected to the penta-valued logic proposed in [12]. This logic uses the following five logical values: *true*, *false*, *undefined* (incomplete or unknown), *contradictory* (inconsistent or overdefined) and *indeterminate* (equidistant or neutral).

If we consider the distance $d(x,y)=|x-y|$ and the similarity $s(x,y)=1-|x-y|$ then we obtain: $d(\mu,\nu)=t+f$.

$$s(\mu,\nu) = u + c + i \ . \tag{3.9}$$

$$|\tau| = d(\mu,\nu) \ . \tag{3.10}$$

$$|\delta| = d(\mu,1-\nu) \ . \tag{3.11}$$

Seeing the formulae (3.10), (3.11), we can generalize the formulae of *net truth* and *definedness*. We consider an arbitrary distance measure $D:[0,1]\times[0,1]\to[0,1]$ that verifies the inequality (3.3):

$$D(\mu,\nu) + D(\mu,1-\nu) \le 1 \ . \tag{3.12}$$

It results the following generalization of τ and δ:

$$\tau^* = \begin{cases} D(\mu,\nu) & \text{if } \mu > \nu \\ 0 & \text{if } \mu = \nu \\ -D(\mu,\nu) & \text{if } \mu < \nu \end{cases} .$$

$$\delta^* = \begin{cases} D(\mu,1-\nu) & \text{if } \mu+\nu > 1 \\ 0 & \text{if } \mu+\nu = 1 \\ -D(\mu,1-\nu) & \text{if } \mu+\nu < 1 \end{cases} .$$

From (3.4) it results: $i^* = 1 - D(\mu,\nu) - D(\mu,1-\nu)$.

Finally, for indexes t^*, f^*, u^*, c^* we will use formulae (3.5), (3.6), (3.7), (3.8) and one obtains:

generalized index of truth: $t^* = \max(\tau^*, 0)$.

generalized index of falsity: $f^* = \max(-\tau^*, 0)$.

generalized index of inconsistency: $c^* = \max(\delta^*, 0)$.

generalized index of incompleteness: $u^* = \max(-\delta^*, 0)$.

It results the following two equalities:

$$D(\mu,\nu) = t^* + f^* \ .$$

$$S(\mu,\nu) = u^* + c^* + i^* \ . \tag{3.13}$$

where $S(\mu,\nu)$ is the negation of $D(\mu,\nu)$ and it represents a similarity measure between the membership μ and the non-membership ν.

4 New Metric Distance on the Interval [a,b]

There are important distances defined on the interval $[0,1]$ and that is because the membership functions are defined on the $[0,1]$. The frequently formula used for the distance on the interval $[0,1]$ is the following:

$$d(x, y) = |x - y| \ . \tag{4.1}$$

The distance (4.1) takes into account only the difference $x - y$. We additionally propose a distance measure that includes also the sum $x + y$. This new distance is defined by:

$$D(x,y)=\frac{2\,|x-y|}{1+|x-y|+|x+y-1|}\ . \tag{4.2}$$

More generally, for the interval $[a,b]$, the distance has the following form:

$$D(x,y)=\frac{2\,|x-y|}{b-a+|x-y|+|x+y-a-b|}\ . \tag{4.3}$$

or

$$D(x,y)=\frac{2\,|x-y|}{b-a+|2x-a-b|\vee|2y-a-b|}\ . \tag{4.4}$$

where $\alpha\vee\beta=\max(\alpha,\beta)$.

The distance defined by (4.3) or (4.4) verifies the following three metric properties:

- $D(x,y)=0 \ \Leftrightarrow\ x=y$.
- $D(x,y)=D(y,x)$.
- $D(x,y)+D(y,z)\geq D(x,z)$.

The first two properties are evident. We will show the proof for the triangle inequality. We will analyze six possibilities. In the first four, we have $y\in\{\min(x,y,z),\max(x,y,z)\}$ and for the last two, we have $y=median(x,y,z)$.

p1) $x\geq z\geq y\ \Rightarrow\ D(x,y)\geq D(x,z)$ and it results $D(x,y)+D(y,z)\geq D(x,z)$.

p2) $y\geq z\geq x\ \Rightarrow\ D(x,y)\geq D(x,z)$ and it results $D(x,y)+D(y,z)\geq D(x,z)$.

p3) $y\geq x\geq z\ \Rightarrow\ D(y,z)\geq D(x,z)$ and it results $D(x,y)+D(y,z)\geq D(x,z)$.

p4) $z\geq x\geq y\ \Rightarrow\ D(z,y)\geq D(x,z)$ and it results $D(x,y)+D(y,z)\geq D(x,z)$.

p5) $x\geq y\geq z$

it results:
$$\begin{cases}|2x-a-b|\leq|2x-a-b|\vee|2z-a-b|\\ |2y-a-b|\leq|2x-a-b|\vee|2z-a-b|\\ |2z-a-b|\leq|2x-a-b|\vee|2z-a-b|\end{cases}.$$

and
$$\begin{cases}|2x-a-b|\vee|2y-a-b|\leq|2x-a-b|\vee|2z-a-b|\\ |2y-a-b|\vee|2z-a-b|\leq|2x-a-b|\vee|2z-a-b|\end{cases}.$$

it consequences:

$$\begin{cases}D(x,y)\geq\dfrac{2\,|x-y|}{b-a+|2x-a-b|\vee|2z-a-b|}\\ D(y,z)\geq\dfrac{2\,|y-z|}{b-a+|2x-a-b|\vee|2z-a-b|}\end{cases}$$

and summing up, one obtains:

$$D(x,y)+D(y,z)\geq\frac{2\big(|x-y|+|y-z|\big)}{b-a+|2x-a-b|\vee|2z-a-b|}\geq D(x,z)$$

p6) $z\geq y\geq x$ This case can be proven similarly to case *p5)*.

The similarity can be defined for $[a,b]$ using the negation of distance, namely:

$$S(x,y)=1-D(x,y)=\frac{b-a-|x-y|+|x+y-a-b|}{b-a+|x-y|+|x+y-a-b|}\ .$$

For the interval [0,1] the similarity has the following particular form:

$$S(x,y)=\frac{1-|x-y|+|x+y-1|}{1+|x-y|+|x+y-1|}\ . \tag{4.5}$$

5 The Entropy and Cardinality of Bifuzzy Sets

We define measures for entropy and cardinality in the framework of penta-valued representation.

5.1 The Entropy of Bifuzzy Sets

For a bifuzzy set, uncertainty results from the imprecise boundaries or it results from the lack of crisp distinction between the elements belonging and not belonging to a set. The uncertainty of fuzzy sets is called fuzziness. A measure of fuzziness is the entropy first mentioned by Zadeh [17]. Kaufman [9] proposed to measure a degree of fuzziness of any fuzzy set by a metric distance between its membership function and the membership function of its nearest crisp set. Another way given by Yager [16] was to view a degree of fuzziness in term of a lack of distinction between the fuzzy set and its complement. Kosko [10] proposed to measure the fuzzy entropy by the ratio between the distance to the nearest crisp element and the distance to the farthest crisp element. But this formula is nothing else then a similarity measure between the considered element and its complement. This is equivalent with the similarity measure between the membership μ and the non-membership ν. In order to give a proper definition of the entropy of bifuzzy sets we have to consider three aspects of uncertainty contained in these sets: the first one is connected with the missing information that results in a "gap" between the membership function and the non-membership function; the second is connected with the supplementary information that results in a "glut" between the membership function and the non-membership function; the third is the lack of distinction between a bifuzy value and its negation. All these are cumulated in the total uncertainty and this represents the bifuzzy entropy.

Let there be a penta-fuzzy value $x=(t,f,u,c,i)$ and its complement $x^c=(f,t,u,c,i)$ [12]. Because $t+f+u+c+i=1$, for these two particular penta-fuzzy values, one can use the Bhattacharyya similarity [3]:

$$S_B(x,x^c)=\sqrt{t\cdot f}+\sqrt{f\cdot t}+\sqrt{u\cdot u}+\sqrt{c\cdot c}+\sqrt{i\cdot i}\ .$$

Having the similarity, one can define the entropy. Therefore we get:

$$e=S_B(x,x^c)\ .$$

Because $t\cdot f=0$ one obtains the sum of these three components of uncertainty: u,c,i

$$e=u+c+i\ . \tag{5.1.1}$$

Incompleteness u, inconsistency c and indeterminacy i represent the sources of total uncertainty or in other words the sources of entropy. From (5.1.1), (3.9) and (3.13) it results that the proposed measures of entropy represents in same time the similarity measure between the membership μ and the non-membership ν. Next, we

describe some requirements for the bifuzzy entropy. In the space (μ,ν) the entropy function $e(\mu,\nu)$ verifies the following conditions:

e1) $e(1,0)=e(0,1)=0$

e2) $e(\mu,\nu)=1 \Leftrightarrow \mu=\nu$

e3) if $\mu_2>\mu_1\geq\nu_1>\nu_2$ then $e(\mu_1,\nu_1)>e(\mu_2,\nu_2)$

e4) if $\mu_2<\mu_1\leq\nu_1<\nu_2$ then $e(\mu_1,\nu_1)>e(\mu_2,\nu_2)$

e5) $e(\mu,\nu)=e(\nu,\mu)=e(1-\nu,1-\mu)=e(1-\mu,1-\nu)$

The properties (*e1, e2, e3, e4, e5*) represent an extension of properties considered by De Luca and Termini for entropy of fuzzy sets [11] and by Szmidt and Kacprzyk for intuitionistic fuzzy sets [14].

In the space (τ,δ) the entropy function $e(\tau,\delta)$ verifies the following conditions:

f1) $e(1,0)=e(-1,0)=0$

f2) $e(0,\delta)=1$

f3) if $|\tau_1|>|\tau_2|$ then $e(\tau_1,\delta)<e(\tau_2,\delta)$

f4) if $|\delta_1|>|\delta_2|$ then $e(\tau,\delta_1)\geq e(\tau,\delta_2)$

f5) $e(\tau,\delta)=e(-\tau,\delta)=e(\tau,-\delta)=e(-\tau,-\delta)$

Also, we can consider a vector representation for entropy of bifuzzy sets, namely:

$$\vec{e}=(u,c,i)\ .$$

For intuitionistic fuzzy sets it results a vector with two components:

$$\vec{e}_{ifs}=(u,i)\ .$$

For paraconsistent fuzzy sets it results a vector with two components:

$$\vec{e}_{pfs}=(c,i)\ .$$

For fuzzy sets it results a scalar:

$$e_{fs}=i\ .$$

Grzegorzewski and Mrówka proposed a vector representation of intuitionistic fuzzy entropy [8]. They considered the following vector:

$$\vec{e}_{GM}=\left(1-|\mu-\nu|,\pi\right)\ . \tag{5.1.2}$$

Using $d(x,y)=|x-y|$, one obtains:

$$e=1-|\mu-\nu|=1-|\tau|\ . \tag{5.1.3}$$

$$u=\max(1-\mu-\nu,0)\ .$$

$$c=\max(\mu+\nu-1,0)\ .$$

$$i=1-|\mu-\nu|-|\mu+\nu-1|\ .$$

For fuzzy sets one obtains:

$$e_{fs}=i_{fs}=1-|2\mu-1|\ . \tag{5.1.4}$$

The entropy (5.1.4) and in the same time the index of indeterminacy represents the Kaufmann index of fuzziness [9]. For intuitionistic fuzzy sets one obtains:

$$e_{ifs} = 1 - |\mu - \nu| \ .$$

$$\vec{e}_{ifs} = (\pi, 1 - \pi - |\mu - \nu|) \ . \tag{5.1.5}$$

Bustince and Burillo considered for intuitionistic fuzzy sets the following measure for entropy [5]: $e_{BB} = \pi$.

This value is a component of the vector approach (5.1.2) or (5.1.5).

Using the metric distance $D(x, y) = \dfrac{2|x - y|}{1 + |x - y| + |x + y - 1|}$, one obtains:

$$e^* = \frac{1 - |\mu - \nu| + |\mu + \nu - 1|}{1 + |\mu - \nu| + |\mu + \nu - 1|} = \frac{1 - |\tau| + |\delta|}{1 + |\tau| + |\delta|} . \tag{5.1.6}$$

$$u^* = \frac{2 \cdot \max(1 - \mu - \nu, 0)}{1 + |\mu - \nu| + |\mu + \nu - 1|} \ .$$

$$c^* = \frac{2 \cdot \max(\mu + \nu - 1, 0)}{1 + |\mu - \nu| + |\mu + \nu - 1|} \ .$$

$$i^* = \frac{1 - |\mu - \nu| - |\mu + \nu - 1|}{1 + |\mu - \nu| + |\mu + \nu - 1|} \ .$$

For fuzzy sets one obtains: $e^*_{fs} = i^*_{fs} = \dfrac{1 - |2\mu - 1|}{1 + |2\mu - 1|}$. (5.1.7)

The entropy and the index of indeterminacy (5.1.7) represent the Kosko entropy [10]. For intuitionistic fuzzy sets one obtains:

$$e^*_{ifs} = \frac{1 - |\mu - \nu| + \pi}{1 + |\mu - \nu| + \pi} \ . \tag{5.1.8}$$

$$u^*_{ifs} = \frac{2\pi}{1 + |\mu - \nu| + \pi} \ . \tag{5.1.9}$$

$$i^*_{ifs} = \frac{1 - |\mu - \nu| - \pi}{1 + |\mu - \nu| + \pi} \ .$$

$$\vec{e}^*_{ifs} = \left(\frac{2\pi}{1 + |\mu - \nu| + \pi}, \frac{1 - |\mu - \nu| - \pi}{1 + |\mu - \nu| + \pi} \right) .$$

Formula (5.1.8) represents the entropy defined by Szmidt and Kacprzyk for intuitionistic fuzzy sets [13]. Also, we can underline that (5.1.9) represents a generalization of Atanassov intuitionistic fuzzy index. This generalization differs from that suggested in [2].

The entropy functions e and e^* defined by (5.1.3) and (5.1.6) verify the conditions (*e1, e2, e3, e4, e5*) in the space (μ, ν) and (*f1, f2, f3, f4, f5*) in the space (τ, δ).

Finally, the bifuzzy entropy for a set A is obtained with the subsequent formula:

$$E(A) = \sum_{x \in X} e(x) \ .$$

5.2 The Cardinality of Bifuzzy Sets

We will define the cardinality for a bifuzzy set A by:

$$card(A) = \sum_{x \in X} n(x) \ .$$

where the cardinality function n can be defined in two coordinate spaces: (μ,ν) or (τ,δ).

In the space (μ,ν) the cardinality function $n(\mu,\nu)$ verifies the following conditions:

c1) $n(1,0)=1,\ n(0,1)=0$

c2) if $\mu_1 < \mu_2$ then $n(\mu_1,\nu) < n(\mu_2,\nu)$

c3) if $\nu_1 > \nu_2$ then $n(\mu,\nu_1) < n(\mu,\nu_2)$

c4) $n(\mu,\nu)+n(\nu,\mu)=1$

c5) $n(\mu,\nu)=n(1-\nu,1-\mu)$

In the space (τ,δ) the cardinality function $n(\tau,\delta)$ verifies the following conditions:

d1) $n(1,0)=1,\ n(-1,0)=0$

d2) if $\tau_1 < \tau_2$ then $n(\tau_1,\delta) < n(\tau_2,\delta)$

d3) if $|\delta_1| > |\delta_2|$ then $n(\tau,\delta_1)\cdot sign(\tau) \le n(\tau,\delta_2)\cdot sign(\tau)$

d4) $n(\tau,\delta)+n(-\tau,\delta)=1$

d5) $n(\tau,\delta)=n(\tau,-\delta)$

We construct the cardinality of bifuzzy set, considering a function with the following structure:

$$n = t + \frac{c+u+i}{2} \ . \tag{5.2.1}$$

Using the distance $d(x,y)=|x-y|$ one obtains:

$$n = \frac{1}{2} + \frac{\mu-\nu}{2} = \frac{1}{2} + \frac{\tau}{2} \ . \tag{5.2.2}$$

For fuzzy sets, one obtains:

$$n_{fs} = \mu \ .$$

Using the proposed distance $D(x,y)=\dfrac{2|x-y|}{1+|x-y|+|x+y-1|}$ one obtains:

$$n^* = \frac{1}{2} + \frac{\mu-\nu}{1+|\mu-\nu|+|\mu+\nu-1|} = \frac{1}{2} + \frac{\tau}{1+|\tau|+|\delta|} \ . \tag{5.2.3}$$

and for fuzzy sets, it results:

$$n^*_{fs} = \frac{1}{2} + \frac{2\mu-1}{1+|2\mu-1|}$$

The cardinality functions n and n^* defined by (5.2.2) and (5.2.3) verify the conditions (*c1, c2, c3, c4, c5*) in the space (μ,ν) and (*d1, d2, d3, d4, d5*) in the space (τ,δ).

6 Conclusions

In this paper, a new metric distance and similarity between elements of bounded interval $[a,b]$ is constructed. Based on this distance, new measures of entropy and cardinality for bifuzzy sets were defined. The new measures were constructed in the framework of representation of bifuzzy sets with positive and negative values. A multivalued representation of entropy is constructed in the framework of penta-valued knowledge representation. Also, some particular forms of cardinality and entropy are obtained for fuzzy sets and intuitionistic fuzzy sets.

References

1. Atanassov, K.: Intuitionistic fuzzy sets. Fuzzy Sets and Systems 20, 87–96 (1986)
2. Barrenechea, E., Bustince, H., Pagola, M., Fernandez, J., Sanz, J.: Generalized Atanassov's Intuitionistic Fuzzy Index. Construction method. In: Proceedings of IFSA-EUSFLAT Conference, Lisbon (July 2009)
3. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society 35, 99–109 (1943)
4. Belnap, N.: A useful four-valued logic. In: Reidel, D. (ed.) Modern uses of multiple-valued logics, Dordrecht-Boston, pp. 8–37 (1977)
5. Burillo, P., Bustince, H.: Entropy of intuitionistic fuzzy sets and on interval-valued fuzzy sets. Fuzzy Sets and Systems 78, 305–316 (1996)
6. Cornelis, C., Deschrijver, G., Kerre, E.: Square and Triangle: Reflections on Two Prominent Mathematical Structures for the Representation of Imprecision. Notes on Intuitionistic Fuzzy Sets 9, 11–21 (2003)
7. Frank, M.J.: On the simultaneous associativity of f(x,y) and x+y-f(x,y). Aeq. Math. 19, 194–226 (1979)
8. Grzegorzewski, P., Mrówka, E.: On the Entropy of Intuitionistic Fuzzy Sets, and Interval-Valued Fuzzy Sets. In: Proceedings of IPMU 2004 Conference, Perugia, Italy, July 4-9 (2004)
9. Kaufmann, A.: Introduction to the Theory of Fuzzy Subsets. Acad. Press, New York (1975)
10. Kosko, B.: Fuzzy entropy and conditioning. Information sciences 40, 165–174 (1986)
11. De Luca, A., Termini, S.: A definition of nonprobabilistic entropy in the setting of fuzzy theory. Information and Control 20, 301–312 (1972)
12. Pătraşcu, V.: A New Penta-valued Logic Based Knowledge Representation. In: Proceedings of IPMU 2008 Conference, Malaga, Spain, June 22-27, pp. 17–22 (2008)
13. Szmidt, E., Kacprzyk, J.: Similarity of Intuitionistic Fuzzy Sets and the Jaccard Coefficient. In: IPMU 2004 Conference, Perugia, Italy, July 4-9 (2004)
14. Szmidt, E., Kacprzyk, J.: A new measure of entropy and its connection with a similarity measure for intuitionistic fuzzy sets. In: EUSFLAT-LFA 2005 Conference, Barcelona, Spain, pp. 461–466 (2005)
15. Smarandache, F.: Neutrosophic set. A generalization of the intuitionistic fuzzy set. Intern. J. Pure Appl. Math. 24, 287–297 (2005)
16. Yager, R.R.: On measures of fuzziness and negation. Part I: Membership in the unit interval. Int. J. Gen. syst. 5, 221–229 (1997)
17. Zadeh, L.: Fuzzy sets. Information and Control 8, 338–353 (1965)

Some Remarks on the Solutions to the Functional Equation $I(x, y) = I(x, I(x, y))$ for D-Operations

Sebastià Massanet and Joan Torrens

Department of Mathematics and Computer Science
University of the Balearic Islands
07122 Palma de Mallorca, Spain
sebasmassanet@gmail.com, dmijts0@uib.es

Abstract. This paper is devoted to the iterative functional equation $I(x, y) = I(x, I(x, y))$ for all $x, y \in [0, 1]$ where I denotes a fuzzy implication. This equation, that comes from a tautology in crisp logic, is revised from the results obtained in A. Xie and F. Qin (2010) Information Sciences, doi:10.1016/j.ins.2010.01.023, clarifying which kinds of continuous t-norms and t-conorms can be used in order to generate D-operations satisfying the mentioned equation.

1 Introduction

Aggregation operations are functions commonly used in the process of merging several inputs into only one representative output. This simple idea has lead to a theory based on such functions with a great number of applications in many fields, like fuzzy control, decision making, data fusion, image processing, measure theory and others. Due to these applications, the theory of aggregation functions is a field of increasing interest as it is reflected by the recent publication of some different but complementary monographs on the topic (see [5], [6], [10] and [18]).

Another application of aggregation functions can be found in fuzzy logic. Specifically, some kinds of aggregation functions have been used in defining fuzzy implications. Not only t-norms and t-conorms, but also copulas, quasi-copulas and even conjunctors in general ([9]), representable aggregation functions ([7]), and mainly uninorms ([1], [4], [8], [13], [15], [16]). Fuzzy implications are essential in fuzzy logic, not only because they are used to model fuzzy conditionals, but also because forward and backward inferences in any fuzzy rules based system are managed through fuzzy implications. There are many equations involving fuzzy implications, usually derived from equivalences in crisp logic, that have been studied in the literature (see [2], [3], [14]), in order to obtain fuzzy implications with specific properties.

One of these equations is given by

$$I(x, y) = I(x, I(x, y)) \quad \text{for all} \quad x, y \in [0, 1] \tag{1}$$

where I denotes a fuzzy implication. This equation comes from the tautology in crisp logic,

$$p \to (p \to q) \equiv p \to q,$$

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 666–675, 2010.

and it was already studied for R, (S,N) and QL-operations in [17]. Equation (1) was also studied for the case of D-operations in the forthcoming paper [19]. The difference is that in case of QL-operations it becomes clear from the results in [17] which continuous t-conorms and t-norms may be used to generate QL-operations satisfying Equation (1). On the contrary, for the case of D-operations characterizations found in [19] are given through conditions that are hard to verify. Thus, except for the case of the maximum t-conorm, the results proved there are not practical in order to decide when a given t-norm and a t-conorm generate a D-operation satisfying Equation (1).

In this paper, we want to deal deeper with this equation in order to clarify these questions, specially for the case of Archimedean t-norms and t-conorms (in fact nilpotent ones), and with the case of ordinal sum t-conorms and t-norms. Additionally, with our study infinitely many examples are derived and the structure of the corresponding D-operations is given in many cases. Moreover, the important case of implications in the sense of Definition 1 below, is not considered separately in [19]. One goal in our study will be to detail such a case and we will prove that one and only one D-implication is obtained satisfying Equation (1) and that this implication is in fact the Kleene-Dienes implication.

The paper is structured as follows. After some preliminaries introducing the necessary notation, we deal with Equation (1) in Section 3, by discussing which kinds of t-norms and t-conorms can be used to generate D-operations satisfying it, specially in cases when S is an Archimedean t-conorm or an ordinal sum. We deal also in this section with the important case of D-implications.

2 Preliminaries

We will suppose the reader to be familiar with basic concepts on continuous t-norms and t-conorms. For any result used but not recalled here see for instance [11]. We will recall only some usual notation in this framework and also some definitions and facts on fuzzy implications.

First of all, recall that nilpotent t-norms and t-conorms are fully characterized as those t-norms and t-conorms that are conjugate with the Łukasiewicz t-norm ($T_{\mathbf{L}}$) and t-conorm ($S_{\mathbf{L}}$), respectively. That is, a t-norm T is nilpotent if and only if there exists an automorphism φ of the unit interval such that

$$T(x,y) = (T_{\mathbf{L}})_{\varphi}(x,y) = \varphi^{-1}(\max(\varphi(x)+\varphi(y)-1,0)).$$

Recall that in this case the zero-region of $(T_{\mathbf{L}})_{\varphi}$ is given by

$$(T_{\mathbf{L}})_{\varphi}(x,y) = 0 \Longleftrightarrow y \leq (N_C)_{\varphi}(x) = \varphi^{-1}(1-\varphi(x))$$

where N_C denotes the classical negation $N_C(x) = 1-x$. And a dual result holds also for t-conorms.

Next we recall the most accepted definition of fuzzy implication.

Definition 1. *A binary operator $I : [0,1]^2 \to [0,1]$ is said to be an implication function, or an implication, if it satisfies:*

(I1) $I(x,z) \geq I(y,z)$ *when* $x \leq y$, *for all* $z \in [0,1]$.
(I2) $I(x,y) \leq I(x,z)$ *when* $y \leq z$, *for all* $x \in [0,1]$.
(I3) $I(0,0) = I(1,1) = 1$ *and* $I(1,0) = 0$.

Note that, from the definition, it follows that $I(0,x) = 1$ and $I(x,1) = 1$ for all $x \in [0,1]$ whereas the symmetrical values $I(x,0)$ and $I(1,x)$ are not derived from the definition.

As we have commented, implication functions can be derived from several classes of aggregation functions. We will deal in this paper with those derived from t-norms and t-conorms. The most usual kinds of such implications are:

- R-implications derived from a t-norm T by

$$I_T(x,y) = \sup\{z \in [0,1] \mid T(x,z) \leq y\} \quad \text{for all} \quad x,y \in [0,1].$$

- (S,N)-implications derived from a t-conorm S and a fuzzy negation N by $I_{S,N}(x,y) = S(N(x),y)$ for all $x,y \in [0,1]$. If N is strong, they are called S-implications.
- QL-operations derived from a t-conorm S, a t-norm T and a fuzzy negation (usually strong) N by $I_{QL}(x,y) = S(N(x),T(x,y))$ for all $x,y \in [0,1]$.
- D-operations derived from a t-conorm S, a t-norm T and a fuzzy negation (usually strong) N by $I_D(x,y) = S(T(N(x),N(y)),y)$ for all $x,y \in [0,1]$.

Note that R and (S,N)-implications are always implications in the sense of Definition 1, whereas QL and D-operations are not implications in general. A characterization of those cases when QL or D-operations are implications is still open (see [3]). However, in both cases a common necessary condition is known.

Proposition 1. ([12], [3]) *Let S be a t-conorm, T a t-norm, N a fuzzy negation and I_D the corresponding D-operation (or I_{QL} the QL-operation). If I_D (I_{QL}) is an implication then $S(N(x),x) = 1$ for all $x \in [0,1]$.*

When S and N are continuous the necessary condition given in the proposition above implies that S must be nilpotent, say $S = (S_{\mathbf{L}})_{\Phi}$, and moreover it must be $N(x) \geq (N_C)_{\Phi}(x)$ for all $x \in [0,1]$. Recall that D-operations satisfy all conditions of Definition 1 except in general condition (I2). Additionally D-operations always satisfy $I_D(x,0) = N(x)$ for all $x \in [0,1]$ and $I_D(1,y) = y$ for all $y \in [0,1]$ (see [12]).

3 Main Results

Our main goal is to study which continuous t-conorms S, continuous t-norms T and strong negations N can be used to derive D-operations satisfying Equation (1). We will begin however with some general results without conditions on continuity. First of all, recall that the least and the greatest fuzzy negations are respectively given by

$$N_{D1}(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x > 0. \end{cases} \qquad \text{and} \qquad N_{D2}(x) = \begin{cases} 1 & \text{if } x < 1 \\ 0 & \text{if } x = 1 \end{cases}$$

When we derive D-operations from these negations, they are always implications and they always satisfy Equation (1) without any further condition on the t-norm T and the t-conorm S.

Proposition 2. *Let S be a t-conorm, T a t-norm, N a fuzzy negation and I_D the corresponding D-operation. Then,*

- *If N is the greatest fuzzy negation, I_D always satisfy Equation (1) and it is the implication given by*

$$I_D(x,y) = \begin{cases} 1 & \text{if } x < 1 \\ y & \text{if } x = 1 \end{cases} \tag{2}$$

- *If N is the smallest fuzzy negation, I_D always satisfy Equation (1), it is never an implication and it is given by*

$$I_D(x,y) = \begin{cases} y & \text{if } (x,y) \neq (0,0) \\ 1 & \text{if } (x,y) = (0,0) \end{cases} \tag{3}$$

Proof. It is easy to see that for any t-norm T and t-conorm S, the corresponding D-operation I_D is given by Equation (2) when N is the greatest fuzzy negation. Hence, it is clearly an implication and satisfies $I(x, I(x,y)) = I(x,y) = 1$ when $x < 1$ and $I(1, I(1,y)) = I(1,y) = y$.

Similarly, it can be proved that I_D is given by Equation (3) when N is the smallest fuzzy negation. It is not an implication because the section $I(0,y)$ is not increasing, but it again satisfies Equation (1) since $I(x, I(x,y)) = I(x,y) = y$ when $(x,y) \neq (0,0)$, whereas $I(0,0) = 1$ and also $I(0, I(0,0)) = I(0,1) = 1$. □

Let us now deal with the case when the t-conorm S is continuous. From the well known structure of continuous t-conorms we can divide our study into three cases, depending on whether the t-conorm S is the maximum, any continuous Archimedean t-conorm, or any ordinal sum of continuous Archimedean t-conorms.

3.1 When $S = S_M$

In this case results from [19] leave totally clear which continuous t-norms and strong negations can be used for our purposes. The result proved there looks as follows.

Theorem 1. ([19]) *If a D-operation I is generated by a t-conorm S_M, a continuous t-norm T, and a strong negation N, then it satisfies Equation (1).*

We only want to make the following remark on this result. The same proof of the theorem above works for any t-norm (not necessarily continuous) and for any fuzzy negation N (not necessarily strong) and thus the theorem remains true in this more general framework.

3.2 When S Is Continuous Archimedean

In this case, after showing that a necessary condition for the satisfaction of Equation (1) by a D-operation generated by a continuous Archimedean t-conorm is that both S and T must be nilpotent, the following result is proved in [19].

Theorem 2. ([19]) *Let I be a D-operation generated by a nilpotent t-conorm $S = (S_L)_\varphi$, a nilpotent t-norm $T = (T_L)_\phi$, and a strong negation N. Then it satisfies Equation (1) if and only if $I(x,y) \geq N(\phi^{-1}(1-\phi(N(x))))$ holds when $0 < (T_L)_\phi(N(x),N(y)) < \varphi^{-1}(1-\varphi(y))$, where both φ and ϕ are order automorphisms on the unit interval.*

As we can see, the necessary and sufficient condition in the above theorem is not intuitive and it is not clear from it which t-norms and t-conorms can be used in order to generate a D-operation satisfying Equation (1). However, several results can be proved enlightening what happens in this case. It is proved in [19] that in this case the following relation must hold

$$(N_C)_\varphi(x) \leq N(x) \leq (N_C)_\phi$$

for all $x \in [0,1]$. Moreover, we have the following remarks.

Remark 1. A straightforward computation shows that the condition $I(x,y) \geq N(\phi^{-1}(1-\phi(N(x))))$ when $0 < (T_L)_\phi(N(x),N(y)) < \varphi^{-1}(1-\varphi(y))$ is equivalent to the condition

$$\phi^{-1}(\phi(N(x)) + \phi(N(y)) - 1) \geq \varphi^{-1}(\varphi(N((N_C)_\phi(N(x)))) - \varphi(y))$$

when $0 < \phi^{-1}(\phi(N(x)) + \phi(N(y)) - 1) < \varphi^{-1}(1-\varphi(y))$. Thus, the condition can be expressed only in terms of N and the automorphisms φ and ϕ, making easier its verification.

Remark 2. The resulting D-operation can be computed in this case and it is easy to see that its expression is given by

$$I(x,y) = \begin{cases} y & \text{if } T(N(x),N(y)) = 0 \\ \varphi^{-1}(\varphi(T(N(x),N(y))) + \varphi(y)) & \text{if } 0 < T(N(x),N(y)) < (N_C)_\varphi(y) \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

Note that there are many cases for which the condition of the previous theorem is satisfied. For instance, we have the following particular cases.

Proposition 3. *Consider a nilpotent t-conorm $S = (S_{\mathbf{L}})_\varphi$, a nilpotent t-norm $T = (T_{\mathbf{L}})_\phi$ and a strong negation N such that $(N_C)_\varphi \leq N = (N_C)_\phi$ and $h = \varphi\phi^{-1}$ is a subadditive function. Then the D-operation I generated by these operators satisfies Equation (1).*

Proof. By Remark 1, it is sufficient to prove that

$$\phi^{-1}(\phi(N(x)) + \phi(N(y)) - 1) \geq \varphi^{-1}(\varphi(N((N_C)_\phi(N(x)))) - \varphi(y))$$

whenever $0 < T(N(x), N(y)) < (N_C)_\varphi(y)$. Since $N = (N_C)_\phi$ we have $\phi(N(x)) = 1 - \phi(x)$ and then the previous condition can be rewritten as

$$\begin{array}{l}\phi^{-1}(\phi(N(x)) + \phi(N(y)) - 1) \geq \varphi^{-1}(\varphi(N(x)) - \varphi(y)) \Leftrightarrow \\ \varphi(\phi^{-1}(\phi(N(x)) - \phi(y))) \geq \varphi(N(x)) - \varphi(y)\end{array}$$

Now, taking $N(x) = \phi^{-1}(a)$, $y = \phi^{-1}(b)$ and $h = \varphi\phi^{-1}$, we obtain that the last inequality becomes $h(a - b) \geq h(a) - h(b)$ for all a, b such that

$$T(\phi^{-1}(a), N(\phi^{-1}(b))) < (N_C)_\varphi(\phi^{-1}(b)).$$

It has to be remarked that $a - b = \phi(N(x)) - \phi(y) = \phi(N(x)) + \phi(N(y)) - 1 > 0$ because $0 < T(N(x), N(y))$. Finally, consider $c = a - b$ and we obtain

$$h(c) \geq h(b + c) - h(b) \Leftrightarrow h(b + c) \leq h(b) + h(c).$$

Thus, if h is subadditive the inequality above holds and consequently I satisfies Equation (1). □

Proposition 4. *Consider a nilpotent t-conorm $S = (S_{\mathbf{L}})_\varphi$, a nilpotent t-norm $T = (T_{\mathbf{L}})_\phi$ and a strong negation N such that $(N_C)_\varphi = N \leq (N_C)_\phi$. Then the D-operation I generated by these operators satisfies Equation (1) if and only if $h = \varphi\phi^{-1}$ is a subadditive function.*

Proof. By Remark 1, I satisfies (1) if and only if

$$\phi^{-1}(\phi(N(x)) + \phi(N(y)) - 1) \geq \varphi^{-1}(\varphi(N((N_C)_\phi(N(x)))) - \varphi(y))$$

whenever $0 < T(N(x), N(y)) < (N_C)_\varphi(y)$. However, since $N = (N_C)_\varphi$ we have $\varphi(N(x)) = 1 - \varphi(x)$ and we can rewrite this condition as follows,

$$\begin{array}{ll}\varphi(\phi^{-1}(\phi(N(x)) + \phi(N(y)) - 1) \geq \varphi(N((N_C)_\phi(N(x)))) - \varphi(y) & \Leftrightarrow \\ \varphi(\phi^{-1}(\phi(N(x)) + \phi(N(y)) - 1) \geq 1 - \varphi((N_C)_\phi(N(x))) - \varphi(y) & \Leftrightarrow \\ \varphi(\phi^{-1}(\phi(N(x)) + \phi(N(y)) - 1) \geq \varphi(N(y)) - \varphi(\phi^{-1}(1 - \varphi(N(x)))) & \end{array}$$

Now, taking $N(x) = \phi^{-1}(a)$, $N(y) = \phi^{-1}(b)$ and $h = \varphi\phi^{-1}$, we obtain that the last inequality becomes

$$h(a + b - 1) \geq h(b) - h(1 - a) \tag{5}$$

whenever $0 < T(\phi^{-1}(a), \phi^{-1}(b)) < \phi^{-1}(b)$. That is, I satisfies (1) if and only if Equation (5) holds for all a, b with $a + b - 1 > 0$. Finally, changing $c = 1 - a$ Equation (5) becomes $h(b - c) \geq h(b) - h(c)$ and then considering $d = b - c$ it derives into

$$h(c + d) \leq h(c) + h(d) \quad \text{for all} \quad c, d \quad \text{with} \quad c + d \leq 1.$$

That is, I satisfies Equation (1) if and only if $h = \varphi\phi^{-1}$ is a subadditive function. □

In Example 3 in [19], it is proved that taking $\varphi(x) = x$, $N(x) = 1 - x$ and $\phi(x) = x^2$, the obtained D-operation satisfies Equation (1) through the verification of the condition of Theorem 2. However, note that in this case $h(x) = \varphi \circ \phi^{-1}(x) = \sqrt{x}$ is a subadditive function and consequently, by the proposition above the corresponding D-operation satisfies Equation (1). Thus, Example 3 in [19] is a particular case of the above proposition. Similarly, in Example 4 in [19] function h agrees with $h(x) = x^2$ and consequently the corresponding I does not satisfy (1) because h is not subadditive.

We note that in order D-operations to be fuzzy implications S must be such that $S(x, N(x)) = 1$. Thus, we can find D-implications satisfying Equation (1) only in the case when S is nilpotent. Unfortunately, there is one and only one D-implication satisfying Equation (1), which agrees with the Kleene-Dienes implication.

Proposition 5. *Consider a nilpotent t-conorm $S = (S_{\mathbf{L}})_{\varphi}$, a nilpotent t-norm $T = (T_{\mathbf{L}})_{\phi}$ and a strong negation N such that the corresponding D-operation I satisfies Equation (1). Then the following conditions are equivalent.*

(i) I is an implication.
(ii) $(N_C)_{\phi} = N = (N_C)_{\varphi}$.
(iii) I is the Kleene-Dienes implication.

Proof. Let us begin with the proof of $(i) \Rightarrow (ii)$. Suppose that I is an implication. Note on one hand that

$$T(N(x), N(y)) = 0 \iff N(y) \leq (N_C)_{\phi}(N(x)) \iff y \geq N((N_C)_{\phi}(N(x))) \tag{6}$$

Thus, suppose first that there is some y_0 such that $(N_C)_{\varphi}(y_0) < N(y_0)$. Since T is continuous there exists an $x_0 > 0$ such that $T(N(x_0), N(y_0)) = (N_C)_{\varphi}(y_0)$ and then $I(x_0, y_0) = 1$ by Remark 2, whereas $I(x_0, y) = y$ for all $y \geq N((N_C)_{\phi}(N(x_0)))$ by Equation (6), violating condition (I2). That is, $N = (N_C)_{\varphi}$.

On the other hand, using Theorem 2 we know that $I(x, y) \geq N((N_C)_{\phi}(N(x)))$ for all $x, y \in [0, 1]$. Thus, fixed any $x \in]0, 1[$ and taking into account Equation (6), it must be $I(x, y) = N((N_C)_{\phi}(N(x)))$ for all $y \leq N((N_C)_{\phi}(N(x)))$ in order to preserve condition (I2). That is, it must be

$$\varphi(T(N(x), N(y))) + \varphi(y)) = \varphi(N((N_C)_{\phi}(N(x)))) \text{ for all } y \leq N((N_C)_{\phi}(N(x)))$$

or equivalently

$$\varphi(\phi^{-1}(\phi(N(x)) + \phi(N(y)) - 1) + 1 - \varphi(N(y))) = 1 - \varphi((N_C)_{\phi}(N(x)))$$

or

$$\varphi(\phi^{-1}(\phi(N(x)) + \phi(N(y)) - 1)) = \varphi(N(y)) - \varphi((N_C)_{\phi}(N(x))).$$

Now, taking $h = \varphi\phi^{-1}$, $N(x) = \phi^{-1}(a)$ and $N(y) = \phi^{-1}(b)$, we obtain

$$h(a + b - 1) = h(b) - h(1 - a) \quad \text{for all} \quad a, b \quad \text{with} \quad a + b \geq 1$$

leading to a Cauchy equation and consequently to $h = Id$. That is, $N = (N_C)_\varphi = (N_C)_\phi$.

Now $(ii) \Rightarrow (iii)$ is straightforward from the expression (4) and $(iii) \Rightarrow (i)$ is trivial. □

3.3 When S Is an Ordinal Sum

In this case the characterization given in [19] is as follows.

Theorem 3. ([19]) *Let I be a D-operation generated by a t-norm T a strong negation N, and a t-conorm $S = (\langle a_m, b_m, S_m\rangle)_{m\in A}$, where T is a continuous Archimedean t-norm or an ordinal sum of the corresponding family $\{[c_m, d_m], T_m\}$, with each T_m being a continuous Archimedean t-norm, S_m a continuous Archimedean and A a finite or countable infinite index set. Then Equation (1) holds if and only if for arbitrary fixed $(x, y) \in [0, 1]^2$, there exists some (a_m, b_m) such that $T(N(x), N(y)) \in (a_m, b_m)$, and the following conditions are satisfied. If $(y \in (a_m, b_m)$ and $I(x, y) < b_m)$ or $y \notin (a_m, b_m)$, then it always holds that $T(N(x), N \circ I(x, y)) \leq a_m$.*

It is obvious that the condition on the previous theorem is hard to verify and it is not clear from it which t-norms and t-conorms can be used. In this case we can give some partial results clarifying the situation in some particular cases.

Proposition 6. *Let I be a D-operation generated by a strong negation N with fixed point e, a continuous t-conorm S satisfying $S(x, x) = x$ for all $x \leq e$ and a t-norm T. Then I satisfies Equation (1).*

Proof. When a continuous t-conorm S satisfies $S(x, x) = x$ for all $x \leq e$ we have $S(x, y) = \max(x, y)$ for all x, y such that $\min(x, y) \leq e$. Let us see that in this case $I(x, y)$ is always given by $I(x, y) = \max(T(N(x), N(y)), y)$. Effectively,

- When $T(N(x), N(y)) \leq y$, since it holds also $T(N(x), N(y)) \leq N(y)$, we have $T(N(x), N(y)) \leq e$ and the result follows.
- When $T(N(x), N(y)) > y$, we have $y < T(N(x), N(y)) \leq N(y)$ and consequently $y < e$ and the result follows again.

Now, Theorem 1 ends the proof. □

Note that in the examples given in [19] after Remark 8 and in Remark 9 both t-conorms are idempotent at least until the fixed point of the negation and consequently the corresponding D-operation satisfies Equation (1) not only for the product t-norm or for an specific ordinal sum t-norm, but also for any t-norm (even not necessarily continuous).

Proposition 7. *Let I be a D-operation generated by a continuous t-conorm S, a strong negation N with fixed point e and a continuous t-norm T satisfying $T(e, e) = e$. Then I satisfies Equation (1) if and only if $S(x, x) = x$ for all $x \leq e$.*

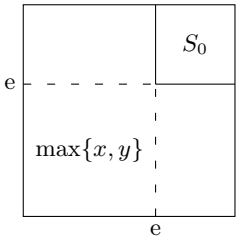


Fig. 1. Structure of the t-conorm given in Proposition 6 with S_0 any t-conorm

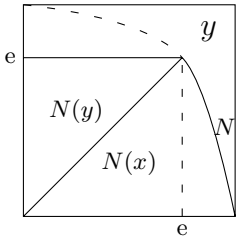


Fig. 2. Structure of the D-operation given in Corollary 1

Proof. If $S(x, x) = x$ for all $x \leq e$ then by the previous proposition, I satisfies Equation (1). Reciprocally, if T is a continuous t-norm satisfying $T(e, e) = e$ then $T(x, y) = \min(x, y)$ whenever $\min(x, y) \geq e$ and, in particular, $T(x, N(x)) = \min(x, N(x))$ for all $x \in [0, 1]$. Thus, taking $y = 0$ and $x \geq e$, we have

$$\begin{aligned} I(x, 0) &= T(N(x), N(0)) = N(x) \\ I(x, I(x, 0)) &= I(x, N(x)) = S(T(N(x), x), N(x)) = S(\min(N(x), x), N(x)) \\ &= S(N(x), N(x)). \end{aligned}$$

Therefore, if I satisfies Equation (1) we obtain $S(N(x), N(x)) = N(x)$ for all $x \geq e$ and so $S(x, x) = x$ for all $x \leq e$. □

Note that this proposition includes Theorem 11 in [19] where the minimum t-norm is considered. Moreover, in this case the structure of the resulting D-operation can be described and can be viewed in Figure 2.

Corollary 1. *Let I be a D-operation generated by a strong negation N with fixed point e, a continuous t-conorm S satisfying $S(x, x) = x$ for all $x \leq e$, and a continuous t-norm T satisfying $T(e, e) = e$. Then I satisfies Equation (1) and it is given by*

$$I(x, y) = \begin{cases} y & \text{if } y \geq e \text{ or } N(x) \leq y < e \\ N(x) & \text{if } y < e \text{ and } y < \min\{x, N(x)\} \\ N(y) & \text{if } y < e \text{ and } y \geq x \end{cases}$$

Acknowledgements. This work has been partially supported by the Spanish Grant MTM2009-10320 with FEDER support.

References

1. Aguiló, I., Suñer, J., Torrens, J.: A characterization of residual implications derived from uninorms. In: Proceedings of IFSA-EUSFLAT 2009, Lisbon, pp. 333–338 (2009)
2. Baczyński, M., Jayaram, B.: (S,N)- and R-implications: A state-of-the-art survey. Fuzzy Sets and Systems 159, 1836–1859 (2008)
3. Baczyński, M., Jayaram, B.: Fuzzy Implications. Studies in Fuzziness and Soft Computing, vol. 231. Springer, Heidelberg (2008)
4. Baczyński, M., Jayaram, B.: (U,N)-implications and their characterizations. Fuzzy Sets and Systems 160, 2049–2062 (2009)
5. Beliakov, G., Pradera, A., Calvo, T.: Aggregation Functions: A Guide for Practicioners. Springer, Heidelberg (2007)
6. Calvo, T., Mayor, G., Mesiar, R. (eds.): Aggregation operators. New trends and applications. Studies in Fuzziness and Soft Computing, vol. 97. Physica-Verlag, Heidelberg (2002)
7. Carbonell, M., Torrens, J.: Continuous R-implications generated from representable aggregation functions. Fuzzy Sets and Systems (accepted)
8. De Baets, B., Fodor, J.C.: Residual operators of uninorms. Soft Computing 3, 89–100 (1999)
9. Durante, F., Klement, E.P., Mesiar, R., Sempi, C.: Conjunctors and their residual implicators: Characterizations and construction methods. Mediterranean Journal of Mathematics 4, 343–356 (2007)
10. Grabisch, M., Marichal, J.L., Mesiar, R., Pap, E.: Aggregation functions. Encyclopedia of Mathematics and its Applications, vol. 127. Cambridge University Press, Cambridge (2009)
11. Klement, E.P., Mesiar, R., Pap, E.: Triangular norms. Kluwer Academic Publishers, London (2000)
12. Mas, M., Monserrat, M., Torrens, J.: QL versus D-implications. Kybernetika 42, 351–366 (2006)
13. Mas, M., Monserrat, M., Torrens, J.: Two types of implications derived from uninorms. Fuzzy Sets and Systems 158, 2612–2626 (2007)
14. Mas, M., Monserrat, M., Torrens, J., Trillas, E.: A survey on fuzzy implication functions. IEEE Transactions on Fuzzy Systems 15(6), 1107–1121 (2008)
15. Ruiz, D., Torrens, J.: Residual implications and co-implications from idempotent uninorms. Kybernetika 40, 21–38 (2004)
16. Ruiz-Aguilera, D., Torrens, J.: Distributivity of residual implications over conjunctive and disjunctive uninorms. Fuzzy Sets and Systems 158, 23–37 (2007)
17. Shi, Y., Ruan, D., Kerre, E.E.: On the characterization of fuzzy implications satisfying $I(x, y) = I(x, I(x, y))$. Information Sciences 117, 2954–2970 (2007)
18. Torra, V., Narukawa, Y.: Modeling decisiona. In: Information fusion and aggregation operators. Cognitive Technologies. Springer, Heidelberg (2007)
19. Xie, A., Qin, F.: Solutions of the functional equation $I(x, y) = I(x.I(x.y))$ for a continuous D-operation. Information Sciences (2010), doi:10.1016/j.ins.2010.01.023

On an Open Problem of U. Höhle - A Characterization of Conditionally Cancellative T-Subnorms

Balasubramaniam Jayaram

Indian Institute of Technology Hyderabad,
Yeddumailaram - 502 205, India
jbala@iith.ac.in

Abstract. In this work we solve an open problem of U.Höhle [Problem 11, Fuzzy Sets and Systems 145 (2004) 471-479]. We show that the solution gives a characterization of all conditionally cancellative t-subnorms. Further, we give an equivalence condition for a conditionally cancellativite t-subnorm to be a t-norm and hence show that conditionally cancellativite t-subnorms whose natural negations are strong are, in fact, t-norms.

1 Introduction

The paper by Klement et al. [6] is a collection of open problems posed during the 24^{th} Linz Seminar on fuzzy set theory. They deal with unsolved problems (as of then) related to fuzzy aggregation operations, especially t-norms and t-subnorms. Since the publication of [6], some problems mentioned therein have been solved - for instance, Problem **1** was solved by Ouyang et al. [8], Problem **5** was solved by Ouyang and Li [8] while for some other problems partial solutions have been given, see for instance, the papers of Viceník [9], [10], [11] relating to Problem **4**(i).

One of the open problems listed therein was posed by Prof. U. Höhle (Problem **11**) which reads as follows:

*Problem 1 (U.Höhle, [6], Problem **11**).* Characterize all left-continuous t-norms T which satisfy

$$I(x, T(x, y)) = \max(n(x), y), \quad x, y \in [0, 1] \,. \tag{1}$$

where I is the residual operator linked to T, i.e.,

$$I(x, y) = \sup\{t \in [0, 1] | T(x, t) \leq y\}, \quad x, y \in [0, 1] \,, \tag{2}$$

$$n(x) = n_T(x) = I(x, 0) \text{ for all } x \in [0, 1]. \tag{3}$$

Further, Prof. U.Höhle goes on to remark the following:

Remark 1. "In the class of continuous t-norms, only nilpotent t-norms fulfill the above property."

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 676–682, 2010.

In this work we deal with two problems. Firstly, we solve the above open problem of U.Höhle and show that the solution gives a characterization of all conditionally cancellative t-subnorms. From the proven result it does follow that the remark of Prof. U.Höhle - Remark 1 - is not always true and give an equivalence condition for it to be true, viz., that the natural negation obtained from the t-norm is strong.

Secondly, this quite naturally leads us to consider conditionally cancellative t-subnorms whose natural negations are involutive. Once again, by proving an equivalence condition for a conditionally cancellative t-subnorm to be a t-norm, we show that conditionally cancellative t-subnorms whose natural negations are involutive, in fact, become t-norms.

2 Preliminaries

Definition 1. *A function* $N\colon [0,1] \to [0,1]$ *is called a* fuzzy negation *if* N *is decreasing and* $N(0) = 1$, $N(1) = 0$.

Definition 2 ([5], Definition 1.7). *A* t-subnorm *is a function* $M\colon [0,1]^2 \to [0,1]$ *such that it is monotonic non-decreasing, associative, commutative and* $M(x,y) \leq \min(x,y)$ *for all* $x, y \in [0,1]$.

Note that for a t-subnorm 1 need not be the neutral element, unlike in the case of a t-norm.

Definition 3 (cf. [5], Definition 2.9 (iii)). *A t-subnorm* M *satisfies the* Conditional Cancellation Law *if, for any* $x, y, z \in (0,1]$,

$$M(x,y) = M(x,z) > 0 \text{ implies } y = z \,. \tag{CCL}$$

Alternately, (CCL) implies that on the positive domain of M, i.e., on the set $\{(x,y) \in (0,1]^2 \mid M(x,y) > 0\}$, M is strictly increasing.

Definition 4 (cf. [1], Definition 2.3.1). *Let* M *be any t-subnorm. Its natural negation* n_M *is given by*

$$n_M(x) = \sup\{t \in [0,1] \mid M(x,t) = 0\}, \quad x \in [0,1] \,. \tag{4}$$

Note that though $n_M(0) = 1$, it need not be a fuzzy negation, since $n_M(1)$ can be greater than 0. However, we have the following result.

Lemma 1 (cf. [1], Proposition 2.3.4). *Let* M *be any t-subnorm and* n_M *its natural negation. Then we have the following:*

(i) $M(x,y) = 0 \Longrightarrow y \leq n_M(x)$.
(ii) $y < n_M(x) \Longrightarrow M(x,y) = 0$.
(iii) If M *is left-continuous then* $y = n_M(x) \Longrightarrow M(x,y) = 0$, *i.e., the reverse implication of (i) also holds.*

3 Solution to the Open Problem of U. Höhle

It should be noted that in the case T is left-continuous - as stated in **Problem 1** - the sup in (2) actually becomes max. It is worth mentioning that the residual can be determined for more generalised conjunctions and the conditions under-which this residual becomes a fuzzy implication can be found in, for instance, [2], [4]. Hence we further generalise the statement of **Problem 1** by considering a t-subnorm instead of a t-norm and also dropping the condition of left-continuity. As we show below the solution characterizes the set of all conditionally cancellative t-subnorms.

Theorem 1. *Let M be any t-subnorm and I the residual operation linked to M by (2). Then the following are equivalent:*

(i) The pair (I, M) satisfies (1).
(ii) M is a Conditionally Cancellative t-subnorm.

Proof. Let M be any t-subnorm, not necessarily left-continuous. Note that we denote n_M simply by n.

(i) $\Longrightarrow$ (ii): Let the adjoint pair (I, M) satisfy (1). On the contrary, let us assume that there exist $x, y, z \in (0, 1)$ such that $M(x, y) = M(x, z) > 0$ but $y < z$. Then we have that

$$\text{LHS (1)} = I(x, M(x, y)) = \sup\{t \in [0, 1] \mid M(x, t) \leq M(x, y)\} \geq z > y \ .$$

However, note that, from Lemma 1 (i) we have that $y \geq n(x)$, since $M(x, y) > 0$. Thus

$$\text{RHS (1)} = \max(n(x), y) = y < \text{LHS (1)} \ ,$$

a contradiction to the fact that the adjoint pair (I, M) satisfies (1). Hence M satisfies (CCL).

(ii) $\Longrightarrow$ (i): Now, let M satisfy (CCL). Consider any arbitrary $x, y \in [0, 1]$. Then either $n(x) > y$ or $n(x) \leq y$.
If $n(x) > y$, then by Lemma 1 (ii) we see that $M(x, y) = 0$ and hence

$$\text{LHS (1)} = I(x, M(x, y)) = I(x, 0) = n(x) = \max(n(x), y) = \text{RHS (1)}.$$

If $n(x) \leq y$ and $M(x, y) = 0$ then by Lemma 1(i) we have that $n(x) \geq y$ and hence $n(x) = y$ and it reduces to the above case. Hence let $M(x, y) > 0$. Then

$$\text{RHS (1)} = \max(n(x), y) = y \ .$$

We claim now that LHS (1) $= I(x, M(x, y)) = y$. If this were not true, then there exists $1 \geq z > y$ ($z \not< y$ by the monotonicity of M) such that

$$I(x, M(x, y)) = \sup\{t \in [0, 1] \mid M(x, t) \leq M(x, y)\} = z.$$

This implies that there exists a $w \in (0,1)$ such that $z > w > y$ and $M(x,w) \leq M(x,y)$, which by the monotonicity of t-subnorm implies that $M(x,w) = M(x,y)$ with $w \gneq y$, a contradiction to the fact that M satisfies (CCL). Thus the adjoint pair (I,M) satisfies (11). □

Example 1. Consider the product t-norm $T_{\mathbf{P}}(x,y) = xy$, which is a strict t-norm and hence continuous and Archimedean, whose residual is the Goguen implication given by

$$I_{\mathbf{GG}}(x,y) = \begin{cases} 1, & \text{if } x \leq y, \\ \dfrac{y}{x}, & \text{if } x > y. \end{cases}$$

It can be easily verified that the pair $(T_{\mathbf{P}}, I_{\mathbf{GG}})$ does indeed satisfy (11) whereas the natural negation of $T_{\mathbf{P}}$ is the Gödel negation

$$n_{T_{\mathbf{P}}}(x) = I_{\mathbf{GG}}(x,0) = \begin{cases} 1, & \text{if } x = 0, \\ 0, & \text{if } x > 0. \end{cases}$$

This example clearly shows that the remark of U.Höhle, Remark 1, is not always true. In the following we give an equivalence condition under which it is true.

Theorem 2. *Let T be a continuous t-norm that satisfies (11) along with its residual. Then the following are equivalent:*

(i) T is nilpotent.
(ii) n_T is strong.

Proof. (i) $\Longrightarrow$ (ii): Obvious.
(ii) $\Longrightarrow$ (ii): If T is continuous and satisfies (11) along with its residual then, from Theorem 1, T is conditionally cancellative and hence necessarily Archimedean by [5], Proposition 2.15 (ii). Thus T is either nilpotent or strict. If T is continuous with a strong natural negation, clearly, T has zero-divisors and hence T is nilpotent. □

4 Conditional Cancellativity and Unit Element

From the above remarks we note that when the natural negation of the underlying conjunction (a continuous t-norm, in the above case) is strong the class of conjunctions that satisfy (11) along with its residual gets restricted. Hence we study the class of t-subnorms M that satisfy (11) along with its residual and whose natural negations are strong. In other words, we seek the characterization of the class of conditionally cancellative t-subnorms with strong natural negations.

Let us recall from the remark following Definition 4 that the natural negation of a t-subnorm n_M need not be a fuzzy negation. If a t-subnorm has 1 as its neutral element, i.e., if it is a t-norm, then we have

$$M(1,y) = 0 \Longleftrightarrow y = 0,$$
$$i.e.,\ y = \sup\{t | M(1,t) = 0\} = n_M(1) = 0.$$

Equivalently, by the monotonicity of M we have that n_M is a fuzzy negation. However, this is only a necessary and not a sufficient condition.

Note that, so far, no general result giving equivalence conditions under which a t-subnorm becomes a t-norm is available. It was Jenei [3] who proposed some suficiency conditions and showed that left-continuous t-subnorms with strong natural negations are t-norms, i.e., 1 does become a neutral element.

In the following we give an equivalence condition for a conditionally cancellative t-subnorm to be a t-norm and show that in the case n_M is a strong negation then M always is a t-norm.

Lemma 2. *Let M be a conditionally cancellative t-subnorm. Let $M(1, y_0) = y_0$, for some $y_0 \in (0, 1]$.*

(i) Then $M(1, y) = y$ for all $y \in [y_0, 1]$.
(ii) Let $y^* = \sup\{t | M(1, t) = 0\} = n_M(1)$. Then $M(1, y) = y$ for all $y \in (y^*, y_0]$.

Proof. Let $M(1, y_0) = y_0$, for some $y_0 \in (0, 1]$.

(i) Let $y_0 < y \leq 1$. Clearly, $y_0 = M(1, y_0) < M(1, y) \leq y$. If $M(1, y) = y' < y$, then by associativity and conditional cancellativity we have

$$\left.\begin{array}{l} M(M(1, y_0), y) = M(y_0, y) \\ \\ M(M(1, y), y_0) = M(y', y_0) \end{array}\right\} \Longrightarrow M(y_0, y) = M(y_0, y') \Longrightarrow y = y' \ ,$$

i.e., $M(1, y) = y$ for all $y \geq y_0$.

(ii) Let $y^* < y \leq y_0$. Clearly, $y_0 = M(1, y_0) > y \geq M(1, y) = y'$. If $M(1, y) = y' < y$, then, once again, by associativity and conditional cancellativity we have

$$\left.\begin{array}{l} M(M(1, y_0), y) = M(y_0, y) \\ \\ M(M(1, y), y_0) = M(y', y_0) \end{array}\right\} \Longrightarrow M(y_0, y) = M(y_0, y') \Longrightarrow y = y' \ ,$$

i.e., $M(1, y) = y$ for all $y \in (y^*, y_0]$. □

Based on the above result, we now have the following equivalence condition for a conditionally cancellative t-subnorm to be a t-norm:

Theorem 3. *Let M be any conditionally cancellative t-subnorm. Then the following are equivalent:*

(i) M is a t-norm.
(ii) n_M is a negation and $M(1, y_0) = y_0$, for some $y_0 \in (0, 1]$.

Proof. Sufficiency is obvious. Necessity follows from the fact that if n_M is a negation then $y^* = 0$ in Lemma 2 above. □

The final result of this work shows that in the case n_M is a strong negation then M always is a t-norm.

Theorem 4. *Let M be any conditionally cancellative t-subnorm. If n_M is a strong natural negation then M is a t-norm.*

Proof. Our approach will be to show that $M(1,1) = 1$ and then the result follows easily from Theorem 3. Note also that since n_M is a strong negation, we have that $n_M(x) = 1 \iff x = 0$ and $n_M(x) = 0 \iff x = 1$. Equivalently, $M(1,x) = 0 \iff x = 0$.

On the contrary, let us assume that $M(1,y) < y$ for all $y \in (0,1]$. In particular, $M(1,1) = z$ such that $0 < z < 1$. Since n_M is strong, there exists a $z' \in (0,1)$ such that $z = n_M(z')$. We claim that $z' = 0$ and hence $z = 1$.

If not, then there exists $0 < z'' < z'$ and by the definition of n_M we have that $M(z,z'') = 0$. Also, by our assumption $0 < M(1,z'') = z^* < z''$. Now, by associativity and conditional cancellativity we have

$$\left.\begin{array}{l} M(M(1,1),z'') = M(z,z'') \\ M(M(1,z''),1) = M(z^*,1) \end{array}\right\} \Longrightarrow M(z,z'') = 0 = M(z^*,1)$$
$$\Longrightarrow z^* = 0\,,$$

a contradiction. Thus $z = 1$ and hence we have the result. □

5 Concluding Remarks

In this work we have solved a more generalised version of an open problem of U.Höhle and shown that the solution gives a characterization of all conditionally cancellative t-subnorms. Further, by proving an equivalence condition for a conditionally cancellative t-subnorm to be a t-norm, we have shown that conditionally cancellative t-subnorms with involutive natural negations are t-norms.

References

1. Baczyński, M., Jayaram, B.: Fuzzy Implications. In: Studies in Fuzziness and Soft Computing, vol. 231. Springer, Heidelberg (2008)
2. Demirli, K., De Baets, B.: Basic properties of implicators in a residual framework. Tatra Mount. Math. Publ. 16, 31–46 (1999)
3. Jenei, S.: Continuity of left-continuous triangular norms with strong induced negations and their boundary condition. Fuzzy Sets and Systems 124, 35–41 (2001)
4. Jayaram, B., Mesiar, R.: I-Fuzzy Equivalence Relations and I-Fuzzy Partitions. Info. Sci. 179, 1278–1297 (2009)
5. Klement, E.P., Mesiar, R., Pap, E.: Triangular norms. Kluwer, Dordrecht (2000)
6. Klement, E.P., Mesiar, R., Pap, E.: Problems on triangular norms and related operators. Fuzzy Sets and Systems 145, 471–479 (2004)
7. Ouyang, Y., Li, J., Fang, J.: A conditionally cancellative left-continuous t-norm is not necessarily continuous. Fuzzy Sets and Systems 157, 2328–2332 (2006)
8. Ouyang, Y., Li, J.: An answer to an open problem on triangular norms. Info. Sci. 175, 78–84 (2005)

9. Viceník, P.: Additive generators of associative functions. Fuzzy Sets and Systems 153, 137–160 (2005)
10. Viceník, P.: Intersections of ranges of additive generators of associative functions. Tatra Mt. Math. Publ. 40, 117–131 (2008)
11. Viceník, P.: Additive generators of border-continuous triangular norms. Fuzzy Sets and Systems 159, 1631–1645 (2008)

Triangular Norms and Conorms on the Set of Discrete Fuzzy Numbers

Jaume Casasnovas and J. Vicente Riera

Department of Mathematics and Computer Science, University of Balearic Islands,
07122 Palma de Mallorca, Spain
{jaume.casasnovas,jvicente.riera}@uib.es

Abstract. In this paper, we study different methods to construct triangular operations (t-norms and t-conorms) on the bounded distributive lattice $\mathcal{A}_1^L$, of discrete fuzzy numbers whose support is a subset of consecutive natural numbers on a finite chain L of consecutive natural numbers. Moreover, we propose a method to compare two t-norms (t-conorms) defined on $\mathcal{A}_1^L$.

1 Introduction

Triangular norms (t-norms in short) on the unit interval were systematically investigated in the sixties by Schweizer and Sklar [12] in the framework of probabilistic metric spaces. Moreover, these authors introduced triangular conorms (t-conorms in short) as dual operations of t-norms. Since then, they are a useful tool for defining the intersection and union of fuzzy sets, and for modelling the logic connectives conjunction and disjunction in fuzzy logic. Due to the close connection between order theory and fuzzy set theory, see e.g. [8], several authors have investigated with t-norms and t-conorms on a general bounded partially ordered set (finite chains, product lattices, real unit square, etc), see e.g [1,9,11,15] and many others.

Voxman [13], introduced the concept of discrete fuzzy number such as a fuzzy subset on $\mathbb{R}$ with discrete support and analogous properties to a fuzzy number. It is well known that, arithmetic and lattice operations between fuzzy numbers are defined using the Zadeh's extension principle [10]. But, in general, for discrete fuzzy numbers this method fails [2,3,4,14]. We have studied this drawback [2,3,4] and we have obtained new closed operations in the set of discrete fuzzy numbers. In particular, we showed [5] that $\mathcal{A}_1$, the set of discrete fuzzy numbers whose support is a sequence of consecutive natural numbers, is a distributive lattice. In this lattice, we considered a partial order, obtained in a usual way, from the lattice operations of this set. So, from this partial order, we investigated [6] the extension of monotone operations defined on a discrete setting to a closed binary operation of discrete fuzzy numbers and also, we investigated different properties such as the monotonicity, commutativity and associativity.

The aim of this article is the exposition of two methods to construct triangular operations (t-norms and t-conorms) defined on the bounded distributive lattice $\mathcal{A}_1^L = \{A \in \mathcal{A}_1 \mid supp(A) \subseteq L\}$, where L denotes the finite chain

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 683–692, 2010.

$L = \{0, 1, \cdots, n\}$. The first method is based on the result obtained in our article [6] and the second method is based on the representation of intuitionistic fuzzy t-norms and t-conorms [7]. Moreover, we propose the comparison of two t-norms(t-conorms) defined on $\mathcal{A}_1^L$ from the comparison of two t-norms(t-conorms) defined on L.

2 Preliminaries

2.1 Triangular Norms and Conorms on Partially Ordered Sets

Let $(P; \leq)$ be a non-trivial bounded partially ordered set (poset) with "e" and "m" as minimum and maximum elements respectively.

Definition 1. *[1,15] A* triangular norm *(briefly t-norm) on P is a binary operation $T : P \times P \to P$ such that for all $x, y, z \in P$ the following axioms are satisfied:*

1. $T(x, y) = T(y, x)$ *(commutativity)*
2. $T(T(x, y), z) = T(x, T(y, z))$ *(associativity)*
3. $T(x, y) \leq T(x', y')$ *whenever* $x \leq x', y \leq y'$ *(monotonicity)*
4. $T(x, m) = x$ *(boundary condition)*

Definition 2. *A* triangular conorm *(t-conorm for short) on P is a binary operation $S : P \times P \to P$ which, for all $x, y, z \in P$ satisfies (1), (2), (3) and (4′): $S(x, e) = x$, as boundary condition.*

2.2 Triangular Norms and Conorms on Discrete Settings

Let L be the totally ordered set $L = \{0, 1, \ldots, n\} \subset \mathbb{N}$. A t-norm(t-conorm) defined on L will be called a discrete t-norm(t-conorm).

Definition 3. *[11] A t-norm(t-conorm) $T(S) : L \times L \to L$ is said to be* smooth *if it satisfies $T(S)(x+1, y) - T(S)(x, y) \leq 1$ and $T(S)(x, y+1) - T(S)(x, y) \leq 1$.*

Definition 4. *[11] A t-norm(t-conorm) $T : L \times L \to L$ is said to be* divisible *if it satisfies: For all $x, y \in L$ with $x \leq y$, there is $z \in L$ such that $x = T(y, z)(y = S(x, z))$.*

Proposition 1. *[11] Given a t-norm(t-conorm) $T(S)$ on L, it is equivalent:*

1. *$T(S)$ is smooth*
2. *$T(S)$ is divisible*

2.3 Discrete Fuzzy Numbers

By a fuzzy subset of $\mathbb{R}$, we mean a function $A : \mathbb{R} \to [0, 1]$. For each fuzzy subset A, let $A^\alpha = \{x \in \mathbb{R} : A(x) \geq \alpha\}$ for any $\alpha \in (0, 1]$ be its α-level set (or α-cut). By $supp(A)$, we mean the support of A, i.e. the set $\{x \in \mathbb{R} : A(x) > 0\}$. By A^0, we mean the closure of $supp(A)$.

Definition 5. *[13] A fuzzy subset A of $\mathbb{R}$ with membership mapping $A : \mathbb{R} \to [0,1]$ is called* discrete fuzzy number *if its support is finite, i.e., there exist real numbers $x_1, ..., x_n \in \mathbb{R}$ with $x_1 < x_2 < ... < x_n$ such that $supp(A) = \{x_1, ..., x_n\}$, and there are natural numbers s, t with $1 \leq s \leq t \leq n$ such that:*

1. *$A(x_i)$=1 for any natural number i with $s \leq i \leq t$ (*core*)*
2. *$A(x_i) \leq A(x_j)$ for each natural number i, j with $1 \leq i \leq j \leq s$*
3. *$A(x_i) \geq A(x_j)$ for each natural number i, j with $t \leq i \leq j \leq n$*

Remark 1. If the fuzzy subset A is a discrete fuzzy number then the support of A coincides with its closure, i.e. $supp(A) = A^0$.

From now on, we will denote the set of discrete fuzzy numbers by DFN and the abbreviation *dfn* will denote a discrete fuzzy number.

Theorem 1. *[14] (Representation of discrete fuzzy numbers) Let A be a discrete fuzzy number. Then the following statements (1)-(4) hold:*

1. *A^α is a nonempty finite subset of $\mathbb{R}$, for any $\alpha \in [0,1]$*
2. *$A^{\alpha_2} \subseteq A^{\alpha_1}$ for any $\alpha_1, \alpha_2 \in [0,1]$ with $0 \leq \alpha_1 \leq \alpha_2 \leq 1$*
3. *For any $\alpha_1, \alpha_2 \in [0,1]$ with $0 \leq \alpha_1 \leq \alpha_2 \leq 1$, if $x \in A^{\alpha_1} - A^{\alpha_2}$ we have $x < y$ for all $y \in A^{\alpha_2}$, or $x > y$ for all $y \in A^{\alpha_2}$*
4. *For any $\alpha_0 \in (0,1]$, there exist some real numbers $\alpha_0^{'}$ with $0 < \alpha_0' < \alpha_0$ such that $A^{\alpha_0'} = A^{\alpha_0}$ (i.e. $A^\alpha = A^{\alpha_0}$ for any $\alpha \in [\alpha_0', \alpha_0]$).*

Theorem 2. *[14] Conversely, if for any $\alpha \in [0,1]$, there exists $A^\alpha \subset \mathbb{R}$ satisfying analogous conditions to the (1)-(4)of Theorem 1, then there exists a unique $A \in DFN$ such that its α-cuts are exactly the sets A^α for any $\alpha \in [0,1]$.*

Throughout this article, if X and Y are subsets of real numbers
$\mathsf{X} \bigwedge \mathsf{Y} = \min(\mathsf{X}, \mathsf{Y}) = \{z = \min(x, y) \mid x \in \mathsf{X}, y \in \mathsf{Y}\}$ and
$\mathsf{X} \bigvee \mathsf{Y} = \max(\mathsf{X}, \mathsf{Y}) = \{z = \max(x, y) \mid x \in \mathsf{X}, y \in \mathsf{Y}\}$.

2.4 Maximum and Minimum of Discrete Fuzzy Numbers

Let A, B be two dfn and $A^\alpha = \{x_1^\alpha, \cdots, x_p^\alpha\}$, $B^\alpha = \{y_1^\alpha, \cdots, y_k^\alpha\}$ their α-cuts respectively.

In [4], for each $\alpha \in [0,1]$, we consider the following sets,

$$min_w(A,B)^\alpha = \{z \in supp(A) \bigwedge supp(B) | \min(x_1^\alpha, y_1^\alpha) \leq z \leq \min(x_p^\alpha, y_k^\alpha)\} \text{ and}$$

$$max_w(A,B)^\alpha = \{z \in supp(A) \bigvee supp(B) | \max(x_1^\alpha, y_1^\alpha) \leq z \leq \max(x_p^\alpha, y_k^\alpha)\}$$

where $supp(A) \bigwedge supp(B) = \{z = min(x,y) | x \in supp(A), y \in supp(B)\}$ and $supp(A) \bigvee supp(B) = \{z = max(x,y) | x \in supp(A), y \in supp(B)\}$.

Proposition 2. *[4] There exist two unique discrete fuzzy numbers, that we will denote by $min_w(A,B)$ and $max_w(A,B)$, such that they have the previous sets $min_w(A,B)^\alpha$ and $max_w(A,B)^\alpha$ as α-cuts respectively.*

The following result is not true, in general, for the set of discrete fuzzy numbers[5].

Proposition 3. *[5] The triplet ($\mathcal{A}_1$,min_w,max_w) is a distributive lattice, where $\mathcal{A}_1$ denotes the set of discrete fuzzy numbers whose support is a sequence of consecutive natural numbers.*

Remark 2. [5] Using these operations, we can define a partial order on $\mathcal{A}_1$ on the usual way:
$A \preceq B$ if and only if $min_w(A, B) = A$, or equivalently, $A \preceq B$ if and only if $max_w(A, B) = B$ for any $A, B \in \mathcal{A}_1$. Equivalently, we can also define the partial ordering in terms of α-cuts:
$A \preceq B$ if and only if $\min(A^\alpha, B^\alpha) = A^\alpha$
$A \preceq B$ if and only if $\max(A^\alpha, B^\alpha) = B^\alpha$

2.5 Discrete Fuzzy Numbers Obtained by Extending Discrete t-Norms(t-Conorms) Defined on a Finite Chain

Let us consider a discrete t-norm(t-conorm) $T(S)$ on the finite chain $L = \{0, 1, \cdots, m\} \subset \mathbb{N}$. Let $\mathcal{D}_L$ be the subset of the discrete fuzzy numbers $\mathcal{D}_L = \{A \in DFN$ such that $supp(A) \subseteq L\}$ and $A, B \in \mathcal{D}_L$. If X and Y are subsets of L, then the subset $\{T(x, y)|x \in \mathsf{X}, y \in \mathsf{Y}\} \subseteq L$ will be denoted by $T(\mathsf{X}, \mathsf{Y})$. Analogously, $S(\mathsf{X}, \mathsf{Y}) = \{S(x, y)|x \in \mathsf{X}, y \in \mathsf{Y}\}$.

So, if we consider the α-cut sets, $A^\alpha = \{x_1^\alpha, ..., x_p^\alpha\}$, $B^\alpha = \{y_1^\alpha, ..., y_k^\alpha\}$, for A and B respectively then $T(A^\alpha, B^\alpha) = \{T(x, y)|x \in A^\alpha, y \in B^\alpha\}$ and $S(A^\alpha, B^\alpha) = \{S(x, y)|x \in A^\alpha, y \in B^\alpha\}$ for each $\alpha \in [0, 1]$, where A^0 and B^0 denotes $supp(A)$ and $supp(B)$ respectively.

Definition 6. *[6]For each $\alpha \in [0, 1]$, let us consider the sets*

$$C^\alpha = \{z \in T(supp(A), supp(B))| \min T(A^\alpha, B^\alpha) \leq z \leq \max T(A^\alpha, B^\alpha)\}$$

$$D^\alpha = \{z \in S(supp(A), supp(B))| \min S(A^\alpha, B^\alpha) \leq z \leq \max S(A^\alpha, B^\alpha)\}$$

Remark 3. [6]From the monotonicity of the t-norm(t-conorm) $T(S)$,

$$C^\alpha = \{z \in T(supp(A), supp(B))|T(x_1^\alpha, y_1^\alpha) \leq z \leq T(x_p^\alpha, y_k^\alpha)\}$$

$$D^\alpha = \{z \in S(supp(A), supp(B))|S(x_1^\alpha, y_1^\alpha) \leq z \leq S(x_p^\alpha, y_k^\alpha)\}$$

For $\alpha = 0$ then $C^0 = T(supp(A), supp(B))$ and $D^0 = S(supp(A), supp(B))$.

Theorem 3. *[6] There exists a unique discrete fuzzy number that will be denoted by $\mathcal{T}(A, B)(\mathcal{S}(A, B))$ such that $\mathcal{T}(A, B)^\alpha = C^\alpha(\mathcal{S}(A, B)^\alpha = D^\alpha)$ for each $\alpha \in [0, 1]$ and $\mathcal{T}(A, B)(z) = sup\{\alpha \in [0, 1] : z \in C^\alpha\}(\mathcal{S}(A, B)(z) = sup\{\alpha \in [0, 1] : z \in D^\alpha\})$*

Remark 4. [6] From the previous theorem, if $T(S)$ is a discrete t-norm(t-conorm) on L, we see that it is possible to define a binary operation on $\mathcal{D}_L = \{A \in DFN|supp(A) \subseteq L\}$,

$$\begin{aligned}\mathcal{T}(\mathcal{S}) : \mathcal{D}_L \times \mathcal{D}_L &\longrightarrow \mathcal{D}_L \\ (A, B) &\longmapsto \mathcal{T}(A, B)(\mathcal{S}(A, B))\end{aligned}$$

that will be called *the extension of the t-norm T (t-conorm S) to* $\mathcal{D}_L$. Moreover, $\mathcal{T}$ and $\mathcal{S}$ are commutative and associative binary operations. Also, if we restrict these operations on the subset $\{A \in \mathcal{A}_1 \mid supp(A) \subseteq L = \{0, 1, \cdots, n\}\} \subseteq \mathcal{D}_L$ we showed that $\mathcal{T}$ and $\mathcal{S}$ are increasing operations as well.

3 Distributive Bounded Lattices on $\mathcal{A}_1$

According to proposition 3, we know that $\mathcal{A}_1$ constitutes a partially ordered set which is a lattice. Now, using this fact, we wish to see that the set $\mathcal{A}_1^L = \{A \in \mathcal{A}_1 \mid supp(A) \subseteq L = \{0, 1, \cdots, n\}\}$ is a bounded distributive lattice with the operations $\min_w$ and $\max_w$, considered in definition 2, as lattice operations.

Proposition 4. *If* $A, B \in \mathcal{A}_1^L$ *then* $\min_w(A, B)$ *and* $\max_w(A, B)$ *belong to* $\mathcal{A}_1^L$.

Proof. According to proposition 3, if $A, B \in \mathcal{A}_1^L \subset \mathcal{A}_1$ then the discrete fuzzy numbers $\max_w(A, B)$ and $\min_w(A, B) \in \mathcal{A}_1$. On the other hand, it is easy to see that the sets $supp(A) \bigwedge supp(B)$ and $supp(A) \bigvee supp(B)$ are subsets of L. So, $\min_w(A, B)^\alpha$ and $\max_w(A, B)^\alpha$ are subsets of L for each $\alpha \in [0, 1]$. Hence, the discrete fuzzy numbers $\min_w(A, B)$ and $\max_w(A, B)$ belong to the set $\mathcal{A}_1^L$. □

Theorem 4. *The triplet* $(\mathcal{A}_1^L, \min_w, \max_w)$ *is a bounded distributive lattice.*

Proof. The distributive lattice structure stems from propositions 2, 3 and 4. Moreover, it is straightforward to see that the natural number n, which is the maximum of the chain L, as a discrete fuzzy number (i.e. it is the discrete fuzzy number N such that it has only the natural number n as support) is the greatest element of the distributive lattice $\mathcal{A}_1^L$. Analogously, the natural number 0, which is the minimum of the chain L, as a discrete fuzzy number (i.e. it is the discrete fuzzy number O such that it has only the natural number 0 as a support) is the least element of the distributive lattice $\mathcal{A}_1^L$. □

4 t-Norms and t-Conorms on $(\mathcal{A}_1^L, \min_w, \max_w)$

In this section, we wish to consider if it is possible to obtain a t-norm(t-conorm) on some subset of $\mathcal{A}_1$ from a discrete triangular norm T(t-conorm S) on L.

Lemma 1. *Let us consider* T *and* S *being a divisible t-norm and a divisible t-conorm on* $L = \{0, 1, \cdots, n\}$ *respectively. If* X *and* Y *are subsets of consecutive natural numbers on* L *then* $T(\mathsf{X}, \mathsf{Y})$ *and* $S(\mathsf{X}, \mathsf{Y})$ *are subsets of consecutive natural numbers of* L *as well, where* $T(\mathsf{X}, \mathsf{Y}) = \{T(a, b),\ a \in \mathsf{X},\ b \in \mathsf{Y}\}$ *and* $S(\mathsf{X}, \mathsf{Y}) = \{S(a, b),\ a \in \mathsf{X},\ b \in \mathsf{Y}\}$.

Proof. We will only show the case of a divisible t-norm because the proof for a divisible t-conorm is analogous. From the monotonicity of T we can write the set $T(\mathsf{X}, \mathsf{Y})$ as

$$T(\mathsf{X},\mathsf{Y}) = \{z = T(a,b),\ a \in X,\ b \in Y \mid T(x,y) \leq z \leq T(x+j,y+k)\}$$

where $x, x+j$ denote the minimum and maximum value of X respectively, and $y, y+k$ denote the minimum and maximum value of Y respectively. It follows that if $T(x,y) = T(x+j,y+k)$ then the property holds because in this case $T(\mathsf{X},\mathsf{Y}) = \{T(x,y)\}$. Now, let us suppose that there exists $c \in L$ such that $c \notin T(\mathsf{X},\mathsf{Y})$ and $T(a,b) < c < T(a',b')$ with $a, a' \in \mathsf{X}$ and $b, b' \in \mathsf{Y}$. Let us consider

$$x_c = \max\{t \in \mathsf{X} \text{ such that } T(t,s) < c, s \in Y\}$$

$$y_c = \max\{s \in \mathsf{Y} \text{ such that } T(t,s) < c, t \in \mathsf{X}\}$$

$$x^c = \min\{t \in \mathsf{X} \text{ such that } T(t,s) > c, s \in \mathsf{Y}\}$$

$$y^c = \min\{s \in \mathsf{Y} \text{ such that } T(t,s) > c, t \in \mathsf{X}\}$$

It is obvious that $T(x,y) \leq T(a,b) \leq T(x_c,y_c) < c < T(x^c,y^c) \leq T(a',b') \leq T(x+j,y+k)$. Now, as T is a divisible t-norm on L we know [11] that T is a smooth t-norm and so, the next inequality $T(x_c+1,y_c) - T(x_c,y_c) \leq 1$ holds. So, from this inequality it is possible to consider two cases:

1. $T(x_c+1,y_c) = T(x_c,y_c)$
2. $T(x_c+1,y_c) = T(x_c,y_c)+1$

In the first case, since $x_c+1 > x_c$ and $T(x_c+1,y_c) = T(x_c,y_c) < c$ we obtain a contradiction with the election of the value of x_c.

In the second case, we have two options. Either $T(x_c+1,y_c) < c$, which is a contradiction with the election of the value x_c or $T(x_c+1,y_z) > c$, which is a contradiction with the existence of the value c, because $T(x_c+1,y_c)$ is the consecutive natural number of $T(x_c,y_c)$.

On the other hand, from the inequality $T(x^c,y^c) - T(x^c-1,y^c) \leq 1$ we obtain similar contradictions as before. □

Lemma 2. *Let us consider T and S being a divisible t-norm and a divisible t-conorm on $L = \{0,1,\cdots,n\}$ respectively. If A and B are discrete fuzzy numbers whose supports are sets of consecutive natural numbers of L then $T(A^\alpha,B^\alpha)$ and $S(A^\alpha,B^\alpha)$ are subsets of consecutive natural numbers of L as well, where $T(A^\alpha,B^\alpha) = \{T(a,b),\ a \in A^\alpha,\ b \in B^\alpha\}$ and $S(A^\alpha,B^\alpha) = \{S(a,b),\ a \in A^\alpha,\ b \in B^\alpha\}$ for all $\alpha \in [0,1]$.*

Proof. It is straightforward from lemma 1. □

Remark 5. Let us point out from the previous lemmas that if X,Y are sets of consecutive natural numbers the set $T(\mathsf{X},\mathsf{Y}) = \{T(x,y) \mid x \in \mathsf{X}\ y \in \mathsf{Y}\}$ can be expressed as $T(\mathsf{X},\mathsf{Y}) = \{z \in \mathbb{N} \mid T(x_1,y_1) \leq z \leq T(x_p,y_k)\}$ where x_1, x_p denote the minimum and the maximum of X respectively and y_1, y_k denote the minimum and the maximum of Y respectively.

Theorem 5. *Let $T(S)$ be a divisible t-norm(t-conorm) on L and let*

$$\begin{aligned}\mathcal{T}(\mathcal{S}) : \mathcal{A}_1^L \times \mathcal{A}_1^L &\to \mathcal{A}_1^L \\ (A, B) &\longmapsto \mathcal{T}(\mathcal{S})(A, B)\end{aligned}$$

be the extension of t-norm(t-conorm) $T(S)$ to $\mathcal{A}_1^L$, defined according to theorem 3. Then, $\mathcal{T}(\mathcal{S})$ is a t-norm(t-conorm) on the bounded set $\mathcal{A}_1^L$.

Proof. First of all, as T is a closed binary operation on L, since lemma 2, the binary operation $\mathcal{T}$ is closed on $\mathcal{A}_1^L$. Moreover, from remark 4 we know that $\mathcal{T}$ is an associative, a commutative and an increasing binary operation. Finally, let us consider the natural number n, which is the maximum of the chain L like a discrete fuzzy number, (i.e., it is the discrete fuzzy number N such that it has only the natural number n as support). Then, it is easy to see that $\mathcal{T}(A, N) = A$, because of $\mathcal{T}(A, N)^\alpha =$

$$= \{z \in T(suppA, n) | T(x_1^\alpha, n) \leq z \leq T(x_p^\alpha, n)\} =$$

$$= \{z \in suppA | x_1^\alpha \leq z \leq x_p^\alpha\} = A^\alpha \text{ for all } \alpha \in [0, 1]. \qquad \square$$

Now, we will see that it is possible to construct a t-norm(t-conorm) on $\mathcal{A}_1^L$ from two divisible t-norms(t-conorms) $T_1, T_2(S_1, S_2)$ defined on L such that $T_1(x, y) \leq T_2(x, y)(S_1(x, y) \leq S_2(x, y))$ for all $(x, y) \in L \times L$.

Proposition 5. *Let $T_1, T_2(S_1, S_2)$ be two discrete t-norms(t-conorms) on L such that $T_1 \leq T_2(S_1 \leq S_2)$, and $A, B \in \mathcal{A}_1^L$. For each $\alpha \in [0, 1]$, let us consider the sets $\mathcal{T}_{T_1,T_2}(A, B)^\alpha = \{z \in L | \min(T_1(A^\alpha, B^\alpha)) \leq z \leq \max(T_2(A^\alpha, B^\alpha))\}$, and $\mathcal{S}_{S_1,S_2}(A, B)^\alpha = \{z \in L | \min(S_1(A^\alpha, B^\alpha)) \leq z \leq \max(S_2(A^\alpha, B^\alpha))\}$. Then, there exist two discrete fuzzy numbers, that will be denoted by $\mathcal{T}_{T_1,T_2}(A, B)$ and $\mathcal{S}_{S_1,S_2}(A, B)$ respectively such that they have the sets $\mathcal{T}_{T_1,T_2}(A, B)^\alpha$ and $\mathcal{S}_{S_1,S_2}(A, B)$ as α-cuts.*

Proof. We wish to show that the subsets $\mathcal{T}_{T_1,T_2}(A, B)^\alpha$ satisfy the conditions 1-4 of theorem 1 of representation of discrete fuzzy numbers because afterwards we can apply theorem 2 and the proposition turns out.

1. $\mathcal{T}_{T_1,T_2}(A, B)^\alpha$ is a nonempty finite set, because A^α and B^α are both nonempty finite sets (the discrete fuzzy numbers are normal fuzzy subsets) and according to lemma 2, $T_1(A^\alpha, B^\alpha)$ and $T_2(A^\alpha, B^\alpha)$ are finite sets of consecutive natural numbers.
2. $\mathcal{T}_{T_1,T_2}(A, B)^\beta \subseteq \mathcal{T}_{T_1,T_2}(A, B)^\alpha$ for any $\alpha, \beta \in [0, 1]$ with $0 \leq \alpha \leq \beta \leq 1$, because if $A, B \in \mathcal{A}_1^L$ and

$$A^\alpha = \{x_1^\alpha, ..., x_p^\alpha\}, A^\beta = \{x_1^\beta, ..., x_r^\beta\},$$
$$B^\alpha = \{y_1^\alpha, ..., y_k^\alpha\}, B^\beta = \{y_1^\beta, ..., y_l^\beta\},$$

then

$$A^\beta \subseteq A^\alpha \text{ implies } x_1^\alpha \leq x_1^\beta \text{ and } x_r^\beta \leq x_p^\alpha \tag{1}$$
$$B^\beta \subseteq B^\alpha \text{ implies } y_1^\alpha \leq y_1^\beta \text{ and } y_l^\beta \leq y_k^\alpha \tag{2}$$

Then, from the monotonicity of the t-norms T_1, T_2, and according to hypothesis $T_1 \leq T_2$ and finally by the relations (11) and (12) we obtain:

$$T_1(x_1^\alpha, y_1^\alpha) \leq T_1(x_1^\beta, y_1^\beta) \leq T_2(x_1^\beta, y_1^\beta) \leq T_2(x_r^\beta, y_l^\beta) \leq T_2(x_p^\alpha, y_k^\alpha)$$

Hence, combining the previous conditions the result holds.

3. If $x \in \mathcal{T}_{T_1,T_2}(A,B)^\alpha$ hence $x \in L$ and x does not belong to $\mathcal{T}_{T_1,T_2}(A,B)^\beta$, then either $x < T_1(x_1^\beta, y_1^\beta)$, which is the minimum of $\mathcal{T}_{T_1,T_2}(A,B)^\beta$, or $x > T_2(x_r^\beta, y_l^\beta)$, which is the maximum of $\mathcal{T}_{T_1,T_2}(A,B)^\beta$.
4. As $A, B \in \mathcal{A}_1^L$, from Theorem 1 of representation of discrete fuzzy numbers, for each $\alpha \in (0,1]$ there exist real numbers α_1' and α_2' with $0 < \alpha_1' < \alpha$ and $0 < \alpha_2' < \alpha$ such that for each $r \in [\alpha_1', \alpha]$, $A^\alpha = A^r$ and moreover $B^\alpha = B^r$, for each $r \in [\alpha_2', \alpha]$. Thus, if $\alpha' = \alpha_1' \vee \alpha_2'$, we can obtain:

$$\min(A^r) = \min(A^\alpha) \text{ and } \max(A^r) = \max(A^\alpha)$$
$$\min(B^r) = \min(B^\alpha) \text{ and } \max(B^r) = \max(B^\alpha)$$

for each $r \in [\alpha', \alpha]$. Therefore

$$T_1(\min(A^r), \min(B^r)) = T_1(\min(A^\alpha), \min(B^\alpha))$$
$$T_2(\max(A^r), \max(B^r)) = T_2(\max(A^\alpha), \max(B^\alpha))$$

Hence,

$$\mathcal{T}_{T_1,T_2}(A,B)^\alpha =$$
$$\{z \in L | T_1(min(A^\alpha), min(B^\alpha)) \leq z \leq T_2(max(A^\alpha), max(B^\alpha))\} =$$
$$\{z \in L | T_1(min(A^r), min(B^r)) \leq z \leq T_2(max(A^r), max(B^r))\} =$$
$$\mathcal{T}_{T_1,T_2}(A,B)^r \text{ for each } r \in [\alpha', \alpha]$$

□

Proposition 6. *The discrete fuzzy numbers $\mathcal{T}_{T_1,T_2}(A,B)$ and $\mathcal{S}_{S_1,S_2}(A,B)$ belong to $\mathcal{A}_1^L$ for all $A, B \in \mathcal{A}_1^L$.*

Proof. According to proposition 5 we know that $\mathcal{T}_{T_1,T_2}(A,B)$ and $\mathcal{S}_{S_1,S_2}(A,B)$ are discrete fuzzy numbers. Moreover, it is obvious that their α-cuts $\mathcal{T}_{T_1,T_2}(A,B)^\alpha$ and $\mathcal{S}_{S_1,S_2}(A,B)^\alpha$ are subsets of consecutive natural numbers for all $\alpha \in [0,1]$. Hence, $\mathcal{T}_{T_1,T_2}(A,B), \mathcal{S}_{S_1,S_2}(A,B) \in \mathcal{A}_1^L$. □

Theorem 6. *Let $T_1 \leq T_2 (S_1 \leq S_2)$ be divisible t-norms(t-conorms) on L and let*

$$\mathcal{T}_{T_1,T_2}(\mathcal{S}_{S_1,S_2}) : \mathcal{A}_1^L \times \mathcal{A}_1^L \longmapsto \mathcal{A}_1^L$$
$$(A,B) \to \mathcal{T}_{T_1,T_2}(\mathcal{S}_{S_1,S_2})(A,B)$$

be a binary operation, where $\mathcal{T}_{T_1,T_2}(\mathcal{S}_{S_1,S_2})(A,B)$ are the discrete fuzzy numbers considered in the previous proposition 5. Then, $\mathcal{T}_{T_1,T_2}(\mathcal{S}_{S_1,S_2})$ is a t-norm(t-conorm) on the bounded set $\mathcal{A}_1^L$.

Proof. The commutativity and associativity properties stem from the commutativity and associativity properties of the discrete t-norms T_1 and T_2. Now, we wish

to show that $\mathcal{T}_{T_1,T_2}$ is an increasing mapping, i.e. for each $A, B, C \in \mathcal{A}_1^L$ such that $B \preceq C$ (where $\preceq$ is the partial order defined in remark 2) then $\mathcal{T}_{T_1,T_2}(A,B) \preceq \mathcal{T}_{T_1,T_2}(A,C)$, i.e. $\min_w(\mathcal{T}_{T_1,T_2}(A,B), \mathcal{T}_{T_1,T_2}(A,C)) = \mathcal{T}_{T_1,T_2}(A,B)$. This last relation is equivalent to see that $min_w(\mathcal{T}_{T_1,T_2}(A,B), \mathcal{T}_{T_1,T_2}(A,C))^\alpha = \mathcal{T}_{T_1,T_2}(A,B)^\alpha$ for all $\alpha \in [0,1]$ where $\mathcal{T}_{T_1,T_2}(A,B)^\alpha = \{z \in L | T_1(x_1^\alpha, y_1^\alpha) \leq z \leq T_2(x_p^\alpha, y_k^\alpha)\}$ and $\mathcal{T}_{T_1,T_2}(A,C)^\alpha = \{z \in L | T_1(x_1^\alpha, w_1^\alpha) \leq z \leq T_2(x_p^\alpha, w_l^\alpha)\}$ with $A^\alpha = \{x_1^\alpha, ..., x_p^\alpha\}$, $B^\alpha = \{y_1^\alpha, ..., y_k^\alpha\}$, $C^\alpha = \{w_1^\alpha, ..., w_l^\alpha\}$ for A, B and C respectively.

Since remark 2, we know that the fact $B \preceq C$, implies that $\min(B^\alpha, C^\alpha) = B^\alpha$ and $y_1^\alpha \leq w_1^\alpha$, $y_k^\alpha \leq w_l^\alpha$ for all $\alpha \in [0,1]$. According to these last relations and from the monotonicity of T_1, T_2 and the condition $T_1 \leq T_2$ we obtain that

$T_1(x_1^\alpha, y_1^\alpha) \leq T_1(x_1^\alpha, w_1^\alpha)$ and $T_2(x_p^\alpha, y_k^\alpha) \leq T_2(x_p^\alpha, w_l^\alpha)$.

Therefore, $\min_w(\mathcal{T}_{T_1,T_2}(A,B), \mathcal{T}_{T_1,T_2}(A,C))^\alpha =$

$$= \{z \in supp(\mathcal{T}_{T_1,T_2}(A,B)) \bigwedge supp(\mathcal{T}_{T_1,T_2}(A,C)) |$$

$$\min(T_1(x_1^\alpha, y_1^\alpha), T_1(x_1^\alpha, w_1^\alpha)) \leq z \leq \min(T_2(x_p^\alpha, y_k^\alpha), T_2(x_p^\alpha, w_l^\alpha))\} =$$

$$= \{z \in supp(\mathcal{T}_{T_1,T_2}(A,B)) \bigwedge supp(\mathcal{T}_{T_1,T_2}(A,C)) | T_1(x_1^\alpha, y_1^\alpha) \leq z \leq T_2(x_p^\alpha, y_k^\alpha)\} =$$

(As $supp(\mathcal{T}_{T_1,T_2}(A,B))$ and $supp(\mathcal{T}_{T_1,T_2}(A,C))$ are subsets of consecutive natural numbers)

$$= \{z \in supp(\mathcal{T}_{T_1,T_2}(A,B)) | T_1(x_1^\alpha, y_1^\alpha) \leq z \leq T_2(x_p^\alpha, y_k^\alpha)\} = \mathcal{T}_{T_1,T_2}(A,B)^\alpha.$$

Finally, let us consider the natural number n, which is the maximum of the chain L like a discrete fuzzy number, (i.e., it is the discrete fuzzy number N such that it has only the natural number n as support). Then, it is easy to see that $\mathcal{T}_{T_1,T_2}(A,N) = A$, because of $\mathcal{T}_{T_1,T_2}(A,N)^\alpha = \{z \in L | T_1(x_1^\alpha, n) \leq z \leq T_2(x_p^\alpha, n)\} = \{z \in L | x_1^\alpha \leq z \leq x_p^\alpha\} = A^\alpha$ for all $\alpha \in [0,1]$. □

4.1 Comparison of t-Norms and t-Conorms on $\mathcal{A}_1^L$

It is well known [11] that the comparison of t-norms(t-conorms) defined on a finite chain L is done in the usual way, i.e., pointwise. So,

Proposition 7. *Let us consider $T_1(S_1)$ and $T_2(S_2)$ two divisible t-norms (t-conorms) on the finite chain $L = \{0, 1, \cdots, n\}$ verifying $T_1(S_1)(x,y) \leq T_2(S_2)(x,y)$ for all $(x,y) \in L \times L$. Then, for all $A, B \in \mathcal{A}_1^L$ the inequality $\mathcal{T}_1(\mathcal{S}_1)(A,B) \preceq \mathcal{T}_2(\mathcal{S}_2)(A,B)$ holds.*

Proof. From remark 2, it is enough to see that

$$\min(\mathcal{T}_1(A,B)^\alpha, \mathcal{T}_2(A,B)^\alpha) = \mathcal{T}_1(A,B)^\alpha$$

for all $\alpha \in [0,1]$, where $(\mathcal{T}_1(A,B))^0$ and $(\mathcal{T}_2(A,B))^0$ denote the supports of $\mathcal{T}_1(A,B)$ and $\mathcal{T}_2(A,B)$ respectively. For this reason, let us consider the α-cuts $A^\alpha = \{x_1^\alpha, \cdots, x_p^\alpha\}$ and $B^\alpha = \{y_1^\alpha, \cdots, y_k^\alpha\}$ for A, B respectively. So, from

lemma 2 $\mathcal{T}_1(A,B)^\alpha = \{z \in \mathbb{N} \mid T_1(x_1^\alpha, y_1^\alpha) \leq z \leq T_1(x_p^\alpha, y_k^\alpha)\}$ and $\mathcal{T}_2(A,B)^\alpha = \{z \in \mathbb{N} \mid T_2(x_1^\alpha, y_1^\alpha) \leq z \leq T_2(x_p^\alpha, y_k^\alpha)\}$ are sets of consecutive natural numbers because T_1 and T_2 are divisible t-norms on L and $A, B \in \mathcal{A}_1^L$, then

$$\min(\mathcal{T}_1(A,B)^\alpha, \mathcal{T}_2(A,B)^\alpha) =$$

$$\{z = \min(x,y) \mid x \in \mathcal{T}_1(A,B)^\alpha, y \in \mathcal{T}_2(A,B)^\alpha\} =$$

$\{z \in \mathbb{N} \mid \min(T_1(x_1^\alpha, y_1^\alpha), T_2(x_1^\alpha, y_1^\alpha)) \leq z \leq \min(T_1(x_p^\alpha, y_k^\alpha), T_2(x_p^\alpha, y_k^\alpha))\} =$ (As $T_1(x,y) \leq T_2(x,y)$ for all $x, y \in L$)$=\{z \in \mathbb{N} \mid T_1(x_1^\alpha, y_1^\alpha) \leq z \leq T_1(x_p^\alpha, y_k^\alpha)\}=$

$$\mathcal{T}_1(A,B)^\alpha. \qquad \square$$

Acknowledgments. This work has been partially supported by the MTM2009-10962 project grant.

References

1. De Baets, B., Mesiar, R.: Triangular norms on product lattices. Fuzzy Sets and Systems 104, 61–75 (1999)
2. Casasnovas, J., Riera, J.V.: On the addition of discrete fuzzy numbers. WSEAS Transactions on Mathematics, 549–554 (2006)
3. Casasnovas, J., Riera, J.V.: Discrete fuzzy numbers defined on a subset of natural numbers. In: Theoretical Advances and Applications of Fuzzy Logic and Soft Computing: Advances in Soft Computing, vol. 42, pp. 573–582. Springer, Heidelberg (2007)
4. Casasnovas, J., Riera, J.V.: Maximum and minimum of discrete fuzzy numbers. In: Frontiers in Artificial Intelligence and Applications: artificial intelligence research and development, vol. 163, pp. 273–280. IOS Press, Amsterdam (2007)
5. Casasnovas, J., Riera, J.V.: Lattice properties of discrete fuzzy numbers under extended min and max. In: Proceedings IFSA-EUSFLAT, Lisbon, pp. 647–652 (2009)
6. Casasnovas, J., Riera, J.V.: Extension of discrete t-norms and t-conorms to discrete fuzzy numbers. In: Fifth international summer school on aggregation operators (AGOP 2009), Palma de Mallorca, pp. 77–82 (2009)
7. Deschrijver, G., Cornelis, C., Kerre, E.E.: On the representation of intuitionistic fuzzy t-norms and t-conorms. IEEE Trans. Fuzzy Systems 12, 45–61 (2004)
8. Goguen, J.: L-fuzzy sets. J. Math. Anal. Appl. 18, 145–174 (1967)
9. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms. Kluwer, Dordrecht (2000)
10. Klir, G., Bo, Y.: Fuzzy sets and fuzzy logic (Theory and applications). Prentice-Hall, Englewood Cliffs (1995)
11. Mayor, G., Torrens, J.: Triangular norms on discrete settings. In: Klement, E.P., Mesiar, R. (eds.) Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms, pp. 189–230. Elsevier, Amsterdam (2005)
12. Schweizer, B., Sklar, A.: Associative functions and statistical triangle inequalities. Publ. Math. Debrecen 8, 169–186 (1961)
13. Voxman, W.: Canonical representations of discrete fuzzy numbers. Fuzzy Sets and Systems 54, 457–466 (2001)
14. Wang, G., Wu, C., Zhao, C.: Representation and Operations of discrete fuzzy numbers. Southeast Asian Bulletin of Mathematics 28, 1003–1010 (2005)
15. Zhang, D.: Triangular norms on partially ordered sets. Fuzzy Sets and Systems 153, 195–209 (2005)

Arity-Monotonic Extended Aggregation Operators

Marek Gągolewski[1,2,⋆] and Przemysław Grzegorzewski[1,2]

[1] Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
[2] Faculty of Mathematics and Information Science, Warsaw University of Technology
pl. Politechniki 1, 00-661 Warsaw, Poland
gagolews@ibspan.waw.pl, pgrzeg@ibspan.waw.pl

Abstract. A class of extended aggregation operators, called impact functions, is proposed and their basic properties are examined. Some important classes of functions like generalized ordered weighted averaging (OWA) and ordered weighted maximum (OWMax) operators are considered. The general idea is illustrated by the Producer Assessment Problem which includes the scientometric problem of rating scientists basing on the number of citations received by their publications. An interesting characterization of the well known h-index is given.

Keywords: aggregation, extended aggregation function, OWA, OWMax, h-index, scientometrics.

1 Introduction

Aggregation plays a central role in many areas of the human activity, including not only statistics, engineering, computer science or physics but also decision making, economy and social sciences. It appears always when the reasoning requires merging several values into a single one which may represent a kind of synthesis for all individual inputs. Such functions projecting multidimensional numerical space of input values into one dimension are generally called *aggregation operators*.

Apart from particular applications the theory of aggregation operators is a rapidly developing mathematical domain (we refer the reader to [6] for the recent state of art monograph).

Classically, aggregation operators are usually considered for a fixed number of arguments. For some applications it may be to restrictive. We face such a situation in the so-called Producer Assessment Problem where given alternatives are rated not only with respect to the quality of delivered items but also to their productivity. As a typical example we may indicate the problem of rating scientists by their publications' citations.

This is the reason that the aggregation operators defined for arbitrary number of arguments are of interest. In the paper we propose the axiomatic approach to such a class of extended aggregation operators, called impact functions, and discuss some interesting properties of such functions for different arities. We also study the properties of the

⋆ Corresponding author.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 693–702, 2010.

generalized versions of some well known classes of aggregation operators like generalized ordered weighted averaging (OWA) and ordered weighted maximum (OWMax) operators. It is worth noting that well-known Hirsch h-index turns out to be a particular example of the latter family.

2 Preliminaries

2.1 Basic Notation

We adopt the notational convention from the recent monograph [6].

Let $\mathbb{I} = [a, b]$ denote any nonempty closed interval of extended real numbers $\bar{\mathbb{R}} = [-\infty, \infty]$. In this paper we assume that $0 \cdot \infty = 0$. Unless stated otherwise, $n, m \in \mathbb{N}$.

Let $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ denote the set of all nonnegative integers. Moreover, let $[n] := \{1, 2, \dots, n\}$.

The set of all vectors of arbitrary length with elements in $\mathbb{I}$, i.e. $\bigcup_{n=1}^{\infty} \mathbb{I}^n$, will be denoted by $\mathbb{I}^{1,2,\cdots}$.

Given any $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{I}^n$, we write $\mathbf{x} \leq \mathbf{y}$ iff $(\forall i \in [n])$ $x_i \leq y_i$. Moreover, $(n * x)$ is equivalent to $(x, x, \dots, x) \in \mathbb{I}^n$.

Let $x_{(i)}$ denote the ith-smallest value of $\mathbf{x} = (x_1, \dots, x_n)$. For simplicity of notation, we assume that $x_{(n+j)} := x_{(n)}$ for $j = 1, 2, \dots$.

For any $\mathbf{x} \in \mathbb{I}^n$ and $\mathbf{y} \in \mathbb{I}^m$ and any function f defined on $\mathbb{I}^{n+m}$ the notation $\mathsf{f}(\mathbf{x}, \mathbf{y})$ stands for $\mathsf{f}(x_1, \dots, x_n, y_1, \dots, y_m)$.

If $\mathsf{f} : X \to Y$ and $X' \subset X$ then a function $\mathsf{f}|_{X'} : X' \to Y$ such that $(\forall x \in X')$ $\mathsf{f}|_{X'}(x) = \mathsf{f}(x)$ is called a *restriction* of f to X'. Furthermore, if $\mathcal{F}$ is a family of functions mapping X to Y, then $\mathcal{F}|_{X'} := \{\mathsf{f}|_{X'} : \mathsf{f} \in \mathcal{F}\}$.

2.2 Aggregation Functions

Let us recall the notion of the aggregation function, which is often considered in the literature. Note that it is a particular sublass of the very broad family of aggregation operators. Here is a slightly modified version of the definition given in [6].

Definition 1. *An **aggregation function** in $\mathbb{I}^n = [a, b]^n$ is any function $\mathsf{a}^{(n)} : \mathbb{I}^n \to \bar{\mathbb{R}}$ which*

(nd) *is nondecreasing in each variable, i.e.* $(\forall \mathbf{x}, \mathbf{y} \in \mathbb{I}^n)$ $\mathbf{x} \leq \mathbf{y} \Rightarrow \mathsf{a}^{(n)}(\mathbf{x}) \leq \mathsf{a}^{(n)}(\mathbf{y})$,

(bl) *fulfills the lower boundary condition:* $\inf_{\mathbf{x} \in \mathbb{I}^n} \mathsf{a}^{(n)}(\mathbf{x}) = a$,

(bu) *fulfills the upper boundary condition:* $\sup_{\mathbf{x} \in \mathbb{I}^n} \mathsf{a}^{(n)}(\mathbf{x}) = b$.

Typical examples of aggregation functions are: sample minimum, maximum, arithmetic mean, median, and OWA operators. On the other hand, generally sample size, sum and constant function are not aggregation functions in the above sense.

It is worth noticing that axioms (nd) and (bl) imply $\mathsf{a}^{(n)}(n * a) = a$. We also have $\mathsf{a}^{(n)}(n * b) = b$ by (nd) and (bu).

From now on, let $\mathcal{E}(\mathbb{I})$ be the set of all functions $\mathsf{F} : \mathbb{I}^{1,2,\cdots} \to \bar{\mathbb{R}}$.

Now let us extend the class of aggregation functions to any number of arguments. Our definition agrees with the one given in [6]. Note that quite a different extension was proposed by Mayor and Calvo in [11].

Definition 2. *An **extended aggregation function** in $\mathbb{I}^{1,2,\cdots}$ is a function $\mathsf{A} \in \mathcal{E}(\mathbb{I})$, whose restriction $\mathsf{a}^{(n)} := \mathsf{A}|_{\mathbb{I}^n}$ to $\mathbb{I}^n$ for any $n \in \mathbb{N}$ is an aggregation function in $\mathbb{I}^n$.*

Note that any extended aggregation function may be regarded as a sequence $(\mathsf{a}^{(n)})_{n\in\mathbb{N}}$ of aggregation functions.

The set of all extended aggregation functions in $\mathbb{I}^{1,2,\cdots}$ will be denoted $\mathcal{E}_{\mathcal{A}}(\mathbb{I})$.

3 The Producer Assessment Problem and Impact Functions

Consider a **producer** (e.g. a writer, scientist, artist, craftsman) and a nonempty set of his **products** (e.g. books, papers, works, goods). Suppose that each product is given a rating (a single number in $\mathbb{I} = [a, b]$), where a denotes the lowest admissible rating. Here are some typical examples:

Producers	Products	Rates	Discipline
Scientists	Scientific articles	Number of citations	Scientometrics (see e.g. [7])
Scientific institutes	Scientists	The h-index	Scientometrics
Web servers	Web pages	Number of in-links	Webometrics
Artists	Paintings	Auction price	Auctions

Each possible state of a producer can be described by a point in $\mathbb{I}^{1,2,\cdots}$. The **Producer Assessment Problem** (or PAP for short) involves constructing and analyzing functions which can be used to rate producers. A family of such functions should take into account two following aspects of producer's quality:

1. the ability to make highly-rated products,
2. overall productivity.

Clearly, the first component can be described well by a very broad class of (extended) aggregation functions. However, in practice we are also interested in comparing entities with different productivity. Therefore, we need some sine qua non conditions for such assessing functions.

Definition 3. *An **impact function** in $\mathbb{I}^{1,2,\cdots}$ is a function $\mathsf{J} \in \mathcal{E}(\mathbb{I}), \mathbb{I} = [a, b]$ which is*

(nd) *nondecreasing in each variable:* $(\forall n)(\forall \mathbf{x}, \mathbf{y} \in \mathbb{I}^n)\ \mathbf{x} \le \mathbf{y} \Rightarrow \mathsf{J}(\mathbf{x}) \le \mathsf{J}(\mathbf{y})$,
(am) *arity-monotonic, i.e.* $(\forall n, m)(\forall \mathbf{x} \in \mathbb{I}^n)(\forall \mathbf{y} \in \mathbb{I}^m)\ \mathsf{J}(\mathbf{x}) \le \mathsf{J}(\mathbf{x}, \mathbf{y})$,
(sy) *symmetric, i.e.* $(\forall n)(\forall \mathbf{x} \in \mathbb{I}^n)\ (\forall \sigma \in \mathfrak{S}_{[n]})\ \mathsf{J}(x_1, \ldots, x_n) = \mathsf{J}(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$, *where $\mathfrak{S}_{[n]}$ denotes the set of all permutations of $[n]$*,
(bl) *fulfills the lower boundary condition:* $\inf_{\mathbf{x}\in\mathbb{I}^{1,2,\cdots}} \mathsf{J}(\mathbf{x}) = a$,
(wbu) *fulfills the weak upper boundary condition:* $\sup_{\mathbf{x}\in\mathbb{I}^{1,2,\cdots}} \mathsf{J}(\mathbf{x}) \le b$.

The set of all impact functions will be denoted by $\mathcal{E}_{\mathcal{I}}(\mathbb{I})$. Note that the set of requirements given in Def. 3 is a generalized version of the axiomatization proposed by Woeginger [15,16] for the scientometric impact indices (for other axiomatizations of the so-called bibliometric impact indices see [8,9,12]).

The property (am) expresses the intuition that in many applications of the PAP, the increase in production should not decrease the overall ranking. However, sometimes it be may be viewed as a weak point, because many aggregation operators are excluded.

Please note that impact functions are not necessarily aggregation functions (in the sense of Def. 1), because axiom (bu) is replaced by its weaker form (wbu). It is so because the upper boundary condition together with (am) seems too tight (if some $\mathsf{J} \in \mathcal{E}_{\mathcal{I}}(\mathbb{I})$ fulfills (bu), then — by $(\mathrm{bu}), (\mathrm{nd}), (\mathrm{am})$ and (sy) — $(\forall \mathbf{x} \in \mathbb{I}^n)(\exists i \in [n])$ $x_i = b$ implies $\mathsf{J}(x_1, \ldots, x_n) = b$).

4 Axiomatic Approach

4.1 Axiomatic Modeling

Below we discuss a set of properties which may be used to describe behavioral aspects of classes of impact functions. Formally, a **property** P of functions in $\mathcal{E}(\mathbb{I})$ is just a subset of $\mathcal{E}(\mathbb{I})$. We denote by $P_{(\mathrm{nd})}, P_{(\mathrm{am})}, \ldots$ the properties that appear in Def. 1 and Def. 3, i.e. some families of functions satisfying axioms $(\mathrm{nd}), (\mathrm{am}), \ldots$.

The concept of axiomatic modeling in decision making dates back as far as the works of Arrow [1] (impossibility theorem in the problem of social states ordering) and May [10] (group decision functions).

One approach considers a **characterization** of functions, i.e. a finite set of properties $P_1, \ldots, P_k \subseteq \mathcal{E}(\mathbb{I})$ such that $(!\exists f)\ f \in \bigcap_{i=1}^{k} P_i$. Moreover, that set should be minimal, i.e. $(\forall j \in [k])\ |\bigcap_{i=1, i\neq j}^{k} P_i| > 1$.

In the other approach a family of properties $\mathcal{P}$ that *seem* to be *sensible* from the *practical* point of view is of interest. Unfortunately quite often some of the properties $P_1, \ldots, P_l \in \mathcal{P}$ are contradictory, i.e. $P_1 \cap \cdots \cap P_l = \emptyset$. Therefore, in such a case there is no perfect (aggregation/impact) function that fulfill all imaginable properties.

4.2 Arity-Free Property

Generally, any two restrictions $\mathsf{A}|_{\mathbb{I}^n}$ and $\mathsf{A}|_{\mathbb{I}^m}$ of the extended aggregation function A, where $n \neq m$, are not necessarily related. However, here we are especially interested in properties which concern relations between restrictions of aggregation operators for different arities. Below we propose a formalism that concerns the above mentioned ideas.

Definition 4. *An **arity-free property** is any $P \subseteq \mathcal{E}(\mathbb{I})$ such that*

$$\left\{ \mathsf{F}|_{\mathbb{I}^m} : \mathsf{F} \in P,\ \mathsf{F}|_{\mathbb{I}^n} = \mathsf{f}^{(n)} \right\} = P|_{\mathbb{I}^m}$$

holds $(\forall n \neq m)$ $(\forall \mathsf{f}^{(n)} \in P|_{\mathbb{I}^n})$.

A family of all arity-free properties will be denoted by $\mathcal{P}_{\mathrm{af}}$. The other properties are called **arity-dependent**. Please notice that depending on the context we implicitly assume some fixed $\mathbb{I}$.

It can be seen easily that four of the axioms in Defs. 1 and 3 can be treated as arity-free properties, i.e. $P_{(\mathrm{nd})}, P_{(\mathrm{bl})}, P_{(\mathrm{bu})}, P_{(\mathrm{sy})} \in \mathcal{P}_{\mathrm{af}}$. However, in general $P_{(\mathrm{am})} \notin \mathcal{P}_{\mathrm{af}}$.

4.3 Ordering Property

Each function in $\mathcal{E}(\mathbb{I})$ implies a **ranking**, i.e. a linear (total) ordering relation in a set of producers' states. If valuation is expressed by a *single* numerical value such result can further be considered as a point-of-reference (e.g. for the author self-improving process). Therefore, it would be interesting to distinguish a class of properties that concern only the relation between the function values regardless of the particular values assumed by these functions.

Definition 5. *An **ordering property** is any $P \subseteq \mathcal{E}(\mathbb{I})$ such that the following condition*

$$\mathsf{F} \in P \Longrightarrow \mathsf{g} \circ \mathsf{F} \in P$$

holds for any nondecreasing function $\mathsf{g} : \bar{\mathbb{R}} \to \bar{\mathbb{R}}$, where $\circ$ marks function composition, i.e. $(\mathsf{g} \circ \mathsf{F})(x) = \mathsf{g}(\mathsf{F}(x))$.

A family of all ordering properties will be denoted by $\mathcal{P}_{\mathrm{ord}}$. Note that $P_{(\mathrm{nd})}, P_{(\mathrm{am})}, P_{(\mathrm{sy})} \in \mathcal{P}_{\mathrm{ord}}$, but generally $P_{(\mathrm{bl})}$ and $P_{(\mathrm{bu})} \notin \mathcal{P}_{\mathrm{ord}}$.

Even though being obvious the following proposition is worth of explicit stating.

Proposition 1. *Any impact function cannot be defined by means of ordering properties only.*

Proof. Assume conversely that F is a unique function such that $\mathsf{F} \in P_1 \cap P_2 \cap \ldots$, for some (possibly finite) sequence $P_1, P_2, \cdots \subseteq \mathcal{P}_{\mathrm{ord}}$. Take any nondecreasing $\mathsf{g} : \bar{\mathbb{I}} \to \bar{\mathbb{R}}$ and let $\mathsf{F}' := \mathsf{g} \circ \mathsf{F}$. We have $\mathsf{F} \in P_1$ and $\mathsf{F}' \in P_1$. For any $i = 2, 3, \ldots$ we have either $\mathsf{F}' \in P_1 \cap \cdots \cap P_i$ or $P_1 \cap \cdots \cap P_i = \emptyset$, which contradicts our assumption. □

Proposition 1 can also be formulated as follows: Any intersection of ordering properties is also an ordering property.

4.4 Further Results

The following lemma allows to check efficiently whether a non-decreasing function is arity-monotonic.

Lemma 1. *For any $\mathsf{F} \in P_{(\mathrm{nd})}$ we have*

$$\mathsf{F} \in P_{(\mathrm{am})} \iff (\forall \mathbf{x} \in \mathbb{I}^{1,2,\cdots})\ \mathsf{F}(\mathbf{x}) \leq \mathsf{F}(\mathbf{x}, \min \mathbb{I}).$$

Proof. $(\Rightarrow)$ Trivial.

$(\Leftarrow)$ Fix $\mathbf{x} \in \mathbb{I}^n$ and $\mathbf{y} \in \mathbb{I}^m$ for some n and m. We have $\mathsf{F}(\mathbf{x}) \leq \mathsf{F}(\mathbf{x}, \min \mathbb{I}) \leq \mathsf{F}(\mathbf{x}, 2 * \min \mathbb{I}) \leq \cdots \leq \mathsf{F}(\mathbf{x}, m * \min \mathbb{I}) \leq \mathsf{F}(\mathbf{x}, \mathbf{y})$ by (nd), since $(\mathbf{x}, m * \min \mathbb{I}) \leq (\mathbf{x}, \mathbf{y})$. □

In addition to axiom (am) some other reasonable arity-dependent conditions could be considered.

It can sometimes be justifiable to treat the value $a = \min \mathbb{I}$ as the "minimal **admissible** quality". Adding new products with such rate (negligible elements) should not affect the overall ranking.

Definition 6. *We say that a function* $\mathsf{F} \in \mathcal{E}(\mathbb{I})$ *is* ***zero-insensitive*** *(denoted* $\mathsf{F} \in P_{(\mathrm{zi})}$*) iff* $\mathsf{F}(\mathbf{x}, a) = \mathsf{F}(\mathbf{x})$ *for any* $\mathbf{x} \in \mathbb{I}^{1,2,\cdots}$.

We have $P_{(\mathrm{zi})} \cap P_{(\mathrm{nd})} \subseteq P_{(\mathrm{am})} \cap P_{(\mathrm{nd})}$ (by Lemma 1) and $P_{(\mathrm{zi})} \in \mathcal{P}_{ord}$.

Definition 7. *We say that a function* $\mathsf{F} \in P_{(\mathrm{bl})} \cap P_{(\mathrm{wbu})}$ *satisfies* ***condition*** $(\mathrm{s}-)$ *(denoted* $\mathsf{F} \in P_{(\mathrm{s}-)}$*) iff* $(\forall \mathbf{x} \in \mathbb{I}^{1,2,\cdots})\ (\forall y \in \mathbb{I})\ \ y \leq \mathsf{F}(\mathbf{x}) \Longrightarrow \mathsf{F}(\mathbf{x}, y) \leq \mathsf{F}(\mathbf{x})$.

Note that if $\mathsf{F} \in \mathcal{E}_{\mathcal{I}}(\mathbb{I})$ then $\mathsf{F} \in P_{(\mathrm{s}-)}$ iff $(\forall \mathbf{x} \in \mathbb{I}^{1,2,\cdots})\ \mathsf{F}(\mathbf{x}, \mathsf{F}(\mathbf{x})) = \mathsf{F}(\mathbf{x})$. We may also see that $\mathsf{F} \in P_{(\mathrm{s}-)} \Rightarrow \mathsf{F} \in P_{(\mathrm{zi})}$.

Definition 8. *We say that a function* $\mathsf{F} \in P_{(\mathrm{bl})} \cap P_{(\mathrm{wbu})}$ *satisfies* ***condition*** $(\mathrm{s}+)$ *(denoted* $\mathsf{F} \in P_{(\mathrm{s}+)}$*) iff* $(\forall y \in \mathbb{I}, y > a)\ \mathsf{F}(y) > a$ *and* $(\forall \mathbf{x} \in \mathbb{I}^{1,2,\cdots})\ (\forall y \in \mathbb{I})\ y > \mathsf{F}(\mathbf{x}) \Longrightarrow \mathsf{F}(\mathbf{x}, y) > \mathsf{F}(\mathbf{x})$.

Please, notice that both $P_{(\mathrm{s}+)}$ and $P_{(\mathrm{s}-)} \notin \mathcal{P}_{\mathrm{ord}}$.

5 Exemplary Impact Functions

Below we examine two classes of important and interesting aggregation operators: ordered linear combination and ordered conditional maximum which generalize OWA and OWMax operators, respectively. From now on, let $\mathbb{I} = [0, \infty]$.

5.1 Ordered Linear Combination

Definition 9. *A* ***triangle of coefficients*** *is a sequence* $\triangle = (c_{i,n} \in \bar{\mathbb{R}} : i \in [n], n \in \mathbb{N})$.

Such object can be represented by

$$\begin{array}{cccc} c_{1,1} & & & \\ c_{1,2} & c_{2,2} & & \\ c_{1,3} & c_{2,3} & c_{3,3} & \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

Definition 10. *Given arbitrary triangle of coefficients* $\triangle = (c_{i,n})_{i \in [n], n \in \mathbb{N}}$ *the* ***ordered linear combination*** *associated with* $\triangle$ *is a function* $\mathsf{OLC}_{\triangle} \in \mathcal{E}(\mathbb{I})$ *such that for any* $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{I}^{1,2,\cdots}$

$$\mathsf{OLC}_{\triangle}(\mathbf{x}) = \sum_{i=1}^{n} c_{i,n}\, x_{(n-i+1)}, \tag{1}$$

where $x_{(n-i+1)}$ *denotes the* i*th-largest value of* $\mathbf{x}$.

A special case of OLC is the class of *ordered weighted averaging functions* (OWA, introduced in [17]) with $\triangle = (w_{i,n})_{i \in [n], n \in \mathbb{N}}$ such that $(\forall n)\ \sum_{j=1}^{n} w_{j,n} = 1$ and $w_{i,n} \in [0, 1]$ for $i \in [n]$. So defined $\triangle$ is called a *weighting triangle* (see e.g. [3,4]) as an extension of OWA for input vectors of arbitrary length).

Lemma 2. *For* $\mathbb{I} = [a, b]$, $a \geq 0$, *any* $n \in \mathbb{N}$ *and given* $\mathbf{c}, \mathbf{c}' \in [0, \infty]^n$ *we have*

$$(\forall \mathbf{x} \in \mathbb{I}^n)\ \sum_{i=1}^{n} c_i\, x_{(n-i+1)} \geq \sum_{i=1}^{n} c'_i\, x_{(n-i+1)} \Longleftrightarrow (\forall k \in [n])\ \sum_{i=1}^{k} c_i \geq \sum_{i=1}^{k} c'_i. \tag{2}$$

Proof. $\sum_{i=1}^{n} c_i\, x_{(n-i+1)} - \sum_{i=1}^{n} c'_i\, x_{(n-i+1)} = \sum_{i=1}^{n-1}(c_i - c'_i)\left(x_{(n-i+1)} - x_{(1)}\right) + x_{(1)}\sum_{i=1}^{n}(c_i - c'_i) = \sum_{k=1}^{n}\left(x_{(n-k+1)} - x_{(n-k)}\right)\sum_{i=1}^{k}(c_i - c'_i) = (*)$, where $x_{(0)} = 0$. For every $j \in [n]$ we have $x_{(n-j+1)} - x_{(n-j)} \geq 0$ because $\min \mathbb{I} \geq 0$. Therefore $(*) \geq 0$ holds for all $\mathbf{x}$ iff $\sum_{i=1}^{k}(c_i - c'_i) \geq 0$ for every $k \in [n]$. □

It can be seen easily that we have equality at the left side of (2) iff $(\forall i \in [n])\ c_i = c'_i$.

Proposition 2. *For* $\mathbb{I} = [0, \infty]$ *and any* $\triangle = (c_{i,n})_{n\in\mathbb{N}, i\in[n]}$

a) $\mathsf{OLC}_\triangle \in \mathcal{E}_{\mathcal{A}}(\mathbb{I})$ *iff* $(\forall n)\ (\forall i \in [n])\ c_{i,n} \geq 0$ *and* $(\exists j \in [n])\ c_{j,n} > 0$.
b) $\mathsf{OLC}_\triangle \in \mathcal{E}_{\mathcal{I}}(\mathbb{I})$ *iff* $(\forall n)\ (\forall i \in [n])\ c_{i,n} \geq 0$ *and* $\sum_{j=1}^{i} c_{j,n+1} \geq \sum_{j=1}^{i} c_{j,n}$.

An easy proof based on Lemma 1 and 2 is omitted. Axiom (nd) is fulfilled due to the condition $c_{i,n} \geq 0$. Such $\mathsf{OLC}_\triangle$ is called an ordered conical combination.

Note that for an interval $\mathbb{I}' = [a, b]$, where $a \geq 0$ and $b < \infty$, under (nd), axioms (bl) and (bu) hold if and only if $\mathsf{OLC}_\triangle$ is an OWA (a.k.a. ordered convex combination). In that case the only aggregation function which is an impact function is the sample maximum $\mathsf{Max}(x_1, \ldots, x_n) := x_{(n)}$.

Let us consider other properties mentioned in Sec. 4.4.

Lemma 3. *For* $\mathbb{I} = [0, \infty]$ *and any* $\triangle = (c_{i,n})_{n\in\mathbb{N}, i\in[n]}$ *such that* $\mathsf{OLC}_\triangle \in \mathcal{E}_{\mathcal{I}}(\mathbb{I})$ *the following holds.*

a) $\mathsf{OLC}_\triangle \in P_{(\mathrm{zi})}$ *iff* $(\forall n)(\forall i \in [n])\ c_{i,n+1} = c_{i,n} \geq 0$.
b) $\mathsf{OLC}_\triangle \in P_{(\mathrm{s}-)}$ *iff* $(\exists j \in \mathbb{N})\ (\forall n)\ c_{j,n} = q$ *for* $j \geq n$ *and* $(\forall i \in [n], i \neq j)\ c_{i,n} = 0$, *where* $q \in (0, 1]$.
c) $\mathsf{OLC}_\triangle \in P_{(\mathrm{s}+)}$ *iff* $c_{1,1} > 0$ *and* $(\forall n)$ *if* $\sum_{i=1}^{n} c_{i,n} < 1$ *then* $c_{n+1,n+1} > 0$.

Note that the triangle of coefficients in a) and b) may be replaced by a (infinite-length) coefficients vector $(c_{1,1}, c_{2,2}, c_{3,3} \ldots)$.

Proof (Sketch). a) Obvious.

b) Let us fix n. We have to consider 3 cases. 1. If $c_{\cdot,n} = (n * 0)$ then condition (s+) holds iff $c_{\cdot,n+1} = (n * 0, q)$ for some $q \geq 0$; 2. If $c_{\cdot,n} = ((n-1) * 0, q)$ for some $q > 0$ then (s+) (for any $\mathbf{x}$) iff $q \in (0, 1]$ and $c_{\cdot,n+1} = ((n-1) * 0, q, 0)$; 3. If $c_{\cdot,n} = ((i-1) * 0, q, (n-i) * 0)$ for some $q \geq 0$ and $i \in [n-1]$ then (s+) iff $c_{\cdot,n} = ((i-1) * 0, q, (n-i+1) * 0)$.

c) Fix n. For any $\mathbf{x} \in \mathbb{I}^n$ if $\sum_{i=1}^{j} c_{i,n} \geq 1$ for some $j \leq n$ then $\sum_{i=1}^{j} c_{i,n} x_{(n-i+1)} + \varepsilon > x_{(n-j+1)}$ for any $\varepsilon > 0$, as $c_{i,n} > 0$ and therefore $\mathsf{OLC}_\triangle(\mathbf{x}, \mathsf{OLC}_\triangle(\mathbf{x}) + \varepsilon) > \mathsf{OLC}_\triangle(\mathbf{x})$. If $\sum_{i=1}^{n} c_{i,n} < 1$ then for $\mathbf{y} = (n * b)$ $\mathsf{OLC}_\triangle(\mathbf{y}) < b$ and then $\mathsf{OLC}_\triangle(\mathbf{y}, \mathsf{OLC}_\triangle(\mathbf{y}) + \varepsilon) > \mathsf{OLC}_\triangle(\mathbf{y})$ iff $c_{n+1,n+1} > 0$ and the proof is complete. □

Note that both (s−) and (s+) holds if and only if $c_{1,n} = 1$ and $c_{j,n} = 0$ for any n and $j > 1$, i.e. $\mathsf{OLC}_\triangle$ is the sample maximum.

5.2 Ordered Conditional Maximum

Definition 11. *The* ***ordered conditional maximum*** *associated with a triangle of coefficients* $\triangle = (c_{i,n} \in \mathbb{I} : i \in [n], n \in \mathbb{N})$ *is a function* $\mathsf{OCM}_\triangle \in \mathcal{E}(\mathbb{I})$ *such that*

$$\mathsf{OCM}_{\triangle}(\mathbf{x}) = \bigvee_{i=1}^{n} c_{i,n} \wedge x_{(n-i+1)} \tag{3}$$

for $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{I}^{1,2,\ldots}$.

A particular case of OCM is the ordered weighted maximum operator (OWMax) introduced in [5], defined for $\mathbb{I} = [0,1]$ and such that $\bigvee_{i=1}^{n} c_{i,n} = 1$ and $c_{i,n} \in [0,1]$.

Furthermore, OCM also generalizes the well-known h-index (see [14]). The h-index was originally defined by Hirsch [7] for ratings in $\mathbb{N}_0$ as a function h such that

$$\mathsf{h}(x_1, \ldots, x_n) = \max\{i = 0, \ldots, n : x_{(n-i+1)} \geq i\}.$$

It was proposed as a method of assessing scientific merit of individual researchers by means of the number of citations received by their scientific papers. Its popularity possibly arose from an appealing interpretation: one has h-index of, say H, if H of his papers gained at least H and the remaining $n - H$ papers at most H citations. Interestingly, a similar object was earlier defined in the context of Bonferroni-type multiple significance testing (see e.g. [2]).

Lemma 4. *If* $\triangle = (c_{i,n})_{i \in [n], n \in \mathbb{N}}$ *such that* $c_{i,n} = i$ *for* $n \in \mathbb{N}$ *and* $i \in [n]$ *then for any* $m \in \mathbb{N}$, $x_1, \ldots, x_m \in \mathbb{I} \cap \mathbb{N}_0$,

$$\mathsf{OCM}_{\triangle}(x_1, \ldots, x_m) = \mathsf{h}(x_1, \ldots, x_m).$$

Proof. Let $H = \max\{i : x_{(n-i+1)} \geq i\}$. We have $\bigvee_{i=1}^{H} i \wedge x_{(n-i+1)} = H$ and $\bigvee_{i=H+1}^{n} i \wedge x_{(n-i+1)} = x_{(n-H)} < H + 1$. However, since $x_{(n-H)} \in \mathbb{N}_0$, then $x_{(n-H)} \leq H$ and therefore $\bigvee_{i=1}^{n} i \wedge x_{(n-i+1)} = H$. □

Note that we have $\mathsf{h}(2, 1.5) = 1$ but $\mathsf{OCM}_{\triangle}(2, 1.5) = 1.5$. Generally, for arbitrary $\mathbf{x}$, $\mathsf{OCM}_{\triangle}(\mathbf{x}) = \max\{H, x_{(n-H)}\} \in [H, H+1)$.

Lemma 5. *For any* $\mathbb{I}$ *and any* $n \in \mathbb{N}$, *given* $\mathbf{c}, \mathbf{c}' \in \mathbb{I}^n$ *we have*

$$(\forall \mathbf{x} \in \mathbb{I}^n) \bigvee_{i=1}^{n} c_i \wedge x_{(n-i+1)} \geq \bigvee_{i=1}^{n} c'_i \wedge x_{(n-i+1)} \iff (\forall k \in [n]) \bigvee_{i=1}^{k} c_i \geq \bigvee_{i=1}^{k} c'_i. \tag{4}$$

The proof is omitted. Note that if $K = \{k = 2, 3, \ldots, n : c_k \leq \bigvee_{i=1}^{k-1} c_i\}$ then $\bigvee_{i=1}^{n} c_i \wedge x_{(n-i+1)} = \bigvee_{i=1, i \notin K}^{n} c_i \wedge x_{(n-i+1)}$. Additionally, it is easily seen that we have equality at the left side of (4) iff $(\forall k \in [n])$ $\bigvee_{i=1}^{k} c_i = \bigvee_{i=1}^{k} c'_i$.

Proposition 3. *For any* $\mathbb{I} = [a, b]$ *and any* $\triangle = (c_{i,n})_{i \in [n], n \in \mathbb{N}}$, $c_{i,n} \in \mathbb{I}$

a) $\mathsf{OCM}_{\triangle} \in \mathcal{E}_{\mathcal{A}}(\mathbb{I})$ *iff* $(\forall n)$ $(\exists j \in [n])$ $c_{j,n} = b$.
b) $\mathsf{OCM}_{\triangle} \in \mathcal{E}_{\mathcal{I}}(\mathbb{I})$ *iff* $(\forall n)$ $(\forall i \in [n])$ $c_{i,n} \geq a$ *and* $\bigvee_{j=1}^{i} c_{j,n+1} \geq \bigvee_{j=1}^{i} c_{j,n}$.

The proof is omitted. Let us consider other properties.

Lemma 6. *For any* $\mathbb{I} = [a, b]$ *and any* $\triangle = (c_{i,n})_{i\in[n],n\in\mathbb{N}}$ *such that* $\mathsf{OCM}_\triangle \in \mathcal{E}_\mathcal{I}(\mathbb{I})$ *the following holds:*

a) $\mathsf{OCM}_\triangle \in P_{(\mathrm{zi})}$ *iff* $(\forall n)(\forall i \in [n])\ \bigvee_{j=1}^{i} c_{j,n+1} = \bigvee_{j=1}^{i} c_{j,n}$.
b) $\mathsf{OCM}_\triangle \in P_{(\mathrm{s}-)}$ *iff* $\mathsf{OCM}_\triangle \in P_{(\mathrm{zi})}$.
c) $\mathsf{OCM}_\triangle \in P_{(\mathrm{s}+)}$ *iff* $c_{1,1} > a$ *and* $(\forall n)$ *if* $c_{1,n} < b$ *then* $c_{1,n+1} > \bigvee_{i\in[n],c_{i,n}<b} c_{i,n}$.

Proof (Sketch). a) It follows from the remark to Lemma 5.

b) Let us fix n. We should only show that $(\forall i \in [n])\ \bigvee_{j=1}^{i} c_{j,n} = \bigvee_{j=1}^{i} c_{j,n+1}$ implies $\mathsf{OCM}_\triangle \in (\mathrm{s}-)$. Let $\mathsf{OCM}_\triangle(\mathbf{x}) = c_{j,n} \wedge x_{(n-j+1)}$ for some j. But as $\mathsf{OCM}_\triangle(\mathbf{x}) \leq x_{(n-j+1)}$ and $x_{(n-j+1)} \wedge \bigvee_{i=j+1}^{n+1} c_{i,n+1} \leq x_{(n-j+1)}$, it holds $\mathsf{OCM}_\triangle(\mathbf{x}, \mathsf{OCM}_\triangle(\mathbf{x})) = c_{j,n+1} \wedge x_{(n-j+2)} = c_{j,n} \wedge x_{(n-j+1)} = \mathsf{OCM}_\triangle(\mathbf{x})$.

c) Let us fix n and let $c_{j,n} = \bigvee_{i\in[n],c_{i,n}<b} c_{i,n}$ for some $j \in [n]$. Take $\mathbf{x} = (n * c_{j,n})$. Then for any $\varepsilon > 0$ $\mathsf{OCM}_\triangle(\mathbf{x}, c_{j,n} + \varepsilon) > c_{j,n} = \mathsf{OCM}_\triangle(\mathbf{x})$ iff $c_{1,n+1} > c_{j,n}$. Now take any $\mathbf{y} \in \mathbb{I}^n$. Let $\mathsf{OCM}_\triangle(\mathbf{y}) = c_{j,n} \wedge y_{(n-j+1)}$ for some j. Then $c_{j,n} \wedge y_{(n-j+1)} < c_{1,n+1} \wedge ((c_{j,n} \wedge y_{(n-j+1)}) + \varepsilon)$, which completes the proof. □

All $\mathsf{OCM}_\triangle$ satisfying both conditions $(\mathrm{s}-)$ and $(\mathrm{s}+)$ are equivalent to the sample maximum (when $c_{1,1} = b$). The extended h-index is an impact function satisfying (zi).

6 Conclusions

In the paper we have considered a class of aggregation operators and discussed their basic properties. The particular attention has been directed to remarkable classes of such functions, i.e. ordered linear combination and ordered conditional maximum operators, which generalize OWA and OWMax operators, respectively. However, extensions of many other classes of aggregation operators would be interesting too.

The problem of ratings based on citations was mentioned to illustrate the need and the importance of such extensions of the aggregation operators. Nowadays the aforementioned h-index is probably the best known scientometric tool. However, many other interesting bibliometric indices exist in the literature and they surely could be also characterized in the framework of the impact functions. This is the topic of our further research.

Finally, we want to stress that the suggested generalization of aggregation operators might have applications not only in scientometrics or — generally — the Producer Assessment Problem but in many other fields. However, one has to be conscious that aggregation performed in some special areas may potentially require other particular requirements that should be expressed by different axioms.

References

1. Arrow, K.J.: A difficulty in the concept of social welfare. Journal of Political Economy 58(4), 328–346 (1950)
2. Benjamini, Y., Hochberg, Y.: Controlling false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B 57(1), 289–300 (1995)

3. Calvo, T., Kolesarova, A., Komornikova, M., Mesiar, R.: Aggregation operators: properties, classes and construction methods. In: Calvo, T., Mayor, G., Mesiar, R. (eds.) Aggregation Operators. New Trends and Applications, Studies in Fuzziness and Soft Computing, vol. 97, pp. 3–104. Physica-Verlag, New York (2002)
4. Calvo, T., Mayor, G.: Remarks on two types of extended aggregation functions. Tatra Mountains Mathematical Publications 16, 235–253 (1999)
5. Dubois, D., Prade, H., Testemale, C.: Weighted fuzzy pattern matching. Fuzzy Sets and Systems 28, 313–331 (1988)
6. Grabisch, M., Pap, E., Marichal, J.L., Mesiar, R.: Aggregation Functions, Cambridge (2009)
7. Hirsch, J.E.: An index to quantify individual's scientific research output. Proceedings of the National Academy of Sciences 102(46), 16569–16572 (2005)
8. Marchant, T.: An axiomatic characterization of the ranking based on the h-index and some other bibliometric rankings of authors. Scientometrics 80(2), 325–342 (2009)
9. Marchant, T.: Score-based bibliometric rankings of authors. Journal of the American Society for Information Science and Technology 60(6), 1132–1137 (2009)
10. May, K.O.: A set of independent necessary and sufficient conditions for simple majority decision. Econometrica 20(4), 680–684 (1952)
11. Mayor, G., Calvo, T.: On extended aggregation functions. In: Proc. IFSA 1997, vol. 1, pp. 281–285. Academia, Prague (1997)
12. Palacios-Huerta, I., Volij, O.: The measurement of intellectual influence. Econometrica 72(3), 963–977 (2004)
13. Torra, V. (ed.): Information Fusion in Data Mining. Studies in Fuzziness and Soft Computing, vol. 123. Springer, Heidelberg (2003)
14. Torra, V., Narukawa, Y.: The h-index and the number of citations: two fuzzy integrals. IEEE Transactions on Fuzzy Systems 16(3), 795–797 (2008)
15. Woeginger, G.J.: An axiomatic analysis of Egghe's g-index. Journal of Informetrics 2(4), 364–368 (2008)
16. Woeginger, G.J.: An axiomatic characterization of the Hirsch-index. Mathematical Social Sciences 56(2), 224–232 (2008)
17. Yager, R.R.: On ordered weighted averaging aggregation operators in multictriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics 18(1), 183–190 (1988)

Some Properties of Multi–argument Distances and Fermat Multidistance

Javier Martín and Gaspar Mayor

University of the Balearic Islands,
Cra. Valldemossa km 7.5, 07122 Palma, Spain
{javier.martin,gmayor}@uib.es

Abstract. The conventional definition of distance between two points has been extended to apply to collections of more than two elements. In this paper some significant properties of those multi–argument distances are analyzed and we study a class of multidistances that we call Fermat multidistances.

Keywords: distance, multidistance, Fermat points, regularity.

1 Introduction

The conventional definition of distance over a space specifies properties that must be obeyed by any measure of "how separated" two points in that space are. However often one wants to measure how separated the members of a collection of more than two elements are. The usual way to do this is to combine the distance values for all pairs of elements in the collection, into an aggregate measure. Thus, given an Euclidean triangle (A, B, C) we can combine the distances AB, AC, BC using, for instance, a 3–dimensional OWA operator, say W. Then, we calculate the distance of A, B, C by means of the formula $D(A, B, C) = W(AB, AC, BC)$. It is clear that we have to choose the weighting vector of W such that the multi–argument distance function D satisfies a group of axioms that extend in some degree to those for ordinary distance functions. Of course we can consider other procedures to measure how separated the vertices A, B, C are: in Euclidean geometry the Fermat point of a triangle (A, B, C) is the point F for which the sum of the distances from F to the vertices is as small as possible; i.e. it is the point F such that $FA + FB + FC$ is minimized. Then we can define $D(A, B, C) = FA + FB + FC$.

For pairwise distances and related distance matrix see for example [1]. A recent paper [6] deals with the problem of aggregating pairwise distance values in order to construct a multidistance function [11,4].

In addition to their intrinsic mathematical interest, multidistances have many potential applications. In [5] the concept of T-indistinguishability relation, which measure the degree of similarity between two elements of a set, has been extended with the aim that it also measures the similarity between all the elements of a finite list. The close relationship between them has been studied and examples

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 703–711, 2010.

of how to obtain T-multi-indistinguishabilities from multidistances has been explained. We think that further deeper development from both theoretical and practical points of view should be carried out. Multidistances can be also directly incorporated into many other domains: distance–based clustering, pattern recognition, etc, where the extension of ordinary (binary) distances to multidimensional collections can be of interest. This is also the case of the so–called Jensen–Shannon divergence (JSD) which is a distance for probability distributions that have been used to treat different problems such as analysis of symbolic sequences, examination of texts in literature or separation of quantum states [3,10].

In a recent paper related to consensus theory, consensus measures based on metrics on weak orders were built by means of the arithmetic mean. As we will see, this is nothing else but a particular multidistance [2]. In this same framework, the Condorcet's method to deal with the Condorcet's effect (the application of the pairwise majority rule to individual preference orders can generate a non transitive collective preference) leads to a difficult combinatorial optimization problem. It can be seen that M is a Condorcet's solution (i.e., M is a ranking with maximum support among all possible rankings) if and only if M is a Fermat point (median) in a certain discrete metric space [9], but now one may find the same problem: the medians can be difficult to calculate. It is in particular the case when the metric space of all considered elements is not a good metric space for medians.

In this paper we study a class of multi–argument distance functions that we call Fermat multidistances. In Section 2 we expose some general preliminaries. Section 3 is devoted to the discussion of some new properties in the study of multidistances. And finally Section 4 deals with multi–argument distances based on Fermat points.

2 Preliminaries

Definition 1. *A function* $D : \bigcup_{n\geqslant 1} \mathbb{X}^n \to [0,\infty)$ *is a* multidistance *on a non empty set* $\mathbb{X}$ *when the following properties hold, for all* n *and* $x_1,\ldots,x_n,y \in \mathbb{X}$*:*

(m1) $D(x_1,\ldots,x_n) = 0$ *if and only if* $x_i = x_j$ *for all* $i,j = 1,\ldots n$.
(m2) $D(x_1,\ldots,x_n) = D(x_{\pi(1)},\ldots,x_{\pi(n)})$ *for any permutation* π *of* $1,\ldots,n$,
(m3) $D(x_1,\ldots,x_n) \leqslant D(x_1,y) + \ldots + D(x_n,y)$,

We say that D *is a* strong *multidistance if it fulfills (m1), (m2) and:*

(m3') $D(\boldsymbol{x}_1,\ldots \boldsymbol{x}_k) \leqslant D(\boldsymbol{x}_1,\boldsymbol{y}) + \ldots + D(\boldsymbol{x}_k,\boldsymbol{y})$ *for all* $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k,\boldsymbol{y} \in \bigcup_{n\geqslant 1} \mathbb{X}^n$.

Expressions like $D(\boldsymbol{x},\boldsymbol{y})$ in (m3'), that is, the function D applied to two lists $\boldsymbol{x} = (x_1,\ldots,x_n) \in \mathbb{X}^n$ and $\boldsymbol{y} = (y_1,\ldots,y_m) \in \mathbb{X}^m$, signify the multidistance of the joined list:

$$D(\boldsymbol{x},\boldsymbol{y}) = D(x_1,\ldots,x_n,y_1,\ldots,y_m). \tag{1}$$

Remark 1.

1) If D is a multidistance on $\mathbb{X}$, then the restriction of D to $\mathbb{X}^2$, $D|_{\mathbb{X}^2}$, is an ordinary distance on $\mathbb{X}$.
2) Any ordinary distance d on $\mathbb{X}$ can be extended in order to obtain a multidistance. For example, we can define a function D_M in this way:

$$D_M(x_1, \ldots, x_n) = \max\{d(x_i, x_j);\ i < j\}. \tag{2}$$

Then, D_M is a strong multidistance on $\mathbb{X}$ such that $D|_{\mathbb{X}^2} = d$. We will call it *maximum multidistance*.

Example 1. Based on the drastic distance, defined by

$$d(x, y) = \begin{cases} 1 \text{ if } x \neq y, \\ 0 \text{ if } x = y, \end{cases} \tag{3}$$

several multidistances can be defined extending it. For instance,

- $D_1(x_1, \ldots, x_n) = \begin{cases} 0 \text{ if } x_i = x_j\ \forall i, j, \\ 1 \text{ otherwise,} \end{cases}$
- $D_2(x_1, \ldots, x_n) = |\{x_1, \ldots, x_n\}| - 1.$

Both of them are strong.

Proposition 1. *Let D and D' be multidistances on $\mathbb{X}$.*

1) $D + D'$ is a multidistance on $\mathbb{X}$.
2) If $k > 0$, then kD is a multidistance on $\mathbb{X}$.
3) $\frac{D}{1+D}$ and $\min\{1, D\}$ are also multidistances on $\mathbb{X}$, with values in $[0, 1]$.

Remark 2. If D is a multidistance on $\mathbb{X}$, then the function $\tilde{D} : \bigcup_{n \geqslant 1} \mathbb{X}^n \times \bigcup_{n \geqslant 1} \mathbb{X}^n \to [0, +\infty)$ defined by

$$\tilde{D}(\boldsymbol{x}, \boldsymbol{y}) = D(x_1, \ldots, x_n, y_1, \ldots, y_m) \tag{4}$$

for all $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{X}^n$ and $\boldsymbol{y} = (y_1, \ldots, y_m) \in \mathbb{X}^m$ is not a distance on the set $\bigcup_{n \geqslant 1} \mathbb{X}^n$.

However, pseudodistances (distance functions allowing $d(x, y) = 0$ for distinct values x, y) on this set can be constructed by means of D and an ordinary distance δ on $\mathbb{R}$ in this way:

$$D_\delta(\boldsymbol{x}, \boldsymbol{y}) = \delta(D(\boldsymbol{x}), D(\boldsymbol{y})). \tag{5}$$

Given a non empty set $\mathbb{X}$, a multidistance D defined on it and a list $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{X}^n$, the following sets can be considered:

$$\begin{aligned} \mathbb{X}_{(\boldsymbol{x}, >)} &= \{y \in \mathbb{X} \colon D(\boldsymbol{x}, y) > D(\boldsymbol{x})\}, \\ \mathbb{X}_{(\boldsymbol{x}, =)} &= \{y \in \mathbb{X} \colon D(\boldsymbol{x}, y) = D(\boldsymbol{x})\}, \\ \mathbb{X}_{(\boldsymbol{x}, <)} &= \{y \in \mathbb{X} \colon D(\boldsymbol{x}, y) < D(\boldsymbol{x})\}. \end{aligned} \tag{6}$$

(similar definitions for $\mathbb{X}_{(\boldsymbol{x}, \geqslant)}$ and $\mathbb{X}_{(\boldsymbol{x}, \leqslant)}$)

These three sets form a partition of $\mathbb{X}$, with possibly one or two of them empty.

Definition 2. *Given a multidistance D and a list $\boldsymbol{x} = (x_1, \ldots, x_n) \in \bigcup_{n\geqslant 1} \mathbb{X}^n$, the* closed ball *of center $\boldsymbol{x}$ and radius $r \in \mathbb{R}$ is the set:*

$$B(\boldsymbol{x}, r) = \{y \in \mathbb{X} \colon D(\boldsymbol{x}, y) \leqslant D(\boldsymbol{x}) + r\}. \tag{7}$$

If $n = 1$ and $r \geqslant 0$, the list reduces to a point x, and the ball is $B(x, r) = \{y \in \mathbb{X} \colon D(x, y) \leqslant D(x) + r\}$. As $D(x) = 0$, $B(x, r)$ is an ordinary closed ball of the metric space $\mathbb{X}$.

Remark 3.

1) The closed ball centered at a list $\boldsymbol{x}$ and radius 0 is the set $\mathbb{X}_{(\boldsymbol{x},\leqslant)}$.
2) There exist non empty balls with negative radius if and only if $\mathbb{X}_{(\boldsymbol{x},<)} \neq \emptyset$.

3 Properties of Multidistances

In this section four remarkable properties for multidistances are introduced. They are the following:

- **Regularity** [7]**:**

$$D(\boldsymbol{x}, y) \geqslant D(\boldsymbol{x}), \tag{8}$$

  for all $\boldsymbol{x} \in \bigcup_{n\geqslant 1} \mathbb{X}^n$, $y \in \mathbb{X}$. That is, the multidistance of a list cannot decrease when adding a new element.
- **Stability:**

$$D(\boldsymbol{x}) = D(\boldsymbol{x}, x_i), \tag{9}$$

  for all $\boldsymbol{x} \in \bigcup_{n\geqslant 1} \mathbb{X}^n$ and any element x_i of $\boldsymbol{x}$. In other words, repeated elements are superfluous regarding the value of the multidistance of a list.
- **Superadditivity:**

$$D(\boldsymbol{x}, \boldsymbol{y}) \geqslant D(\boldsymbol{x}) + D(\boldsymbol{y}), \tag{10}$$

  for all $\boldsymbol{x}, \boldsymbol{y} \in \bigcup_{n\geqslant 1} \mathbb{X}^n$
- **Homogeneity:**

$$D(\overbrace{\boldsymbol{x}, \ldots, \boldsymbol{x}}^{k}) = kD(\boldsymbol{x}), \tag{11}$$

  for all $\boldsymbol{x} \in \bigcup_{n\geqslant 1} \mathbb{X}^n$.

Stability and homogeneity are incompatible because the first means, in particular, $D(\boldsymbol{x}, \ldots, \boldsymbol{x}) = D(\boldsymbol{x})$. The relationship between superadditivity and regularity is given in the following result.

Proposition 2. *Any superadditive multidistance is regular.*

Proof. Superadditivity means that $D(\boldsymbol{x}, y) \geqslant D(\boldsymbol{x}) + D(y)$ for a list $\boldsymbol{x}$ and an element y; but $D(y) = 0$. □

The converse is not true, as the next example shows.

Example 2. Consider the multidistance D defined in this way: $D(\boldsymbol{x})$ is the diameter of the ball with minor radius containing all the elements of the list $\boldsymbol{x}$. Clearly, D is regular, because the diameter can not decrease when a new element is added, but it is not superadditive. For instance, in the Euclidean line, $D(0,1,0,2) = 2 \not\geqslant D(0,1) + D(0,2) = 1 + 2$.

The problem of finding the smallest enclosing ball (SEB) of a set of points is a well–studied problem with a large number of applications. In the two–dimensional case, the Minimal Enclosing Circle Problem is, simply stated, the problem of finding the smallest circle that completely contains a set of points. Formally, given a set S of n planar points, find the circle C of smallest radius such that all points in S are contained in either C or its boundary. The minimal enclosing circle is used in planning the location of a shared facility (for example, a hospital servicing a community). In 1983, N. Meggido showed that the minimal enclosing circle problem can be solved in $O(n)$ time using the prune–and–search techniques for linear programming [8].

The relationship with strongness can also be established.

Proposition 3. *Any strong multidistance is regular, stable, superadditive and non–homogeneous.*

Proof. Condition (8) of regularity is a particular case of (m3'), with $k = 1$ and added element y. The superadditivity can be proven in this way:

$$D(\boldsymbol{x}, \boldsymbol{y}) \geqslant D(\boldsymbol{x}, \boldsymbol{z}) + D(\boldsymbol{y}, \boldsymbol{z}) \geqslant D(\boldsymbol{x}) + D(\boldsymbol{y}),$$

the first inequality due to the strongness and the second one, to the regularity.

To prove the stability we can see that $D(\boldsymbol{x}) \geqslant D(\boldsymbol{x}, x_i)$, because D is regular. On the other hand, for an appropriate application of strongness:

$$D(\boldsymbol{x}, x_i) \leqslant D(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, x_i) + D(x_i, x_i, x_i) = D(\boldsymbol{x}),$$

and so D is stable, hence non homogeneous. □

The four properties, together with strongness, are depicted as a Venn diagram in Figure 1. The Fermat multidistance is also placed; it is regular, superadditive, homogeneous and non–stable, as we will see in Section 4. Regularity and stability have a precise meaning in terms of the balls defined in Section 2.

Proposition 4. *Let D be a multidistance.*

1. *D is regular if and only if the balls with negative radius are empty.*
2. *If D is stable then the elements of any list belong to any ball with non–negative radius centered at the list.*

Proof.

1. Regularity means that $D(\boldsymbol{x}, y) - D(\boldsymbol{x}) \geqslant 0$, that is, $B(\boldsymbol{x}, r) \neq \emptyset$ when $r < 0$.
2. If D is stable then $D(\boldsymbol{x}, x_i) - D(\boldsymbol{x}) = 0$ and so $x_i \in B(\boldsymbol{x}, r)$ when $r \geqslant 0$. □

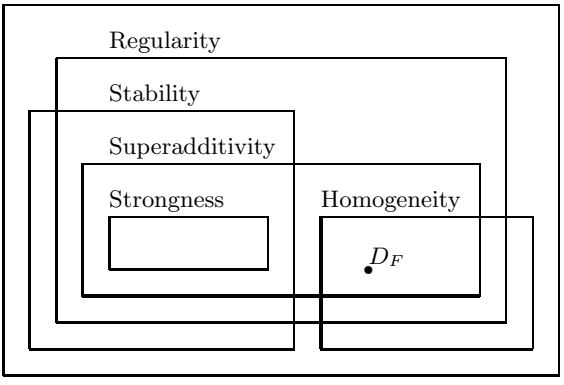


Fig. 1. Properties of multidistances

Remark 4. The second statement of the proposition can be converted into a characterization in this way: D is regular and stable if and only if the elements of any list belong to any ball with non–negative radius centered at the list.

Now we study how these properties are fulfilled for two families of multidistances based on an ordinary distance d defined on a set $\mathbb{X}$: the ones defined with OWAs and those based on the sum of the ordinary pairwise distances of the elements of the list.

OWA–Based Multidistances. Let $\mathcal{W} = \{W_n; n \geq 2\}$ be a family of OWAs [12], where the weights $\omega_1^n, \ldots, \omega_{\binom{n}{2}}^n$ of the $\binom{n}{2}$–dimensional OWA W_n, are applied to the list of the $\binom{n}{2}$ pairwise distances arranged in an increasing order.

The function $D_W : \bigcup_{n \geqslant 1} \mathbb{X}^n \to [0, \infty)$ defined by:

$$D_W(\boldsymbol{x}) = \begin{cases} 0 & \text{if } n = 1, \\ W_n(\overbrace{d(x_1, x_2), d(x_1, x_3), \ldots, d(x_{n-1}, x_n)}^{\binom{n}{2}}) & \text{if } n \geqslant 2, \end{cases} \tag{12}$$

for all $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{X}^n$, is a multidistance if and only if, for all $n \geqslant 3$:

$$\omega_1^n + \ldots + \omega_{n-1}^n > 0. \tag{13}$$

Special cases are when $W_n = \max$, with lists of weights $(1, 0, \ldots, 0)$ for all n, obtaining the maximum multidistance D_M given in (2), or when $W_n = \left(\frac{1}{\binom{n}{2}}, \ldots, \frac{1}{\binom{n}{2}}\right)$, whose corresponding multidistance is the arithmetic mean of the pairwise distances.

– If D_W is regular then, for all $n \geqslant 3$:

$$\omega_1^n + \ldots + \omega_{n-1}^n = 1, \tag{14}$$

- D_W is stable if and only if $\omega_1^n = \alpha$ and $\omega_{n-1}^n = 1 - \alpha$.
- D_M is the only strong multidistance of the family.
- There are neither homogeneous nor superadditive OWA–based multidistances.

Sum–Based Multidistances. We can define a function $D_\lambda : \bigcup_{n\geqslant 1} \mathbb{X}^n \to [0,\infty)$ in this way:

$$D_\lambda(\boldsymbol{x}) = \begin{cases} 0 & \text{if } n = 1, \\ \lambda(n) \sum_{i<j} d(x_i, x_j), & \text{if } n \geqslant 2, \end{cases} \tag{15}$$

Such a function is a multidistance if and only if:

(i) $\lambda(2) = 1$,
(ii) $0 < \lambda(n) \leq \frac{1}{n-1}$ for any $n > 2$.

As an example, if $\lambda(n) = \frac{1}{\binom{n}{2}}$ then D_λ is again the arithmetic mean.

With respect to the properties:

- D_λ is regular if and only if $\lambda(n) = \frac{1}{n-1}$ for all $n \geqslant 2$.
- There are no stable, homogeneous or strong multidistances within this family.

4 Fermat Multidistance

Let $(\mathbb{X}, d)$ be a metric space and $(x_1, \ldots, x_n) \in \mathbb{X}^n$.

Proposition 5. *The function $f : \mathbb{X} \to [0, +\infty)$ defined by $f(x) = \sum_{i=1}^n d(x_i, x)$ is continuous and it reaches a minimum.*

Sketch of the proof. Each $d(x_i, x)$, $x \in \mathbb{X}$, is a continuous, non–negative and lower bounded function, such that $d(x_i, x_i) = 0$. Therefore, the function f also reaches a minimum value. □

Definition 3. *We will call* Fermat set *of a list $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{X}^n$ and we will denote by $F_{\boldsymbol{x}}$, the set of points where the function f above reaches the minimum value:*

$$F_{\boldsymbol{x}} = \Big\{ x \in \mathbb{X} : \sum_{i=1}^n d(x_i, x) \leqslant \sum_{i=1}^n d(x_i, x'), \forall x' \in \mathbb{X} \Big\}. \tag{16}$$

Proposition 6. *Let $D_F : \bigcup_{n\geqslant 1} \mathbb{X}^n \to [0,\infty)$ be the function defined by:*

$$D_F(x_1, \ldots, x_n) = \min_{x\in\mathbb{X}} \Big\{ \sum_{i=1}^n d(x_i, x) \Big\}. \tag{17}$$

1) D_F is a multidistance on $\mathbb{X}$.
2) D_F is maximum in the sense that any other multidistance D such that $D|_{\mathbb{X}^2} = D_F|_{\mathbb{X}^2}$ takes values not greater than D_F: $D(\boldsymbol{x}) \leqslant D_F(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \bigcup_{n\geqslant 1} \mathbb{X}^n$.

3) D_F is not in general a strong multidistance.

This function D_F will be called the *Fermat multidistance* associated with an ordinary distance d. It was defined in [4], based on the idea of the Fermat point explained in the introduction, and it fulfills three of the properties defined in Section 3.

Example 3. In the metric space $(\mathbb{R}^m, d)$, where d is the Manhattan distance:

$$d((a_1,\ldots,a_m),(b_1,\ldots,b_m)) = \sum_{i=1}^{m} |a_i - b_i|,$$

the set F associated to the points $\boldsymbol{x}_i = (a_{i1},\ldots,a_{im})$, $i = 1,\ldots,n$ is:

$$F_{(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)} = \prod_{i=1}^{m} [a_{(\lfloor \frac{n+1}{2} \rfloor)i}, a_{(\lfloor \frac{n+2}{2} \rfloor)i}],$$

where the parentheses in the subindexes signify that $a_{(1)i},\ldots,a_{(m)i}$ is a non–decreasing rearrangement of the components i of the points, $\lfloor x \rfloor$ is the *floor* function and $\prod$ is the usual cartesian product of sets.

If m is an odd number, the extreme points of the intervals coincide and the set $F_{(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)}$ reduces to a point, whose coordinates are the medians of the respective coordinates of the points $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$: see Figure 2.

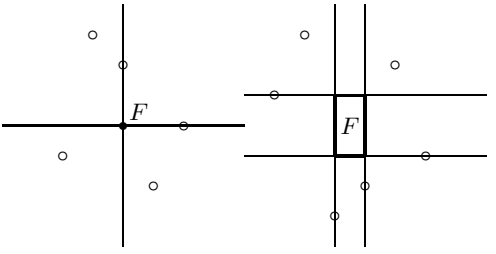


Fig. 2. Fermat sets F associated with two families of points in $\mathbb{R}^2$

Proposition 7. *The multidistance of Fermat D_F defined on any metric space $(\mathbb{X}, d)$ is superadditive, and hence regular, and homogeneous.*

Proof. The homogeneity of D_F follows from the fact that the Fermat sets of $\overbrace{\boldsymbol{x},\ldots,\boldsymbol{x}}^{k}$ and $\boldsymbol{x}$ are the same. The superadditivity can be proven as follows. Let z_1, z_2, z_3 be points in the Fermat sets of $\boldsymbol{x} \in \mathbb{X}^n$, $\boldsymbol{y} \in \mathbb{X}^m$ and the join list $\boldsymbol{xy} \in \mathbb{X}^{n+m}$, respectively. Then,

$$\begin{aligned} D_F(\boldsymbol{x},\boldsymbol{y}) &= \textstyle\sum_{i=1}^{n} d(x_i, z_3) + \sum_{i=1}^{m} d(y_i, z_3) \\ &\geqslant \textstyle\sum_{i=1}^{n} d(x_i, z_1) + \sum_{i=1}^{m} d(y_i, z_2) \\ &= D_F(\boldsymbol{x}) + D_F(\boldsymbol{y}). \end{aligned}$$

□

The remain property of stability is not fulfilled. For example, in the Euclidean line the multidistance of the list $(0,1,2)$ is $D(0,1,2)=2$, because $F_{(0,1,2)}=\{1\}$, and on the other hand the multidistance $(0,0,1,2)$ has Fermat set $F_{(0,0,1,2)}=\{0\}$, hence $D(0,0,1,2)=3$.

The sets defined by (6) can be expressed in terms of the Fermat sets:

$$\mathbb{X}_{(\boldsymbol{x},>)}=\mathbb{X}\setminus F_{\boldsymbol{x}},\quad \mathbb{X}_{(\boldsymbol{x},=)}=F_{\boldsymbol{x}},\quad \mathbb{X}_{(\boldsymbol{x},<)}=\emptyset. \tag{18}$$

Acknowledgments. The authors acknowledge the support of the Govern Balear grant PCTIB2005GC1–07 and the Spanish DGI grant MTM2009–10962.

References

1. Blumenthal, L.M.: Theory and applications of distance geometry, 2nd edn. Chelsea Publishing Co., New York (1970)
2. García–Lapresta, J.L., Pérez–Román, D.: Some measures of consensus generated by distances on weak orders. In: XIV Congreso Español sobre Tecnologías y Lógica Fuzzy ESTYLF, Asturias, Spain (2008)
3. Lamberti, P.W., Majtey, A.P., Borras, A., Casas, M., Plastino, A.: Metric character of the quantum Jensen–Shannon divergence. Phys. Rev. A 77 (2008)
4. Martín, J., Mayor, G.: How separated Palma, Inca and Manacor are? In: International Workshop on Aggregation Operators AGOP, Palma, Spain (2009)
5. Martín, J., Mayor, G.: An axiomathical approach to T–multi-indistinguishabilities. In: IFSA–EUSFLAT, Lisbon (2009)
6. Martín, J., Mayor, G.: Aggregating pairwise distance values. In: Workshop Preference Modelling and Decision Analysis EUROFUSE, Pamplona, Spain (2009)
7. Martín, J., Mayor, G.: Regular multidistances. In: XV Congreso Español sobre Tecnologías y Lógica Fuzzy ESTYLF (2010)
8. Meggido, N.: Linear–time algorithms for linear programming in $\mathbb{R}^3$ and related problems. SIAM J. Comput. 12 (1983)
9. Monjardet, B.: "Mathématique Social" and Mathematics. A case study: Condorcet's efect and medians. Electronic Journal for History of Probability and Statistics 4, 1 (2008)
10. Ré, M.A., Prato, D.P., Lamberti, P.W.: Discriminación de distribuciones de probabilidad usando distancias entrópicas. Anales AFA 19 (2007)
11. Wolpert, D.H.: Metrics for more than two points at once. In: International Conference on Complex Systems ICCS, Boston (2004)
12. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE T. Syst. Man Cyb. 18 (1988)

Mixture Utility in General Insurance

Jana Špirková

Department of Quantitative Methods
and Information Systems,
Faculty of Economics,
Matej Bel University,
Tajovského 10,
975 90 Banská Bystrica, Slovakia

Abstract. Aggregation functions and utility functions belong to very interesting parts of modern decision making theory. We develop basic concept of the connection of aggregation functions theory and utility theory to determine gross annual premium in general insurance. We introduce specific values of the gross annual premium on the basis of aggregation of the person's utility functions which were determined empirically based on a short personal interview. Moreover, by specific utility function we determine minimal gross annual premium acceptable for the insurer.

Keywords: Utility function, Mixture utility operator, Risk averse, Risk loving.

1 Introduction

This paper was inspired by the books Modern Actuarial Risk Theory [2] and Actuarial Models - The Mathematics of Insurance [9] where the authors introduce a model for determination of maximal and minimal premium in insurance. Moreover, in [9] Rotar introduces many types of classical utility functions, for example, positive-power function $u(x) = x^a$ for all $x \geq 0$ and some $a > 0$; negative-power function $u(x) = \frac{1}{x^a}$ for all $x > 0$ and some $a > 0$; logarithmic function $u(x) = lnx$ for $x > 0$, etc.

However, in real life people do not behave according to theoretical utility functions. There is a psychological problem rather than a mathematical one. Seriousness and also uncertainty of respondent's answers depend on situation, on form of questions asked, on time which respondents have, and on many psychological and social factors. We can find a very interesting approach about utility functions in [3].

In our paper we introduce a possibility of the determination of the person's utility function on the basis of personal interview with virtual money. First, we develop one type of aggregation operators, so-called mixture utility operators - MU_g, generalized mixture utility operators - $GMU_{\mathbf{g}}$ and ordered generalized mixture utility operators - $OGMU_{\mathbf{g}}$. These operators represent some extension of

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 712–721, 2010.

mixture operators, generalized mixture operators and ordered generalized mixture operators [1], [4], [5], [10], [11], [12]. Moreover, we aggregate a number of the person's utility functions obtained from the interview using the mixture utility operators above. Consequently, we determine maximal value of the gross annual premium in general insurance by means of final aggregated utility function. Because it might be difficult to determine utility function for insurance company, we determine minimal premium acceptable for insurer by classic model of utility function $U(x) = lnx$.

Our paper is organized as follows: in Section 2 we recall basic properties of utility functions and their applications in insurance [2], [9]. We recall properties of mixture operators, generalized mixture operators and ordered generalized mixture operators mainly focused on their monotonicity. In Section 3 we develop MU_g operator, $GMU_{\mathbf{g}}$ operator and $OGMU_{\mathbf{g}}$ operator and sufficient conditions for their non-decreasingness. In Section 4 we introduce a specific application of the aggregated utility function on the determination of gross annual premium. Moreover, we introduce corresponding minimal premium acceptable for insurer. Finally, in Section 5 some conclusions and indications of our next investigation about mentioned topic are included.

2 Preliminaries

In this section we recall you some basic features of utility function and expected utility, too. We recall some definitions of different types of mixture operators and some sufficient conditions for their non-decreasingness.

2.1 Utility Function

Utility function may be used as a basis for describing individual approaches to risk. Three basic approaches have been characterized. Opposite cases refer to *risk loving* and *risk averse* who accepts favorable gambles only. There is risk-neutral between these two extremes. Risk-neutral behavior is typical of persons who are enormously wealthy. Many people may be both risk averse and loving, depending on the range of monetary values being considered.

The theorem below describes properties of the utility function and its expected value.

Theorem 1. *(Jensen's inequality)[2], [9] Let X be a random variable (with a finite expectation). Then, if $u(x)$ is concave,*

$$E\left[u\left(X\right)\right] \leq u\left(E\left[X\right]\right). \tag{1}$$

If $u(x)$ is convex,

$$E\left[u\left(X\right)\right] \geq u\left(E\left[X\right]\right). \tag{2}$$

Equality holds if and only if $u\left(x\right)$ is linear with respect to X or $var\left(X\right) = 0$.

In this part we illustrate whether to buy insurance by evaluating an individual's decision.

Now, suppose that our respondent has two alternatives - to buy insurance or not. Suppose that he owns a capital w and that he values wealth by the utility function u. Let's assume he is insured against a loss X for a premium P. If he is insured that means a certain alternative. This decision gives us the utility value $u\,(w-P)$. If he is not insured that means an uncertain alternative. In this case the expected utility is $E\,[u\,(w-X)]$. Based on Jensen's inequality (1) we get

$$E\,[u\,(w-X)] \leq u\,(E\,[w-X]) = u\,(w-E\,[X]) = u\,(w-P)\,. \tag{3}$$

Since utility function u is a non-decreasing continuous function, this is equivalent to $P \leq P^{max}$, where P^{max} denotes the maximum premium to be paid. This so-called *zero utility premium* is the solution to the following utility equilibrium equation

$$E\,[u\,(w-X)] = u\,(w-P^{max})\,. \tag{4}$$

The difference $(w-P^{max})$ is also called *certainty equivalent.*

The insurer with utility function $U(x)$ and capital W, with insurance of loss X for a premium P must satisfy the inequality

$$E\,[U\,(W+P-X)] \geq U\,(W)\,, \tag{5}$$

and hence for the minimal accepted premium P^{min}

$$U\,(W) = E\left[U\left(W+P^{min}-X\right)\right]. \tag{6}$$

2.2 Mixture Operator

In this part we review some mixture operators introduced in [4], [6], [8]. Suppose that each alternative $\mathbf{x}$ is characterized by a score vector $\mathbf{x} = (x_1, \ldots, x_n) \in [0,1]^n$, where $n \in N - \{1\}$ is the number of applied criteria. A mixture operator can be defined as follows:

Definition 1. *Mixture operator* $M_g : [0,1]^n \to [0,1]$ *is the arithmetic mean weighted by a continuous weighting function* $g : [0,1] \to]0,\infty[$ *given by*

$$M_g(x_1,\ldots,x_n) = \frac{\sum\limits_{i=1}^{n} g(x_i)\cdot x_i}{\sum\limits_{i=1}^{n} g(x_i)}, \tag{7}$$

where $(x_1,\ldots,x_n)$ *is an input vector.*

Observe that due to the continuity of weighting function g, each mixture operator M_g is continuous. Evidently, M_g is an idempotent operator, [6], [8].

Note that sometimes different continuous weighting functions are applied for different criteria score what leads to a generalized mixture operator, see [6], [8].

Definition 2. *Generalized mixture operator* $M_{\mathbf{g}} : [0,1]^n \to [0,1]$ *is given by*

$$M_{\mathbf{g}}(x_1, \ldots, x_n) = \frac{\sum_{i=1}^{n} g_i(x_i) \cdot x_i}{\sum_{i=1}^{n} g_i(x_i)}, \tag{8}$$

where $(x_1, \ldots, x_n)$ *is an input vector and* $\mathbf{g} = (g_1, \ldots, g_n)$ *is a vector of continuous weighting functions.*

Obviously, generalized mixture operators are continuous and idempotent.

Generalized mixture operator based on the ordinal approach can be defined as follows.

Definition 3. *Ordered generalized mixture operator* $M_{\mathbf{g}}^{'} : [0,1]^n \to [0,1]$ *is given by*

$$M_{\mathbf{g}}^{'}(x_1, \ldots, x_n) = \frac{\sum_{i=1}^{n} g_i(x_{(i)}) \cdot x_{(i)}}{\sum_{i=1}^{n} g_i(x_{(i)})}, \tag{9}$$

where $\mathbf{g} = (g_1, \ldots, g_n)$ *is a vector of continuous weighting functions and* $(x_{(1)}, \ldots, x_{(n)})$ *is a non-decreasing permutation of an input vector.*

An ordered generalized mixture operator is a generalization of an OWA operator [13], corresponding to constant weighting functions $g_i = w_i$, $w_i \in [0,1]$, $\sum_{i=1}^{n} w_i = 1$.

However, a mixture operator need not be non-decreasing. Marques-Pereira and Pasi [4] stated first sufficient condition for a weighting function g in order mixture operator (7) is non-decreasing. It can be defined as follows:

Proposition 1. *Let* $g : [0,1] \to]0, \infty[$ *be a non-decreasing smooth weighting function which satisfies the next condition:*

$$0 \leq g'(x) \leq g(x) \tag{10}$$

for all $x \in [0,1]$.

Then $M_g : [0,1]^n \to [0,1]$ *is an aggregation operator for each* $n \in N$, $n > 1$.

We have generalized sufficient condition (10) in our previous work. In the next part we recall more general sufficient conditions mentioned in [5], [12].

Proposition 2. *Let* $g : [0,1] \to]0, \infty[$ *be a non-decreasing smooth weighting function which satisfies the condition:*

$$0 \leq g'(x)(1-x) \leq g(x) \tag{11}$$

for all $x \in [0,1]$.

Then $M_g : [0,1]^n \to [0,1]$ *is an aggregation operator for each* $n \in N$, $n > 1$.

Moreover, we have improved sufficient condition (11), but constrained by n.

Proposition 3. *For a fixed $n \in N$, $n > 1$, let $g : [0,1] \to]0,\infty[$ be a non-decreasing smooth weighting function satisfying the condition:*

$$\frac{g^2(x)}{(n-1)g(1)} + g(x) \geq g'(x)(1-x) \tag{12}$$

for all $x \in [0,1]$.

Then $M_g : [0,1]^n \to [0,1]$ is an aggregation operator.

In the next proposition we introduce a sufficient condition for non-decreasingness of generalized mixture operators.

Proposition 4. *For a fixed $n \in N$, $n > 1$, $i = 1, \ldots, n$, let $g_i : [0,1] \to]0,\infty[$ be a non-decreasing smooth weighting functions, such that*

$$\frac{g_i^2(x)}{\sum\limits_{j \neq i} g_j(1)} + g_i(x) \geq g_i'(x) \cdot (1-x) \tag{13}$$

for all $x \in [0,1]$.

Then $M_{\mathbf{g}} : [0,1]^n \to [0,1]$, where $\mathbf{g} = (g_1, \ldots, g_n)$, is an aggregation operator.

3 Mixture Utility Operator

In this part we develop mixture utility operators, generalized mixture utility operators and ordered generalized mixture utility operatorss.

Suppose that each alternative $\mathbf{u}$ is characterized by utility vector $\mathbf{u} = (u_1, \ldots, u_n) \in [0,1]^n$, where $n \in N - \{1\}$ is the number of aggregated utility values.

Definition 4. *Mixture utility operator $MU_g : [0,1]^n \to [0,1]$ is the arithmetic mean weighted by a continuous weighting function $g : [0,1] \to]0,\infty[$ given by*

$$MU_g(u_1(x), \ldots, u_n(x)) = \frac{\sum\limits_{i=1}^{n} g(u_i(x)) \cdot u_i(x)}{\sum\limits_{i=1}^{n} g(u_i(x))}, \tag{14}$$

where $(u_1(x), \ldots, u_n(x))$ is a vector of utility values for fixed x, $x \in R$.

Observe that due to the continuity of weighting function g, each mixture utility function MU_g is continuous and idempotent.

On the basis of Definition 2. a generalized mixture utility operator can be defined as follows:

Definition 5. *Generalized mixture utility operator* $GMU_{\mathbf{g}} : [0,1]^n \to [0,1]$ *is given by*

$$GMU_{\mathbf{g}}(u_1(x), \ldots, u_n(x)) = \frac{\sum_{i=1}^{n} g_i(u_i(x)) \cdot u_i(x)}{\sum_{i=1}^{n} g_i(u_i(x))}, \tag{15}$$

where $(u_1(x), \ldots, u_n(x))$ *is a vector of utility values for fixed* x, $x \in R$ *and* $\mathbf{g} = (g_1, \ldots, g_n)$ *is a vector of weighting functions.*

Clearly, generalized mixture utility functions are continuous and idempotent.

Definition 6. *Ordered generalized mixture utility function* $OGMU_{\mathbf{g}} : [0,1]^n \to [0,1]$ *is given by*

$$OGMU_{\mathbf{g}}(u_1, \ldots, u_n) = \frac{\sum_{i=1}^{n} g_i(u_{(i)}(x)) \cdot u_{(i)}(x)}{\sum_{i=1}^{n} g_i(u_{(i)}(x))}, \tag{16}$$

where $\mathbf{g} = (g_1, \ldots, g_n)$ *is a vector of continuous weighting functions and* $(u_{(1)}(x), \ldots, u_{(n)}(x))$ *is a non-decreasing permutation of a vector of utility values for fixed* x, $x \in R$.

However, mixture utility operators do not have to be non-decreasing. On the basis of sufficient conditions (10), (11), (12) and (13), we can rewrite similar sufficient conditions for mixture utility functions. If operators $MU_g : [0,1]^n \to [0,1]$, $GMU_{\mathbf{g}} : [0,1]^n \to [0,1]$ and $OGMU_{\mathbf{g}} : [0,1]^n \to [0,1]$ satisfy conditions

$$0 \le g'(u(x)) \le g(u(x)) \quad \text{for all} \quad u(x) \in [0,1], \tag{17}$$

$$0 \le g'(u(x))(1-u(x)) \le g(u(x)) \quad \text{for all} \quad u(x) \in [0,1], \tag{18}$$

$$\frac{g^2(u(x))}{(n-1)g(1)} + g(u(x)) \ge g'(u(x))(1-u(x)) \quad \text{for all} \quad u(x) \in [0,1], \tag{19}$$

$$\frac{g_i^2(u(x))}{\sum_{j \ne i} g_j(1)} + g_i(u(x)) \ge g_i'(u(x)) \cdot (1-u(x)) \quad \text{for all} \quad u(x) \in [0,1] \tag{20}$$

then they are aggregation functions.

4 Maximal and Minimal Premium

In practice, the utility function can be determined empirically by a personal interview made by the decision maker. This function may be constructed from the information gleaned from the short interview. The respondent may use this function in any personal decision analysis in which the payoff falls between 0 and 30000€. We recall the interview which is compiled as follows:

Suppose that you are owner of an estate that has the possible loss of 30000€ in the future. However, you have a possibility to withdraw from this possible loss under the penalty the amount: 1000€, 5000€, 10000€, 15000€, 25000€. Your portfolio manager can provide you with information expressing the probability of the loss the 30000€.
What would be the biggest probability of the loss, to retain risk of the possible loss above?

Only a few proportioned graphic points are required. We have information (data) from three respondents. They have very similar approach to risk, that means they are risk averse for larger losses and risk loving for smaller losses. Utility function for the first respondent is determined by points $(0, 1)$, $(-1000, 0.85)$, $(-5000, 0.75)$, $(-10000, 0.60)$, $(-15000, 0.60)$, $(-25000, 0.20)$, $(-30000, 0.00)$.

From the second respondent we obtain points as follows: $(0, 1)$, $(-1000, 0.85)$, $(-5000, 0.75)$, $(-10000, 0.60)$, $(-15000, 0.50)$, $(-25000, 0.40)$, $(-30000, 0.00)$, and from the last one $(0, 1)$, $(-1000, 0.85)$, $(-5000, 0.70)$, $(-10000, 0.75)$, $(-15000, 0.60)$, $(-25000, 0.40)$, $(-30000, 0.00)$.

Observe that these utility functions are for larger losses concave and for smaller losses convex, as shown in Figure 1.

In this case we have three utility functions which we aggregate by MU_g operator and $OGMU_{\mathbf{g}}$ operator in order to get only one *collective* utility function for people with similar approach to risk.

We decided to aggregate utility values by MU_g operator because we can observe as weights rise continuously according to the utility value. Also, we aggregate individual utility values by means of $OGMU_{\mathbf{g}}$ operator to allocate higher weight to the higher utility value. That means we give the higher weight to the more risk averse utility value. Moreover, we can observe a modification of weights continuously. We use weighting function $g(u(x)) = 0.75u^2(x) + 0.25$ for aggregation with MU_g function and vector of weighting functions $\mathbf{g} = (g_1, g_2, g_3)$, where $g_1(u(x)) = 0.75u^2(x) + 0.25$, $g_2(u(x)) = 0.5u^2(x) + 0.5$, $g_3(u(x)) = 0.2u^2(x) + 0.8$ for aggregation with $OGMU_{\mathbf{g}}$ operator. All used weighting functions fulfill monotonicity conditions (17) - (20) .

We decided to use such aggregation in order to see and be able to investigate weight by means of continuous mixture utility function.

Different person's utility functions for our three respondents and mixture utility functions were created by the SPSS system and they are presented in Table 1. This table also provides the values of Adjusted R square, F statistics and significance level for individual utility functions and for aggregated mixture utility operators.

Remark 1. Expected utility is calculated by the well-known formula

$$E\left[u\left(X\right)\right] = \sum_{i=1}^{n} u(x_i) \cdot p_i, \tag{21}$$

where $X = (x_1, x_2, \ldots, x_n)$ is a vector of the possible alternatives and p_i is the probability of alternative x_i.

Expected utilities can be calculated by linear function, too which is determined uniquely by points $[-30000, , u(-30000)]$ and $[0, u(0)]$. In both cases we get the same values of the expected utilities.

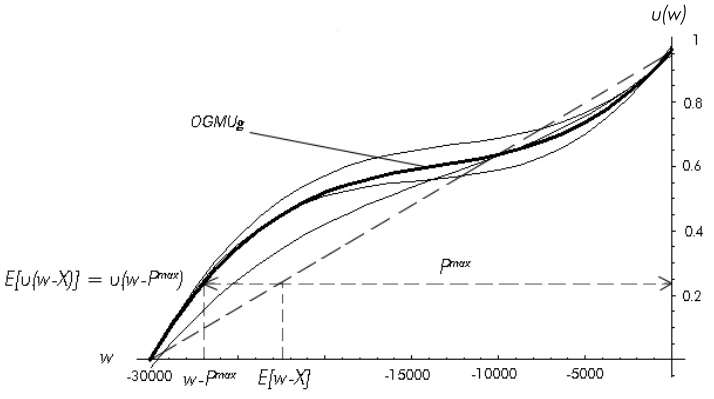


Fig. 1. Utility functions of selected respondents (functions from Table 1)

Table 1. Person's utility functions and mixture utility functions

Utility Function	Adjusted R square	F	Sig.
$u_1(x) = 5.549 \cdot 10^{-14}x^3 + 2.183 \cdot 10^{-9}x^2 + 4.784 \cdot 10^{-5}x + 0.948$	0.964	54.379	0.004
$u_2(x) = 1.191 \cdot 10^{-13}x^3 + 5.108 \cdot 10^{-9}x^2 + 7.801 \cdot 10^{-5}x + 0.977$	0.966	57.871	0.004
$u_3(x) = 1.003 \cdot 10^{-13}x^3 + 3.762 \cdot 10^{-9}x^2 + 5.425 \cdot 10^{-5}x + 0.955$	0.965	55.515	0.004
$MU_g(x) = 9.301 \cdot 10^{-14}x^3 + 3.726 \cdot 10^{-9}x^2 + 6.004 \cdot 10^{-5}x + 0.960$	1	1.698E6	0.000
$OGMU_{\mathbf{g}}(x) = 9.969 \cdot 10^{-14}x^3 + 4.025 \cdot 10^{-9}x^2 + 6.281 \cdot 10^{-5}x + 0.962$	1	1.030E4	0.000

In the figure below are the individual utility functions and final $OGMU\mathbf{g}$ function created by the SPSS system.

We determine minimal premium by the means of (6) with respect to utility function for insurer $U(x) = lnx$ with his basic capital $W = 2655513.51€$ and loss $X = 30000€$.

Equation (6) can be rewritten as follows:

$$U\left(W\right) = p \cdot U(W + P^{min} - X) + (1 - p) \cdot U(W + P^{min}), \tag{22}$$

Table 2. Expected Utility, Maximal and Minimal Premium

probability of loss p	$E[MU_g]$	P^{max} with respect to MU_g	$E[OGMU_g]$	P^{max} with respect to $OGMU_{\mathbf{g}}$	P^{min}
0.00	0.960000	0.00	0.962000	0.00	0.00
0.01	0.950409	161.35	0.952466	153.29	301.69
0.05	0.912047	841.73	0.914329	799.08	1508.10
0.1	0.864093	1786.65	0.866657	1694.17	3015.34
0.2	0.768186	4155.05	0.771314	3928.81	6027.24
0.3	0.672279	7904.69	0.675971	7463.95	9035.69
0.4	0.576372	15656.20	0.580628	15854.90	12040.70
0.5	0.480465	21018.60	0.485285	21358.90	15042.40
0.6	0.384558	23877.00	0.389942	24116.10	18040.60
0.7	0.288651	25896.30	0.294599	26055.30	21035.50
0.8	0.192744	27494.30	0.199256	27590.20	24027.00
0.9	0.096837	28835.00	0.103913	28879.00	27015.20
1.0	0.00093	30000.00	0.008570	30000.00	30000.00

and hence

$$W = (W + P^{min} - X)^p \cdot (W + P^{min})^{(1-p)}. \tag{23}$$

We determine individual minimal premiums with corresponding probability by the Mathematica 5 system.

Remark 2. Utility functions are used to compare investments mutually. For this reason, we can scale a utility function by multiplying it by any positive constant and/or translate it by adding any other constant(positive or negative). This kind of transformation is called a positive affine transformation. All our results would be the same.

5 Conclusion

We have shown how to construct person's utility function and we have calculated maximal premium for loss of 30000€ according to the mixture utility function. On the basis of mixture utility function and ordered generalized mixture utility function we have determined maximal premium for persons having approximately similar approach to risk. Another utility function would be required if evaluating a decision with more extreme payoffs or if our respondent's attitudes change because of a new job or lifestyle. Moreover, the utility function must be revised from time to time. In our future work we would like to investigate mixture utility operators with other weighting functions, insurer's utility function and to extend and in more detail investigate mentioned model with different interviews and fictive games.

Acknowledgments. This work was supported by grant VEGA 1/0539/08.

References

1. Calvo, T., Mesiar, R., Yager, R.R.: Quantitative Weights and Aggregation, Fellow. IEEE Transaction on fuzzy systems 2(1), 62–69 (2004)
2. Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M.: Modern Actuarial Risk Theory. Kluwer Academic Publishers, Boston (2001)
3. Khurshid, A.: The end of rationality? Abstracts. In: Tenth International Conference on Fuzzy Set Theory and Applications, FSTA 2010, Slovak Republic, pp. 10–11 (2010)
4. Marques-Pereira, R.A., Pasi, G.: On non-monotonic aggregation: Mixture operators. In: Proceedings of the 4th Meeting of the EURO Working Group on Fuzzy Sets (EUROFUSE 1999) and 2nd International Conference on Soft and Intelligent Computing (SIC 1999), Budapest, Hungary, pp. 513–517 (1999)
5. Mesiar, R., Špirková, J.: Weighted means and weighting functions. Kybernetika 42(2), 151–160 (2006)
6. Ribeiro, R.A., Marques-Pereira, R.A.: Generalized mixture operators using weighting functions: A comparative study with WA and OWA. European Journal of Operational Research 145, 329–342 (2003)
7. Ribeiro, R.A., Marques-Pereira, R.A.: Weights as functions of attribute satisfaction values. In: Proceedings of the Workshop on Preference Modelling and Applications (EUROFUSE), Granada, Spain, pp. 131–137 (2001)
8. Ribeiro, R.A., Marques-Pereira, R.A.: Aggregation with generalized mixture operators using weighting functions. Fuzzy Sets and Systems 137, 43–58 (2003)
9. Rotar, V.I.: Actuarial Models - The Mathematics of Insurance. Chapman & Hall/CRC Press, Boca Raton (2007)
10. Špirková, J.: Weighting Functions For Aggregation Operator. In: 3rd International Summer School On Aggregation Operators, Universita Della Svizzera Italiana, Lugano, Switzerland, pp. 127–130 (2005)
11. Špirková, J.: Mixture and quasi-mixture operators. In: IPMU 2006, France, pp. 603–608 (2006)
12. Špirková, J.: Dissertation thesis, Weighted aggregation operators and their applications, Bratislava (2008)
13. Yager, R.R.: Generalized OWA Aggregation Operators. Fuzzy Optimization and Decision Making 3, 93–107 (2004)

Application of Evolutionary Algorithms to the Optimization of the Flame Position in Coal-Fired Utility Steam Generators

W. Kästner[1], R. Hampel[1], T. Förster[1], M. Freund[1], M. Wagenknecht[1] D. Haake[2], H. Kanisch[2], U.-S. Altmann[3], and F. Müller[3]

[1] University of Applied Sciences Zittau/Goerlitz, IPM, Germany
[2] Vattenfall Europe Generation AG, Cottbus, Germany
[3] CombTec GmbH, Zittau, Germany

Abstract. The construction and operation of modern power station units on coal base are subject to considerable economic, ecological and political constraints enforcing the system's elements to operate at their technical upper limits. The control and prevention of fuel-caused harmful effects is crucial for profits and losses with respect to availability, efficiency and maintenance expenses. Advanced power plant technologies are an indefensible high-risk without adequate developed monitoring and optimization systems. A real-time optimization method of the flame position in a multi-burner firing system that is affected by corrosion processes at the evaporative heating surfaces is presented. The base is a model of the flame position generated by artificial neural networks. Evolutionary algorithms are used to optimize the fuel distribution which depends on the feeding of the coal dust burners. By analyzing the current plant status and technical restrictions for the respective components the solution set is reduced to a single solution which is realized as control action.

Keywords: combustion chamber, corrosion, optimization, artificial neural network, evolutionary algorithm.

1 Introduction

1.1 Background

To meet the technical, economical and ecological challenges in the operation of a conventional power plant a modern process monitoring and optimal process controlling are essential. The combustion chamber as the central and most complex element of a coal-fired power plant is of particular interest. Its functioning is decisive to meet performance criteria like efficiency, pollutant emissions, component load or resource management. The control and reduction of high-temperature corrosion at the evaporative heating surfaces are of crucial importance. By symmetric positioning the flame, damage-causing processes in the wall area like particle composition, oxygen deficiency and thermal load can be reduced [2]. At present the operator can control the flame position in the combustion chamber by trimming the mill feeders (being the correcting variable for the burner firing rate).

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 722–730, 2010.

1.2 Reference Unit and Operational Measuring Technique

The reference steam generator in the Jänschwalde power plant is characterized by a two-point firing (Fig. 1, left hand). Under normal conditions, five of the six mills which blow the pulverised brown coal via main and vapour burners into the combustion chamber are operating [3], one stays in reserve or is under maintenance. Hence, for a uniform split-up of the overall fuel quantity to the mills the flame position in the combustion chamber is asymmetric. The asymmetry is the main cause for high-temperature corrosion (Fig. 1, right hand). By an appropriate distribution of the fuel (resulting in a distribution of the respective air quantity to the operating burners) a shift of the flame to the centre of the combustion chamber can be performed. Hereby, the current plant status restricts the degrees of freedom for the optimization.

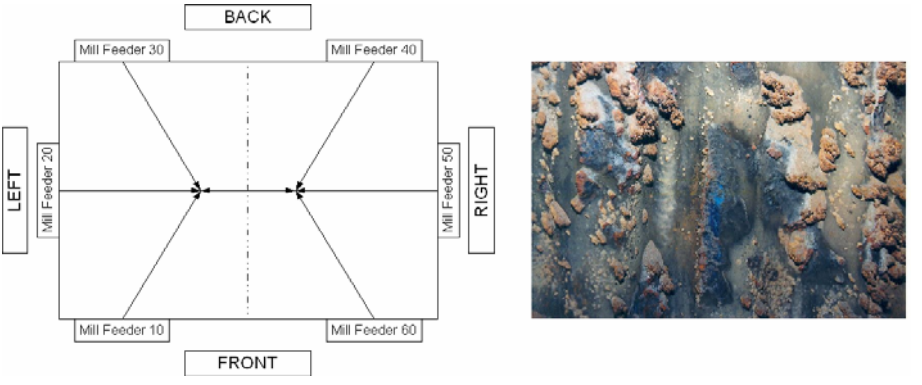


Fig. 1. Schematic depiction of the two-point firing (left hand) and the evaporative heating surface with corrosion (right hand)

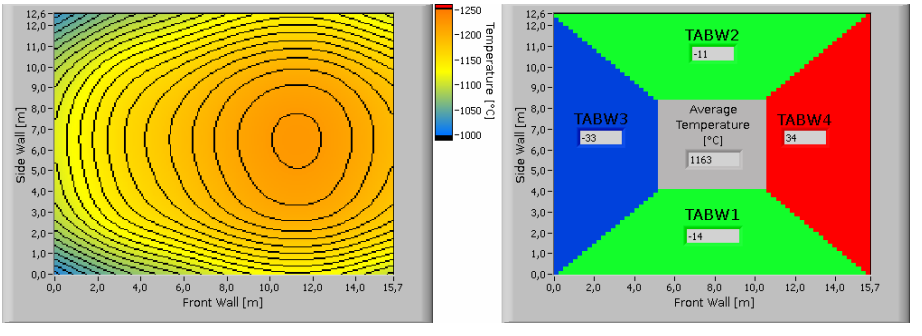


Fig. 2. Isothermal line display of the temperature distribution (left hand) and abstracted 4-Zone-Trapezoidal-Display (right hand)

For a correct estimation of the current flame position the measurements of the installed diagnostic system CombPyr-2D$^+$ established by the CombTec Company [1] were used. At this a temperature distribution over the furnace cross-sectional area is determined using 12 pyrometers at a level above the burner belt. This distribution can

be depicted by the 4-Zone-Trapezoidal-Display (Fig. 2). The 4 zones under consideration show the difference between the mean zone temperatures ($TABW_{1...4}$) and the overall cross section mean. The unsymmetrical flame position to be optimized can be described by these 4 values in qualitative and quantitative ways.

2 Process Modelling

The physical processes of the local effects are rather complex and an analytic description for an on-line application is not possible. Thus, a process model based on artificial neural networks (ANN) was built [8]. In doing so, expert knowledge [4] and operational measuring data were used. The 4 zone deviations determined by the CombPyr system are modelled using the parameters considered as significant for the flame positioning (amongst others the mass and volume flows, coal quality attributes, mill temperatures after classifier, oxygen regime). For the ANN a multilayer perceptron technique [5] was used. This was of special advantage, because no apriori known functional dependencies could be used (as mentioned above). They had to be approximated by characteristic process patterns (learn data). Connecting the inputs and outputs via hidden layers and weight matrices you get an abstract and generalization able model. Figure 3 shows a principal network structure. To improve the model quality, for each combination of 5 operating mills an adapted ANN was created.

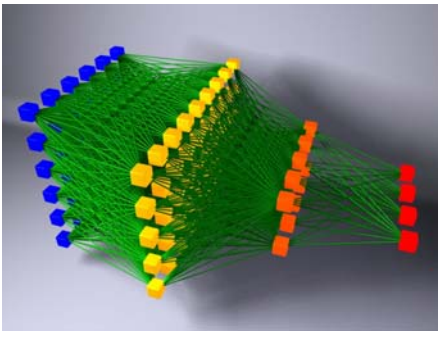


Fig. 3. Example of an ANN network structure

3 Process Optimization

3.1 Optimization with Evolutionary Algorithms

Evolutionary algorithms (EA) are optimization methods which are based on evolutionary processes in the nature [6]. The performance criteria arguments are coded as individuals (solution candidates). The values of the performance criteria (respectively of the fitness function) evaluate the individual's quality. By the fitness a selection is performed and unsuited individuals are dropped. By crossover and mutation the individuals are varied. Thus, a goal-oriented search to get appropriate solutions can be realized.

Figure 4 shows the principal structure of an EA. After initializing a starting population of individuals and their evaluation by the performance criteria and fitness function the artificial evolution is realized by a loop improving the individuals iteratively. From the population parents are selected and new children are born by crossover and mutation with corresponding evaluations. By a replacement strategy the new population is built from the previous (parents and new born children). If a stop criterion is fulfilled the evolution is terminated and the found individuals are output.

The main advantage of EA is the fact that they do not require a special structure of the optimization problem. Particularly, no gradients are necessary. Thus, rather general problems (nonlinear, discontinuous, multimodal) can be solved. In the projects context it is of special interest that they are also suited for multicriterial problems [7].

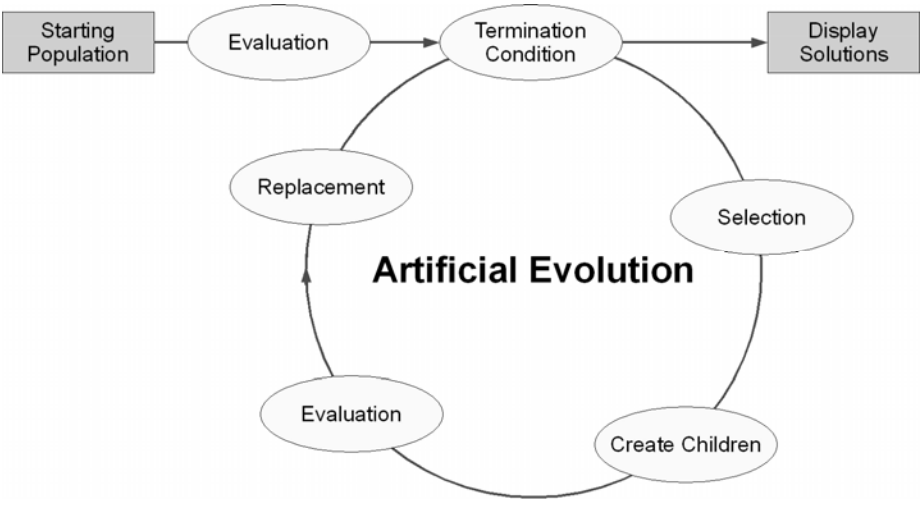


Fig. 4. Principle of artificial evolution

3.2 The Optimization Problem

The speeds of the mill feeders $DRZ_{10...60}$ are the variables to be optimized. They are proportional to the coal quantity and hence to the burner firing rate. From the technological and unit-dependent general requirements in the operating point you get the restrictions for the optimization [8]. As limiting factors the changing coal quality, the wear of the coal mills and the current pollutant emission are mentioned. This results in lower and upper bounds for the speeds of the mill feeder which are derived from measurements or given by the operator. These bounds are strict (1).

$$DRZ_{i,\min} \le DRZ_{i,\text{opt}} \le DRZ_{i,\max} \quad . \tag{1}$$

The overall fuel quantity of the steam generator has to be maintained by the single individuals as well (2).

$$\sum_{i=10...60} DRZ_i = const \quad . \tag{2}$$

The ANN inputs which are coupled with the mill feeder speeds by the instrumentation and control system are adapted to the changed speeds by the implementation of the control characteristics [3] in order to enable a correct prediction and evaluation of the changed flame position. The remaining inputs are assumed to be kept constant by trimming (accounting the large delay times of the control processes).

The performance criteria are constituted by the temperature deviations in the zones of the combustion chamber $TABW_{1\ldots4}$. A central flame position is characterized by an even temperature deviation in the opposite zones. Hence, the algorithm aims at minimizing the differences of temperature deviations between zones 1 and 2 (ΔT_1) and respectively zones 3 and 4 (ΔT_2) in the trapezoid model (Fig.2). The corresponding bicriterial problem is given by (3).

$$\begin{aligned} \Delta T_1 &= \left| TABW_1 - TABW_2 \right| \rightarrow \min \\ \Delta T_2 &= \left| TABW_3 - TABW_4 \right| \rightarrow \min \end{aligned} \quad . \tag{3}$$

3.3 The Optimization Algorithm

For the optimization the created ANN models are at disposal. They yield nonlinear and multimodal criteria. The time for the calculation of the criteria for one individual was 1.6 milliseconds. Hence, with a stipulated optimization-duration of maximum 20 seconds, online application is possible (supposing an effective implementation of the algorithm). The developed optimization algorithm approximates the Pareto set of the given bicriterial problem, therefore it indicates a Pareto EA. To keep the computation time small some simplifications were realized. No selection of parents is made. Actually all individuals are used to create children who are built merely by mutation. The crossover is omitted, since it did not show a significant improvement.

The individual c contains the values of the variables, performance criteria and fitness. For the function evaluations the ANN are used and for the solution candidates (3) is applied. Afterwards, the fitness $fitness^*(c)$ is calculated by Pareto ranking (Fig. 5). It has the advantage (over varying the weights of a scalar surrogate criterion) of a computing time reduction. In the ranking process, all individuals get the initial fitness 0. Then they are compared. If an individual c_2 has increased values for both criteria in comparison with an individual c_1 it is dominated by the latter and gets a higher fitness value (i.e., a lower fitness value corresponds to a higher quality). If an individual c_3 has a smaller value for one criterion and a larger for the other one with respect to individual c_1, both individuals are incomparable and equipollent. In this case, the fitness values of c_1 and c_3 remain unchanged.

In correspondence to the population size *POPSIZE*, copies of the current plant status (ANN inputs) are merged in one population. The plant status itself is saved at position 0 of the population (*Population*[0]) and will not be changed by the evolution process. This ensures the access to the current status. Uniform mutation was applied. For each mill feeder speed $DRZ_{10\ldots60}$ it is separately decided whether a variation is necessary. To it, a uniformly distributed random number z corresponding to the mutation rate *MUTRATE* is determined (4).

$$z \in [-MUTRATE; MUTRATE] \ . \tag{4}$$

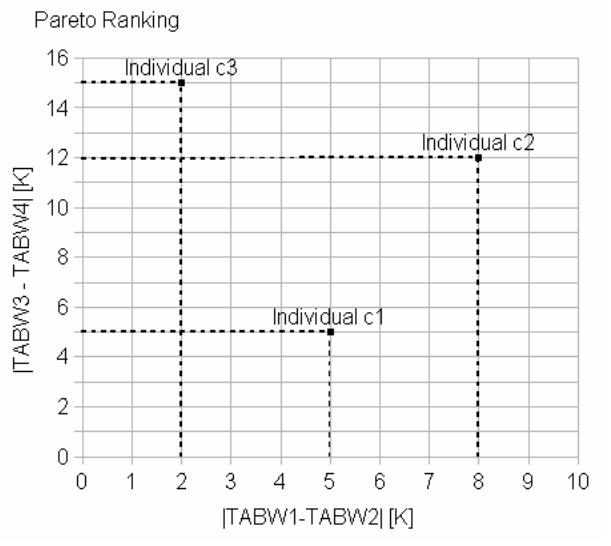


Fig. 5. Principle of Pareto-Ranking

By the mutation only the constant sum of the mill feeder speeds (2) is not ensured. Hence, a repair function has to be used. Each mill feeder of the individual is modified by (5).

$$DRZ_{10...60}(c) = \frac{\sum_{i=10...60} DRZ_i(Population[0]) - DRZ_i(c)}{5} \quad . \tag{5}$$

For the creation of the starting population for the next iteration, the individual's fitness is varied by a niching method (fitness sharing). To it, all individuals are compared with respect to their variable values. As similarity measure the Manhattan distance d_{Man} is used (6).

$$d_{\text{Man}}(c;c_i) \sum_{j=10...60} \left| DRZ_j(c) - DRZ_j(c_i) \right| \quad . \tag{6}$$

If individuals have similar speeds (using a fixed threshold *DIST*), the fitness of the individual with higher fitness is further increased. The fitness of the individual with lower fitness remains unchanged. Therefore, individuals in one niche with higher fitness are punished and get a lower initial probability for the next population. For each pair of individuals a sharing value *sh* is determined by (7).

$$sh(c;c_i) = \begin{cases} DIST - d_{Man}(c;c_i) & \text{wenn} \quad d_{Man}(c;c_i) < DIST \\ 0 & \text{sonst} \end{cases} \quad . \tag{7}$$

The presented sharing-operator prevents the concentration of the population in only few niches and uses the explorative character of the algorithm investigating the target area. The fitness *fitness*(*c*) is given by the following (8).

$$fitness(c) = fitness^*(c) + \sum_{\substack{c_i \in Population \text{ mit} \\ fitness^*(c_i) < fitness^*(c)}} sh(c;c_i)^2 \quad . \tag{8}$$

The replacement strategy consists in the selection of individuals with the lowest fitness from the set of solution candidates of the previous population and the current children. The maximal number of iterations is given by *MAXGEN*.

The values for the population size *POPSIZE*, the maximal number of iterations *MAXGEN*, the mutation rate *MUTRATE* and the sharing distance *DIST* were determined by statistical experiments with different parameter values where the convergence of the population and a small computing time of the algorithm served as criteria.

3.4 The Optimization Output

The algorithm yields 80 solution candidates (individuals) which approximate the Pareto set. The shape of the set may vary with regard to the operational and plant status. In Figure 6 exemplarily two results are depicted. The left-hand graphic shows a good approximation of the Pareto set. The approximation in the second case is altered for a different (more complex) plant status.

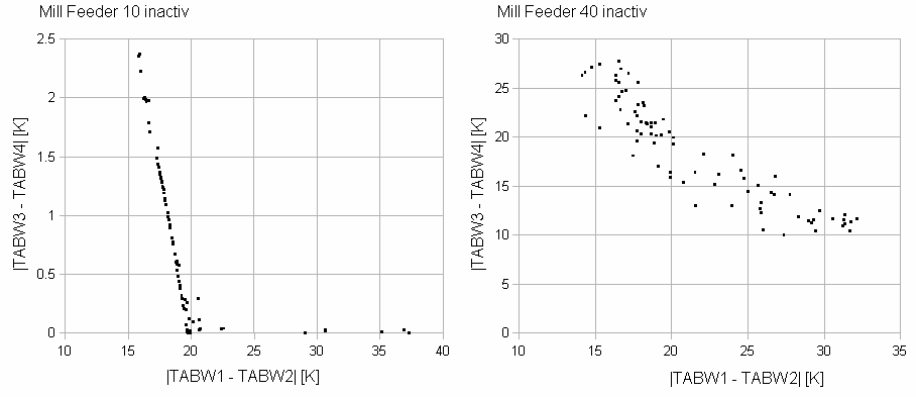


Fig. 6. Results of the optimization

4 Choice of a Final Solution

The reduction of the solution set provided by the EA to a solution matching best the current plant status is carried out by appropriate indicators. These indicators can be weighted with respect to given operational priorities. The smallest overall evaluation (weighted sum) characterizes the most preferable solution to be realized. The indicators evaluate not only the fitness of the candidate, but also the complex firing based and operational consequences. Among others, the expected pollution, the necessary mill loads and the resulting dynamic of the flame must be taken into account. Since different solutions can lead to similar flame positions, a solution can be found that is

optimal concerning the current state. Alternatively, a re-computation or re-evaluation may be necessary to new situations (e.g. the steady varying operating parameters). This leads to a drastic unburden of the operating personnel and to an objectifying of the decision-making process.

5 Functional Verification

The optimization algorithm (including the ANN) and the solution choice module have been implemented in a software system. It also realizes the data processing, the plant state diagnosis and the communication with the operator. In several test runs the functionality of the developed method and the capability of the EA with respect to the improvement of the flame position could be verified at the reference unit [8]. Figure 7 shows the temperature distribution before (upper graphics) and after (lower graphics) the optimization process. An extended version of the program including an automatic mill feeder trimming with an integration of the control signals in the instrumentation and control system is being realized and will be tested in the first half of 2010.

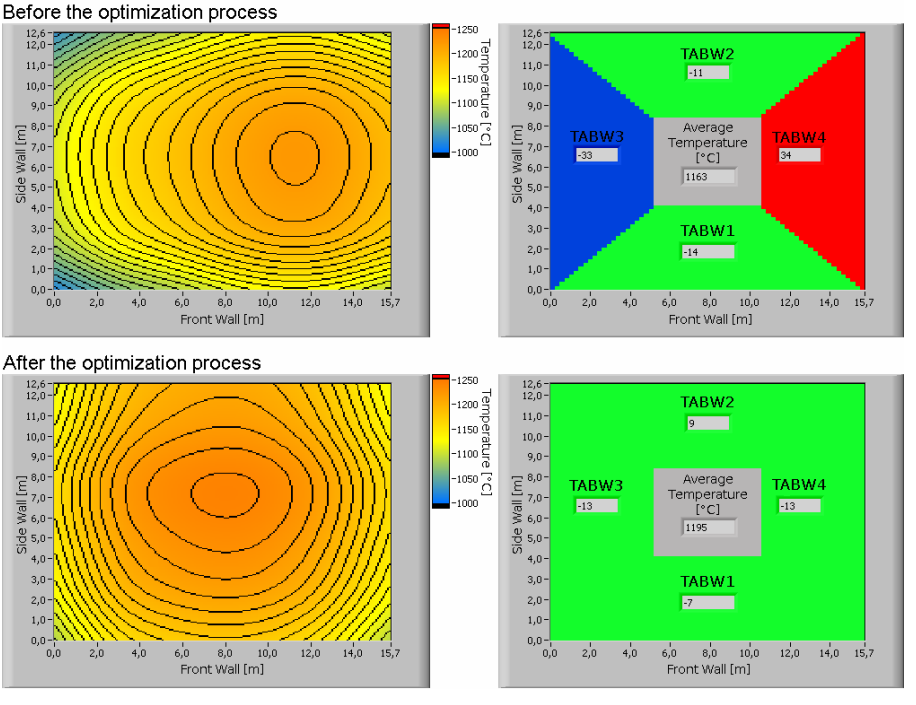


Fig. 7. Example of a realized solution

Acknowledgement. This research project was funded by the Federal Ministry of Economics and Technology of the FRG (BMWi grant no. 1718X07).

References

1. Altmann, U.-S., Förster, T.: Entwicklung eines Feuerungsdiagnosesystems für Kraftwerks-Dampferzeuger unter Anwendung von Soft-Computing-Verfahren (Development of a firing diagnosis system for utility steam generator using soft-computing methods). Final Report (aFuE-project, FMER grant no. 1744X04), Univ. Appl. Sc. Zittau/Goerlitz, IPM, Zittau (2007)
2. Altmann, U.-S., Grusla, S., Müller, F., Marschner, C., Bischoff, H., Sturm, A.: Verfahren zur Beurteilung des Ansatzbildungspotentials von Kraftwerkskohlen und Gewährleistung einer anlagenschonenden Feuerführung (Evaluation methods of the slag formation potential for power plant coal and a guaranteed equipment friendly firing). In: Proceedings 40th Kraftwerkstechnisches Kolloquium 2008, vol. 2, pp. 103–114. TU Dresden, Dresden (2008)
3. Vattenfall Europe: User Manual for the Jänschwalde Power Plant (BV 500). Issue 04/2004. Vattenfall Europe (2004)
4. Müller, F.: Diagnose des Betriebszustandes eines Dampferzeugers (Diagnosis of the operational status of a steam generator. Diploma Thesis, CombTec GmbH/Univ. Appl. Sc. Zittau/Goerlitz, IPM, Zittau (2008)
5. Förster, T.: Untersuchungen zu hierarchischen Netzstrukturen mit Multilayer Perzeptren (Analysis of hierarchical net structures with multilayer perceptrons). Univ. Appl. Sc. Zittau/Goerlitz, IPM, Zittau (2005)
6. Weicker, K.: Evolutionäre Algorithmen (Evolutionary Algorithms). B. G. Teubner Verlag/ GWV Fachverlage GmbH, Wiesbaden (2007)
7. Freund, M.: Entwicklung Genetischer Algorithmen zur mehrkriteriellen Optimierung einer Braunkohlestaubfeuerung (Development of genetic algorithms for the multicriterial optimization of a pulverised brown coal firing). Diploma Thesis, Univ. Appl. Sc. Zittau/Goerlitz, IPM, Zittau (2008)
8. Hampel, R., Förster, T., Freund, M., Kästner, W.: Modell- und Regelbasierte Prozessführung der Feuerlage und Verbrennungsgüte in Großdampferzeugern (Model- and rule-based process control of the flame position and combustion efficiency in utility steam generators). Final Report (aFuE-project, FMET grant no. 1718X07), Univ. Appl. Sc. Zittau/Goerlitz, IPM, Zittau (2009)

Measurement of Ground-Neutral Currents in Three Phase Transformers Using a Genetically Evolved Shaping Filter

Luciano Sánchez[1] and Inés Couso[2]

[1] Universidad de Oviedo, Departamento de Informática
luciano@uniovi.es
[2] Universidad de Oviedo, Departamento de Estadística e I.O. y D.M.
couso@uniovi.es

Abstract. Measuring the current in the neutral-grounding resistor is needed for monitoring resistance-grounded three phase transformers. This current is limited to hundreds of amperes in case of a fault, and are almost negligible otherwise. The current transformer that senses the current must be rated for the fault conditions, thus it is difficult to obtain a precise measurement of the current when there is not a ground-fault in the system.

In this paper we propose a computer-based method for filtering the output of the current transformer and improving its accuracy for small currents. This processing is complicated, as the amount of noise is very high, and this noise is strongly correlated with the useful signal. We propose to use Kalman filtering, based on a model of the system, and augment the state of this model with a shaping filter, whose frequency response, when fed with white Gaussian noise, reproduces our measurements of the ambient noise. In particular, since the Power Spectral Density (PSD) of the noise changes with time, we propose to use a possibilistic description of the PSD of the noise, and search for a model whose PSD is between the soft margins defined by the possibilistic model. We will use a state-space based representation and a genetic algorithm, guided by a fuzzy fitness function, for evolving the shaping filter that best matches the ambient noise. The proposed method has been evaluated with field data captured at a 130KV substation transformer at La Corredoria (Asturias, Spain).

1 Introduction: Ground-Neutral Monitoring Systems

Neutral-grounding resistors (NGR) provide current for ground-fault detection and selective coordination. Its failure is dangerous, as it converts a resistance-grounded system into an ungrounded system. While this is not a common problem in Europe, the probability of the NGR becoming open is higher in those parts of the world where these elements are subject to higher thermal stresses and continuous monitoring of the NGR is needed. Else, there is no indication that the system has become ungrounded, and the risk of transient overvoltages or the presence of voltages on otherwise safe conductors exists.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 731–740, 2010.

In Figure 1 a 130KV tree phase substation transformer and its NGR (yellow circle in the left part of the figure) are shown. The NGR is connected to the neutral of the transformer (center part of the same figure) and to earth (right part). This particular resistor is designed for limiting the current to 500A in case of a fault. A suitable current transformer (CT) measures the current through the NGR and its output is constantly monitored.

Fig. 1. La Corredoria substation transformer and its NGR (left part). Detail of the connection between the NGR and the neutral (center) and ground (right), showing the current transformer used in this paper.

The absence of current at the CT means that either the transformer is balanced or the NGR is open, thus the monitoring of this current does not prevent by itself the risks mentioned before. A NGR monitoring system involves more elements [8,11], as shown in Figure 2 (this system will be further explained in Section 2). In this schema, a resistor is used for measuring the voltage at the neutral of the transformer; this value, divided by the current through the NGR, is the ohmic value of the NGR, as desired.

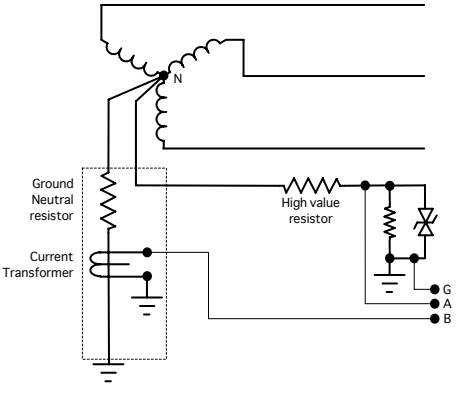


Fig. 2. Schema of a NGR monitoring system

There is still a caveat with this scheme: in case the system is perfectly balanced, no current flows by the NGR and the voltage at the neutral is also null, thus we cannot tell whether the NGR is open or not. There are two possibilities for overcoming this problem:

- We can inject DC or low frequency current into the NGR path and measure the increase in the voltage of the neutral, or
- We can amplify the output of the current transformer and measure the residual current. We expect that this residual current has components in the first odd harmonics, because of the stray capacities of the line, and also at very high frequencies ($>$ 1KHz) because of the use of Power Line Carrier [3], thus it is potentially usable for detecting open NGRs.

In this study we will explore the second option. The main difficulty of this approach is in the level of amplification need, which will be affected by the important ambient noise near a substation transformer. That is to say, we need to separate the measurement of the actual current from those stray currents induced in the transmission line. The problem is, these stray currents are strongly correlated with the signal, and their Power Spectral Density (PSD) also varies with time within certain limits.

We have decided to make a Kalman filtering of the signal, taking into account the spectrum of the noise by mean of a shaping filter [5] that is designed to produce a signal with the same PSD as the ambient noise. This shaping filter will be used to augment the state of the system, as we will explain in Section 2. For obtaining the model of such a shaping filter we propose to use Genetic Algorithms (see Section 3). It is remarked that the novelty of this approach, apart from the definition of specialized genetic operators and the representation of the individuals, is in the fact that the desired PSD is not completely known, thus the GA will be guided by a fuzzy fitness function. In Section 4 we show the results of the application of this system to real-world data. The paper finishes with the concluding remarks in Section 5.

2 Description of the Measurement System

The measuring system that will be analyzed in this paper is shown in the left part of Figure 3. The output of the current transformer is connected to a resistive load R_L (the burden resistor of the current transformer has not been shown in the schema). Alas, if the cable is not perfectly shielded then there will be stray currents flowing through it; we have represented them as a coupled coil L_{mutual} in the figure. When studying the noise, we replace the current transformer by a resistance of high value, and measure the voltage dropped in R_L. The voltage drop in R_L depends on the sum of both currents: the current in the secondary of the CT and the stray currents we have mentioned.

According to our own experimentation, at low frequencies, the current in the NGR is comprised of the first odd harmonics (50Hz, 150Hz, 250Hz and 350Hz), while at high frequencies the contribution of the DLC produces a more complex

profile. In this work we limit ourselves to low frequencies, and propose that the current y_C by the NGR is modeled by the state space discrete system

$$x_C(kh+h) = \Phi_C x_N(kh) \tag{1}$$
$$y_C(xk) = C_C x_C(kh) \tag{2}$$

where Φ_C will have all its poles at the unit circle, at frequencies that match the mentioned harmonics. Observe that such a system is unstable and the input is not needed.

Secondly, we want to devise a similar model for the noise, and combine it with the preceding model as shown in the right part of Figure 3. The noise model is a shaping filter whose input is white Gaussian noise and whose output mimics the frequency behavior of the noise, measured as described before. The model of the shaping filter will be expressed in observable canonical form, because of reasons that will be made clear in Section 3:

$$x_N(kh+h) = \Phi_N x_N(kh) + \Gamma_N v(k) \tag{3}$$
$$y_N(xk) = C_N x_N(kh) \tag{4}$$

where $v(k)$ is Gaussian white noise with mean 0 and variance 1, and $C_N = (1, \ldots, 0)$.

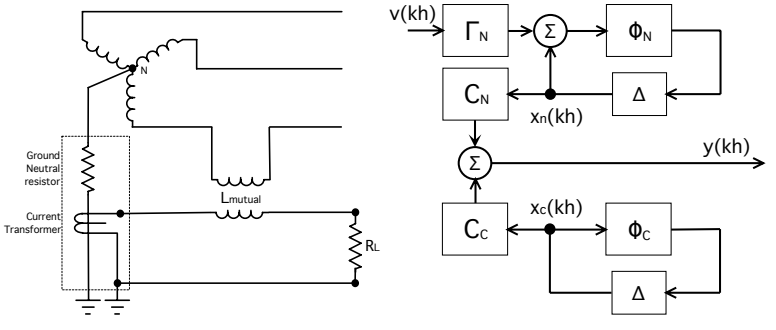


Fig. 3. Block diagram of the augmented system

The augmented system (see Figure 3) is

$$x(kh+h) = \Phi x(kh) + e(k) \tag{5}$$
$$y(xk) = C_N x_N(kh) \tag{6}$$

where $x = (x_C^T | x_N^T)^T$, and $e(k)$ is white noise with covariance R_1

$$R_1 = \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & \Gamma_N \Gamma_N^T \end{array}\right) \tag{7}$$

and the combination of the two state space matrices we have proposed is

$$\Phi = \left(\begin{array}{c|c} \Phi_C & 0 \\ \hline 0 & \Phi_N \end{array}\right) = \left(\begin{array}{cc|c} \begin{matrix}\cos 2\pi f_1 h & \sin 2\pi f_1 h \\ -\sin 2\pi f_1 h & \cos 2\pi f_1 h\end{matrix} & 0 & \\ & \vdots & 0 \\ 0 & \begin{matrix}\cos 2\pi f_4 h & \sin 2\pi f_4 h \\ -\sin 2\pi f_4 h & \cos 2\pi f_4 h\end{matrix} & \\ \hline 0 & & \begin{matrix} c_1\ 1\ 0 \cdots 0\ 0 \\ c_2\ 0\ 1 \cdots 0\ 0 \\ \vdots \\ c_8\ 0\ 0 \cdots 0\ 0 \end{matrix} \end{array}\right) \tag{8}$$

$$C = (C_C \mid C_N) = (C_C \mid 1 \ldots 0) \tag{9}$$

thus particularizing the Kalman filter [5] to this case, we obtain the equations

$$\hat{x}(k+1 \mid k) = \Phi\hat{x}(k \mid k-1) + K(k)(Y(k) - C\hat{x}(k \mid k-1)) \tag{10}$$

$$K(k) = \Phi P(k) C^T (CP(k)C^T)^{-1} \tag{11}$$

$$P(k+1) = \Phi P(k)\Phi^T + R_1 - \Phi P(k) C^T (CP(k)C^T)^{-1} CP(k)\Phi^T. \tag{12}$$

3 Genetic Optimization of a Shaping Filter

This section describes the use of Genetic Algorithms for evolving the three matrices Φ_N, Γ_N and C_N described before. Observe that we need to restrict our search to stable systems, i.e. to those matrices Φ_N whose eigenvalues are in the unit circle. We want that the PSD of this system, when the input $v(k)$ is Gaussian white noise, is similar to the PSD of the ambient noise.

There are two challenges in this search: (1) the PSD is time-varying, and (2) it is not trivial to determine the complex eigenvalues of a general matrix [9]. We have solved them as follows:

1. We have divided the spectrum of frequencies between 50 and 500Hz in 10 bands, and computed five confidence intervals at levels 0.50, 0.25, 0.10, 0.05, 0.01 for the energy of the noise for each band. These five intervals are regarded as α-cuts of fuzzy values describing our knowledge about the PSD of the noise.
2. We do not generate random matrices and discard those corresponding to unstable processes. We represent instead Φ_N by means of the roots of its characteristic polynomial, and use the observable canonical form of the system.

In the following paragraphs we discuss both subjects: the computation of the PSD and the representation of the matrices that comprise the model of the noise.

3.1 Fuzzy PSD

Let $y_N(kh)$ the output of the noise model, when the input is $v(kh)$, white Gaussian noise. Let us assume a zero-th order hold ($y_N(t)$ constant in $t \in [kh, kh+h)$) and let the autocorrelation of this time series be

$$E(y_N(t) \cdot y_N(t+\tau)) = \psi_{y_N}(\tau). \tag{13}$$

Then, the PSD of the noise is the Fourier transform of the autocorrelation,

$$\Psi_x(\omega) = \int_{-\infty}^{\infty} \psi_{y_N}(\tau) e^{-j\omega\tau} d\tau \tag{14}$$

and the energy of the band $[\omega_k - \Delta/2, \omega_k + \Delta/2]$ is

$$\text{energy}(\omega_k) = \int_{\omega_k - \Delta/2}^{\omega_k + \Delta/2} \Psi_x(\omega) d\omega. \tag{15}$$

We have defined 10 bands of width 50Hz each, at frequencies $\omega_k = 100k\pi$, $k = 1, \ldots, 10$. We want to obtain 10 fuzzy sets describing the expected energy at each band, and its dispersion. First, the energy at each band ω is measured N independent times, and the values $\text{energy}_i(\omega)$, $i = 1 \ldots N$ are obtained. Second, following the interpretation in [1,10,2], we compute the intervals

$$I_\alpha(\omega) = [I_\alpha^-(\omega), I_\alpha^+(\omega)] \tag{16}$$

where $I_\alpha(\omega)$ is the smallest interval for which

$$\#\{i \mid I_\alpha^-(\omega) \leq \text{energy}_i(\omega) \leq I_\alpha^+(\omega)\} \geq N(1-\alpha). \tag{17}$$

Lastly, we define the fuzzy PSD of the noise as the fuzzy set with membership function

$$\widetilde{\text{PSD}}_\omega(x) = \sup\{\alpha \mid x \in I_\alpha(\omega)\}. \tag{18}$$

3.2 Representation of the State, Input and Output Matrices

We have mentioned that the state matrix will be codified by mean of its poles, or roots of the characteristic polynomial of Φ_N, which is

$$\sum_{i=0}^{8} c_i \lambda^i = \prod_{i=1}^{8} (\lambda - p_i) \tag{19}$$

where p_i are the complex poles of the system. It is remarked that the poles must be conjugated in pairs or be real numbers for the c_i to be real numbers.

The matrix Γ_N is codified as a vector of 8 real numbers. C has only one term different than zero, and this value is not represented in the genetic chain but solved so that the average energy of the individual is the same as the modal point of the average energy of the noise.

3.3 Generation of the Initial Population

The individuals are generated at random, but taking into account that the complex poles are paired with their conjugates, so the state matrix of the noise is real. The probability of appearance of each pole is not uniform, but biased towards unstable poles of the form $\cos 100k\pi \pm j \sin 100k\pi$.

3.4 Genetic Operators

The genetic operators are, for the most part, standard two-point crossover and arithmetic mutation [7]. The particularities of both are:

Crossover. Since complex poles must be accompanied by their conjugates, we represent both together and prevent the point between them from being an split in the crossover. We also do an arithmetic combination of the cut points, in order to introduce genetic diversity.

Mutation. If a pole is mutated, then its conjugate must be modified in accordance. The mutation is defined as the convex combination between the value and a randomly generated parameter.

3.5 Fitness Function

The fitness is the degree of compatibility between the fuzzy PSD of the individual (obtained by simulation of the candidate model in 10 random sequences, as explained in Section 3.1) and the fuzzy PSD of the ambient noise,

$$\sum_k \log \frac{\widetilde{\mathrm{PSD}_{\mathrm{model}}}(100k\pi)}{\widetilde{\mathrm{PSD}_{\mathrm{noise}}}(100k\pi)} \tag{20}$$

where the "log" and "quotient" operators are the extensions of the conventional operators to fuzzy aritmetic [4].

3.6 Generational Scheme

The generational scheme is steady state, with a tournament of size 5, where the offspring of the two winners replaces the last two individuals in the tournament. The probabilities of mutation and crossover are 0.05 and 1. In the tournament we have used the uniform dominance defined in [6].

4 Numerical Results

In this section we will compare first the use of crisp and fuzzy fitness functions for finding the model of the ambient noise, then we will apply it to estimate the current in the NGR of a 130KV transformer, as mentioned.

4.1 Crisp and Fuzzy Fitness Functions

In the left part of Figure 4 we have superimposed some experimental measurements of the PSD of the ambient noise, taken at the same point but at different

times. Observe that there is a significant dispersion of the energy of each harmonic, mainly at high frequencies. In this paper we claim that it is preferable to fit a model to a fuzzy description of this data than fitting a model to the average values of the same data.

To illustrate this result, we have learned 30 crisp models and 30 fuzzy models, following the methodology explained in the preceding section. The fitness of a crisp model was computed with the same equation (20) than the fuzzy model, but replacing the fuzzy PSD by the average of the PSDs. The 60 models were tested against an independent set of 10 PSDs of the actual noise, measured at a later time. The results are shown in the right part of Figure 4 and in Table 1.

Table 1. Train and test error (30 repetitions) of crisp and fuzzy approaches

	Train			Test		
	Best	Average	Worst	Best	Average	Worst
Crisp-GA		5.073		3.430	5.135	9.230
Fuzzy-GA	2.538	4.248	8.133	2.510	4.312	8.477

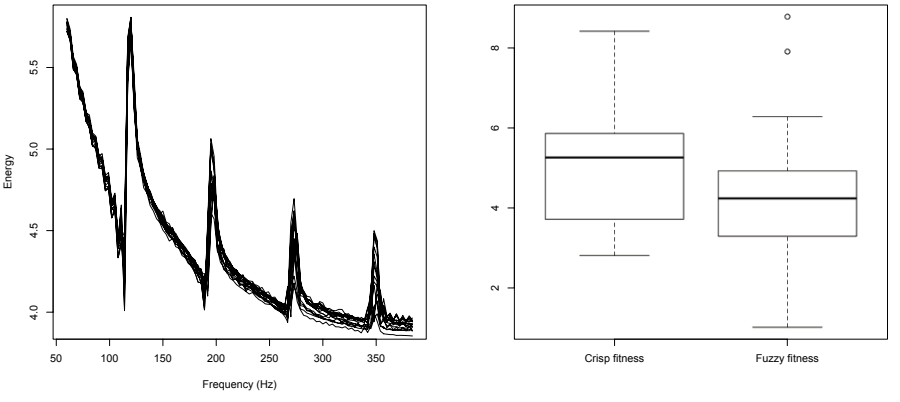


Fig. 4. Left: Dispersion of the ambient PSD. Right: Boxplots of the average error model with crisp and fuzzy fitness functions.

The boxplots show the dispersion of the average error of the model, and the table shows the mean values of the best, average and worst models. The p-value of a Wilcoxon test between the two samples of test errors ("Average" column) is 0.0399, thus there is a significant difference between crisp and fuzzy methodologies, at a confidence level of 95%.

4.2 Real-World Data

We have sampled the output of the current transformer (CT) in the NGR, and the results are displayed in the red curve in the left part of Figure 5. Observe that

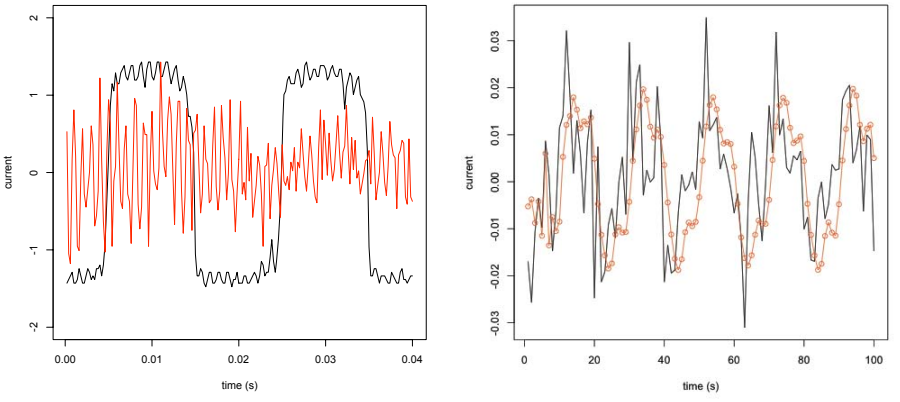


Fig. 5. Left: signal (red) and noise (black) in the CT. Right: Kalman filtering of the CT output (red line with circles).

there is a high amount of high-frequency noise, that makes hard to perceive the actual shape of the current; we have filtered out these components with a low-pass filter before applying the Kalman filtering. We have superimposed (black trace) a capture of the ambient noise taken elsewhere, that is not synchronized with this current. In the right part of the same figure we have displayed again the preprocessed output of the CT, along with the Kalman filtering of this signal (red curve with dots); observe that this filtered signal is a nearly periodical curve comprising the first four odd harmonics, as expected.

5 Concluding Remarks and Future Work

In this paper we have proposed a method for obtaining the state-space equations describing a shaping filter that mimics the ambient noise at a substation transformer. This model is intended to augment the state of a model of the current flowing through the NGR of the transformer, in order to obtain a model whose input is white noise and therefore suitable for Kalman filtering. We have shown that the use of a possibilistic representation of the dispersion of the PSD of the noise with time can be exploited by a fuzzy fitness function-driven Genetic Algorithm, producing models with improved generalization capabilities.

In future works we intend to extend this study to high frequencies, that have been filtered out in this paper. Frequencies above 1KHz are seldom considered in the determination of the NGR continuity, however we think that the use of the power line for digital transmission causes a measurable effect that could be captured with different sensors (i.e. Rogowski coils) and compared to the measurement in the neutral of the transformer, allowing a finer monitoring of the health of the NGR.

Acknowledgements

This work was supported by the Spanish Ministry of Education and Science, grant TIN2008-06681-C06-04.

References

1. Couso, I., Sánchez, L.: Higher order models for fuzzy random variables. Fuzzy Sets and Systems 159, 237–258 (2008)
2. Goodman, I.R., Nguyen, H.T.: Uncertainty Models for Knowledge-based Systems. North-Holland, Amsterdam (1985)
3. Franklin, G.A.: A practical guide to harmonic frequency interference affecting high-voltage power-line carrier coupling systems. IEEE Trans. Power Delivery 24(2), 630–642 (2009)
4. Kaufmann, A., Gupta, M.: Introduction to Fuzzy Arithmetic, Theory and Applications. Van Nostrand Reinhold (1985)
5. Grewal, M.S., Andrews, A.P.: Kalman filtering: theory and practice using MATLAB, 3rd edn. Wiley, Chichester (2008)
6. Limbourg, P.: Multi-objective optimization of problems with epistemic uncertainty. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 413–427. Springer, Heidelberg (2005)
7. Michalewicz, Z.: Genetic algorithms + Data Structures = Evolution Programs. Springer, Heidelberg (1992)
8. Paulson, G.E.: Monitoring neutral-grounding resistors. In: Pulp and Paper Industry Technical Conference, pp. 238–241 (1999)
9. Press, W., et al.: Numerical Recipes in C. The Art of Scientific Computing. Cambridge University Press, Cambridge (1992)
10. Sánchez, L., Couso, I., Casillas, J.: Genetic learning of fuzzy rules based on low quality data. Fuzzy Sets and Systems 160(17), 2524–2552 (2009)
11. Selkirk, D., Savostianik, M., Crawford, K.: Why neutral-grounding resistors need continuous monitoring. In: Petroleum and Chemical Industry Technical Conference, PCIC 2008, pp. 1–7 (2008)

A Genetic Algorithm for Feature Selection and Granularity Learning in Fuzzy Rule-Based Classification Systems for Highly Imbalanced Data-Sets

Pedro Villar[1], Alberto Fernández[2], and Francisco Herrera[2]

[1] Department of Software Engineering,
[2] Department of Computer Science and Artificial Intelligence,
E.T.S. Ing. Informática y de Telecomunicación, University of Granada
pvillarc@ugr.es, alberto@decsai.ugr.es, herrera@decsai.ugr.es

Abstract. This contribution proposes a Genetic Algorithm for jointly performing a feature selection and granularity learning for Fuzzy Rule-Based Classification Systems in the scenario of data-sets with a high imbalance degree. We refer to imbalanced data-sets when the class distribution is not uniform, a situation that it is present in many real application areas. The aim of this work is to get more compact and precise models by selecting the adequate variables and adapting the number of fuzzy labels for each problem.

Keywords: Fuzzy Rule-Based Classification Systems, imbalanced data-sets, Genetic Algorithms, feature selection, granularity level.

1 Introduction

The problem of imbalanced data-sets [1] for binary classification occurs when the number of instances for each class are very different between them, and usually the less representative class is the one which has more interest from the point of view of the learning task. We develop an experimental analysis in the context of imbalance classification for binary data-sets when the class imbalance ratio is high. In this study, we will make use of linguistic Fuzzy Rule Based Classification Systems (FRBCSs), a very useful tool in the framework of computational intelligence, since they provide a very interpretable model for the end user [2]. The good behavior of FRBCS when dealing with imbalanced data-sets has been recently analysed in [3].

An FRBCS presents two main components: the Inference System and the Knowledge Base (KB). The KB is composed of the Rule Base (RB) constituted by the collection of fuzzy rules, and of the Data Base (DB), containing the membership functions of the fuzzy partitions associated to the linguistic variables. The composition of the KB of an FRBCS directly depends on the problem being solved. If there is no expert information about the problem under solving, an automatic learning process must be used to derive the KB from examples.

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 741–750, 2010.

In many classification problems, a large number of features can originate RBs with a high number of rules, thus presenting a low degree of interpretability and a possible overfitting (the error over the training data set is very low but the FRBCS present a significative decrease on the prediction ability). This problem can be tackled from a double perspective: a reduction of the rule set, minimising the number of fuzzy rules included in the RB or a feature selection process that reduces the number of features used by the FRBCS. Notice that, for high dimensional problems and problems where a high number of instances is available, it is difficult for rule reduction approaches to get small rule sets, and therefore the system comprehensibility and interpretability may not be as good as desired. For high dimensionality classification problems, a feature selection process, that determines the most relevant variables before or during the FRBCS inductive learning process, must be considered. It increases the efficiency and accuracy of the learning and classification stages.

The number of labels per linguistic variable (granularity) is an information that has not been considered to be relevant for the majority of FRBCS learning methods. However, the fuzzy partition granularity of a linguistic variable can be viewed as a sort of context information with a significative influence in the FRBCS behavior. Considering a specific label set for a variable, some labels can result irrelevant, that is, they can contribute nothing and even can cause confusion. In other cases, it would be necessary to add new labels to appropriately differentiate the values of the variable. In a previous work [4], we analyse the influence of granularity learning in the performance of FRBCSs for imbalanced data sets, and the results obtained show that is possible an significant improvement in the classification ability only by learning an adequate number of labels per variable although the complexity of the model was lightly increased.

Our objective is to propose a genetic learning process to improve the prediction ability of the FRBCSs for imbalanced data-sets joint with a significative reduction of the model complexity in order to increase the FRBCS interpretability. Our proposal uses a Genetic Algorithm (GA) for jointly perform a feature selection and a granularity learning, and considers a classical FRBCS learning method to derive the rule base, the Chi et al.'s approach [5]. In order to show the influence of choosing a good set of features and an adequate granularity level, we compare the results obtained with the ones obtained by Chi et al.'s method with all the variables selected with and without an adequate granularity level. We also want to check the performance of our method compared with a non-FRBCS classification model, C4.5 [6], a decision tree algorithm that has been used as a reference in the imbalanced data-sets field [7].

We have selected a large collection of data-sets with high imbalance from UCI repository [8] for developing our experimental analysis. In order to deal with the problem of imbalanced data-sets we will make use of a preprocessing technique, the "Synthetic Minority Over-sampling Technique" (SMOTE) [9], to balance the distribution of training examples in both classes. Furthermore, we will perform a statistical study using non-parametric tests [10] to find significant differences among the obtained results.

This contribution is organized as follows. First, Section 2 introduces the problem of imbalanced data-sets, describing its features, how to deal with this problem and the metric we have employed in this context. Next, in Section 3 we will expose the characteristics of our proposal, a GA for feature selection and granularity learning. Section 4 contains the experimental study. Finally, in Section 5, some conclusions will be pointed out.

2 Imbalanced Data-Sets in Classification

Learning from imbalanced data is an important topic that has recently appeared in the Data Mining community [1]. This problem is very representative since it appears in a variety of real-world applications including, but not limited to, medical applications, finance, telecommunications, biology and so on. We refer to imbalanced data when the class distribution is not uniform. In this situation, the number of examples that represents one of the classes of the data-set (usually the concept of interest) is much lower than that of the other classes. We will use the imbalance ratio (IR) [11] as a threshold to categorize the different imbalanced scenarios, which is defined as the ratio of the number of instances of the majority class and the minority class. We consider that a data-set presents a high degree of imbalance when its IR is higher than 9 (less than 10% of positive instances).

Standard classifier algorithms have a bias towards the majority class, since the rules that predicts the higher number of examples are positively weighted during the learning process in favour of the accuracy metric. Consequently, the instances that belongs to the minority class are misclassified more often than those belonging to the majority class [12].

In a previous work on this topic [3], we analysed the cooperation of some preprocessing methods with FRBCSs, showing a good behaviour for the oversampling methods, specially in the case of the SMOTE methodology [9]. According to this, we will employ in this contribution the SMOTE algorithm in order to deal with imbalanced data-sets. In short, its main idea is to form new minority class examples by interpolating between several minority class examples that lie together.

Most of proposals for automatic learning of classifiers use some kind of accuracy measure like the classification percentage over the example set. However, these measures can lead to erroneous conclusions working with imbalanced data-sets since it doesn't take into account the proportion of examples for each class. Therefore, in this work we use the Area Under the Curve (AUC) metric [13], which can be defined as $(1 + TP_{rate} - FP_{rate})/2$, where TP_{rate} is the percentage of positive cases correctly classified as belonging to the positive class and FP_{rate} is the percentage of negative cases misclassified as belonging to the positive class.

3 Genetic Algorithm for the Data Base Learning

In this section, we propose an standard generational GA for the DB that allows us to select a set of variables (feature selection) and learn an adequate number

of labels for each selected variable (granularity learning). Once the granularity for each selected feature are determined, the DB is built. Uniform partitions with triangular membership functions are considered due to its simplicity. Next, we use a quick method that derives the fuzzy classification rules and then the chromosome can be evaluated. The RB derivation algorithm used in this work is the method proposed in [5], that we have called the Chi et al.'s method.

We denote our proposal as GA-FS-GL (Genetic Algorithm for Feature Selection and Granularity Learning). The main purpose of GA-FS-GL is to obtain FRBCSs with good accuracy and reduced complexity taking the feature selection and granularity learning as a base. Unfortunately, FRBCSs with good performance have a high number of rules, thus presenting a low degree of readability. On the other hand, as mentioned before, the KB design methods sometimes lead to a certain overfitting to the training data-set used for the learning process. In order to avoid that problem, our genetic process try to design a compact and interpretable KB by penalizing FRBCSs with high number of selected variables and/or high granularity average as it will be explained in this Section. Next, we describe the main components of GA-FS-GL.

Encoding the DB. For a classification problem with N variables, each chromosome will be composed of two parts to encode the relevant variables and the number of linguistic terms for variable (i.e. the granularity):

- Relevant variables (C_1): the selected features are stored in a binary coded array of length N. In this array, an 1 indicates that the correspondent variable is selected for the FRBCS.
- Granularity level (C_2): the number of labels per variable is stored in an integer array of length N. In this contribution, the possible values considered are taken from the set $\{2, \ldots, 7\}$.

If v_i is the bit that represents whether the variable i is selected and g_i is the granularity of variable i, a representation of the chromosome is shown next:

$$C_1 = (v_1, v_2, \ldots, v_N) \quad C_2 = (g_1, g_2, \ldots, g_N) \qquad C = C_1 C_2$$

Initial Gene Pool. The initial population is composed of six groups with a different number of selected variables. Next, we describe its generation:

- In the first group all the chromosomes have all the features selected. It is composed of two parts. In the first part all the chromosomes have the same granularity in all its variables and it is composed of g chromosomes, with g being the cardinality of the significant term set, in our case $g = 6$, corresponding to the six possibilities for the number of labels, $2 \ldots 7$. For each granularity level, one individual is created. The second part is composed of 10 chromosomes and the granularity level is randomly selected.
- The next four groups have the same structure than the first group but each one of them with a different percentage of randomly selected variables (75%, 50%, 25% and 10%). So, each group has $g+10$ chromosomes (16 in our case).

– The last group is composed for the remaining chromosomes, and all of their components are randomly selected.

The minimum number of individuals is the sum of the chromosomes of the five first groups: $(g + 10) \times 5$.

Evaluating the Chromosome. There are three steps that must be done to evaluate each chromosome:

– Generate the DB using the information contained in the chromosome. For all the selected variables ($v_i = 1$), a uniform fuzzy partition with triangular membership functions is built considering the number of labels of that variable (g_i).
– Generate the RB by running the the Chi et al.'s method.
– Calculate the value of the evaluation function: The usual way to proceed in this type of genetic learning is to choose a kind of accuracy measure over the training data-set, like the AUC metric. However, as mentioned before, we will lightly penalize FRBCSs with high number of selected variables and/or high granularity levels in order to avoid the possible overfitting, thus improving the generalization capability of the final FRBCS. To do that, once the RB has been generated and its AUC over the training set has been calculated, the fitness function to be minimized is:

$$F_C = \omega_1 \cdot (1 - AUC) + \omega_2 \cdot (Ng/N)$$

being Ng the sum of the granularity levels of all the selected variables. In order to normalize these two values, we calculate ω_2 taking two values as a base: the AUC of the FRBCS obtained with the RB generation method considering the DB with all the variables selected, the maximum number of labels (max_g) per variable and uniform fuzzy partitions:

$$\omega_2 = \alpha_{\omega_2} \cdot \frac{AUC_{max\_g}}{max\_g}$$

with α_{ω_2} being a weighting percentage.

Genetic Operators

– **Selection:** we will employ the tournament selection with k = 2, in which two chromosomes are selected at random from the population, and the one with highest fitness is taken to be included in the next population, after the application of the genetic operators.
– **Crossover:** the crossover works in the two parts of the chromosome at the same time. Therefore, an standard crossover operator is applied over C_1 and C_2. This operator performs as follows: a crossover point p is randomly generated in C_1 and the two parents are crossed at the p-th variable in C_1 (the possible values for p are $\{2, \ldots, N\}$). The crossover is developed this way in the two chromosome parts, C_1 and C_2, thereby producing two meaningful descendants.

– **Mutation:** two different operators are used, each one of them acting on different chromosome parts. A brief description of them is given below:
  - *Mutation on* C_1: As this part of the chromosome is binary coded, a simple binary mutation is developed, flipping the value of the gene.
  - *Mutation on* C_2: The mutation operator selected for C_2 performs a slight change in the selected variable. Once a granularity level is randomly selected to be muted, a local modification is developed by changing the number of labels of the variable to the immediately upper or lower value (the decision is made at random). When the value to be changed is the lowest (2) or highest one (7), the only possible change is developed.

4 Experimental Study

We will study the performance of GA-FS-GL employing a large collection of imbalanced data-sets with a high imbalance ratio ($IR > 9$). Specifically, we have considered twenty-two data-sets from UCI repository [8] with different IR, as shown in Table 1, where we denote the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority), class attribute distribution and IR. This table is in ascendant order according to the IR. Multi-class data-sets are modified to obtain two-class imbalanced problems, defining the joint of one or more classes as positive and the joint of one or more classes as negative. In order to reduce the effect of imbalance, we will employ the SMOTE preprocessing method [9] for all our experiments, considering only the 1-nearest neighbour to generate the synthetic samples, and balancing both classes to the 50% distribution.

We will analyse the influence of feature selection and granularity learning by means of a comparison between the performance of GA-FS-GL and two FRBCS models obtained by Chi et al.'s method with all the variables selected:

– The original Chi et al.'s method, that needs of the existence of a previous definition for the DB, normally uniform fuzzy partitions with the same number of labels in all the variables. So, it is necessary to choose a number of labels. The usual values employed for Chi et al.'s approach in the specialized literature are 3 and 5 labels per variable. Previous experiments [4] showed that the FRBCSs with three labels for variable obtain better results in prediction ability (less value in AUC for the test data set) and interpretability (less number of rules) so we choose this granularity level for the comparison. In the latter, we will refer that method as G3-Chi.
– The method proposed in [4] (denoted GA-GL), that uses a GA for granularity learning and the Chi et al.'s method to derive the RB.

As mentioned before, we also compare the results of GA-FS-GL with C4.5 [6], a method of reference in the field of classification with imbalanced data-sets [7]. The configuration for the FRBCSs approaches, GA-FS-GL, GA-GL and Chi et al.'s, is presented below. This parameter selection has been carried out according to the results achieved by the Chi et al.'s method in our former studies on imbalanced data-sets [3]:

Table 1. Summary Description for Imbalanced Data-Sets

Data-set	#Ex.	#Atts.	Class (min.; maj.)	%Class(min., maj.)	IR
Yeast2vs4	514	8	(cyt; me2)	(9.92, 90.08)	9.08
Yeast05679vs4	528	8	(me2; mit,me3,exc,vac,erl)	(9.66, 90.34)	9.35
Vowel0	988	13	(hid; remainder)	(9.01, 90.99)	10.10
Glass016vs2	192	9	(ve-win-float-proc; build-win-float-proc, build-win-non_float-proc,headlamps)	(8.89, 91.11)	10.29
Glass2	214	9	(Ve-win-float-proc; remainder)	(8.78, 91.22)	10.39
Ecoli4	336	7	(om; remainder)	(6.74, 93.26)	13.84
Yeast1vs7	459	8	(vac; nuc)	(6.72, 93.28)	13.87
Shuttle0vs4	1829	9	(Rad Flow; Bypass)	(6.72, 93.28)	13.87
Glass4	214	9	(containers; remainder)	(6.07, 93.93)	15.47
Page-blocks13vs2	472	10	(graphic; horiz.line,picture)	(5.93, 94.07)	15.85
Abalone9vs18	731	8	(18; 9)	(5.65, 94.25)	16.68
Glass016vs5	184	9	(tableware; build-win-float-proc, build-win-non_float-proc,headlamps)	(4.89, 95.11)	19.44
Shuttle2vs4	129	9	(Fpv Open; Bypass)	(4.65, 95.35)	20.5
Yeast1458vs7	693	8	(vac; nuc,me2,me3,pox)	(4.33, 95.67)	22.10
Glass5	214	9	(tableware; remainder)	(4.20, 95.80)	22.81
Yeast2vs8	482	8	(pox; cyt)	(4.15, 95.85)	23.10
Yeast4	1484	8	(me2; remainder)	(3.43, 96.57)	28.41
Yeast1289vs7	947	8	(vac; nuc,cyt,pox,erl)	(3.17, 96.83)	30.56
Yeast5	1484	8	(me1; remainder)	(2.96, 97.04)	32.78
Ecoli0137vs26	281	7	(pp,imL; cp,im,imU,imS)	(2.49, 97.51)	39.15
Yeast6	1484	8	(exc; remainder)	(2.49, 97.51)	39.15
Abalone19	4174	8	(19; remainder)	(0.77, 99.23)	128.87

- Conjunction operator to compute the compatibility degree of the example with the antecedent of the rule: Product T-norm.
- Rule Weight: Penalized Certainty Factor [14].
- Conjunction operator between the compatibility degree and the rule weight: Product T-norm.
- Fuzzy Reasoning Method: Winning Rule.

To develop the different experiments we consider a *5-folder cross-validation model*, i.e., 5 random partitions of data with a 20%, and the combination of 4 of them (80%) as training and the remaining one as test. Since a GA is a probabilistic method, three runs with different seeds for the pseudo-random sequence are made for each data partition. For each data-set we consider the average results of the five partitions per three executions. Furthermore, Wilcoxon's Signed-Ranks Test [15] is used for statistical comparison of our experimental results. The specific parameters setting for the GA of GA-FS-GL is listed below, being N the number of variables:

-Number of evaluations: $500 \cdot N$ -Population Size: 100 individuals
-Crossover Probability P_c : 0.6 -Mutation Probability P_m : 0.2
-Parameters of the evaluation function (Section 3): (ω_1 : 0.7 , α_{ω_2} : 0.3)

Table 2 shows the results in performance (using the AUC metric) for GA-FS-GL and the algorithms employed for comparison, that is, G3-Chi, GA-GL and C4.5, being AUC_{Tr} the AUC over the training data-set and AUC_{Tst} the AUC over the test data-set. The final line of the table shows the mean of the number of rules (NR) of the classifiers.

Table 2. Detailed results table for the problems considered

Data-set	G3-Chi		GA-GL		GA-FS-GL		C4.5	
	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
Yeast2vs4	89.68	87.36	93.79	90.84	94.38	94.52	98.14	85.88
Yeast05679vs4	82.65	79.17	86.11	81.78	83.37	78.97	95.26	76.02
Vowel0	98.57	98.39	99.59	99.07	96.58	96.49	99.67	94.94
Glass016vs2	62.71	54.17	85.96	60.54	78.23	56.07	97.16	60.62
Glass2	66.54	55.30	83.71	57.42	79.42	56.88	95.71	54.24
Ecoli4	94.06	91.51	98.14	90.90	93.20	92.31	97.69	83.10
Yeast1vs7	82.00	80.63	82.43	75.79	77.37	70.75	93.51	70.03
Shuttle0vs4	100.00	99.12	100.00	99.42	100.00	99.97	99.99	99.97
Glass4	95.27	85.70	98.71	87.92	95.02	85.20	98.44	85.08
Page-Blocks13vs4	93.68	92.05	99.59	99.10	98.25	96.99	99.75	99.55
Abalone9vs18	70.23	64.70	82.38	73.68	78.63	68.18	95.31	62.15
Glass016vs5	90.57	79.71	98.21	85.43	95.50	84.57	99.21	81.29
Shuttle2vs4	95.00	90.78	99.73	94.25	99.09	98.78	99.90	99.17
Yeast1458vs7	71.25	64.65	85.69	65.47	76.00	74.67	91.58	53.67
Glass5	94.33	83.17	98.03	79.92	94.57	79.15	99.76	88.29
Yeast2vs8	78.61	77.28	84.57	79.32	81.69	79.46	91.25	80.66
Yeast4	83.58	83.15	86.90	80.66	84.47	80.31	91.01	70.04
Yeast1289vs7	74.70	77.12	80.27	70.98	76.00	74.67	94.65	68.32
Yeast5	94.68	93.58	96.48	94.73	95.58	93.54	97.77	92.33
Ecoli0137vs26	93.96	81.90	97.69	81.36	97.22	80.99	96.78	81.36
Yeast6	88.48	88.09	91.09	86.06	89.37	87.01	92.42	82.80
Abalone19	71.44	63.94	80.28	69.03	77.40	73.16	85.44	52.02
Mean	85.09	80.52	91.33	81.98	88.39	81.42	95.93	78.25
NR mean	68.67		82.36		37.31		22.45	

As it can be observed, the performance obtained by GA-FS-GL is higher than the one for G3-Chi, both in AUC_{Tr} and AUC_{Tst}, showing the significative influence of the feature selection and granularity level in the behaviour of the classifier. GA-FS-GL obtain results very similar to GA-GL in AUC (Table 3 shows no significative differences between them in AUC_{Tst}) but the number of rules is very much lower in GA-FS-GL by the feature selection process, reducing the complexity of the model. Therefore, the interpretability of the FRBCSs generated by GA-FS-GL is greater than the other methods. Furthermore, GA-FS-GL present better results than C4.5 in AUC_{Tst}. This situation is represented statistically by means of a Wilcoxon test (Table 3, with R^+ corresponds to GA-FS-GL and R^- to the other method).

Table 3. Wilcoxon test to compare the methods according to their performance

Comparison	R^+	R^-	p-value
GA-FS-GL vs. G3-Chi	150.5	102.5	0.436
GA-FS-GL vs. GA-GL	95.0	158.0	0.306
GA-FS-GL vs. C4.5	198.5	54.5	0.019

GA-FS-GL obtain precise and interpretable models by selecting a reduced set of features and finding an appropriate granularity level in each selected variable. Thus, we show in Table 4 the mean of selected variables (SV) in the first column. The remaining columns show two values for each feature of the problem, the first is the selection ratio of the variable, that is, the relation between the number of

Table 4. Mean of number of selected variables and labels learned by GA-FS-GL

		Variables									
Data-set	SV	1	2	3	4	5	6	7	8	9	10
Yeast2vs4	2.0	1.0/3.0	.00/0.0	1.0/4.0	.00/0.0	.00/0.0	.00/0.0	.00/0.0	.00/0.0	-	-
Yeast05679vs4	2.8	1.0/2.4	.40/2.5	.60/2.0	.00/0.0	.80/2.0	.00/0.0	.00/0.0	.00/0.0	-	-
Glass016vs2	2.6	.40/4.5	.20/3.0	.00/0.0	.60/5.0	.60/5.7	.40/4.0	.20/7.0	.00/0.0	.20/3.0	-
Glass2	2.6	.40/5.0	.00/0.0	.40/2.0	.40/4.5	1.0/4.6	.00/0.0	.00/0.0	.00/0.0	.40/6.5	-
Ecoli4	2.0	.00/0.0	.60/2.0	.00/0.0	.00/0.0	1.0/3.0	.20/3.0	.20/2.0	-	-	-
Shuttle0vs4	2.0	.20/3.0	.20/4.0	.20/3.0	.00/0.0	.00/0.0	.20/3.0	.80/3.8	.40/4.5	.00/0.0	-
Yeast1vs7	2.2	.60/2.7	.00/0.0	1.0/2.6	.00/0.0	.00/0.0	.00/0.0	.20/2.0	.40/3.5	-	-
Glass4	2.4	.00/0.0	.00/0.0	.40/4.0	.60/3.7	.20/2.0	.00/0.0	.60/3.0	.60/3.3	.00/0.0	-
Pageblocks13vs4	2.0	1.0/4.4	.00/0.0	.00/0.0	.00/0.0	1.0/4.4	.00/0.0	.00/0.0	.00/0.0	.00/0.0	.00/0.0
Abalone9vs18	2.2	.40/2.0	.00/0.0	.00/0.0	.20/2.0	.00/0.0	.60/6.7	.00/0.0	1.0/5.8	-	-
Glass016vs5	2.8	.20/6.0	.40/3.0	1.0/3.6	.20/2.0	.00/0.0	.00/0.0	.20/3.0	.60/3.3	.20/3.0	-
Shuttle2vs4	2.8	.60/3.0	.00/0.0	1.0/3.0	.00/0.0	.00/0.0	.00/0.0	1.0/2.2	.00/0.0	.20/3.0	-
Yeast1458vs7	4.0	.60/5.3	.80/5.0	1.0/4.6	.60/5.3	.00/0.0	.00/0.0	.00/0.0	1.0/3.2	-	-
Glass6	2.4	.00/0.0	.40/3.5	1.0/3.2	.00/0.0	.00/0.0	.00/0.0	.20/3.0	.60/3.0	.20/4.0	-
Yeast2vs8	2.2	.80/4.0	.40/2.0	.00/0.0	.00/0.0	.00/0.0	1.0/2.0	.00/0.0	.00/0.0	-	-
Yeast4	2.6	1.0/3.0	.40/2.0	.80/3.0	.00/0.0	.40/2.0	.00/0.0	.00/0.0	.00/0.0	-	-
Yeast1289vs7	3.2	1.0/2.2	.00/0.0	1.0/3.2	.00/0.0	.00/0.0	.00/0.0	.20/5.0	1.0/2.2	-	-
Yeast5	2.8	1.0/3.2	.80/2.3	.60/2.0	.20/2.0	.20/2.0	.00/0.0	.00/0.0	.00/0.0	-	-
Yeast6	2.8	.80/3.0	.80/2.3	.00/0.0	.00/0.0	.40/2.0	.00/0.0	.20/2.0	.60/2.7	-	-
Ecoli0137vs26	3.2	1.0/3.6	.40/3.5	1.0/2.8	.00/0.0	.00/0.0	.40/4.5	.40/5.0	-	-	-
Abalone19	2.0	.00/0.0	.00/0.0	.20/3.0	.00/0.0	.00/0.0	.60/6.7	.20/3.0	1.0/5.8	-	-
Vowel0	2.2	.00/0.0 11 .00/0.0	.00/0.0 12 .00/0.0	.00/0.0 13 .00/0.0	.80/4.0	1.0/7.0	.20/3.0	.00/0.0	.20/2.0	.00/0.0	.00/0.0

occasions in that the variable was selected and the number of total executions for each problem. The second value is the average of the number of labels for the cases in which that variable was selected.

As it can be observed in Table 4, the number of selected variables is very low. In all the problems the number of selected features is reduced, at least, to the half of the original. Moreover, in nineteen problems, less than three variables are selected in the average of the 15 executions. Regarding to the granularity level mean, there are significant differences among the variables of each data-set. This situation is caused by the advantage of increasing or decreasing the granularity for a good data representation in the fuzzy partition. Therefore, GA-FS-GL obtain FRBCSs with high prediction ability and very reduced complexity, that was the main purpose of this contribution.

5 Conclusions

This contribution has proposed a method to design FRBCS with good accuracy and interpretability for imbalanced data-sets with a high imbalance ratio. A GA is used for feature selection and granularity learning, which is combined with an efficient fuzzy classification rule generation method to obtain the complete KB of the FRBCS. We must remark one advantage of our proposal, the GA can be combined with any rule generation method. We have used a simple algorithm for efficiency but another more accurate one can be used. Our future work will be focused on applying a multi-objective genetic algorithm in order to obtain a set of solutions with different trade-off between accuracy (high AUC) and

interpretability (low number of rules), eliminating the problem of the choice of weights in the fitness function.

Acknowledgments. This work had been supported by the Spanish Ministry of Science and Technology under Project TIN2008-06681-C06-01.

References

1. Chawla, N.V., Japkowicz, N., Kolcz, A.: Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations 6(1), 1–6 (2004)
2. Ishibuchi, H., Nakashima, T., Nii, M.: Classification and modeling with linguistic information granules: Advanced approaches to linguistic Data Mining. Springer, Heidelberg (2004)
3. Fernández, A., García, S., Del Jesus, M.J., Herrera, F.: A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. Fuzzy Sets and Systems 159(18), 2378–2398 (2008)
4. Villar, P., Fernández, A., Herrera, F.: A Genetic Learning of the Fuzzy Rule-Based Classification System Granularity for highly Imbalanced Data-Sets. In: 2009 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2009), pp. 1689–1694 (2009)
5. Chi, Z., Yan, H., Pham, T.: Fuzzy algorithms with applications to image processing and pattern recognition. World Scientific, Singapore (1996)
6. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
7. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A Study of the Behaviour of Several Methods for Balancing Machine Learning Training Data. SIGKDD Explorations 6(1), 20–29 (2004)
8. Asuncion, A., Newman, D.J.: UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, http://www.ics.uci.edu/~mlearn/MLRepository.html
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligent Research 16, 321–357 (2002)
10. García, S., Herrera, F.: An Extension on "Statistical Comparisons of Classifiers over Multiple data sets" for all Pairwise Comparisons. Journal of Machine Learning Research 9, 2607–2624 (2008)
11. Orriols-Puig, A., Bernadó-Mansilla, E.: Evolutionary rule-based systems for imbalanced datasets. Soft Computing 13(3), 213–225 (2009)
12. Weiss, G.M.: Mining with rarity: a unifying framework. SIGKDD Explorations 6(1), 7–19 (2004)
13. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering 17(3), 299–310 (2005)
14. Ishibuchi, H., Yamamoto, T.: Rule Weight Specification in Fuzzy Rule-Based Classification Systems. IEEE Transactions on Fuzzy Systems 13, 428–435 (2005)
15. Sheskin, D.: Handbook of parametric and nonparametric statistical procedures, 2nd edn. Chapman & Hall/CRC, Boca Raton (2006)

Learning of Fuzzy Rule-Based Meta-schedulers for Grid Computing with Differential Evolution

R.P. Prado, S. García-Galán, J.E. Muñoz Expósito, A.J. Yuste, and S. Bruque

Telecommunication Engineering Department in University of Jaén,
Alfonso X el Sabio, 28 Linares, Jaén, Spain
{rperez,sgalan,jemunoz,ajyuste,sbruque}@ujaen.es

Abstract. Grid computing has arisen as the next-generation infrastructure for high demand computational applications founded on the collaboration and coordination of a large set of distributed resources. The need to satisfy both users and network administrators QoS demands in such highly changing environments requires the consideration of adaptive scheduling strategies dealing with inherent dynamism and uncertainty. In this paper, a meta-scheduler based on Fuzzy Rule-Based Systems is proposed for scheduling in grid computing. Moreover, a new learning strategy inspired by stochastic optimization algorithm Differential Evolution (DE), is incorporated for the evolution of expert system knowledge or rules bases. Simulation results show that knowledge acquisition process is improved in terms of convergence behaviour and final result in comparison to other evolutionary strategy, genetic Pittsburgh approach. Also, the fuzzy meta-scheduler performance is compared to other extended scheduling strategy, EASY-Backfilling in diverse criteria such as *flowtime*, *tardiness* and *machine usage*.

Keywords: Evolutionary Algorithms, Knowledge Acquisition, Fuzzy Rule-Based Systems, Grid Computing.

1 Introduction

Grid computing is an emerging framework providing computational capabilities for the solving of large-scale problems in science and engineering previously restricted by local resources [1]. A grid is made up of a set of heterogeneous resources, geographically distributed and interconnected through high speed networks that cooperate together with the aim of overcoming actual technological limitations of machines as independent components nowadays. Grids are *fully dynamic environments with uncertainties* [2] where new resources become available, existing machines fall down or decrease performance in response to local policies with time. In this sense, a major challenge is given by the efficient assignment of jobs to existing resources or grid scheduling problem, which is known to be a NP-hard problem [3].

Fuzzy Rule-Based Systems are expert systems incorporating knowledge in the form of IF-THEN fuzzy rules and fuzzy sets [4] extensively used in diverse areas such as intelligent control of elevator systems [5] or classification in speech/music

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 751–760, 2010.

discrimination applications [6]. Also, scheduling is increasingly arising as an application field for FRBSs [7,8]. However, FRBS-based schedulers performance highly depends on the quality of theirs knowledge bases and so with the learning strategy. In this regard, the role of Genetic Algorithms (GAs) must be highlighted [4]. Specifically, there exist two successful strategies for the learning of fuzzy rules, namely, Pittsburgh and Michigan approaches [4,9]. Pittsburgh approach is a well-known genetic learning strategy for FRBSs founded on the definition of RBs as chromosomes or individuals of the population to be evolved. Thus, in Pittsburgh approach, genetic operators, i.e., crossover, mutation and selection are applied at the level of sets of rules. As a consequence, Pittsburgh strategy provides a RB by competition and cooperation (i.e., crossover) with other RBs in contrast to other extended strategies such as Michigan approach where individual rules must compete and cooperate to be incorporated in the final RB. In this work, a new learning strategy is suggested to evolve rule bases (RBs) of fuzzy schedulers based on the well-known evolutionary strategy Differential Evolution (DE) [10].

DE is a stochastic real-parameter optimization algorithm derived from Evolutionary Algorithms (EAs) extensively used in recent applications including location management in mobile computing [11], optimization of non-linear chemical processes [12] and reactive power optimization [13]. The algorithm follows the classical EAs computational steps. However, DE suggests the modification of population individuals by weighted difference of other randomly selected individuals in such a way that no separated probability distribution is required, thus resulting in a self-organized procedure. Also, DE is a very simple and straightforward strategy. DE is driven by only a few control parameters and can be implemented in a few lines of code in most programming languages. Further, DE has proven its efficiency in optimizing a range of multi-dimensional objective functions in terms of convergence speed, robustness and final accuracy [10]. Hence, in this work, DE algorithm is adapted for the evolution of fuzzy rules. Simulations results prove this strategy outperforms other widely used learning strategy, Pittsburgh approach, considering final evaluation and convergence behaviour. Moreover, the fuzzy meta-scheduler performance is compared to an extended scheduling strategy, EASY-Backfilling [14]. The rest of the paper is summarized as follows. Section 2 provides an overview of the scheduling problem in grids and introduces DE main features. In Section 3, the suggested fuzzy scheduler with evolutionary learning is presented. Finally, Section 5 concludes the paper.

2 Background

A computational grid is generally seen as set of H_j heterogeneous computational machines, distributed within diverse administrative or resources domains, $RD_j = \{r_{j,1}, r_{j,2}, \ldots, r_{j,H_j}\}$, each imposing its own access and sharing policies [3]. Also, the association of RDs makes up a global domain so-called Virtual Organization, $VO = \{RD_1, RD_2, \ldots, RD_G\}$. Hence, the scheduling problem in grids can be viewed as a hierarchical problem concerning two levels. On the one hand, a grid meta-scheduler, assigns L users jobs $J = \{J_1, J_2, \ldots, J_L\}$ to available RDs within

a grid VO, whereas local schedulers are responsible for scheduling jobs within its own RD.

Many efforts have been made to provide solutions to this problem. Many of today's production systems such as Condor [15] and meta-scheduling systems like Grid Service Broker [16], are based on *queued-based* strategies such as EASY-Backfilling or EDF [14]. On the other hand, as stated in [17] any scheduling strategy for grids aiming to provide a certain level of QoS must concern grid state, resulting in *scheduled-based strategies*. In this regard, adaptive scheduling suggests the consideration of future and present grid state to avoid or prevent performance deterioration [3]. FRBS models are flexible and are increasingly used as scheduling systems for scheduling jobs in large-scale distributed systems. However, the learning of these systems is critical for performance and so new strategies are demanded to improve knowledge acquisition process. In this regard, learning strategies based on GAs are extensively used. In GFRBSs [4] rules or sets of rules are considered as individuals that cooperate and compete as to be included in the next generation and genetic operators are applied to evolve population. Specifically, Pittsburgh, Michigan and hybrid approaches [4,9] are some of the most successful strategies in this area. However, in the light of the relevance of the learning strategies in FRBS-schedulers, a new learning strategy based on DE is proposed for the learning of fuzzy rules.

DE is a stochastic, population-based optimization algorithm introduced by Storn and Price in 1996 [10]. It was initially intended to optimize real parameter, real valued functions. In DE, a set of NP D-dimensional vector parameters, $x^i_G = 1, 2, ..., NP$, are considered in every generation, G, and the goal is to find vector x^*, in a way that fitness $f : X \subseteq \mathbb{R}^N \rightarrow \mathbb{R}$ is optimum; $x^* \in X \, such \, that \, (f(x^*) \leq f(x) \, \forall x \in X)$. Following the general procedure in EAs, DE generates new vector parameters at every generation by means of mutation, crossover and selection procedures. Concretely, a mutant vector, so-called donor vector, v^i_{G+1}, is constructed for every target vector, x^i_G, through the weighted difference of randomly selected parameters vectors to a third vector and its components are included to trial vector, u^i_{G+1}, with crossover (CR) probability. Next, trial vector performance is compared to target vector performance and the one with best object function evaluation is included for the next generation $G+1$. DE has proven to be effective in many research areas for optimization of objective functions that are non-linear, multi-dimensional or presenting many local minima or stochasticity and new application areas are continually coming out for this evolutionary strategy as shown by a number of recent publications [11,12,13]. In this paper, we suggest the adaptation of DE to knowledge acquisition in grid computing.

3 Proposed Meta-scheduler with Differential Evolution-Based Learning Strategy

3.1 Fuzzy Rule-Based Meta-scheduler for Grid Computing

The suggested meta-scheduler bases its strategy in the fuzzy description of the grid state and the application of Mamdani type rules to infer the best selection

of resources. Rules are made up of antecedents and a consequent, representing activation condition and corresponding RD state evaluation, respectively. Hence, rules have the form

$$R_i = IF\, x_1 is\, A_{1l}\, and/or \ldots\, x_n is\, A_{nl}\, THEN\, x_o\, is\, B_l \tag{1}$$

where $(x_1, \ldots, x_n)$ indicates RDs input features, A_{nl} and B_l denote the associated fuzzy sets for feature x_n and output for rule i, respectively, and l is limited to the number of fuzzy sets for input and output, NF_{in} and NF_{out}. Concretely, in this work, three gaussian membership functions are suggested to describe the condition of every input feature in each RD, $NF_{in} = 3$, in such a way that there exist seven possible sets, A, for each antecedent element. On the other hand, five sets are considered for rules consequents, $NF_{out} = 5$, and so eleven values, B, can be associated to x_o.

There exist many grid features, $(x_1, \ldots, x_n)$, that may be relevant to be considered in jobs scheduling for grid computing. In this work, we suggest the use of seven variables to describe every RD state. Table 1 shows a brief description of these variables. It must be underlined here that accuracy in RDs status description is critical for the fuzzy meta-scheduler performance and this way, a complex genome is required. Note other variables may be included to provide a more precise description of every RD state. However, as it can be derived from rules encoding, the search space highly increases with the number of considered features. Hence, this description tries to reconcile both accuracy in resources state characterization and complexity of fuzzy rules learning.

Fig. 1 shows the general structure for the fuzzy meta-scheduler within a grid environment. As it can be derived, it follows the general structure for Mamdani fuzzy logic systems (FLSs): *fuzzification system*, *inference system*, *defuzzification system* and *knowledge base*. The goal is to obtain a performance index or RD selector, y_o, describing the suitability of the RD under analysis to be selected for the next schedule. First, *fuzzification system* is responsible for associating a linguistic label to every crisp input value featuring a RD state. Linguistic labels correspond to fuzzy sets and so a certain degree of uncertainty is tolerated in the characterization of resources conditions. On the other hand, *inference system* applies system knowledge (rules bases) to obtain a fuzzy linguistic label for the output. Finally, *defuzzification system* translates this output linguistic label into

Table 1. Grid system input features

Feature	Description
Number of free processing elements (**FPE**)	Number of free processing element within RD_i.
Previous Tardiness (**PT**)	Sum of tardiness of all finished jobs in RD_i.
Resource Makespan (**RM**)	Current makespan for RD_i.
Resource Tardiness (**RT**)	Current tardiness of jobs within RD_i.
Previous Score (**PS**)	Previous deadline score of already finished jobs in RD_i.
Resource Score (**RS**)	Number of non delayed jobs so far in RD_i.
Resources In Execution (**RE**)	Number of Resources currently executing jobs within RD_i.

a crisp value, RD selector, y_o. Thus, in every schedule, the RD with higher y_o is selected as target domain for the considered unscheduled job. However, a job may require specific machine characteristics and so only those RDs satisfying jobs constraints participate in each schedule. Note that the role of RBs is critical for the whole scheduling system performance. Thereby, next section is focused on providing high quality RBs by applying a learning strategy based on well-known DE optimization algorithm.

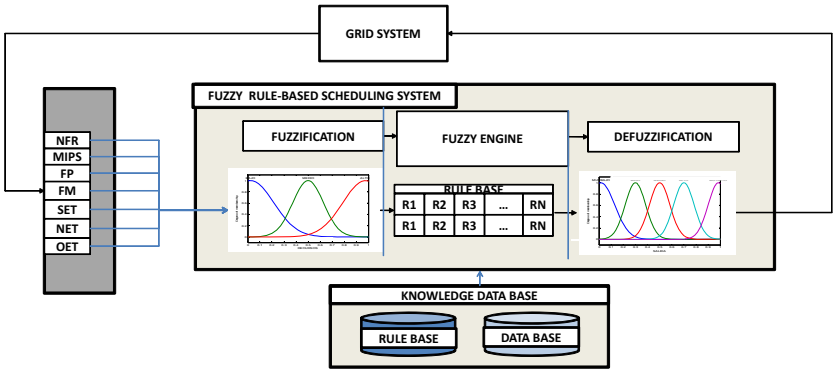


Fig. 1. Fuzzy Rule-Based Meta-scheduler structure

3.2 Learning Strategy Based on Differential Evolution

The proposed DE-based learning strategy follows the general procedure of evolutionary algorithms, i.e., initialization, mutation, crossover and selection. First, in the initialization stage, the cost function to be optimized with D real parameters must be specified. There exist a number of criteria to be selected for the construction of cost function in grid computing depending on both users and administrators (i.e., *flowtime*, *makespan*, *machine usage*, *slowdown, average weighted response time* or *tardiness*). However, we will focus on single criteria optimization. Also, in this work, function parameters are represented by antecedents, consequent and connectives of rules. Hence, on the basis of rules encoding, population individuals have the form:

$$RB_G^i = \begin{bmatrix} x^i_{1,1,G} & x^i_{1,2,G} & \cdots & x^i_{1,D,G} \\ x^i_{2,1,G} & x^i_{2,2,G} & \cdots & x^i_{2,D,G} \\ \cdots & \cdots & \cdots & \cdots \\ x^i_{NR,1,G} & x^i_{NR,2,G} & \cdots & x^i_{NR,D,G} \end{bmatrix}, \; i = 1, 2, ..., N. \tag{2}$$

where NR denotes the number of fuzzy rules and G is the generation number. Note that canonical DE strategy has been modified to evolve FRBSs and every individuals row represents the codification of a fuzzy rule. Thus, parameters vectors are extended to [NR, D] dimension matrices representing sets of rules. Concretely, with seven input features for grid characterization (see Table 1)

$D = 10$. Also, in this stage, for every RB^i, each parameter is randomly initialized considering each individual lower and upper bounds for antecedents, consequents and connectives of the rules:

$$x^i_{j,k,G} \in [-NF_{in},\, NF_{in}]\,,\; j \in \{1, 2, ..., NR\},\; k \in \{1, 2, ..., NI\} \tag{3}$$

$$x^i_{j,k,G} \in [-NF_{out},\, NF_{out}]\,,\; j \in \{1, 2, ..., NR\},\; k = NI + 1 \tag{4}$$

$$x^i_{j,k,G} \in \{1, 2\},\; j \in \{1, 2, ..., NR\},\; k = D \tag{5}$$

where NI denotes the number of input variables and NF_{in} and NF_{out} represent the number of input features inputs and outputs features, respectively. Note two possible connectives are considered ("AND" and "OR" represented by 1 and 2, respectively) and rules weights are equal to unity.

Once population has been initialized, every candidate, RB^i_G, undergoes mutation, crossover and selection processes. Mutation is typically employed in EAs as a way to avoid local optimums and thus to expand search space. With this aim, for every target base RB^i_G, three individuals, $RB^{r_1}_G$, $RB^{r_2}_G$ and $RB^{r_3}_G$ are randomly selected in a way that r_1, r_2 and r_3 are different to i and the weighted difference of $RB^{r_2}_G$ and $RB^{r_3}_G$ is added to the third base $RB^{r_1}_G$:

$$DB^i_{G+1} = RB^{r_1}_G + F(RB^{r_2}_G,\, RB^{r_3}_G) \tag{6}$$

where F is the mutation factor, $F \in [0, 2]$, and DB^i_{G+1} is the donor base for RB^i_G. Next, crossover is considered. Within this stage, the trial base TB^i_{G+1} is constructed though the elements of both target, RB^i_G and donor bases, DB^i_{G+1}. Specifically, elements of DB^i_{G+1} are incorporated to TB^i_{G+1} with probability CR:

$$TB^i_{j,k,G+1} = \begin{cases} DB^i_{j,k,G+1} & if\, rand^i_{j,k} \leq CR \\ RB^i_{j,k,G+1} & if\, rand^i_{j,k} > CR \end{cases} \tag{7}$$

with $j \in \{1, 2, ..., NR\}$, $k \in \{1, 2, ..., D\}$. It must be highlighted that uniform distributions are considered for random selections in this approach. Finally, selection process is responsible for deciding which bases are chosen for the next generation. To this end, both target bases and trial bases are evaluated and the ones with higher scores are selected:

$$RB^i_{G+1} = \begin{cases} TB^i_{G+1} & if\, f(TB^i_{G+1}) \leq f(RB^i_G) \\ RB^i_G & otherwise \end{cases} \tag{8}$$

Mutation, crossover and selection processes are repeated until the stopping condition is met. Concretely, in this work, a fixed number of generations is considered as stopping condition. It is to be mentioned there exist other versions for DE, including exponential crossover or x/rand differential variation [10]. However, in this work, a standard DE-based strategy is proposed for rules evolution.

4 Simulations Results and Discussion

The suggested learning strategy and meta-scheduler performance are evaluated with Alea [18]. Alea is a simulation software based on GridSim toolkit that allows the analysis of scheduling strategies in computational grids. In this work, the grid scenario and jobs traces are obtained from Czech National Grid Infrastructure Metacentrum [19]. Metacentrum is a CESNET (operator of academic network of the Czech Republic -National Research and Education Network, NREN) project whose goal is to cooperate in the development of a high performance computational infrastructure by coordinating a large set of institutions resources worldwide. Specifically, the grid is made up of 210 machines, integrating 806 CPUs of heterogeneous types (i.e., Opteron and Xeon) and speed (i.e., 1500-3200 MHz), running Linux and distributed in 14 RDs. On the other hand, jobs are collected from Metacentrum traces from January to May 2009 (available at [19]). Concretely, 2000 and 2400 jobs are considered in training and validation scenario, respectively. Also, *makespan* [3] is suggested as training index for the learning process.

With the aim of providing a fair step by step comparison in terms of computational effort or functions evaluations (FEs) through generations, Pittsburgh strategy is taken into account [4,9]. As mentioned before, Pittsburgh population is made up of a set of RBs and genetic operators are applied at this level. Hence, as it occurs in suggested DE-based strategy, every generation requires a number of selected population FEs. This way, both strategies can be compared at every generation bearing in mind the same computational effort. To be precise, DE is configured with binomial crossover with probability CR=0.5, mutation factor F=0.8 with 1/rand differential variation, population size -18 NP- and RB size set to 10 rules. On the other hand, Pittsburgh approach considers two-point crossover (i.e., two random cut points are selected for the combination of RBs [4]), exponential decreasing mutation factor (i.e., random modification of genome elements as to avoid local optima), elitist selection $\lambda = 0.9$ (i.e., those individuals achieving the best accuracy are selected with rate λ), population size of 20 RBs in a way that 18 particles are evaluated at every generation and initial maximum RB size fixed to 20 rules. Simulations are conducted for 120 generations and 30 experiments for every learning strategy. Note that both evolutionary strategies behaviour can be compared considering the same genome, number of generations and total FEs. Also, in order to further analyze Pittsburgh approach in comparison to DE, we conducted simulations with Pittsburgh approach in an alternative configuration. Specifically, Pittsburgh behaviour is also studied with $\lambda = 0.8$ and initial maximum RB size set to 10 rules. Fig. 2 illustrates convergence behaviour for best particle at every generation for DE and considered configurations for Pittsburgh strategy. It is shown that DE-based learning strategy converges faster than Pittsburgh approach from tenth generation and achieves a more accurate final result than Pittsburgh approach in both settings. Concretely, DE improves best Pittsburgh configuration solution final training fitness by 1.54% on average, as presented in Table 2. Further, Table 2 summarizes statistics on the 30 runs regarding distribution (average, best and worse). It is shown that DE best solution outperforms best Pittsburgh

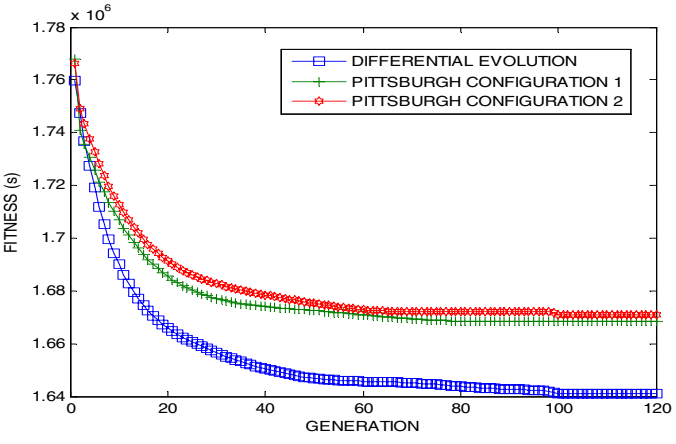


Fig. 2. DE and Pittsburgh approach best individual average fitness (*makespan*)

Table 2. Training results

Strategy/fitness statistic (hours)	Average fitness	Best fitness	Worst fitness
DIFFERENTIAL EVOLUTION	456.0620	433.8595	459.5848
Pittsburgh configuration 1	463.2184	451.4051	467.8432
Pittsburgh configuration 2	465.0112	457.8231	467.8431

Table 3. Scheduling strategies results in validation scenario

Metric/Strategy	Fuzzy-DE	Fuzzy-Pittsburgh	EASY-BF
Makespan (hours)	456.0620	461.0906	485.9961
Flow-time (hours)	24.6542	24.6849	24.3032
Weighted usage (%)	46.97	46.42	44.47
Classic usage (%)	58.28	56.52	47.01
Tardiness (hours)	1.3377	1.3215	0.8987

solution by 3.89%. Also, DE worst solution is 1.77% reducer than Pittsburgh one. Moreover, Table 3 presents results for DE, Pittsburgh and EASY schedulers in validation scenario. It is observed that DE performance in *makespan* outperforms Pittsburgh (in its best configuration -configuration 1-) by 1.10% and EASY-BF by 6.16%. Further, DE-based scheduler succeeds in improving *machine weighted* (1.17% and 5.32% in comparison to Pittsburgh and EASY-BF, respectively) and *classic usage* (3.02% with respect to Pittsburgh and 19.34% to EASY-BF). However, as expected, metrics such as *flowtime* and *tardiness* present a more deteriorated performance. Note that the proposed learning strategy has been trained to improve *makespan*, which may have conflicting interest with these two criteria.

5 Conclusions and Future Work

In this work, a meta-scheduler based on FRBSs for grid computing has been suggested. The meta-scheduler obtains a fuzzy characterization of every participating RD state in order to consider system inherent uncertainty and infers an associated performance index showing the suitability to be selected for the next schedule. The inference process is subject to the system knowledge quality and thus with the learning process. That is, RBs quality is relevant for the whole meta-scheduler performance and a new evolutionary strategy for the evolution of fuzzy rules inspired on DE has been proposed in this work. DE has proven to be an efficient optimization algorithm in diverse areas of science and engineering and its adaptation to fuzzy meta-schedulers learning has been presented. Simulations results show DE-based learning strategy improves other classic learning strategy, Pittsburgh approach, in training fitness (*makespan*) final result (1.54%) and convergence behaviour. Also, it has been shown that DE-based scheduler outperforms *EASY-BF* scheduling strategy in terms of *makespan* by 6.16% and *machine weighted* and *classic usage* by 5.32% and 19.34%, respectively. Thus, accuracy in training results and convergence properties together with algorithm simplicity supports the use of DE for learning of FRBS schedulers for grid computing. In future work, new versions of DE and multi-objective techniques will be studied as to further improve the fuzzy meta-scheduler performance.

Acknowledgments

This work has been financially supported by the Andalusian Government (Research Project P06-SEJ-01694).

References

1. Foster, I., Kesselman, C.: The Grid 2: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers Inc., San Francisco (2003)
2. Klusacek, D.: Dealing with Uncertainties in Grids through the Event-based Scheduling Approach. In: Fourth Doctoral Workshop on Mathematical and Engineering Methods in Computer Science (MEMICS 2008), vol. 1, pp. 978–980 (2008)
3. Xhafa, F., Abraham, A.: Meta-heuristics for grid scheduling problems. In: Metaheuristics for Scheduling: Distributed Computing Environments. Studies in Computational Intelligence. Springer, Germany (2008), ISBN: 978–3-540-79437-0
4. Cordón, O., Herrera, F., Hoffmann, F., Magdalena, L.: Genetic fuzzy systems: Evolutionary tuning and learning of fuzzy knowledge bases. World Scientific Pub. Co. Inc., Singapore (2001)
5. Jamaludin, J., Rahim, N., Hew, W.: Development of a self-tuning fuzzy logic controller for intelligent control of elevator systems. Engineering Applications of Artificial Intelligence 22(8), 1167–1178 (2009)
6. Muñoz-Expósito, J.E., García-Galán, S., Ruiz-Reyes, N., Vera-Candeas, P.: Adaptive network-based fuzzy inference system vs. other classification algorithms for warped lpc-based speech/music discrimination. Eng. Appl. Artif. Intell. 20(6), 783–793 (2007)

7. Franke, C., Hoffmann, F., Lepping, J., Schwiegelshohn, U.: Development of scheduling strategies with genetic fuzzy systems. Appl. Soft Comput. 8(1), 706–721 (2008)
8. Prado, R.P., Galán, S.G., Yuste, A.J., Expósito, J.E.M., Santiago, A.J.S., Bruque, S.: Evolutionary fuzzy scheduler for grid computing. In: Cabestany, J., Sandoval, F., Prieto, A., Corchado, J.M. (eds.) IWANN 2009. LNCS, vol. 5517, pp. 286–293. Springer, Heidelberg (2009)
9. Ishibuchi, H., Yamamoto, T., Nakashima, T.: Hybridization of fuzzy gbml approaches for pattern classification problems. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 35(2), 359–365 (2005)
10. Storn, R., Price, K.: Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J. of Global Optimization 11(4), 341–359 (1996)
11. Wang, W.-H., Wang, F.-R., Pan, Q.-K., Zuo, F.-C.: Improved differential evolution algorithm for location management in mobile computing. In: International Workshop on Intelligent Systems and Applications, ISA 2009, pp. 1–5 (2009)
12. Yüzgeç, U.: Performance comparison of differential evolution techniques on optimization of feeding profile for an industrial scale baker's yeast fermentation process. ISA Transactions 49(1), 167–176 (2010)
13. Zhang, X., Chen, W., Dai, C., Cai, W.: Dynamic multi-group self-adaptive differential evolution algorithm for reactive power optimization. International Journal of Electrical Power and Energy Systems (2009) (in Press, Corrected Proof)
14. Klusacek, D., Rudova, H., Baraglia, R., Pasquali, M., Capannini, G.: Comparison of multi-criteria scheduling techniques. In: Grid Computing: Achievements and prospects, pp. 173–184. Springer, Heidelberg (2008)
15. Thain, D., Tannenbaum, T., Livny, M.: Distributed computing in practice: The Condor experience. Concurrency and Computation Practice and Experience 17(2-4), 323–356 (2005)
16. Venugopal, S., Buyya, R., Winton, L.: A grid service broker for scheduling distributed data-oriented applications on global grids. In: Proceedings of the 2nd workshop on Middleware for grid computing, pp. 75–80. ACM, New York (2004)
17. Klusacek, D., Rudova, H.: Improving QoS in computational Grids through schedule-based approach. In: Scheduling and Planning Applications Workshop at the Eighteenth International Conference on Automated Planning and Scheduling (ICAPS 2008), Sydney, Australia (2008)
18. Klusacek, D., Matyska, L., Rudova, H.: Alea - Grid scheduling simulation environment. In: Wyrzykowski, R., Dongarra, J., Karczewski, K., Wasniewski, J. (eds.) PPAM 2007. LNCS, vol. 4967, pp. 1029–1038. Springer, Heidelberg (2008)
19. C. N. G. Infrastructure, Metacentrum data sets, http://www.fi.muni.cz/~xklusac/index.php?page=meta2009

Author Index