Vicenç Torra
Yasuo Narukawa
Aïda Valls
Josep Domingo-Ferrer (Eds.)

LNAI 3885

Modeling Decisions for Artificial Intelligence

Third International Conference, MDAI 2006
Tarragona, Spain, April 2006
Proceedings

Springer

Lecture Notes in Artificial Intelligence 3885

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Vicenç Torra Yasuo Narukawa
Aïda Valls Josep Domingo-Ferrer (Eds.)

Modeling Decisions for Artificial Intelligence

Third International Conference, MDAI 2006
Tarragona, Spain, April 3-5, 2006
Proceedings

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Vicenç Torra
Institut d'Investigació en Intel.ligència Artificial
Consejo Superior de Investigaciones Cientificas
Campus UAB, 08193 Bellaterra, Catalonia, Spain
E-mail: vtorra@iiia.csic.es

Yasuo Narukawa
Toho Gakuen
3-1-10, Naka, Kunitachi
Tokyo, 186-0004, Japan
narukawa@fz.dis.titech.ac.jp

Aïda Valls
Josep Domingo-Ferrer
Universitat Rovira i Virgili
ETSE, Departament d'Enginyeria Informàtica i Matemàtiques
Campus Sescelades, Avinguda Països Catalans, 26, 43007 Tarragona, Catalonia, Spain
E-mail: avalls@urv.net, josep.domingo@urv.net

Library of Congress Control Number: 2006921542

CR Subject Classification (1998): I.2, F.4.1, F.1, H.2.8, I.6

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-32780-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-32780-6 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11681960 06/3142 5 4 3 2 1 0

Preface

This volume contains papers presented at the 3rd International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2006), held in Tarragona, Catalonia, Spain, April 3-5, 2006. This conference followed MDAI 2004 (held in Barcelona, Catalonia) and MDAI 2005 (held in Tsukuba, Japan) both with proceedings also published in the LNAI series (Vols. 3131 and 3558).

The aim of this conference was to provide a forum for researchers to discuss theory and tools for modeling decisions, as well as applications that encompass decision-making processes and information fusion techniques.

The organizers received 97 papers from 21 different countries, 31 of which are published in this volume. Each submission received at least two reviews from the Program Committee and a few external reviewers. We would like to express our gratitude to them for their work. The plenary talks presented at the conference are also included in this volume.

The conference was supported by the Universitat Rovira i Virgili (School of Engineering and the Department of Computer Engineering and Mathematics), the CSIC, the Department of Risk Engineering (U. Tsukuba), the Catalan Association for Artificial Intelligence (ACIA), the European Society for Fuzzy Logic and Technology (EUSFLAT), the Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT), the Generalitat de Catalunya (AGAUR 2004XT 0004), the Ajuntament de Tarragona, the Diputació de Tarragona and the Spanish Ministry of Science and Education.

December, 2005

Vicenç Torra
Yasuo Narukawa
Aïda Valls
Josep Domingo-Ferrer

Organization

Modeling Decisions for Artificial Intelligence – MDAI 2006

General Chairs

Aïda Valls, Universitat Rovira i Virgili, Catalonia, Spain
Josep Domingo-Ferrer, Universitat Rovira i Virgili, Catalonia, Spain

Program Chairs

Vicenç Torra, IIIA-CSIC, Catalonia, Spain
Yasuo Narukawa, Toho Gakuen, Japan

Program Committee

G. Beliakov (Australia)
U. Bodenhofer (Austria)
T. Calvo (Spain)
J. Dujmovic (USA)
M. Grabisch (France)
E. Herrera-Viedma (Spain)
F. Herrera (Spain)
K. Hirota (Japan)
M. Inuiguchi (Japan)
J. Kacprzyk (Poland)
Z.-Q. Liu (Hong Kong, China)
L. Magdalena (Spain)
J.-L. Marichal (Luxemburg)
R. Meo (Italy)
R. Mesiar (Slovakia)
A. Moreno (Catalonia, Spain)
S. Miyamoto (Japan)
T. Murofushi (Japan)
M. Ng (Hong Kong, China)
T. Onisawa (Japan)
G. Pasi (Italy)
C. Sierra (Catalonia, Spain)
R. R. Yager (USA)
N. Zhong (Japan)

Local Organizing Committee Chairs

Antonio Moreno, Universitat Rovira i Virgili, Catalonia, Spain
David Riaño, Universitat Rovira i Virgili, Catalonia, Spain

Finance Chair

Antonio Moreno, Universitat Rovira i Virgili, Catalonia, Spain

Organizing Committee

Montserrat Batet, John Bohada, Jordi Castellà, David Isern, Xavier Mallafré, Antoni Martínez, Pere Millan, David Sànchez, Francesc Sebé, Agustí Solanas

Additional Referees

A. Solanas, J. M. Garrell-Guiu, J. Nin, J. Herrera, J. Jacas, E. Okamoto, Y. Maeda, K. Fujimoto, Y. Endo, G. Escalada, J. Castro, A. Valls, J. Osawa, Y. Yoshida, Y. Ohtsubo, E. Szmidt, L. Xie, J. Zeng, W. Feng, U. Rothblum, T. Denoeux, J. Schiltz, F. Klijn, M. Yasuda, J. Nakagami, M. Kurano

Supporting Institutions

Universitat Rovira i Virgili (URV)
Departament d'Enginyeria Informàtica i Matemàtiques (DEIM, URV)
Escola Tècnica Superior d'Enginyeria (ETSE, URV)
Consejo Superior de Investigaciones Científicas (CSIC)
Department of Risk Engineering (U. Tsukuba)
Catalan Association for Artificial Intelligence (ACIA)
European Society for Fuzzy Logic and Technology (EUSFLAT)
Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT)
Generalitat de Catalunya (AGAUR 2004XT 0004)
Ajuntament de Tarragona
Diputació de Tarragona
Spanish Ministry of Science and Education

Table of Contents

Invited Talks

Regular Papers

Asymmetric and Compound Preference Aggregators

Jozo J. Dujmović

Department of Computer Science, San Francisco State University,
1600 Holloway Ave, San Francisco, CA 94132, USA
jozo@sfsu.edu

Extended Abstract. Choosing among options and selecting the best alternative is a fundamental component of human decision-making. The best alternative is the result of a mental process called *system evaluation*. The evaluated system can be any collection of interrelated components. The system as a whole and its components are expected to have some desired features and satisfy specific requirements. Consequently, system evaluation is a process of determining the extent to which a system satisfies a given set of requirements.

The process of intuitive system evaluation has all characteristics of fuzzy reasoning and computing with words in the sense of Zadeh:

- The process occurs in an environment of imprecision, uncertainty and partial truth.
- The process is based on manipulating perceptions of value, which we call preferences. Preferences reflect degrees of satisfaction of requirements (0=no satisfaction, 1=complete satisfaction, and values between 0 and 1 reflect a partial satisfaction). The global satisfaction of a set of requirements (the value of a complex system) is also a matter of degree.
- Intuitive evaluation is based on manipulating perceptions of value without any measurements and any computation. It is imprecise, granular (based on fuzzy information granulation), and reflects the limited human ability to distinguish detail and store information.

The mental process of system evaluation has observable properties. These properties can be used to develop mathematical models of the intuitive system evaluation. Our framework for mathematical modeling of system evaluation is called the Continuous Preference Logic (CPL). The goals of CPL are to explain, organize, refine, enhance, and extrapolate the intuitive evaluation process through mathematical modeling and building of software tools.

CPL models must be developed in concordance with needs and limitations of human decision process, avoiding both oversimplifications and exaggerated complexity. We differentiate two almost independent components present in all CPL models:

- Formal logic components
- Semantic components

Formal logic components specify and manipulate simultaneity (andness), replaceability (orness) and negation to build and to structure compound evaluation criteria. Semantic components manipulate perceptions of human goals, by differentiating and adjusting the levels of relative importance of satisfying individual system

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 1 – 4, 2006.

requirements. The overlap of formal logic and semantic components is visible when we define and compare the compensatory power of individual inputs in system evaluation models. The compensatory power is affected both by selecting the level of andness/orness and by specifying relative importance (weights). The level of overlap of formal logic and semantic components can be identified and exemplified, but its deeper understanding requires more research.

CPL models quantify the process of system evaluation performed by human brain. In this case the brain manipulates perceptions of degrees of satisfaction of requirements. The manipulation is aggregative and its goal is to determine a perception of the global quality of the system by aggregating degrees of satisfaction of component requirements. Consequently, the fundamental goal of CPL is to develop mathematical models of the process of aggregation of preferences. The fundamental component of these models is the generalized conjunction/disjunction (GCD) function that is used to model simultaneity and replaceability of evaluation criteria. The main parameter of GCD is andness/orness that can be adjusted to generate a spectrum of aggregators in the range from AND to OR. Andlike operators are called the partial conjunction, and orlike operators are called the partial disjunction. The properties of basic operators of full conjunction, partial conjunction, conjunctive/disjunctive neutrality, partial disjunction, full disjunction, and negation are summarized in Table 1.

The concepts of andness, orness, continuous transition from conjunction to disjunction and models of partial conjunction and partial disjunction were introduced in 1973. For more than 30 years many research efforts were devoted to modeling simultaneity and replaceability using a spectrum of partial conjunction and partial disjunction models, as well as t-norms and t-conorms. The area that is beyond partial conjunction and partial disjunction attracted much less attention.

In the area of building complex system evaluation criteria we are interested in compound aggregators based on superposition of basic functions ANDOR, ORAND, AM, AND, OR, and NOT. The basic functions shown in Table 1 are symmetric, where all inputs have equal role, both semantically and in the formal logic sense.

The first step in increasing the flexibility of aggregators is to introduce weights that reflect different relative importance of inputs and create asymmetric aggregators. The corresponding asymmetric weighted aggregators have a semantic component, but their logical structure can still be as simple as the partial conjunction or the partial disjunction. The next step is to develop fundamental aggregators whose logic structure is asymmetric and more complex.

Two most frequently used fundamental compound aggregators are the conjunctive partial absorption (CPA) and the disjunctive partial absorption (DPA). They are also summarized in Table 1. CPA is an asymmetric aggregator that combines a mandatory input and a desired input. This aggregator is asymmetric at the formal logic level: if the requirement expressed as the mandatory input is not satisfied, then the whole criterion is not satisfied. This is not the case with the desired input. If the requirement of the desired input is not satisfied, the compound criterion can still be partially satisfied. DPA is an aggregator that is dual to CPA: it combines a sufficient input and a desired input. The aggregator reflects the asymmetric situation where the criterion can be fully (or substantially) satisfied by satisfying only the sufficient requirement. The same effect cannot be achieved by satisfying the desired input.

Table 1. Fundamental operators in Continuous Preference Logic

Name	Description	Symbol	Sample implementation
NOT	**Negation**	$\overline{x}$	$1-x$
AND	**Full conjunction**, the maximum level of simultaneity.	$x_1 \wedge x_2$	$\min(x_1, x_2)$
ANDOR (PC)	**Partial conjunction**, a spectrum of simultaneity levels. All input preferences must be to some extent simultaneously satisfied.	$x_1 \Delta x_2$	$(0.5x_1^r + 0.5x_2^r)^{1/r}$ $-\infty < r < 1$
AM (CDN)	**Arithmetic mean**, or conjunctive/ disjunctive neutrality. AM models a perfect balance of simultaneity and replaceability. All inputs are desired, but no one is mandatory or sufficient.	$x_1 \ominus x_2$	$0.5x_1 + 0.5x_2$
ORAND (PD)	**Partial disjunction**, a spectrum of replaceability levels. Each input can be used to partially compensate the lack of remaining inputs.	$x_1 \nabla x_2$	$(0.5x_1^r + 0.5x_1^r)^{1/r}$ $1 < r < +\infty$
OR	**Full disjunction**, the maximum level of replaceability.	$x_1 \vee x_2$	$\max(x_1, x_2)$
CPA	**Conjunctive Partial Absorption**. A combination of mandatory and desired inputs. The mandatory input (m) must be (at least partially) satisfied. Assuming m>0, the desired input (d) can partially compensate an insufficient level of m.	$m \trianglerighteq d$	$m\Delta(m\nabla d)$ Extreme examples: $m \wedge (0.5m + 0.5d)$ $\sqrt{m(0.5m + 0.5d)}$
DPA	**Disjunctive Partial Absorption**. A combination of sufficient and desired inputs. The sufficient input (s) can fully (or substantially) compensate the lack of desired input (d). The desired input can partially compensate the lack of sufficient input.	$s \overline{\triangleright} d$	$s\nabla(s\Delta d)$ Extreme examples: $s \vee (0.5s + 0.5d)$ $\sqrt{\frac{s^2 + (0.5s + 0.5d)^2}{2}}$

Asymmetric operators of CPA and DPA can be compound by nesting. This process yields aggregators that have three distinct logical levels: mandatory, desired and optional (by nesting CPA), and sufficient, desired, and optional (by nesting DPA). This process could be continued to any number of logical levels, but it has no sense and

applicability because the resulting granularity would be beyond the granularity that is reachable by human brain. Nesting of partial absorption aggregators is a typical example that illustrates natural limits of complexity of preference aggregators.

Fundamental preference aggregators can be used to create complex evaluation criteria using a process of stepwise aggregation of preferences. This process is a systematic way to linearly increase the complexity of compound aggregators by always aggregating a small number (e.g. up to 5) of intermediate preferences. Each aggregation creates a granule of similar requirements. The stepwise aggregation of preferences systematically increases the size of granules, enabling justifiable decision-making with compound requirements of increasing complexity.

Asymmetric and compound preference aggregators are one of ways towards computer-enhanced decision making (CEDM). CEDM can be interpreted as an advanced proactive form of decision support, that extends natural human abilities using processes and tools that *explain*, *organize*, *refine*, *enhance*, and *extrapolate* human reasoning in the area of (mostly professional) system evaluation.

Basic preference aggregators reflect observable properties of human decision making through abstraction and quantification. This *explains* the components of human decision process and provides a way to understand and justify the selection of preference aggregators.

The process of stepwise aggregation of preferences provides a systematic way to *organize* and structure the decision process. Without a clear organizational structure the decision process can become chaotic and unreliable.

At the root of the aggregation tree we can aggregate only several global granules. For example, in the case of computer evaluation models the global granules are the subsystems of hardware, software, performance, and vendor support. Each of these granules can be further independently decomposed (e.g., hardware = processor + memory + external memory + I/O + communication units). By continuing this process we provide a *stepwise refinement* of decision models.

Previous steps provide infrastructure that can be used to build tools and techniques that *enhance* human decision models beyond the level of intuitive human reasoning. For example, the complex quantitative criteria can be verified using the sensitivity analysis, tradeoff analysis, and reliability analysis. The sensitivity analysis investigates the effects of changes of inputs or parameters on the value of selected preferences. The tradeoff analysis investigates compensatory features of evaluation criteria, i.e. the possibility to compensate deficiencies in some inputs by improving selected other inputs. The goal of reliability analysis is to compute the confidence levels for ranking of each pair of competitive systems. It enables evaluators to select the best system knowing the level of confidence that corresponds to the proposed decision.

All fundamental preference aggregators are fully specified using a small number of adjustable parameters (e.g. 3 parameters specify CPA and DPA). However, the aggregators give valid aggregation results in all points of the input preference space *extrapolating* human abilities beyond natural limitations. Another example of extrapolation is the process of system optimization, where a quantitative criterion is used to find optimum systems that maximize preference for constrained cost.

The presented techniques show the role of asymmetric and compound aggregators in the area of system evaluation, as well as a promising example of CEDM.

Computational Models of Language Toward Brain-Style Computing

Michio Sugeno

Faculty of Culture and Information Science,
Doshisha University,
1-3 Miyakodani Tatara, Kyotanabe,
Kyoto 610-0394, Japan
msugeno@mail.doshisha.ac.jp

Abstract. The human brain consists of a neural system as hardware and a language system as software. It is, therefore, possible to take two approaches to create the human brain. While the hardware-centered approach is based on computational neuroscience, it is possible to base the software-centered approach on linguistics.

Brain-style computing is considered as one of the main research areas in creating the brain. We take a language-based approach to brain-style computing. To this aim, we have adopted as the basic theory Systemic Functional Linguistics (SFL) initiated by Halliday.

Following Halliday's four principles in the design of human language, we have implemented the computational model of language in context, called the Semiotic Base, and we have developed a set of algorithms of text understanding and generation using this model. The language used in this study is Japanese.

As an application of the models, we are developing Brain-Style Computing System under which we can manage and execute all kinds of computing through meanings. The idea is to verbalize computers by constructing linguistic models of software and hardware applications. Brain-Style Computing System consists of Everyday Language Interface with a Secretary Agent, Semiotic Base, Language Applications, Language Communication Protocol and Language Operating System.

In this talk, I shall discuss some linguistic issues in creating the brain. There are three higher-order functions of the brain concerned with language: processing, utilizing, and learning language. Processing language such as understanding and generation is a basic function with the internal models of language itself and its processing. SFL could reveal what the internal models must be like. SFL could also play an essential role in elucidating the brain functions of language such as thinking with language and learning language.

I shall also show some clinical evidence obtained from studies on aphasia which support the SFL perspective on the system of language. I shall also refer to the brain internal models for motor control and some learning mechanisms in the brain which might be related with language functions.

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 5–6, 2006.

References

1. Halliday, M.A.K. and Matthiessen, C.M.I.M.: Construing experience through meaning, Cassel, 1999
2. Ito N., Sugimoto T., and Sugeno M: "A systemic-functional approach to Japanese text understanding", Computational Linguistics and Intelligent Text Processing, Proc. 5th Int. Conf.(CICLing 2004), Seoul, Korea, pp. 26-37, 2004
3. Takahashi Y., Kobayashi I., Iwazume M., Ito N., and Sugeno M.: "The contents and structure of the Context Base, and its application", Advances in Natural Language Processing, Proc. 4th Int. Conf.(EsTAL 2004), Alicante, Spain, pp. 103-114, 2004

Dominance-Based Rough Set Approach to Case-Based Reasoning

Salvatore Greco[1], Benedetto Matarazzo[1], and Roman Slowinski[2]

[1] Faculty of Economics, University of Catania,
Corso Italia, 55, 95129 – Catania, Italy
[2] Institute of Computing Science, Poznan University of Technology,
60-965 Poznan, and Institute for Systems Research,
Polish Academy of Sciences, 01-447 Warsaw, Poland

Abstract. Case-based reasoning is a paradigm in machine learning whose idea is that a new problem can be solved by noticing its similarity to a set of problems previously solved. We propose a new approach to case-based reasoning. It is based on rough set theory that is a mathematical theory for reasoning about data. More precisely, we adopt Dominance-based Rough Set Approach (DRSA) that is particularly appropriate in this context for its ability of handling monotonicity relationship between ordinal properties of data related to monotonic relationships between attribute values in the considered data set. In general terms, monotonicity concerns relationship between different aspects of a phenomenon described by data: for example, "the larger the house, the higher its price" or "the closer the house to the city centre, the higher its price". In the perspective of case-based reasoning, we propose to consider monotonicity of the type "the more similar is y to x, the more credible is that y belongs to the same set as x". We show that rough approximations and decision rules induced from these approximations can be redefined in this context and that they satisfy the same fundamental properties of classical rough set theory.

1 Introduction

Case-based reasoning (for a general introduction to case-based reasoning see e.g. [10]; for a fuzzy set approach to case-based reasoning see [3]) regards the inference of some proper conclusions related to a new situation by the analysis of similar cases from a memory of previous cases. It is based on two principles [11]:

a) similar problems have similar solutions;
b) types of encountered problems tend to recur.

Gilboa and Schmeidler [4] observed that the basic idea of case-based reasoning can be found in the following sentence of Hume [9]: "From causes which appear *similar* we expect similar effects. This is the sum of all our experimental conclusions." Rephrasing Hume, one can say that "the more similar are the causes, the more similar one expects the effects." Therefore, measuring similarity is the

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 7–18, 2006.

essential point of all case-based reasoning and, particularly, of fuzzy set approach to case-based reasoning [3]. This explains the many research problems that measuring similarity generates within case-based reasoning. Problems of modelling similarity are relative to two levels:

- at level of similarity with respect to single features: how to define a meaningful similarity measure with respect to a single feature?
- at level of similarity with respect to all features: how to properly aggregate the similarity measure with respect to single features in order to obtain a comprehensive similarity measure?

Taking this into account we propose an approach to case-based reasoning which tries to be possibly "neutral" and "objective" with respect to similarity relation, in the sense that at level of similarity concerning single features, we consider only ordinal properties of similarity, and at level of aggregation, we do not impose any specific functional aggregation based on very specific axioms (see for example [4]), but we consider a set of decision rules based on the very general monotonicity property of comprehensive similarity with respect to similarity of single features. Therefore our approach to case-based reasoning is very little "invasive", comparing to the many other existing approaches.

Our approach to case-based reasoning is based on rough set theory ([12, 13]). Rough set theory relies on the idea that some knowledge (data, information) is available about elements of a set. For example, knowledge about patients suffering from a certain disease may contain information about body temperature, blood pressure, etc. All patients described by the same information are indiscernible in view of the available knowledge and form groups of similar cases. These groups are called elementary sets and can be considered as elementary building blocks of the available knowledge about patients. Elementary sets can be combined into compound concepts. Any union of elementary sets is called crisp set, while other sets are referred to as rough set. Each rough set has boundary line cases, i.e. objects which, in view of the available knowledge, cannot be classified with certainty as members of the set or of its complement. Therefore, in the rough set approach, any set is associated with a pair of crisp sets called the lower and the upper approximation. Intuitively, the lower approximation consists of all objects which certainly belong to the set and the upper approximation contains all objects which possibly belong to the set. The difference between the upper and the lower approximation constitutes the boundary region of the rough set.

In our approach to case-based reasoning we do not consider classical rough set theory but its extension called Dominance-based Rough Set Approach (DRSA) [5, 6] that has been proposed to handle ordinal properties of data related to preferences in decision problems. The monotonicity, which is crucial for DRSA, is also meaningful for problems where preferences are not considered. Generally, monotonicity concerns relationship between different aspects of a phenomenon described by data. More specifically, it concerns mutual trends between different variables like distance and gravity in physics or inflation rate and interest rate in economics. Whenever we discover a relationship between different aspects of a phenomenon, this relationship can be represented by a monotonicity

with respect to some specific measures of the considered aspects. So, in general, the monotonicity is a property translating in a formal language a primitive intuition of interaction between different concepts in our knowledge domain. As discovering is an inductive process, it is illuminating to remember the following Proposition 6.363 of Wittgenstein [17]: "[illegible]". We claim that this simplest law is just monotonicity and, therefore, each data analysis method can be seen as a specific way of dealing with monotonicity.

Let us observe that monotonicity is also present in classical rough set theory. In fact, rough set philosophy employs approximation for describing relationships between concepts. For example, coming back to above example of medical diagnosis, the concept of "disease Y" can be represented in terms of such concepts as "low blood pressure and high temperature" or "muscle pain and headache". The approximation is based on a very coarse representation in the sense that, for each aspect characterizing concepts ("low blood pressure", "high temperature", "muscle pain", etc.), only its presence or its absence is considered relevant. Therefore, rough approximation within classical rough set theory involves a very primitive idea of monotonicity related to a scale with only two values: "presence" and "absence".

Monotonicity gains importance when a finer representation of the concepts is considered. A representation is finer when for each aspect characterizing concepts, not only its presence or its absence is taken into account, but also the [illegible] of its presence or absence is considered relevant. Due to graduality, the idea of monotonicity can be exploited in the whole range of its potential. Graduality is typical for fuzzy set philosophy [18] and, therefore, a joint consideration of rough sets and fuzzy sets is worthwhile. In fact, rough set and fuzzy set capture the two basic complementary features of the idea of monotonicity: rough set deals with relationships between different aspects and fuzzy sets deal with expression of different dimensions representing the considered concepts. For this reason, many approaches have been proposed to combine fuzzy sets with rough sets (see for example [2, 16]).

Greco, Matarazzo and Slowinski [7] showed how the framework of DRSA can be very naturally extended to represent any relationship of monotonicity in reasoning about data. In this context one can envisage a knowledge representation model composed of a set of decision rules with the following syntax:

"if object y presents feature f_{i1} in degree at least h_{i1}, and feature f_{i2} in degree at least h_{i2} ..., and feature f_{im} in degree at least h_{im}, then object y belongs to set X in degree at least α".

Greco, Matarazzo and Slowinski [7] proved also that the classical rough set approach [12, 13] can be seen as specific case of the general DRSA model. This is important for several reasons; in particular, this interpretation of DRSA gives an insight into fundamental properties of the classical rough set approach and permits to further generalize the rough set approach.

In this paper, we show that in the framework of DRSA a rough set approach to case-based reasoning can be developed very naturally. Here, the monotonicity concerns the relationships between similarity to some reference objects and membership to some specific sets. In this context we envisage a knowledge representation model composed of a set of decision rules with the following syntax:

"if object y is similar to object x w.r.t. feature f_{i1} in degree at least h_{i1}, and w.r.t. feature f_{i2} in degree at least h_{i2}, and ..., and w.r.t. feature f_{im} in degree at least h_{im}, then object y belongs to set X in degree at least α",
where w.r.t. means "with respect to".

These decision rules are similar to the gradual decision rules [1] being statements of the form "the more object z is X, the more it is Y" or, equivalently, but more technically,

$$\mu_X(z) \geq \alpha \Rightarrow \mu_Y(z) \geq \alpha$$

where X and Y are fuzzy sets whose membership functions are μ_Y and μ_X, and $\alpha \in [0, 1]$.

Within the context of case-based reasoning gradual decision rules assume the following syntax [3]:

"the more object z is similar to a referent object x w.r.t. condition attribute s, the more z is similar to a referent object x w.r.t. decision attribute t"

or, equivalently, but more technically,

$$s(z, x) \geq \alpha \Rightarrow t(z, x) \geq \alpha$$

where s and t measure the credibility of similarity with respect to condition attribute and decision attribute, respectively.

When there is a plurality of condition attributes and decision attributes, functions s and t aggregate similarity with respect to these attributes.

Let us observe that the decision rules we propose do not need the aggregation of the similarity with respect to different features in one comprehensive similarity. This is important because it permits to avoid using aggregation operators (weighted average, min, etc.) which are always arbitrary to some extend. Moreover, the decision rules we propose permit to consider different thresholds for degrees of credibility in the premise and in the conclusion. This is not considered in the gradual decision rules, where the threshold is the same, α, in the premise and in the conclusion.

This article is organized as follows. Section 2 introduces DRSA approach to case-based similarity. Section 3 contains conclusions.

2 Rough Approximation for Case Based Reasoning

In this section, we consider rough approximation of a fuzzy set using a similarity relation in the context of case-based reasoning. The introduced rough approximation is inspired by the rough approximation of a pairwise comparison table within the Dominance-based Rough Set Approach (DRSA).

Let us consider a [illegible] being the 3-tuple

$$\boldsymbol{B} = < U,F,\sigma >,$$

where U is a finite set of [illegible] (universe), $F=\{f_1,f_2,...,f_m\}$ is a finite set of [illegible], and $\sigma:U \times U \times F \rightarrow [0,1]$ is a function such that $\sigma(x,y,f_h) \in [0,1]$ expresses the credibility that object x is similar to object y w.r.t. to feature f_h. The minimal requirement function σ must satisfy is that, for all $x \in U$ and for all $f_h \in F$, $\sigma(x,x,f_h)=1$. Therefore, each pair of objects $(x,y) \in U \times U$ is described by a vector

$$[illegible]_F(x,y)=[\sigma(x,y,f_1), \ldots, \sigma(x,y,f_m)]$$

called [illegible] of (x,y) in terms of the evaluations of the attributes from F; it represents the available information about similarity between x and y. Obviously, similarity between x and y, $x,y \in U$, can be described in terms of any non-empty subset $E \subseteq F$ and in this case we have

$$[illegible]_E(x,y)=[\sigma(x,y,f_h),\ f_h \in E].$$

With respect to any $E \subseteq F$ we can define the dominance relation D_E on $U \times U$ as follows: for any $x,y,w,z \in U$, (x,y) dominates (w,z) with respect to E (denotation $(x,y)D_E(w,z)$) if for any $f_h \in E$

$$\sigma(x,y,f_h) \geq \sigma(w,z,f_h).$$

Given $E \subseteq F$ and $x,y \in U$, let

$$D_E^+(y,x) = \{w \in U : (w,x)D_E(y,x)\},$$
$$D_E^-(y,x) = \{w \in U : (y,x)D_E(w,x)\}.$$

In the pair (y,x), x is considered as [illegible], while y can be called [illegible] because it is conditioning the membership of w in $D_E^+(y,x)$ and in $D_E^-(y,x)$.

Let us also consider a fuzzy set X in U, characterized by the membership function $\mu_X : U \rightarrow [0,1]$. For each cutting level $\alpha \in [0,1]$, the following sets can be defined:

$$X^{\geq\alpha}=\{y \in U : \mu_X(y) \geq \alpha\}, \quad X^{>\alpha}=\{y \in U : \mu_X(y) > \alpha\},$$
$$X^{\leq\alpha}=\{y \in U : \mu_X(y) \leq \alpha\}, \quad X^{<\alpha}=\{y \in U : \mu_X(y) < \alpha\}.$$

For each $\alpha \in [0,1]$ and $* \in \{\geq, >\}$, we can define the E-lower approximation of $X^{*\alpha}$, $\underline{E}_\sigma(X^{*\alpha})$, and the E-upper approximation of $X^{*\alpha}$, $\overline{E}_\sigma(X^{*\alpha})$, based on similarity σ with respect to $E \subseteq F$, respectively, as:

$$\underline{E}_\sigma(X^{*\alpha}) = \{(y,x) \in U \times U : D_E^+(y,x) \subseteq X^{*\alpha}\},$$
$$\overline{E}_\sigma(X^{*\alpha}) = \{(y,x) \in U \times U : D_E^-(y,x) \cap X^{*\alpha} \neq \emptyset\}.$$

For the sake of simplicity, in the following we shall consider $\underline{E}_\sigma(X^{\geq\alpha})$ and $\overline{E}_\sigma(X^{\geq\alpha})$. Of course, analogous considerations hold for $\underline{E}_\sigma(X^{>\alpha})$ and $\overline{E}_\sigma(X^{>\alpha})$. Let us remark that the lower approximation of $X^{\geq\alpha}$ contains all the pairs $(y,x) \in U \times U$ such that any object w being similar to x at least as much as y is similar to x w.r.t. all the considered features $E \subseteq F$ also belongs to $X^{\geq\alpha}$. Thus, on the basis of the data from the fuzzy pairwise information base $\boldsymbol{B}$, if the similarity of an object w to x is not smaller than the similarity of y to x w.r.t. all the considered features $E \subseteq F$, then w belongs to $X^{\geq\alpha}$. In other words, in each pair $(y,x) \in \underline{E}_\sigma(X^{\geq\alpha})$, x is a reference object and y is a limit object which belongs "certainly" to set X with credibility at least α; the limit is understood such that all objects w that are similar to x w.r.t. considered features at least as much as y is similar to x, also belong to X with credibility at least α.

Analogously, the upper approximation of $X^{\geq\alpha}$ contains all the pairs $(y,x) \in U \times U$ such that there is at least one object w being similar to x at least as much as y is similar to x w.r.t. all the considered features $E \subseteq F$ which belongs to $X^{\geq\alpha}$. Thus, on the basis of the data from the fuzzy pairwise information base $\boldsymbol{B}$, if the similarity of an object w to x is not smaller than the similarity of y to x w.r.t. all the considered features $E \subseteq F$, then it is possible that w belongs to $X^{\geq\alpha}$. In other words, in each pair $(y,x) \in \overline{E}_\sigma(X^{\geq\alpha})$, x is a reference object and y is a limit object which belongs "possibly" to set X with credibility at least α; the limit is understood such that there is at least one object w that is similar to x w.r.t. considered features at least as much as y is similar to x and has membership in set X with credibility at least α.

For each $\alpha \in [0,1]$ and $* \in \{\leq, <\}$, we can define the E-lower approximation of $X^{*\alpha}$, $\underline{E}_\sigma(X^{*\alpha})$, and the E-upper approximation of $X^{*\alpha}$, $\overline{E}_\sigma(X^{*\alpha})$, based on similarity σ with respect to $E \subseteq F$, respectively, as:

$$\underline{E}_\sigma(X^{*\alpha}) = \{(y,x) \in U \times U : D_E^-(y,x) \subseteq X^{*\alpha}\},$$

$$\overline{E}_\sigma(X^{*\alpha}) = \{(y,x) \in U \times U : D_E^+(y,x) \cap X^{*\alpha} \neq \emptyset\}.$$

For the sake of simplicity, in the following we shall consider $\underline{E}_\sigma(X^{\leq\alpha})$ and $\overline{E}_\sigma(X^{\leq\alpha})$. Of course, analogous considerations hold for $\underline{E}_\sigma(X^{<\alpha})$ and $\overline{E}_\sigma(X^{<\alpha})$. Let us remark that the lower approximation of $X^{\leq\alpha}$ contains all the pairs $(y,x) \in U \times U$ such that any object w being similar to x at most as much as y is similar to x w.r.t. all the considered features $E \subseteq F$ also belongs to $X^{\leq\alpha}$. Thus, on the basis of the data from the fuzzy pairwise information base $\boldsymbol{B}$, if the similarity of an object w to x is not greater than the similarity of y to x with respect to all the considered features $E \subseteq F$, then w belongs to $X^{\leq\alpha}$. In other words, in each pair $(y,x) \in \underline{E}_\sigma(X^{\leq\alpha})$, x is a reference object and y is a limit object which belongs "certainly" to set X with credibility at most α; the limit is understood such that all objects w that are similar to x w.r.t. considered features at most as much as y is similar to x, also belong to X with credibility at most α.

Analogously, the upper approximation of $X^{\leq\alpha}$ contains all the pairs $(y,x) \in U \times U$ such that there is at least one object w being similar to x at most as much as y is similar to x with respect to all the considered features $E \subseteq F$

which belongs to $X^{\leq\alpha}$. Thus, on the basis of the data from the fuzzy pairwise information base $\boldsymbol{B}$, if the similarity of an object w to x is not greater than the similarity of y to x w.r.t. all the considered features $E \subseteq F$, then it is possible that that w belongs to $X^{\leq\alpha}$. In other words, in each pair $(y,x) \in \underline{E}_\sigma(X^{\leq\alpha})$, x is a reference object and y is a limit object which belongs "possibly" to set X with credibility at most α; the limit is understood such that there is at least one object w that is similar to x w.r.t. considered features at most as much as y is similar to x and has membership in set X with credibility at most α.

Let us remark that we can rewrite the rough approximations $\underline{E}_\sigma(X^{\geq\alpha})$, $\overline{E}_\sigma(X^{\geq\alpha})$, $\underline{E}_\sigma(X^{\leq\alpha})$ and $\overline{E}_\sigma(X^{\leq\alpha})$ as follows:

$$\underline{\ }_\sigma(X^{\geq\alpha})=\{(y,x) \in U \times U\colon \forall w \in U,\ (w,x)D_E(y,x) \Rightarrow w \in X^{\geq\alpha}\},$$

$$\overline{\ }_\sigma(X^{\geq\alpha}) = \{(y,x) \in U \times U\colon \exists w \in U \text{ such that } (w,x)D_E(y,x) \text{ and } w \in X^{\geq\alpha}\},$$

$$\underline{\ }_\sigma(X^{\leq\alpha})=\{(y,x) \in U \times U\colon \forall w \in U,\ (y,x)D_E(w,x) \Rightarrow w \in X^{\leq\alpha}\},$$

$$\overline{\ }_\sigma(X^{\leq\alpha}) = \{(y,x) \in U \times U : \exists w \in U \text{ such that } (y,x)D_E(w,x) \text{ and } w \in X^{\leq\alpha}\}.$$

This formulation of the rough approximation is concordant with the syntax of the decision rules induced by means of DRSA in a pairwise fuzzy information base. More precisely:

- $\underline{\ }_\sigma(X^{\geq\alpha})$ is concordant with decision rules of the type:
  "if object w is similar to object x w.r.t. feature f_{i1} in degree at least h_{i1} and w.r.t. feature f_{i2} in degree at least h_{i2} and ... and w.r.t. feature f_{im} in degree at least h_{im}, then object w belongs to set X in degree at least α",
- $\overline{\ }_\sigma(X^{\geq\alpha})$ is concordant with decision rules of the type:
  "if object w is similar to object x w.r.t. feature f_{i1} in degree at least h_{i1} and w.r.t. feature f_{i2} in degree at least h_{i2} and ... and w.r.t. feature f_{im} in degree at least h_{im}, then object w could belong to set X in degree at least α",
- $\underline{\ }_\sigma(X^{\leq\alpha})$ is concordant with decision rules of the type:
  "if object w is similar to object x w.r.t. feature f_{i1} in degree at most h_{i1} and w.r.t. feature f_{i2} in degree at most h_{i2} and ... and w.r.t. feature f_{im} in degree at most h_{im}, then object w belongs to set X in degree at most α",
- $\overline{\ }_\sigma(X^{\leq\alpha})$ is concordant with decision rules of the type:
  "if object w is similar to object x w.r.t. feature f_{i1} in degree at most h_{i1} and w.r.t. feature f_{i2} in degree at most h_{i2} and ... and w.r.t. feature f_{im} in degree at least h_{im}, then object w could belong to set X in degree at most α",

where $\{i1,\ldots,im\} = E$ and $h_{i1},\ldots,h_{im} \in [0,1]$.

The above definitions of rough approximations and the syntax of decision rules are based on ordinal properties of similarity relations only. In fact, no algebraic operations, such as sum or product, involving cardinal properties of function σ measuring credibility of similarity relations is considered. This is an important characteristic of our approach in comparison with alternative approaches to case-based reasoning.

Let us remark that in the above approximations, even if for two fuzzy sets X and Y we have $X^{\geq\alpha}= Y^{\leq\alpha}$, their approximations are different due to the different directions of cutting the membership function of sets X and Y. Of course, a similar remark holds also for $X^{<\alpha}$ and $Y^{>\alpha}$.

The following theorem states some properties of the rough approximations in a pairwise fuzzy information base.

Theorem. Given a [illegible] $\boldsymbol{B}=< U, F,\sigma >$ and a fuzzy set X in U with membership function $\mu_X(\cdot)$, the following properties hold for any $E \subseteq F$:

1. For any α, $0 \leq \alpha \leq 1$,
   $\underline{E}_\sigma(X^{\leq\alpha}) \subseteq X^{\leq\alpha}\times X^{\leq\alpha} \subseteq \overline{E}_\sigma(X^{\leq\alpha}), \underline{E}_\sigma(X^{\geq\alpha}) \subseteq X^{\geq\alpha}\times X^{\geq\alpha} \subseteq \overline{E}_\sigma(X^{\geq\alpha}),$
   $\underline{E}_\sigma(X^{<\alpha}) \subseteq X^{<\alpha}\times X^{<\alpha} \subseteq \overline{E}_\sigma(X^{<\alpha}), \underline{E}_\sigma(X^{>\alpha}) \subseteq X^{>\alpha}\times X^{>\alpha} \subseteq \overline{E}_\sigma(X^{>\alpha}).$

2. For any α, $0 \leq \alpha \leq 1$,
   $\underline{E}_\sigma(X^{\leq\alpha}) = U \times U - \overline{E}_\sigma(X^{>\alpha}), \underline{E}_\sigma(X^{\geq\alpha}) = U \times U - \overline{E}_\sigma(X^{<\alpha}).$

3. For any α, β, $0\leq\alpha\leq\beta\leq 1$,

$$\begin{array}{ll} \underline{E}_\sigma(X^{\leq\alpha}) \subseteq \underline{E}_\sigma(X^{\leq\beta}), & \underline{E}_\sigma(X^{<\alpha}) \subseteq \underline{E}_\sigma(X^{<\beta}), \\ \underline{E}_\sigma(X^{\geq\alpha}) \supseteq \underline{E}_\sigma(X^{\geq\beta}), & \underline{E}_\sigma(X^{>\alpha}) \supseteq \underline{E}_\sigma(X^{>\beta}), \\ \overline{E}_\sigma(X^{\leq\alpha}) \subseteq \overline{E}_\sigma(X^{\leq\beta}), & \overline{E}_\sigma(X^{<\alpha}) \subseteq \overline{E}_\sigma(X^{<\beta}), \\ \overline{E}_\sigma(X^{\geq\alpha}) \supseteq \overline{E}_\sigma(X^{\geq\beta}), & \overline{E}_\sigma(X^{>\alpha}) \supseteq \overline{E}_\sigma(X^{>\beta}). \end{array}$$

4. For any $x,y,w,z \in U$ and for any α, $0\leq\alpha\leq 1$,

$$[(y,x)D_E(w,x) \text{ and } (w,x) \in \underline{E}_\sigma(X^{\geq\alpha})] \Rightarrow (y,x) \in \underline{E}_\sigma(X^{\geq\alpha}),$$

$$[(y,x)D_E(w,x) \text{ and } (w,x) \in \underline{E}_\sigma(X^{>\alpha})] \Rightarrow (y,x) \in \underline{E}_\sigma(X^{>\alpha}),$$

$$[(y,x)D_E(w,x) \text{ and } (w,x) \in \overline{E}_\sigma(X^{\geq\alpha})] \Rightarrow (y,x) \in \overline{E}_\sigma\ (X^{\geq\alpha}),$$

$$[(y,x)D_E(w,x) \text{ and } (w,x) \in \overline{E}_\sigma\ (X^{>\alpha})] \Rightarrow (y,x) \in \overline{E}_\sigma\ (X^{>\alpha}),$$

$$[(w,x)D_E(y,x) \text{ and } (w,x) \in \underline{E}_\sigma(X^{\leq\alpha})] \Rightarrow (y,x) \in \underline{E}_\sigma(X^{\leq\alpha}),$$

$$[(w,x)D_E(y,x) \text{ and } (w,x) \in \underline{E}_\sigma(X^{<\alpha})] \Rightarrow (y,x) \in \underline{E}_\sigma(X^{<\alpha}),$$

$$[(w,x)D_E(y,x) \text{ and } (w,x) \in \overline{E}_\sigma\ (X^{\leq\alpha})] \Rightarrow (y,x) \in \overline{E}_\sigma\ (X^{\leq\alpha}),$$

$$[(w,x)D_E(y,x) \text{ and } (w,x) \in \overline{E}_\sigma\ (X^{<\alpha})] \Rightarrow (y,x) \in \overline{E}_\sigma\ (X^{<\alpha}).$$

5. For any $E_1 \subseteq E_2 \subseteq F$ and for any α, $0\leq\alpha\leq 1$,

$$\begin{array}{l} \underline{E}_{1\sigma}(X^{\leq\alpha}) \subseteq \underline{E}_{2\sigma}(X^{\leq\alpha}), \underline{E}_{1\sigma}(X^{<\alpha}) \subseteq \underline{E}_{2\sigma}(X^{<\alpha}), \\ \underline{E}_{1\sigma}(X^{\geq\alpha}) \subseteq \underline{E}_{2\sigma}(X^{\geq\alpha}), \underline{E}_{1\sigma}(X^{>\alpha}) \subseteq \underline{E}_{2\sigma}(X^{>\alpha}), \\ \overline{E}_{1\sigma}(X^{\leq\alpha}) \supseteq \overline{E}_{2\sigma}(X^{\leq\alpha}), \overline{E}_{1\sigma}(X^{<\alpha}) \supseteq \overline{E}_{2\sigma}(X^{<\alpha}), \\ \overline{E}_{1\sigma}(X^{\geq\alpha}) \supseteq \overline{E}_{2\sigma}(X^{\geq\alpha}), \overline{E}_{1\sigma}(X^{>\alpha}) \supseteq \overline{E}_{2\sigma}(X^{>\alpha}). \end{array}$$

Proof. 1. For all $x, y \in U$,

$$x \in D_E^+(y,x). \quad (i)$$

Thus, $D_E^+(y,x) \subseteq X^{\geq\alpha}$ implies $x \in X^{\geq\alpha}$ and $y \in X^{\geq\alpha}$. For the definition of $\underline{E}_\sigma(X^{\geq\alpha})$, we have that $(y,x) \in \underline{E}_\sigma(X^{\geq\alpha})$ if $D_E^+(y,x) \subseteq X^{\geq\alpha}$, thus we conclude that, $\forall(y,x) \in U \times U$,

$$(y,x) \in \underline{E}_\sigma(X^{\geq\alpha}) \Rightarrow (y,x) \in X^{\geq\alpha} \times X^{\geq\alpha}$$

i.e.

$$\underline{E}_\sigma(X^{\geq\alpha}) \subseteq X^{\geq\alpha} \times X^{\geq\alpha}.$$

Moreover, from (i) we get that for all $(y,x) \in X^{\geq\alpha} \times X^{\geq\alpha}$, $y \in D_E^-(y,x)$. For the definition of $\overline{E}_\sigma(X^{\geq\alpha})$ we have that $(y,x) \in \overline{E}_\sigma(X^{\geq\alpha})$ if $D_E^-(y,x) \cap X^{\geq\alpha} \neq \emptyset$, thus we conclude that, $\forall(y,x) \in U \times U$,

$$(y,x) \in X^{\geq\alpha} \times X^{\geq\alpha} \Rightarrow (y,x) \in \overline{E}_\sigma(X^{\geq\alpha})$$

i.e.

$$X^{\geq\alpha} \times X^{\geq\alpha} \subseteq \overline{E}_\sigma(X^{\geq\alpha}).$$

Consequently, we proved that

$$\underline{E}_\sigma(X^{\geq\alpha}) \subseteq X^{\geq\alpha} \times X^{\geq\alpha} \subseteq \overline{E}_\sigma(X^{\geq\alpha}).$$

Other cases can be proved analogously.

2. Remembering that $X^{<\alpha} = U - X^{\geq\alpha}$ and observing that

$$D_E^+(y,x) \subseteq X^{\geq\alpha} \Leftrightarrow D_E^+(y,x) \cap (U - X^{\geq\alpha}) = \emptyset \Leftrightarrow D_E^+(y,x) \cap X^{<\alpha} = \emptyset$$

we get

$$\begin{aligned} \underline{E}_\sigma(X^{\geq\alpha}) &= \{(y,x) \in U \times U : D_E^+(y,x) \subseteq X^{\geq\alpha}\} = \\ &= U \times U - \{(y,x) \in U \times U : D_E^+(y,x) \cap X^{<\alpha} \neq \emptyset\} = \\ &= U \times U - \overline{E}_\sigma(X^{<\alpha}). \end{aligned}$$

Analogous proof holds for $\underline{E}_\sigma(X^{\geq\alpha}) = U \times U - \overline{E}_\sigma(X^{<\alpha})$.

3. Let us observe that for any α, β, $0 \leq \alpha \leq \beta \leq 1$

$$X^{\geq\alpha} = \{x \in U : \mu(x) \geq \alpha\} \supseteq \{x \in U : \mu(x) \geq \beta\} = X^{\geq\beta}.$$

Taking this into account, we get

$$\{(y,x) \in U \times U : D_E^+(y,x) \subseteq X^{\geq\alpha}\} \subseteq \{(y,x) \in U \times U : D_E^+(y,x) \subseteq X^{\geq\beta}\}$$

i.e.

$$\_\__\sigma(X^{\geq\alpha}) \subseteq \_\__\sigma(X^{\geq\beta}).$$

Moreover, we also obtain

$$\{(y,x) \in U \times U : D_E^-(y,x) \cap X^{\geq\alpha} \neq \emptyset\} \subseteq \{(y,x) \in U \times U : D_E^+(y,x) \cap X^{\geq\beta} \neq \emptyset\}$$

i.e.

$$\overline{\ }_{\sigma}(X^{\geq\alpha}) \subseteq \overline{\ }_{\sigma}(X^{\geq\beta}).$$

Other cases can be proved analogously.

4. Let us observe that for the transitivity of D_E, for any $x,y,w,z \in U$ and for any $E \subseteq F$

$$\begin{gathered}[(z,x)D_E(y,x) \text{ and } (y,x)D_E(w,x) \Rightarrow (z,x)D_E(w,x)] \\ \Leftrightarrow \\ [(z,x) \in D_E^+(y,x) \text{ and } (y,x)D_E(w,x) \Rightarrow (z,x) \in D_E^+(w,x)] \\ \Leftrightarrow \\ [(y,x)D_E(w,x) \Rightarrow D_E^+(y,x) \subseteq D_E^+(w,x)].\end{gathered}$$

From this we get that if $(y,x)D_E(w,x)$, then

$$D_E^+(w,x) \subseteq X^{\geq\alpha} \Rightarrow D_E^+(y,x) \subseteq X^{\geq\alpha}$$

i.e.

$$[(y,x)D_E(w,x) \text{ and } (w,x) \in \underline{\ }_{\sigma}(X^{\geq\alpha})] \Rightarrow (y,x) \in \underline{\ }_{\sigma}(X^{\geq\alpha}).$$

Other cases can be proved analogously.

5. For any $E_1 \subseteq E_2 \subseteq F$ and for any $x,y,w,z \in U$

$$(x,y)D_{E_2}(w,z) \Rightarrow (x,y)D_{E_1}(w,z)$$

and thus

$$D_{E_1}^+(x,y) \supseteq D_{E_2}^+(x,y) \text{ and } D_{E_1}^-(x,y) \supseteq D_{E_2}^-(x,y).$$

From this we get that for all α, $0 \leq \alpha \leq 1$,

$$D_{E_1}^+(y,x) \subseteq X^{\geq\alpha} \Rightarrow D_{E_2}^+(y,x) \subseteq X^{\geq\alpha}$$

and

$$D_{E_2}^-(y,x) \cap X^{\geq\alpha} \neq \emptyset \Rightarrow D_{E_1}^-(y,x) \subseteq X^{\geq\alpha} \neq \emptyset,$$

which give, respectively,

$$\underline{\ }_{1\sigma}(X^{\geq\alpha}) \subseteq \underline{\ }_{2\sigma}(X^{\geq\alpha}) \text{ and } \overline{\ }_{1\sigma}(X^{\geq\alpha}) \supseteq \overline{\ }_{2\sigma}(X^{\geq\alpha}).$$

Other cases can be proved analogously. ◇

3 Conclusions

We presented a model of case-based reasoning using Dominance-based Rough Set Approach (DRSA). This model is based only on ordinal properties of similarity relations and membership functions of fuzzy sets. Moreover, we did not impose any specific aggregation functional based on specific axioms (see for example [4]), but we considered a set of decision rules based on the very general monotonicity

property of comprehensive similarity with respect to similarity of single features. From this viewpoint our approach to case-based reasoning is as much "neutral" and "objective" as possible and it is very little "invasive" comparing to many other existing approaches. Future research on rough set approach to case-based reasoning can be focused on

- comparison of our approach with other case-based reasoning methodologies and
- the use of our approach for extension of other concepts, results and methodologies of rough set theory.

With respect to comparison of our approach with other case-based reasoning methodologies, an important future research concerns axiomatic considerations. As observed by Gilboa and Schmeidler [4] the interest of axiomatic consideration can be summarized in the following points:

1) meta-scientific reasons: axiomatization provides a link between theoretical terms and observable terms in order to rend the latter meaningful;
2) descriptive reasons: it supplies the basis for testing the empirical validity of the theory because axioms permit to conceive experiments able to falsify the theory rendering the theory falsifiable as requested by Popper [15];
3) normative reasons: a simple set of axioms is often more understandable than the mathematical formulation of the theory and from this viewpoint can be the basis for a discussion with a decision maker about acceptance or rejection of the theory.

With respect to our approach to case-based reasoning, the axiomatic considerations have the further merits of permitting a comparison with the axiomatization of Gilboa and Schmeidler [4] and of pointing out the fact that only monotonic properties of similarity measures are considered.

The research field seems very promising also with respect to rough set theory and we envisage interesting developments with respect to three following issues:

1) generalizations of other rough set fundamental concepts such as reducts and core [13];
2) algebraic properties of the proposed rough approximations (for a general introduction of algebraic properties of classical rough set approach see [14]);
3) application of the absolute and relative rough membership concept (see [8]) in a generalized variable precision model based on the proposed rough approximations in order to admit decision rules with a limited number of counterexamples, which is particularly useful when dealing with large data sets.

Acknowledgements. The research of the first two authors has been supported by the Italian Ministry of Education, University and Scientific Research (MIUR). The third author wishes to acknowledge financial support from the Polish Ministry of Education and Science.

References

1. Dubois, D., Prade, H., Gradual inference rules in approximate reasoning, *Information Sciences*, 61 (1992) 103-122
2. Dubois, D., Prade, H., Putting rough sets and fuzzy sets together, in: R. Slowinski (ed.), *Intelligent Decision Support: Handbook of Applications and Advances of the Sets Theory*, Kluwer, Dordrecht, 1992, pp. 203-232
3. Dubois, D., Prade, H., Esteva, F., Garcia, P., Godo, L., Lopez de Mantara, R., Fuzzy Set Modelling in Case-based Reasoning, *International Journal of Intelligent Systems*, 13 (1998) 345-373
4. Gilboa, I., Schmeidler, D., *A Theory of Case-Based Decisions*, Cambridge University Press, 2001
5. Greco, S., Matarazzo, B., Slowinski R., Rough set theory for multicriteria decision analysis, *European Journal of Operational Research*, 129 (2001) 1-47
6. Greco, S., Matarazzo, B., Slowinski R., Decision rule approach, in: J. Figueira, S. Greco and M. Ehrgott (eds.) *Multiple Criteria Decision Analysis: State of the Art Surveys*, Springer, Berlin, 2005, pp. 507-563
7. S. Greco, B. Matarazzo, R. Slowinski. Generalizing Rough Set Theory through Dominance-based Rough Set Approach. In D. Slezak, J. Yao, J. Peters, W. Ziarko, X. Hu (eds.), *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, LNAI 3642, Springer, Berlin, 2005, pp. 1-11
8. Greco, S., Matarazzo, B., Slowinski R., Rough membership and Bayesian confirmation measures for parametrized rough sets, In D. Slezak, J. Yao, J. Peters, W. Ziarko, X. Hu (eds.), *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, LNAI 3641, Springer, Berlin, 2005, pp. 314-324
9. Hume, D., *An Enquiry concerning Human Understanding*, Oxford, Clarendon Press, 1748
10. Kolodner, J., *Case-Based Reasoning*, Morgan Kaufmann, San Mateo, CA, 1993
11. Leake, D. B., CBR in Context: The Present and Future, in: D. Leake (ed.) *Case-Based Reasoning: Experiences, Lessons, and Future Directions*, AAAI Press/MIT Press, Menlo Park, 1996, pp. 1-30
12. Pawlak, Z., Rough sets, *International Journal of Computer and Information Sciences*, 11 (1982) 341-356
13. Pawlak, Z., *Rough Sets*, Kluwer, Dordrecht, 1991
14. Polkowski, L., *Rough Sets: Mathematical Foundations*, Physica-Verlag, 2002
15. Popper, K. R., *The Logic of Scientific Discovery*, Hutchinson and Co., London, 1958
16. Slowinski, R., Rough set processing of fuzzy information, in: T. Y. Lin, A. Wildberger (eds.), *Soft Computing: Rough Sets, Fuzzy Logic, Neural Networks, Uncertainty Management, Knowledege Discovery*, Simulation Councils, Inc., San Diego, CA, 1995, pp. 142-145
17. Wittgenstein, L., *Tractatus Logico-Philosophicus*, Routledge and Kegan Paul, London, 1922; fifth impression 1951
18. Zadeh, L., Fuzzy Set, *Information and Control*, 8 (1965) 338-353

Towards the Next Generation of Computational Trust and Reputation Models

Jordi Sabater-Mir

Artificial Intelligence Research Institute (IIIA-CSIC),
Campus UAB, Bellaterra, Catalonia, Spain
jsabater@iiia.csic.es

The scientific research in the area of computational trust and reputation mechanisms for virtual societies is a recent discipline oriented to increase the reliability and performance of electronic communities by introducing in such communities these well known human social control mechanisms.

Computer science has moved from the paradigm of an isolated machine to the paradigm of a network of systems and of distributed computing. Likewise, artificial intelligence is quickly moving from the paradigm of an isolated and non-situated intelligence to the paradigm of a situated, social and collective intelligence. This new "social" dimension is the main responsible of the increasing interest on computational trust and reputation mechanisms applied to electronic societies.

Computational trust and reputation systems have been recognized as key factors for successful electronic commerce adoption. These systems are used by intelligent software agents both as a mechanism of search for trustworthy exchange partners and as an incentive in decision-making about whether or not to honour contracts. Reputation is also used in electronic markets as a trust-enforcing, deterrent, and incentive mechanism to avoid cheaters and frauds. Another area of application in agent technology is teamwork and cooperation.

However, in spite of the obvious utility of using a trust and reputation system, they are still not a usual element you can find in an agent architecture like it is the case for example of planners or communication modules. Trust and reputation systems still belong to the set of "not so important" elements of an agent architecture. This is something it has to change if we really want artificial socially intelligent entities.

If we analyze which computational trust and reputation models are used nowadays, we will notice that only the simplest models are really used (see for example e-Bay[1]). Till now, the use of complex trust and reputation models has been somehow questioned and reduced to academic environments. Why an artificial agent has to use one of these complex models if it is evolving in a simple community with 10 or 15 agents that have limited interaction capabilities? In other words, the applications and environments are too simple to justify a complex trust and reputation model that only has sense in a socially complex environment.

However this situation is changing very quickly. The increase of the global connectivity (everything is inter-connected anytime and everywhere) is bringing

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 19–21, 2006.

us a new concept of environment where virtual entities need to interact among them and with humans in a complex way. It is clear then, that in a few years we will start to see the virtual societies we have been theorizing about (and waiting) for a long time in the agents community. In this kind of environment, it will be a must for virtual entities to use complex trust and reputation models if they want to be successful.

Now the question is, are the current trust and reputation models ready to fulfill the requirements of these new virtual societies? We think the answer is no. And this negative answer is because of two reasons: the internals of the model and how the model is being integrated with the rest of the elements of the agent architecture.

Up to now, almost all the efforts have been directed to build trust and reputation models based on a pure numerical approach. There are plenty of these models[2, 3] and the mechanisms they use to calculate the trust and reputation values go from simple aggregation of values[4] to the use of probability theory[5], fuzzy logic[6] or the use of entropy[7] just to put some examples. At the end, each model manipulates the input data in a different way trying to obtain the most accurate trust and reputation values for a given subject. However, if we want to undertake the problems found in socially complex virtual societies, we need theoretically more sophisticated trust and reputation systems. The internal process the trust and reputation system is following to arrive to a final trust and reputation value is as important as the result itself. By ignoring this, we are losing a lot of information that is crucial in order to be successful in the complex world of social relations. We think the solution for that is to use a cognitive approach supported by a solid cognitive theory behind.

The second problem is the integration of the model with the rest of the elements that compound the agent architecture. At this moment, current models are purely reactive black boxes. They receive different inputs (that vary from model to model: direct experiences, witness information, social information...) and, when queried, return the current trust and/or reputation value that has been calculated using the inputs that the model can deal with.

If we really want useful trust and reputation systems, we have to transform them in a proactive element of the agent. The immediate consequence of this proactiveness is that the trust and reputation system could participate in the decision making process by suggesting actions, strategies or complete plans that could help to improve the reliability of the trust and reputation values. But perhaps more important, this new approach opens the possibility of something that till now has been completely ignored, the possibility for the agent to manage its own credibility and reputation in front of the rest of the community. This can only be achieved if the trust and reputation system is at the same level that the other elements that compound the agent architecture and can influence the decision making process by proposing actions and plans to influence the perception others have towards it.

Our current work is going in these two directions: improve the internals of the trust and reputation systems by using a cognitive approach and at the same

time analyze how the trust and reputation system can be integrated in the agent architecture to provide the functionality required by a socially intelligent virtual entity. We think these are the two keystones for the next generation of computational trust and reputation models. Our first step on this direction is RepAge[8], a computational reputation model based on a cognitive theory of reputation[9] and a previous trust and reputation model called ReGreT[10].

References

1. eBay: eBay, http://www.eBay.com. (2002)
2. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial Intelligence Review **24** (2005) 33–60
3. Ramchurn, S., Hunyh, D., Jennings, N.: Trust in multi-agent systems. The Knowledge Engineering Review **1** (2004) 1—25
4. Zacharia, G.: Collaborative reputation mechanisms for online communities. Master's thesis, Massachusetts Institute of Technology (1999)
5. Schillo, M., Funk, P., Rovatsos, M.: Using trust for detecting deceitful agents in artificial societites. Applied Artificial Intelligence (2000) 825—848
6. Carbo, J., Molina, J., Davila, J.: Trust management through fuzzy reputation. Int. Journal in Cooperative Information Systems **12** (2002) 135—155
7. Sierra, C., Debenham, J.: An information-based model for trust. In: Proceedings of the fourth international joint conference on autonomous agents and multiagent systems (AAMAS-05), Utrecht, Netherlands. (2005) 497—504
8. Paolucci, M., Sabater-Mir, J., Conte, R.: What if? dealing with uncertainty in repage's mental landscape. In: Proceedings International Conference on Intelligent Agent Technology (IAT-05), Compigne, France. (2005) 372—378
9. Conte, R., Paolucci, M.: Reputation in artificial societies: Social beliefs for social order. Kluwer Academic Publishers (2002)
10. Sabater, J.: Trust and reputation for agent societies. PhD thesis, Universitat Autonoma de Barcelona (UAB) (2003)

Preference Modeling by Rectangular Bilattices

Ofer Arieli[1], Chris Cornelis[2], and Glad Deschrijver[2]

[1] Department of Computer Science,
The Academic College of Tel-Aviv, Israel
oarieli@mta.ac.il
[2] Fuzziness and Uncertainty Modeling Research Unit,
Department of Mathematics and Computer Science,
Ghent University, Belgium
{chris.cornelis, glad.deschrijver}@UGent.be

Abstract. Many realistic decision aid problems are fraught with facets of ambiguity, uncertainty and conflict, which hamper the effectiveness of conventional and fuzzy preference modeling approaches, and command the use of more expressive representations. In the past, some authors have already identified Ginsberg's/Fitting's theory of bilattices as a naturally attractive candidate framework for representing uncertain and potentially conflicting preferences, yet none of the existing approaches addresses the real expressive power of bilattices, which lies hidden in their associated truth and knowledge orders. As a consequence, these approaches have to incorporate additional conventions and 'tricks' into their modus operandi, making the results unintuitive and/or tedious. By contrast, the aim of this paper is to demonstrate the potential of (rectangular) bilattices in encoding not just the problem statement, but also its generic solution strategy.

1 Introduction

The notion of preference is common in various contexts involving decision or choice. Preference modeling provides declarative means for choosing among alternatives, including different solutions to problems, answers to database queries, decisions of a computational agent, etc. This topic is gaining increasing attention in diverse areas of artificial intelligence such as nonmonotonic reasoning, qualitative decision theory, configuration, and AI planning. More recently, preference modeling has also been used in constraint satisfaction and constraint programming, for treating soft constraints, for describing search heuristics, and for reducing search effort (see, e.g. [9] and [13] for recent collections of papers on these topics).

Conventional preference modeling (see e.g. [25]) is centered on the notion of classical preference structures $\langle P, I, R\rangle$, consisting of three fundamental binary relations (strict preference P, indifference I, and incomparability R) that may hold among the alternatives; usually the evidence in favour of these relations is captured by a so-called relation S that describes, for each couple (u, v) of alternatives, whether u is (known to be) at least as good as v. In

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 22–33, 2006.

practice, it is common to encounter situations where these relationships hold up to a certain [illegible], which gives rise to the study of [illegible] (see e.g. [20, 31, 32]).

Fuzziness, however, cannot adequately cover all the imperfections inherent to real-life data, since the 'one-dimensional' measurements induced by the ordering of membership degrees in fuzzy sets have difficulties coping with information-deficient data. As Tsoukiàs and Vincke noted in [29], fuzzy sets and logic per se do not provide "[illegible]". Indeed, stating that $P(u, v) = 0$ may either mean that u (definitely) is not preferred to v, or simply that there is no information to establish a preference of u over v, and there is no unambiguous way for a decision maker to distinguish between the two situations. For this reason, several researchers have considered more elaborate means of eliciting and representing preferences. In particular, Belnap's logic $\mathcal{FOUR}$ [7, 8], and some of its extensions, built around the truth values 'true', 'false', 'unknown' and 'contradiction', had immediate and intuitive appeal, and were taken as the basis for the approaches in [21, 24, 27, 28, 29]. However, we found that many of these approaches lack a proper way of [illegible] the preferences, and as a consequence no solid analysis tools nor clear strategies for decision making under incomplete and/or conflicting information are available to the reasoner in such cases.

The goal of this paper is to overcome this shortcoming. For this, we consider a certain family of algebraic structures, called [illegible] [18, 22] that encapsulate and refine Belnap's $\mathcal{FOUR}$, and that serve here as a representation platform. We demonstrate the real expressive power of these structures, and in particular of the 'two-dimensional' measurements induced by their dual orderings, in describing and modeling imprecise preferences. As such, the material presented in this paper is not a 'new' approach to preference modeling, but rather a clarification, simplification and streamlining of existing ones.

The remainder of this paper is organized as follows: in Section 2, we recall important preliminary notions about bilattices and their role in uncertainty modeling. Section 3 contains our novel analysis of preference modeling by rectangular bilattices; it exhibits the drawbacks of existing approaches, and describes how they can be mended. In Section 4 we conclude.

2 Preliminaries

2.1 Bilattices

Definition 1. A [illegible] [22] is a triple $\mathcal{B} = (B, \leq_t, \leq_k)$, where B is a nonempty set containing at least two elements, and $(B, \leq_t)$, $(B, \leq_k)$ are complete lattices.[1]

[1] Structures that meet this definition are sometimes called *pre-bilattices*. In such cases the notion 'bilattices' is reserved for some particular type of pre-bilattices which is determined according to the way the two partial orders are related; see Definition 2.

The two partial orders $\leq_t$ and $\leq_k$ of bilattices intuitively represent differences in the degree of truth and in the amount of knowledge/information (respectively), conveyed by the assertions. In the sequel, following the usual notations for the basic bilattice operations, we shall denote by $\wedge$ (respectively, by $\vee$) the $\leq_t$-meet (the $\leq_t$-join) and by $\otimes$ (respectively, by $\oplus$) the $\leq_k$-meet (the $\leq_k$-join) of $\mathcal{B}$. While the meaning of $\wedge$ and $\vee$ corresponds to the standard logical role of these operators, the intuition behind $\otimes$ and $\oplus$ is somewhat less transparent. Fitting [19] calls them and operations, respectively, to indicate that $x \otimes y$ is the most information 'agreed' upon by x and y, while $x \oplus y$ includes everything accepted by at least one of x and y.

We denote by f and t the $\leq_t$-extreme elements, and $\bot$, $\top$ denote the $\leq_k$-extreme elements of $\mathcal{B}$. Intuitively, these elements can be perceived as 'false', 'true', 'unknown' (i.e., neither true nor false) and 'contradictory' (both true and false), respectively. Thus, for instance, $f \leq_t \bot$ since the 'degree of truth' of a statement which is known to be false is smaller than that of a statement about which there is no information whatsoever. On the other hand, $\bot \leq_k f$, since knowing that a statement is false is more informative than knowing nothing at all about it.

Clearly, the more interesting forms of bilattices are those in which the two partial orders are related in one way or another. Below are some common types of such relations:

Definition 2. Let $\mathcal{B} = (B, \leq_t, \leq_k)$ be a bilattice.

- $\mathcal{B}$ is called [22] if all the (twelve) possible distributive laws concerning $\wedge$, $\vee$, $\otimes$, and $\oplus$ hold (for instance, $a \wedge (b \oplus c) = (a \wedge b) \oplus (a \wedge c)$).
- $\mathcal{B}$ is called [17] if each one of $\wedge$, $\vee$, $\otimes$, and $\oplus$, is monotonic with respect to both $\leq_t$ and $\leq_k$ (for instance, if $a \leq_k b$ then $a \wedge c \leq_k b \wedge c$).
- $\mathcal{B}$ is a bilattice [22] if there exists a unary operation $\neg$ satisfying, for every x, y in B, (1) $\neg\neg x = x$, (2) if $x \leq_t y$ then $\neg x \geq_t \neg y$, and (3) if $x \leq_k y$ then $\neg x \leq_k \neg y$.

Originally, Ginsberg considered bilattices with negations. In this case a negation is an involution with respect to the lattice $(B, \leq_t)$ and an order preserving operation of the lattice $(B, \leq_k)$. In such cases it is easy to see that $\neg f = t$, $\neg t = f$, $\neg\bot = \bot$, and $\neg\top = \top$. Following Ginsberg, Fitting introduced the family of interlaced bilattices and showed their usefulness in the context of logic programming (see e.g. [17, 18, 19]). It is easy to verify that distributive bilattices are also interlaced. In the context of fuzzy sets, interlaced bilattices have been considered, e.g., in [11].

Figure 1 in Section 2.3 depicts double-Hasse diagrams of a four-valued bilattice and a nine-valued bilattice. It is easy to verify that both these bilattices are distributive, interlaced, and each one has a negation operator obtained by switching the components of the truth values, that is: $\neg(x, y) = (y, x)$.

2.2 Rectangular Bilattices

Definition 3. Let $\mathcal{L} = (L, \leq_L)$ and $\mathcal{R} = (R, \leq_R)$ be two complete lattices. A [illegible], shortly [illegible], is a structure $\mathcal{L} \odot \mathcal{R} = (L \times R, \leq_t, \leq_k)$, where, for every $x_1, y_1 \in L$ and $x_2, y_2 \in R$,

(1) $(x_1, x_2) \leq_t (y_1, y_2) \Leftrightarrow x_1 \leq_L y_1$ and $x_2 \geq_R y_2$,
(2) $(x_1, x_2) \leq_k (y_1, y_2) \Leftrightarrow x_1 \leq_L y_1$ and $x_2 \leq_R y_2$.

We say that a structure is rectangular if it is isomorphic to a rectangular bilattice. An element (x_1, x_2) of a rectangle $\mathcal{L} \odot \mathcal{R}$ may intuitively be understood such that x_1 represents the amount of belief [illegible] some assertion, and x_2 is the amount of belief [illegible] it. In the context of fuzzy sets, this corresponds to Atanassov's theory of intuitionistic fuzzy sets [5], which extends standard fuzzy set theory so that any element u in a universe U is assigned not only a membership degree, $\mu_A(u)$, but also a non-membership degree $\nu_A(u)$, where both degrees are drawn from the unit interval $[0, 1]$ and satisfy the condition $\mu_A(u) + \nu_A(u) \leq 1$. Rectangular bilattices generalize this idea by not imposing the latter condition, by considering [illegible] lattices (not only the unit interval), and by defining the membership function and the non-membership function over potentially [illegible].

Denote the join and meet operations of a complete lattice $\mathcal{L} = (L, \leq_L)$ by $\wedge_L$ and $\vee_L$, respectively. Then, for every x_1, y_1 in L and x_2, y_2 in R, we have

$$(x_1, x_2) \wedge (y_1, y_2) = (x_1 \wedge_L y_1, x_2 \vee_R y_2),$$
$$(x_1, x_2) \vee (y_1, y_2) = (x_1 \vee_L y_1, x_2 \wedge_R y_2),$$
$$(x_1, x_2) \otimes (y_1, y_2) = (x_1 \wedge_L y_1, x_2 \wedge_R y_2),$$
$$(x_1, x_2) \oplus (y_1, y_2) = (x_1 \vee_L y_1, x_2 \vee_R y_2),$$

Moreover, denoting $0_{\mathcal{L}} = \inf L$ and $1_{\mathcal{L}} = \sup L$, it holds that

$$\bot_{\mathcal{L} \odot \mathcal{R}} = (0_{\mathcal{L}}, 0_{\mathcal{R}}), \quad \top_{\mathcal{L} \odot \mathcal{R}} = (1_{\mathcal{L}}, 1_{\mathcal{R}}), \quad t_{\mathcal{L} \odot \mathcal{R}} = (1_{\mathcal{L}}, 0_{\mathcal{R}}), \quad f_{\mathcal{L} \odot \mathcal{R}} = (0_{\mathcal{L}}, 1_{\mathcal{R}}).$$

It is easy to verify that a rectangular bilattice is indeed a bilattice (in the sense of Definition 1). The next proposition summarizes some basic properties of rectangular bilattices and shows their central role in the theory of bilattices:

Proposition 1.

a) [17] [illegible]
b) [6] [illegible]
c) [22] [illegible] $\mathcal{L}$ [illegible] $\mathcal{R}$ [illegible] $\mathcal{L} \odot \mathcal{R}$ [illegible]
d) [17, 22] [illegible] $\mathcal{L} \odot \mathcal{R}$ [illegible] $\mathcal{L}$ [illegible] $\mathcal{R}$

In the context of item (b) of the proposition above, it is interesting to note that every interlaced bilattice $\mathcal{B} = (B, \leq_t, \leq_k)$ is isomorphic to $\mathcal{L} \odot \mathcal{R}$, where $\mathcal{L} = (\{x \mid x \geq_t \bot\}, \leq_k)$ and $\mathcal{R} = (\{x \mid x \leq_t \bot\}, \leq_k)$. These lattices are unique up to an isomorphism (see [6]). The same lattices may be used for item (d) of the proposition, together with the observation that if $\mathcal{B}$ is a distributive bilattice, then $\mathcal{L}$ and $\mathcal{R}$ are necessarily distributive lattices.

2.3 Squares

An important family of rectangular bilattices are those in which $\mathcal{L}$ and $\mathcal{R}$ coincide. These bilattices are called [3, 4, 12, 15] and $\mathcal{L} \odot \mathcal{L}$ is abbreviated by $\mathcal{L}^2$. The squares that are obtained by the two-valued and the three-valued chains are shown in Figure 1. In the literature, these structures are commonly referred to as $\mathcal{FOUR}$ (after Belnap's [7, 8] original four-valued logic) and $\mathcal{NINE}$ (see e.g. [1, 2]), respectively. An example of a square with an infinite amount of elements is $([0,1], \leq)^2$. In the context of fuzzy set theory, the $\leq_t$–ordering of this square is studied in [12, 15] and its $\leq_k$–ordering is considered in [14].

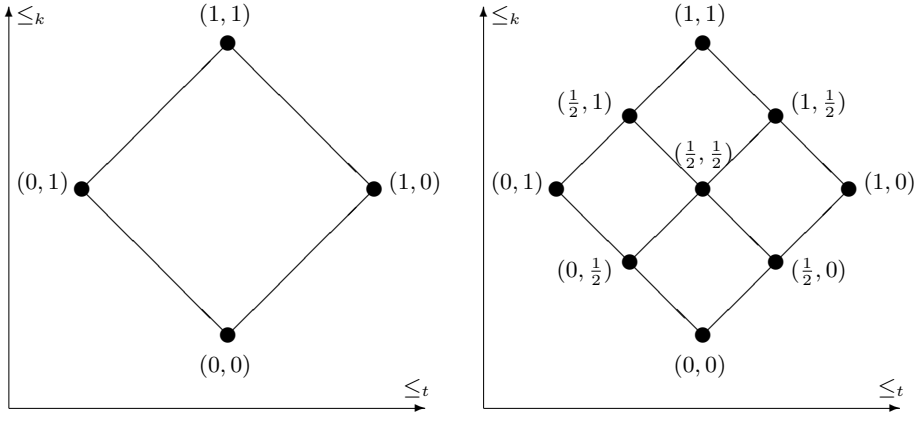


Fig. 1. The squares $\{0,1\}^2$ and $\{0, \frac{1}{2}, 1\}^2$

Again, it is easy to verify that every square $\mathcal{L}^2$ is interlaced, and that it is distributive when $\mathcal{L}$ is distributive. The following proposition shows that the converse is also true.

Proposition 2. [6] $\neg(x, y) = (y, x)$ [2]

A detailed investigation of squares and the graded versions of the logical connectives that can be defined on them appears in [4, 10]. As shown in [3, 4], the evaluation structure of intuitionistic fuzzy sets is equal to the substructure of the of the square $([0,1], \leq)^2$, i.e., the elements (x, y) that satisfy the condition $x + y \leq 1$. In [3, 4] it is also shown that squares are a generalization to complete lattices (not only the unit interval) of interval-valued fuzzy sets [16, 23, 26, 30], an alternative method of extending fuzzy set theory, motivated by the need to replace crisp, $[0,1]$-valued membership degrees by intervals in $[0,1]$ that approximate the (unknown) membership degrees.

[2] Note that by the fact that every distributive lattices is also interlaced, this proposition holds in particular for distributive bilattices with a negation.

3 Modeling Imprecise Preference Information

In a number of recent papers (e.g. [21, 24, 27, 28, 29]), the use of a four-valued logic called DDT (derived from Belnap's original proposal) and some of its graded extensions has been advocated as a means of dealing with the task of preference modeling under incomplete and/or conflicting information. In all of the mentioned papers, bilattice theory per se plays only a subservient role as the convenient 'language' for modeling positive and negative preference arguments separately, and for representing the associated epistemic states of truth, falsity, ignorance and contradiction. By contrast, the aim of this section is to demonstrate and exploit the full expressive power of rectangular bilattices, and of squares in particular, for preference modeling.

3.1 Encoding the Evidence

The problem at hand is that of ranking a (finite) set U of alternatives from the best to the worst, with respect to a number of given criteria. In order to do this, we assume that partial information is available regarding the pairwise comparison of alternatives. In binary preference modeling, it is common to express such information by means of a two-valued [illegible] S in U (see e.g. [25]), where $S(u, v) = 1$ is read as "(there is evidence that) u is at least as good as v". Such an approach can be criticized for lack of expressivity, since explicit evidence that u is [illegible] at least as good as v could only be captured by imposing $S(v, u) = 1$.[3] Yet, as Fortemps and Słowiński argue in [21], arguments in disfavour of a sentence are not necessarily identical to arguments in favour of the opposite sentence!

For this reason, in [29] Tsoukiàs and Vincke propose to distinguish between positive and negative arguments regarding the claim 'u is at least as good as v' ($u \geq v$, for short). Essentially, this amounts to defining the outranking relation S as a mapping from U^2 to $\{0, 1\}^2$, where the value of the first (respectively, the second) component of $S(u, v)$ reveals the [illegible] of arguments in favour (respectively, in disfavour) of $u \geq v$. Clearly, this intuition fits our framework, and Belnap's square $\mathcal{FOUR}$ can be used to endow $\{0, 1\}^2$ with an attractive epistemic structure in terms of truth-hood (the $\leq_t$-ordering: from only evidence against, to only evidence for the claim) and of available information (the $\leq_k$-ordering: from ignorance to conflict).

Definition 4. For ease of notation, in what follows we shall abbreviate T for $(1, 0)$, F for $(0, 1)$, U for $(0, 0)$, and K for $(1, 1)$, to be read as [illegible] and [illegible], respectively.

Of course, nothing stands in the way of generalizing this framework by allowing for graded evidence. For instance, in [21] and [24] the square induced by the unit interval $\mathcal{L} = ([0, 1], \leq)$ was investigated. In general, S can be a mapping

[3] Note that $S(u, v) = 0$ means that there is *no evidence* that u is at least as good as v, which is obviously different than claiming that u is not at least as good as v.

from U^2 to some rectangular bilattice $\mathcal{L}\odot\mathcal{R}$, reflecting that positive and negative arguments may be evaluated according to two different scales.

3.2 Representing the Preferences

Once the various outranking arguments have been provided, the objective then is to present the decision maker with as close to reality and transparent as possible a rendering of the actual state of affairs. In conventional preference modeling (i.e., when $S(u,v) \in \{0,1\}$), a 'decision' concerning two alternatives u and v can take four forms:

1. u is (strictly) [illegible] over v if $S(u,v)=1$ and $S(v,u)=0$,
2. v is (strictly) [illegible] over u if $S(u,v)=0$ and $S(v,u)=1$,
3. u and v are [illegible] if $S(u,v)=1$ and $S(v,u)=1$,
4. u and v are [illegible] if $S(u,v)=0$ and $S(v,u)=0$.

Evidently, all possible situations are covered in this way. Accordingly, one can build three binary relations P ([illegible], corresponding to case 1 and 2), I ([illegible], corresponding to case 3) and R ([illegible], corresponding to case 4), such that $U^2 = P \cup P^{-1} \cup I \cup R$. It is also said that $\langle P, I, R\rangle$ is a [illegible]; it is easy to see that it determines S unequivocally, and vice versa; weakened versions emerge when S becomes a fuzzy relation, a theme explored in e.g. [20, 31, 32]. In what follows, we study the bilattice-valued generalizations of this framework.

A crisp four-valued approach. Let first S be a mapping from U^2 to $\{0,1\}^2$. Each couple of alternatives (u,v) corresponds to a couple $(S(u,v), S(v,u))$ in $(\{0,1\}^2)^2$. For notational ease, and in order to enhance the clarity of the exposition, we shall abbreviate these couples by simply juxtaposing the two letters corresponding to their evaluations. For instance, FK represents the element $((0,1),(1,1))$ that exhibits a situation in which there are only negative arguments for $u \geq v$ and conflicting (both positive and negative) arguments for $v \geq u$.

[illegible] In [27, 28, 29], essentially the same representation, albeit in a more complicated form, is obtained by defining, for every u, v in U,

$$\Delta S(u,v) = 1 \;\Leftrightarrow\; S(u,v) = (1,x) \text{ for some } x \text{ in } \{0,1\}$$

read as, "there is presence of truth in saying that u is at least as good as v", and consequently introducing the so-called [illegible] and [illegible][4] of the formula $S(u,v)$ by, respectively,

$$\mathbf{T}S(u,v) = 1 \Leftrightarrow \Delta S(u,v) = 1 \text{ and } \Delta S(v,u) = 0 \quad (1)$$

$$\mathbf{F}S(u,v) = 1 \Leftrightarrow \Delta S(u,v) = 0 \text{ and } \Delta S(v,u) = 1 \quad (2)$$

$$\mathbf{U}S(u,v) = 1 \Leftrightarrow \Delta S(u,v) = 0 \text{ and } \Delta S(v,u) = 0 \quad (3)$$

$$\mathbf{K}S(u,v) = 1 \Leftrightarrow \Delta S(u,v) = 1 \text{ and } \Delta S(v,u) = 1 \quad (4)$$

In our notations FK denotes the case where $\mathbf{F}S(u,v) = 1$ and $\mathbf{K}S(v,u) = 1$.

[4] These are actually two-valued predicates; in [21] **T**, **F**, **U** and **K** are called strong unary operators.

Thus, a decision maker is confronted with any of sixteen (instead of four) possible situations involving the alternatives u and v. As the prime determination is to try to rank the alternatives, it is worthwhile to endow those various situations with some meaningful structure, and it turns out that bilattices can go a long way in doing just that.

Indeed, starting from the $\leq_t$-ordering on $\mathcal{FOUR}$, we can construct a bilattice-based square on top of $(\{0,1\}^2)^2$ with the following two orderings:

- $(x_1, x_2) \leq_t (y_1, y_2) \Leftrightarrow x_1 \leq_t y_1$ and $x_2 \geq_t y_2$.
  Intuitively, if $(x_1, x_2) = (S(u, v), S(v, u))$ and $(y_1, y_2) = (S(u', v'), S(v', u'))$, then $(x_1, x_2) \leq_t (y_1, y_2)$ expresses that the extent to which u is preferred over v is less than the extent to which u' is preferred over v'. The smallest element is FT (it is not true that $u \geq v$, while it is true that $u \leq v$) and the biggest one is TF ($u \geq v$ and not $v \geq u$).
- $(x_1, x_2) \leq_k (y_1, y_2) \Leftrightarrow x_1 \leq_t y_1$ and $x_2 \leq_t y_2$.
  This ordering ranges between a state of incomparability (FF) and one of indifference (TT).

Starting from the $\leq_k$-ordering on $\mathcal{FOUR}$ we can define two other orderings on $(\{0,1\}^2)^2$ as follows:

- $(x_1, x_2) \leq'_t (y_1, y_2) \Leftrightarrow x_1 \leq_k y_1$ and $x_2 \geq_k y_2$.
  Intuitively, if $(x_1, x_2) = (S(u, v), S(v, u))$ and $(y_1, y_2) = (S(u', v'), S(v', u'))$, then $x_1 \leq_k y_1$ means that we know less about $u \geq v$ than about $u' \geq v'$, and $x_2 \geq_k y_2$ means that we know more about $u \leq v$ than about $u' \leq v'$. So, the bigger (x_1, x_2) according to this ordering, the more we know about $u \geq v$ and the less we know about $u \leq v$.
- $(x_1, x_2) \leq'_k (y_1, y_2) \Leftrightarrow x_1 \leq_k y_1$ and $x_2 \leq_k y_2$.
  This ordering marks the amount of information at our disposition: from a shortage of information (UU) to an excess (KK).

In [28, 29], the authors present a dictionary-style solution to discriminate among the sixteen states, giving concrete names and explanations to each one of them. For instance, TF is called 'strict preference of u over v', KF in their terms is 'weak preference of u over v', etc. This approach, apart from being tedious, is also misleading. As an example, in their approach (as in ours) FF means that u and v are incomparable, whereas UU is read as "u and v are semi incomparable", and FU as "u and v are weakly incomparable". Such terminology implies an inaccurate description of the state of affairs, since

a) the element UU bears no mark of incomparability whatsoever, and
b) referring to $\leq_k$, the elements UF, FK and KF could claim the status of representing 'weak incomparability' with just as much justification as FU.

By contrast, the four order relations considered above serve to discriminate much more naturally, and without bias, among the sixteen states, positioning each state along four scales of measurement.

Extensions to arbitrary (possibly continuous) rectangular bilattices. Another important advantage of our approach is that it can be straightforwardly generalized to graded evidence [illegible]. Indeed, the four orderings $\leq_t$, $\leq_k$, $\leq'_t$, and $\leq'_k$ can equally be defined on $\mathcal{L} \odot \mathcal{R}$ for any complete lattices $\mathcal{L} = (L, \leq_L)$ and $\mathcal{R} = (R, \leq_R)$. The orderings present the decision maker with a rather complete picture of the situation; depending on the underlying goals and attitudes, he or she may exploit the information in various ways.

Consider, for instance, the bilattice $([0,1]^2)^2$ together with, e.g., the normalized Euclidean distance function. For any value $(S(u,v), S(v,u))$ one can measure its distance to the external elements of each order. Such distances give graded information which is often more helpful for the decision maker than just the orderings themselves. For example, when $(S(u,v), S(v,u)) = ((0.1, 0.77), (0.25, 0.41))$, the distance to FT is 0.44 and the distance to TF is 0.67, which indicates a preference of v over u. Likewise, the distances 0.62 and 0.52 to UK and KU respectively may indicate that the amount of available information is greater for "$v \geq u$" than for "$u \geq v$".

Note also that, as shown in Figure 2 (see the diagram on the bottom-left side), the distance to FT (respectively, to TF) of each one of KT, UT, FU, FK, is $1/2$ (respectively, $\sqrt{3}/2$), while the distance to FT (to TF) of TK, TU, UF, KF, is $\sqrt{3}/2$ (respectively, $1/2$). This can be interpreted as follows: the elements on the middle layer do not give any evidence that $u \geq v$ or $u \leq v$, the elements on the second layer from below give more evidence that $u \leq v$, and the elements on the fourth layer provide more evidence that $u \geq v$. As Figure 2 shows, similar layered structures and distance values are also induced by the other orders (see the bottom-right side of this figure for $\leq_k$, the top-left side for $\leq'_t$, and the top-right side for $\leq'_k$).

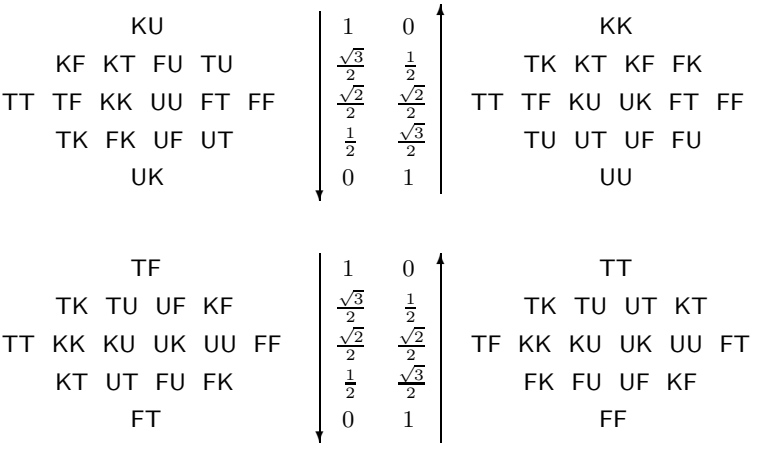


Fig. 2. Euclidean distances to the extreme elements of $\leq_t$ (bottom left), $\leq_k$ (bottom right), $\leq'_t$ (top left), and $\leq'_k$(top right)

Figure 2 reveals a nice symmetry among the four diagrams: there are eight external elements each corresponding to a 'definite' state of affairs (TF and FT: strict preference; TT: indifference; FF: incomparability; KK, UU, UK, KU: information defect) and the eight remaining ones which float somewhat between the extremes (they are always in second or the fourth layer). Note also that the middle layer of each diagram always contains the six other external elements.

As the next proposition shows, the four order relations considered above preserve these distance considerations for every element of the underlying bilattice:

Proposition 3. $\preceq$ $\leq_t$ $\leq_k$ $\leq'_t$ $\leq'_k$ $([0,1]^2)^2$ d 0 1 $\preceq$ u, v $([0,1]^2)^2$ $u \preceq v$ $d(0,u) \leq d(0,v)$ $d(u,1) \geq d(v,1)$

We shall show the claim for $\leq_t$ and its minimal element FT; the other cases are similar.

Let $u = (x_1, x_2)$ and $v = (y_1, y_2)$. If $u \leq_t v$ then $x_1 \leq_t y_1$ and $x_2 \geq_t y_2$, which means that $d(\mathsf{F}, x_1) \leq d(\mathsf{F}, y_1)$ and $d(\mathsf{T}, x_2) \leq d(\mathsf{T}, y_2)$. Thus, $d(\mathsf{FT}, (x_1, x_2)) = \frac{1}{2}\sqrt{d(\mathsf{F}, x_1)^2 + d(\mathsf{T}, x_2)^2} \leq \frac{1}{2}\sqrt{d(\mathsf{F}, y_1)^2 + d(\mathsf{T}, y_2)^2} = d(\mathsf{FT}, (y_1, y_2))$. □

The above representation stands in sharp contrast to existing work relying on the conventions described in Note 1. Indeed, devising graded versions of the predicates **T**, **F**, **U** and **K** requires an explicit choice of how to model the conjunction in the right-hand sides of their defining equalities (1)–(4). In [21] and [24], two different choices involving different t-norms on the unit interval are put forward, each elaborately justified in its own terms. As our exposition reveals, however, this effort is altogether superfluous since it can be avoided by working with the original outranking information. As we have shown, rectangular bilattices offer a simple and natural way of encoding this information, even in cases that the argument in favour of a certain preference and the argument in disfavour of that preference are specified in terms of different ranges.

4 Conclusion

In this paper we introduced a simple and generic solution strategy for modeling imprecise preference information, taking advantage of the new opportunities offered by bilattice-based structures. The 'traditional' approach of evaluating membership functions by values that are arranged in one (and usually total) order, is replaced here by more expressive 'two-dimensional' measurements that reflect different interpretations of the underlying orderings, which may be applied simultaneously. Our approach exploits the order-theoretical ingredients of bilattice theory, and puts existing approaches of preference modeling into a simple and unified perspective. This work therefore demonstrates the applicative aspects of our study on bilattice-based fuzzy sets [3, 4, 10] and vindicates our claim that these structures provide a natural and attractive framework for the representation of uncertain and potentially conflicting information.

References

1. O. Arieli and A. Avron. Reasoning with logical bilattices. *Journal of Logic, Language, and Information*, 5(1):25–63, 1996.
2. O. Arieli and A. Avron. Bilattices and paraconsistency. In D. Batens, C. Mortensen, G. Priest, and J. Van Bendegem, editors, *Frontiers of Paraconsistent Logic*, pages 11–27. Research Studies Press, 2000.
3. O. Arieli, C. Cornelis, G. Deschrijver, and E. E. Kerre. Relating intuitionistic fuzzy sets and interval-valued fuzzy sets through bilattices. In D. Ruan, P. D'Hondt, M. De Cock, M. Nachtegael, and E. E. Kerre, editors, *Applied Computational Intelligence*, pages 57–64. World Scientific, 2004.
4. O. Arieli, C. Cornelis, G. Deschrijver, and E. E. Kerre. Bilattice-based squares and triangles. In L. Godo, editor, *Proc. 8th European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'05)*, volume 3571 of *Lecture Notes on Artificial Intelligence*, pages 563–575. Springer, 2005.
5. K. T. Atanassov. Intuitionistic fuzzy sets, 1983. VII ITKR's Session, Sofia (deposed in Central Sci.-Technical Library of Bulg. Acad. of Sci., 1697/84) (in Bulgarian).
6. A. Avron. The structure of interlaced bilattices. *Journal of Mathematical Structures in Computer Science*, 6:287–299, 1996.
7. N. D. Belnap. How a computer should think. In G. Ryle, editor, *Contemporary Aspects of Philosophy*, pages 30–56. Oriel Press, 1977.
8. N. D. Belnap. A useful four-valued logic. In G. Epstein and J. M. Dunn, editors, *Modern Uses of Multiple-Valued Logic*, pages 7–37. Reidel Publishing Company, 1977.
9. R. Brafman and U. Junker, editors. *Multidisciplinary IJCAI-05 Workshop on Advances in Preference Handling*. 2005. The proceedings is available online at http://wikix.ilog.fr/wiki/bin/view/Preference05/WebHome.
10. C. Cornelis, O. Arieli, G. Deschrijver, and E. E. Kerre. Uncertainty modeling by bilattice-based squares and triangles. *IEEE Transactions on Fuzzy Sets and Systems*, 2006. Accepted.
11. C. Cornelis, K. T. Atanassov, and E. E. Kerre. Intuitionistic fuzzy sets and interval-valued fuzzy sets: a critical comparison. In M. Wagenknecht and R. Hampel, editors, *Proc. 3rd Int. Conf. in Fuzzy Logic and Technology (EUSFLAT'03)*, pages 159–163, 2003.
12. C. Cornelis, G. Deschrijver, and E. E. Kerre. Square and triangle: reflections on two prominent mathematical structures for the representation of imprecision. *Notes on Intuitionistic Fuzzy Sets*, 9(3):11–21, 2003.
13. B. De Baets, M. Delgado, J. Fodor, F. Herrera, E. Herrera-Viedma, and L. Martinez, editors. *International Journal of Intelligent Systems*, volume 18(7). 2003. Special issue on preference modeling and applications.
14. B. De Baets and R. Mesiar. Triangular norms on product lattices. *Fuzzy Sets and Systems*, 104(1):61–75, 1999.
15. G. Deschrijver, C. Cornelis, and E. E. Kerre. Square and triangle: a comparison. In *Proc. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'04)*, pages 1389–1396, 2004.
16. F. Esteva, P. Garcia, and L. Godo. Enriched interval bilattices and partial many-valued logics: an approach to deal with graded truth and imprecision. *Uncertainty, Fuzziness and Knowledge-based Systems*, 2(1):37–54, 1994.
17. M. Fitting. Bilattices in logic programming. In G. Epstein, editor, *Proc. 20th International Symposium on Multiple-Valued Logic*, pages 238–246, 1990.

18. M. Fitting. Kleene's logic, generalized. *Journal of Logic and Computation*, 1:797–810, 1992.
19. M. Fitting. Kleene's three-valued logics and their children. *Fundamenta Informaticae*, 20:113–131, 1994.
20. J. Fodor and M. Roubens. *Fuzzy Preference Modelling and Multicriteria Decision Support*. Kluwer, 1994.
21. P. Fortemps and R. Słowiński. A graded quadrivalent logic for ordinal preference modelling: Loyola-like approach. *Fuzzy Optimization and Decision Making*, 1:93–111, 2002.
22. M. L. Ginsberg. Multi-valued logics: A uniform approach to reasoning in artificial intelligence. *Computer Intelligence*, 4:256–316, 1988.
23. M. B. Gorzałczany. A method of inference in approximate reasoning based on interval-valued fuzzy sets. *Journal Math. Anal. Appl.*, 21(1):1–17, 1987.
24. P. Perny and A. Tsoukiàs. On the continuous extension of a four valued logic for preference modelling. In *Proc. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'98)*, pages 302–309, 1998.
25. M. Roubens and P. Vincke. Preference modelling. *Lecture Notes in Economics and Mathematical Systems*, 250, 1985.
26. R. Sambuc. Fonctions Φ-floues. Application à l'aide au diagnostic en pathologie thyroidienne, 1975. Ph. D. Thesis, University of Marseille.
27. A. Tsoukiàs, P. Perny, and P. Vincke. From concordance/discordance to the modelling of positive and negative reasons in decision aiding. In D. Bouyssou, E. Jacquet-Lagrèze, P. Perny, R. Slowiński, and D. Vanderpooten, editors, *Aiding Decisions with Multiple Criteria: Essays in Honor of Bernard Roy*, pages 147–174. Kluwer, 2002.
28. A. Tsoukiàs and P. Vincke. A new axiomatic foundation of partial comparability. *Theory and Decision*, 39:79–114, 1995.
29. A. Tsoukiàs and P. Vincke. Extended preference structures in MCDA. In J. Clímaco, editor, *Multi-criteria Analysis*, pages 37–50. Springer–Verlag, 1997.
30. I. B. Türkşen. Interval valued fuzzy sets based on normal forms. *Fuzzy Sets and Systems*, 20(2):191–210, 1986.
31. B. Van de Walle, B. De Baets, and E. E. Kerre. A plea for the use of Łukasiewicz triplets in the definition of fuzzy preference structures. (I). General argumentation. *Fuzzy Sets and Systems*, 97(3):349–359, 1998.
32. B. Van de Walle, B. De Baets, and E. E. Kerre. A plea for the use of Łukasiewicz triplets in the definition of fuzzy preference structures. (II). The identity case. *Fuzzy Sets and Systems*, 99(3):303–310, 1998.

Strategies to Manage Ignorance Situations in Multiperson Decision Making Problems

S. Alonso[1], E. Herrera-Viedma[1], F. Chiclana[2], F. Herrera[1], and C. Porcel[1]

[1] Dept. of Computer Science and Artificial Intelligence,
University of Granada, 18071 Granada, Spain
{salonso, viedma, herrera}@decsai.ugr.es, cgporcel@yahoo.com
[2] Centre for Computational Intelligence,
De Montfort University, Leicester LE1 9BH- UK
chiclana@dmu.ac.uk

Abstract. Multiperson decision making problems involve using the preferences of some experts about a set of alternatives in order to find the best of those alternatives. However, sometimes experts cannot give all the information that they are required. Particularly, when dealing with fuzzy preference relations they can avoid giving some of the preference values of the relation. In the literature these incomplete information situations have been faced giving procedures which are able to compute missing information from the preference relations. However, these approaches usually need at least a piece of information about every alternative in the problem. In this paper, several strategies to manage *total ignorance* situations, that is, situations where an expert does not provide *any* information on at least one alternative are presented, and their advantages and disadvantages analised.

Keywords: Ignorance, Incomplete Information, Consistency, Multiperson Decision Making, Fuzzy Preference Relations.

1 Introduction

Multiperson decision-making (MPDM) consists of multiple individuals (usually experts) $E = \{e_1, ..., e_m\}$ interacting to reach a decision. Each expert may have unique motivations or goals and may approach the decision process from a different angle, but have a common interest in reaching eventual agreement on selecting the . , solution(s) to the problem to be solved [4, 12]. Fuzzy preference relations are commonly used to represent decision makers' preferences over the set of possible alternative solutions $X = \{x_1, ..., x_n\}$, $(n \geq 2)$ [2, 5, 6, 7, 16, 17].

In many cases, some experts may not have a "perfect" knowledge of the problem to be solved [3, 9, 10, 11, 18]. For example, an expert might not possess a precise or sufficient level of knowledge of part of the problem or might be unable to discriminate the degree to which some options are better than others. In such cases, an expert would not be able to efficiently express any kind of preference degree between two or more of the available options, an therefore the fuzzy preference relation provided is , , , . [3, 18]. Therefore, it would be

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 34–45, 2006.

of great importance to provide these experts with appropriate tools that allow them to overcome this lack of knowledge in their opinions.

Two different kinds of incomplete information in a MPDM can be identified:

- In this case at least one expert does not provide all possible preference degrees over the set of alternatives, but provides information on his/her preferences in which every alternative is at least compared once against one of the rest of alternatives.
- In this case at least one expert does not provide all possible preference degrees over the set of alternatives, and provides information on his/her preferences in which at least one alternative is not compared against any one of the rest of alternatives. We call this an

Some attention has been paid to the case of partial incomplete information [1, 3, 18]. However, as far as we know, no study has been yet published on MPDM problem with total incomplete information. This paper presents several possible strategies to manage situations in MPDM problems: and . We analyse both their advantages and disadvantages and illustrate their application by examples. To model the consistency property we use the additive transitivity property proposed by Tanino in [16].

The rest of the paper is set out as follows. Section 2 presents notation and concepts needed throughout the papers. In section 3 we present a general consistency based procedure to estimate unknown preferences values in an incomplete fuzzy preference relation. Section 4 presents several strategies to manage ignorance situations in MPDM problems. Advantages and disadvantages associated to each one of these strategies are discussed in section 5. Finally, our concluding remarks will be pointed out in Section 6.

2 Preliminaries

Fuzzy preference relations are commonly used to represent decision makers' preferences over the set of possible alternative solutions $X = \{x_1, ..., x_n\}$, $(n \geq 2)$ [2, 5, 6, 7, 8, 13, 16, 17].

Definition 1. A Fuzzy Preference Relation (FPR) P on a set of alternatives X is a fuzzy set on the product set $X \times X$, i.e., it is characterized by a membership function $\mu_P : X \times X \longrightarrow [0, 1]$.

When cardinality of X is small, the preference relation may be conveniently represented by the $n \times n$ matrix $P = (p_{ik})$, being $p_{ik} = \mu_P(x_i, x_k)$ $(\forall i, k \in \{1, \ldots, n\})$ interpreted as the preference degree or intensity of the alternative x_i over x_k: $p_{ik} = 1/2$ indicates indifference between x_i and x_k $(x_i \sim x_k)$, $p_{ik} = 1$ indicates that x_i is absolutely preferred to x_k, and $p_{ik} > 1/2$ indicates that x_i is preferred to x_k $(x_i \succ x_k)$. Based on this interpretation we have that $p_{ii} = 1/2 \ \ \forall i \in \{1, \ldots, n\}$ $(x_i \sim x_i)$.

Since each expert is characterized by his/her own personal background and experience of the problem to be solved, experts' opinions may differ substantially

(there are plenty of educational and cultural factors that influence an expert's preferences). This diversity of experts could lead to situations where some of them would not be able to efficiently express any kind of preference degree between two or more of the available options. Indeed, this may be due to an expert not possessing a precise or sufficient level of knowledge of part of the problem, or because that expert is unable to discriminate the degree to which some options are better than others. We must note that an expert which is not able to provide a particular preference value p_{ik} does not necessarily imply that he/she is indifferent between both x_i and x_k alternatives, that is, we cannot directly suppose that $p_{ik} = 0.5$.

2.1 Incomplete Fuzzy Preference Relations

Usually, we assume that experts are always able to provide all the preferences required, that is, to provide all p_{ik} values. However, this may not always be the case, and experts end providing an [illegible] [1, 18]. In the following definitions we express the concept of an incomplete fuzzy preference relation:

Definition 2. A function $f: X \longrightarrow Y$ is [illegible] when not every element in the set X necessarily maps onto an element in the set Y. When every element from the set X maps onto one element of the set Y then we have a [illegible] function.

Definition 3. [1] An [illegible] P on a set of alternatives X is a fuzzy set on the product set $X \times X$ that is characterized by a [illegible] membership function.

When a particular preference value p_{ik} is not given by an expert we will note $p_{ik} = x$ and we will call it a [illegible].

From a particular incomplete fuzzy preference relation P_h we define the following sets [1]:

$$A = \{(i,j) \mid i,j \in \{1,\ldots,n\} \wedge i \neq j\}$$

$$MV_h = \left\{(i,j) \in A \mid p_{ij}^h = x\right\}$$

$$EV_h = A \setminus MV_h$$

$$EV_h^i = \{(a,b) \mid (a,b) \in EV_h \ \wedge \ (a = i \vee b = i)\}$$

where MV_h is the set of pairs of alternatives for which the preference degree of the first alternative over the second one is not given by expert e_h, that is, the set of [illegible] of the expert e_h, EV_h is the set of pairs of alternatives for which the expert e_h provides preference values (we call it the [illegible] for e_h) and EV_h^i is the set of preferences about pairs of alternatives given by an expert e_h involving alternative x_i.

2.2 Consistency Property

The definition of a preference relation does not imply any kind of consistency property. In fact, the values of a preference relation may be contradictory. Consistency is usually characterised by [illegible], which represents the idea that

the preference value obtained by directly comparing two alternatives should be equal to or greater than the preference value between these two alternatives obtained using an indirect chain of alternatives.

One of the properties suggested to model the concept of transitivity in the case of fuzzy preference relations is the additive transitivity property [16]:

$$(p_{ij} - 0.5) + (p_{jk} - 0.5) = (p_{ik} - 0.5) \quad \forall i, j, k \in \{1, \ldots, n\}$$

or equivalently:

$$p_{ik} = p_{ij} + p_{jk} - 0.5 \quad \forall i, j, k \in \{1, \ldots, n\} \tag{1}$$

As shown in [6], additive transitivity for fuzzy preference relations can be seen as the parallel concept of Saaty's consistency property for multiplicative preference relations [14].

This kind of transitivity has the following interpretation: suppose we want to establish a ranking between three alternatives x_i, x_j and x_k, and that the information available about these alternatives suggests that we are in an indifference situation, i.e. $x_i \sim x_j \sim x_k$. When giving preferences this situation would be represented by $p_{ij} = p_{jk} = p_{ik} = 0.5$. Suppose now that we have a piece of information that says $x_i \prec x_j$, i.e. $p_{ij} < 0.5$. This means that p_{jk} or p_{ik} have to change, otherwise there would be a contradiction, because we would have $x_i \prec x_j \sim x_k \sim x_i$. If we suppose that $p_{jk} = 0.5$ then we have the situation: x_j is preferred to x_i and there is no difference in preferring x_j to x_k. We must then conclude that x_k has to be preferred to x_i. Furthermore, as $x_j \sim x_k$ then $p_{ij} = p_{ik}$, and so $(p_{ij} - 0.5) + (p_{jk} - 0.5) = (p_{ij} - 0.5) = (p_{ik} - 0.5)$. We have the same conclusion if $p_{ik} = 0.5$. In the case of $p_{jk} < 0.5$, then we have that x_k is preferred to x_j and this to x_i, so x_k should be preferred to x_i. On the other hand, the value p_{ik} has to be equal to or lower than p_{ij}, being equal only in the case of $p_{jk} = 0.5$ as we have already shown. Interpreting the value $p_{ji} - 0.5$ as the intensity of preference of alternative x_j over x_i, then it seems reasonable to suppose that the intensity of preference of x_i over x_k should be equal to the sum of the intensities of preferences when using an intermediate alternative x_j, that is, $p_{ik} - 0.5 = (p_{ij} - 0.5) + (p_{jk} - 0.5)$. The same reasoning can be applied in the case of $p_{jk} > 0.5$.

3 Consistency Based Procedure to Estimate Missing Values in Incomplete Fuzzy Preference Relations

Given a complete fuzzy preference relation, expression 1 can be used to calculate an estimated value cp_{ik} for every p_{ik} as follows:

$$cp_{ik} = \frac{\sum_{j=1; i \neq k \neq j}^{n} cp_{ik}^{j1} + cp_{ik}^{j2} + cp_{ik}^{j3}}{3(n-2)} \tag{2}$$

where cp_{ik}^{j1}, cp_{ik}^{j2}, cp_{ik}^{j3} are directly obtained from expression 1, and the fact that additive transitivity implies reciprocity ($p_{ik} = 1 - p_{ki} \quad \forall i, k$):

$$cp_{ik}^{j1} = p_{ij} + p_{jk} - 0.5, \tag{3}$$

$$cp_{ik}^{j2} = p_{jk} - p_{ji} + 0.5, \tag{4}$$

$$cp_{ik}^{j3} = p_{ij} - p_{kj} + 0.5 \tag{5}$$

When working with an incomplete fuzzy preference relation, the previous expressions cannot be directly applied, as some of the preference values used in the expressions may be unknown. However, an iterative procedure to estimate these unknown or missing values can be derived from the above expressions. The following two different tasks have to be carried out:

A) Establish the elements that can be estimated in each step of the procedure, and
B) produce the particular expression that will be used to estimate a particular missing value.

A) Elements to be estimated in step *h*. The subset of missing values MV that can be estimated in step h of our procedure is denoted by EMV_h ([illegible]) and defined as follows:

$$EMV_h = \left\{ (i,k) \in MV \setminus \bigcup_{l=0}^{h-1} EMV_l \mid i \neq k \ \wedge \ \exists j \in \{H_{ik}^1 \cup H_{ik}^2 \cup H_{ik}^3\} \right\}$$

with

$$H_{ik}^1 = \left\{ j \mid (i,j),(j,k) \in \{EV \bigcup_{l=0}^{h-1} EMV_l\} \right\}$$

$$H_{ik}^2 = \left\{ j \mid (j,i),(j,k) \in \{EV \bigcup_{l=0}^{h-1} EMV_l\} \right\}$$

$$H_{ik}^3 = \left\{ j \mid (i,j),(k,j) \in \{EV \bigcup_{l=0}^{h-1} EMV_l\} \right\}$$

and $EMV_0 = \emptyset$ (by definition). When $EMV_{maxIter} = \emptyset$ with $maxIter > 0$ the procedure will stop as there will not be any more missing values to be estimated. Moreover, if $\bigcup_{l=0}^{maxIter} EMV_l = MV$ then all missing values are estimated, and consequently, the procedure is said to be successful in the completion of the incomplete fuzzy preference relation.

B) Expression to estimate a particular value p_{ik} in step *h*. In order to estimate a particular value p_{ik} with $(i,k) \in EMV_h$, we propose the application of the following function:

function estimate_p(i,k)

1. $cp_{ik}^1 = 0,\ cp_{ik}^2 = 0,\ cp_{ik}^3 = 0,\ \mathcal{K} = 0$
2. $if\ \#H_{ik}^1 \neq 0\ \Rightarrow\ cp_{ik}^1 = \frac{\sum_{j \in H_{ik}^1} cp_{ik}^{j1}}{\#H_{ik}^1}\ ;\ \mathcal{K}++.$
3. $if\ \#H_{ik}^2 \neq 0\ \Rightarrow\ cp_{ik}^2 = \frac{\sum_{j \in H_{ik}^2} cp_{ik}^{j2}}{\#H_{ik}^2}\ ;\ \mathcal{K}++.$
4. $if\ \#H_{ik}^3 \neq 0\ \Rightarrow\ cp_{ik}^3 = \frac{\sum_{j \in H_{ik}^3} cp_{ik}^{j3}}{\#H_{ik}^3}\ ;\ \mathcal{K}++.$
5. Calculate $cp_{ik} = \frac{1}{\mathcal{K}} \left(cp_{ik}^1 + cp_{ik}^2 + cp_{ik}^3 \right)$

end function

The function $estimate\_p(i,k)$ computes the final estimated value of the missing value, cp_{ik}, as the average of all estimated values that can be calculated using all the possible intermediate alternatives x_j and using the three possible expressions (3–5).

Then, the [illegible] is as follows:

ITERATIVE ESTIMATION PROCEDURE

0. $EMV_0 = \emptyset$
1. $h = 1$
2. while $EMV_h \neq \emptyset$ {
3. for every $(i,k) \in EMV_h$ {
4. estimate_p(i,k)
5. }
6. $h++$
7. }

This procedure is able to estimate all the missing values for a given incomplete fuzzy preference relation if a set of $n-1$ non-leading diagonal preference values where each one of the alternatives is compared at least once is known [1]. That means that partial incomplete MPDM problems can be successfully solved using this procedure. However, the only application of this procedure does not solve MPDM problem with total incomplete information. The rest of the paper is devoted to the study of some possible strategies to tackle these situations.

4 Strategies to Manage Ignorance Situations in Decision Making Problems

As per the notation introduced in section 2, an ignorance situation in MPDM problems is defined as follows:

Definition 4. In a MPDM problem with a set of alternatives $X = \{x_1, ..., x_n\}$ and a group of experts $E = \{e_1, ..., e_m\}$ which provide a set of incomplete fuzzy preference relations $\{P_1, ..., P_m\}$, we have a [illegible] if

$$\exists\ (h, i) \mid EV_h^i = \emptyset,$$

that is, at least one of the experts (e_h) does not provide any preference value involving a particular alternative (x_i). We will call x_i the "unknown alternative" for the expert e_h.

4.1 Ad-Hoc Strategies to Manage Ignorance Situations

These strategies estimate missing values in ignorance situations by ad-hoc procedures which are not based in any particular basic principle or property associated to the set of alternatives, experts or relations. Two simple examples of this kind of strategies are the following:

Strategy 1: Assume Indifference Values in the Missing Values

In this case, because an expert does not provide information on an alternative relating it to the rest of alternatives, we may model this situation as a total indifference one and therefore each missing values for the ignored alternative can be replace with a value of 0.5. In this case, the estimation procedure of missing values is as follows:

[illegible] If an incomplete fuzzy preference relation P_h has an ignored alternative x_i, this strategy will compute all its associated missing value as:

$$p_{ik}^h = 0.5\ ;\ p_{ki}^h = 0.5\ \ \forall k \in \{1, ..., n\}, k \neq i.$$

Example 1: We have to solve a decision making problem to find the best of 4 different alternatives: $X = \{x_1, x_2, x_3, x_4\}$. An expert gives the following incomplete fuzzy preference relation

$$P = \begin{pmatrix} - & 0.7 & x & 0.68 \\ 0.4 & - & x & 0.7 \\ x & x & - & x \\ 0.6 & 0.75 & x & - \end{pmatrix},$$

that is, he gives no information about alternative x_3, and thus, we are in a [illegible]. The first estimation procedure assumes that the expert is indifferent with respect to x_3, and the reconstructed fuzzy preference relation is:

$$P = \begin{pmatrix} - & 0.7 & 0.5 & 0.68 \\ 0.4 & - & 0.5 & 0.7 \\ 0.5 & 0.5 & - & 0.5 \\ 0.6 & 0.75 & 0.5 & - \end{pmatrix}.$$

Strategy 2: Assume Random Values in the Missing Values

This strategy estimates the missing values for an ignored alternative as random values within the range of preference values provided by that particular experts, i.e, an unknown preference value will be computed randomly between the maximum and minimum preference degrees of its corresponding column and row. In this case, the estimation procedure of missing values is as follows:

If an incomplete fuzzy preference relation P_h has an ignored alternative x_i, this strategy will compute every missing value as:

$$p_{ik}^h = rand(min(\{p_{jk}^h\}), max(\{p_{jk}^h\})) \; ; \; p_{ki}^h = rand(min(\{p_{kj}^h\}), max(\{p_{kj}^h\}))$$

$$\forall j, k \in \{1, ..., n\}, j \neq k \neq i$$

where $rand(a, b)$ means a random value between a and b and $max(...)$ and $min(...)$ are the usual maximum and minimum operators.

Example 2: We part from the previously presented problem (in example 1). In this case, the estimation procedure reconstructs the missing values with random values between the maximum and minimum preference degrees provided by the expert. For example, $p_{13} \in [0.68, 0.7]$ and $p_{32} \in [0.7, 0.75]$. An example of a possible reconstructed preference relation is:

$$P = \begin{pmatrix} - & 0.7 & 0.69 & 0.68 \\ 0.4 & - & 0.47 & 0.7 \\ 0.53 & 0.71 & - & 0.7 \\ 0.6 & 0.75 & 0.72 & - \end{pmatrix}.$$

4.2 Consistency Based Strategies

These strategies are guided by a basic principle, the consistency property of the incomplete fuzzy preference relations represented by the additive transitivity property. To do so, these strategies use the estimation procedure presented in section 3.

As aforementioned, that procedure needs at least a preference value involving the ignored alternative to be able to estimate the rest of missing preference values of the ignored alternative. Therefore, we need a 'seed' value to initiate the estimation procedure. Depending on the computation of that seed value we can define the following two consistency based strategies:

Strategy 3: Consistency Based Strategies with Indifference Seed Values

Similarly, as in the first strategy, we can start by assuming indifference on the preference values for the ignored alternative, followed by the application of the estimation procedure to complete the rest of missing values of the alternative. Thus, in this case the estimation procedure of missing values is as follows:

Suppose an incomplete fuzzy preference relation P with an ignored alternative x_i, and assume $p_{ij} = 0.5$ for a particular $j \in \{1, \ldots, n\}$

(initial indifference). The preference degrees $\{p_{ik}\}, \forall k \neq i \neq j$ can be estimated via the alternative x_j by means of two of the three possible estimation equations (3–5): $cp_{ik}^{j1} = p_{ij} + p_{jk} - 0.5$ and $p_{ij} = cp_{ik}^{j3} + p_{kj} - 0.5$, which result in $cp_{ik}^{j1} = p_{jk}$ and $cp_{ik}^{j3} = 1 - p_{kj}$, respectively. Because the indifference of a preference value can be assumed for any of the possible values of $j \in \{1, \ldots, n\}$ with $j \neq i \neq k$, then the final estimated values for the i-th row of the incomplete fuzzy preference relation are:

$$\begin{aligned} cp_{ik} &= \frac{1}{2}\left(\frac{\sum_{j=1;j\neq i\neq k}^{n} cp_{ik}^{j1}}{n-2} + \frac{\sum_{j=1;j\neq i\neq k}^{n} cp_{ik}^{j3}}{n-2}\right) \\ &= \frac{1}{2}\left(\frac{\sum_{j=1\ ;\ j\neq i\neq k}^{n} p_{jk}}{n-2} + \frac{\sum_{j=1\ ;\ j\neq i\neq k}^{n}(1-p_{kj})}{n-2}\right) \\ &= 0.5 + \frac{SC_k - SR_k}{2} \end{aligned}$$

with SC_k and SR_k representing the average of the k-th column and k-th row of the complete $(n-1) \times (n-1)$ fuzzy preference relation that is obtained without taking into account the alternative x_i. The parallel application of the above assumption for the preference values p_{ki} provides the following estimation of the values of the i-th column:

$$cp_{ki} = 0.5 + \frac{SR_k - SC_k}{2}. \tag{6}$$

Example 3: If we apply this strategy to the previously mentioned problem (examples 1 and 2), we obtain the following values for p_{13} and p_{32}:

$$p_{13} = 0.5 + \frac{(0.7 + 0.68)/2 - (0.4 + 0.6)/2}{2} = 0.6$$

and

$$p_{32} = 0.5 + \frac{(0.7 + 0.75)/2 - (0.4 + 0.7)/2}{2} = 0.59$$

In this case, the complete reconstructed preference relation is:

$$P = \begin{pmatrix} - & 0.7 & 0.6 & 0.68 \\ 0.4 & - & 0.41 & 0.7 \\ 0.4 & 0.59 & - & 0.51 \\ 0.6 & 0.75 & 0.49 & - \end{pmatrix}.$$

Strategy 4: Consistency Based Strategies with Random Seed Values

This strategy, similarly as in the second strategy, is based on obtaining just one 'seed' random value followed by the application of the procedure to estimate the rest of missing values for the ignored alternative. Thus, in this case the estimation procedure of missing values is as follows:

Suppose an incomplete fuzzy preference relation P_h with an ignored alternative x_i. The estimation procedure is drawn in the following scheme:

1. *do* {
2. $k = irand(1, n)$ // Choose random k
3. } *while*$(k \neq i)$
4. *if* $(rand(0,1) < 0.5)$ { // Place it in missing row
5. $p^h_{ik} = rand(min(\{p^h_{jk}\}), max(\{p^h_{jk}\}))$
   $\forall j \in \{1, ..., n\}, j \neq k \neq i$
6. } *else* { // Place it in missing column
7. $p^h_{ki} = rand(min(\{p^h_{kj}\}), max(\{p^h_{kj}\}))$
   $\forall j \in \{1, ..., n\}, j \neq k \neq i$
8. }
9. Apply the estimation procedure

where $irand(a, b)$ means an integer random value between a and b.

Example 4: From the problem presented in the previous examples, we are going to apply this strategy to reconstruct the missing values. First of all, we obtain a random $k \neq i$. For example $k = 2$. We obtain a random value between $[0, 1]$ to determine if we are going to calculate a seed value for p_{32} or p_{23}. Suppose that the random value is 0.34, so we are going to obtain a random value for $p_{32} \in [0.7, 0.75]$, for example, $p_{32} = 0.74$. Then, we apply the estimation procedure:

$$\begin{pmatrix} - & 0.7 & x & 0.68 \\ 0.4 & - & x & 0.7 \\ x & 0.74 & - & x \\ 0.6 & 0.75 & x & - \end{pmatrix} \rightarrow \begin{pmatrix} - & 0.7 & 0.46 & 0.68 \\ 0.4 & - & x & 0.7 \\ 0.59 & 0.74 & - & 0.61 \\ 0.6 & 0.75 & 0.51 & - \end{pmatrix} \rightarrow \begin{pmatrix} - & 0.7 & 0.46 & 0.68 \\ 0.4 & - & 0.42 & 0.7 \\ 0.59 & 0.74 & - & 0.61 \\ 0.6 & 0.75 & 0.51 & - \end{pmatrix}$$

5 Analysis of the Advantages and Disadvantages of Each Strategy

In this section we analyze some advantages and disadvantages of the proposed strategies, identifying situations where some of the strategies may be more adequate than the others.

- [illegible] is a very simple approach to solve ignorance situations. Although it is not always adequate to assume that not giving preference values for one alternative implies indifference between the unknown alternative and the rest of them, in some situations could be an acceptable option. In fact, its easiness application can be a very appealing factor for its use, specially in problems where there are no other sources of information (neither information about the alternatives or other experts). Particularly, decision making problems with only one expert or criterion are good candidates to apply this strategy.
- [illegible] is also a simple approach, but it can produce a higher level of diversity in the opinions given by the experts. However, it is important to remark that this strategy can produce a decrease in the consistency of the

fuzzy preference relations, because the random values will not usually comply with any kind of transitivity property. This strategy can be a good one to apply in decision problems with a high number of experts or criteria which do not differ too much between them (because it can introduce some diversity in the problem).

- : This strategy improves strategy 1, as it adjusts the estimated preference values to make the preference relation more consistent with the previously existing information. Moreover, the initial indifference supposed for every preference value for the unknown alternative is softened according to the existing information in the preference relation. This approach is interesting when there are no external sources of information about the problem and when a high consistency level is required in the experts' preference relations.
- tries to unify the advantages of strategies 2 and 3: it tries to maintain a high consistency degree in the fuzzy preference relations (with the application of the estimation procedure) whilst it gives a slightly higher level of diversity than strategy 3 (with the generation of the random seed for the estimation procedure).

6 Conclusions

In this paper we have presented several different strategies to solve ignorance situations in Decision Making problems. We have presented some ad-hoc strategies and some consistency guided strategies and have analysed their advantages and disadvantages.

In the future, we will study other possibilities to deal with ignorance situations using different criteria to the consistency one as it could be the use of consensus and/or proximity measures to provide a management system of ignorance situations.

Acknowledgements

This work has been supported by the Research Project TIC2003-07977.

References

1. Alonso, S., Chiclana, F., Herrera, F., Herrera-Viedma, E., Alcalá-Fdez, J.: A Consistency Based Procedure to Estimate Missing Pairwise Preference Values. *International Journal of Intelligent Systems*, in press (2005).
2. Chiclana, F., Herrera, F., Herrera-Viedma, E.: Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations. *Fuzzy Sets and Systems* **97** (1998) 33–48
3. Efe A. Ok.: Utility Representation of an Incomplete Preference Relation, *Journal of Economic Theory* **104** (2002) 429–449

4. Evangelos, T.: *Multi-Criteria Decision Making Methods: A comparative Study*, Kluwer Academic Publishers, Dordrecht (2000)
5. Fodor, J., Roubens, M.: *Fuzzy Preference Modelling and Multicriteria Decision Support*, Kluwert Academic Publichers, Dordrecht (1994).
6. Herrera-Viedma, E., Herrera, F., Chiclana, F., Luque, M.: Some issues on consistency of fuzzy preference relations. *European Journal of Operational Research* **154** (2004) 98–109
7. Kacprzyk, J.: Group decision making with a fuzzy linguistic majority. *Fuzzy Sets and Systems* **18** (1986) 105–118
8. Kahraman, C., Ruan, D., Dogan, I.: Fuzzy group decision-making for facility location selection. *Infomation Sciences* **157** (2003) 135–153
9. Kim, S. H., Ahn B. S.: Group Decision Making Procedure Considering Preference Strenght Under Incomplete Information, *Computers and Operations Research* **24** (1997) 1101-1112
10. Kim, S. H., Ahn, B. S.: Interactive Group Decision Making Procedure Under Incomplete Information, *European Journal Of Operational Research* **116** (1999) 498–507
11. Kim, S. H., Choi, S. H., Kim, J. K.: An interactive procedure for multiple attribute group decision making with incomplete information: Range-based approach, *European Journal of Operational Research* **118** (1999) 139–152
12. Marakas, G. H.: *Decision support systems in the 21th century, 2nd edition*, Pearson Education, Inc., New Jersey, 2003
13. Orlovski, S. A.: Decision-making with fuzzy preference relations. *Fuzzy Sets and Systems* **1** (1978) 155–167
14. Saaty, Th. L.: *Fundamentals of Decision Making and Priority Theory with the AHP*, RWS Publications, Pittsburg, 1994.
15. Szmidt, E., Kacprzyk, J.: A consensus reaching process under intuitionistic fuzzy preference relations, *International Journal of Intelligent Systems* **18:7** (2003) 837–852.
16. Tanino, T.: Fuzzy preference orderings in group decision making. *Fuzzy Sets and Systems* **12** (1984) 117–131
17. Tanino, T.: Fuzzy preference relations in group decision making. In: J. Kacprzyk, M. Roubens (Eds.). *Non-Conventional Preference Relations in Decision Making.* Springer-Verlag, Berlin, (1988) 54–71.
18. Xu, Z.S.: Goal programming models for obtaining the priority vector of incomplete fuzzy preference relation. *International Journal of Approximate Reasoning*, **36:3** (2004) 261–270

An Agent Negotiation Engine for Collaborative Decision Making

Tom Wanyama and Behrouz Homayoun Far

Department of Electrical and Computer Engineering, University of Calgary,
2500 University Drive, N.W. Calgary, Alberta, Canada, T2N 1N4
twanyama@ucalgary.ca, far@ucalgary.ca

Abstract. Negotiation engines are major components of autonomous agents, because negotiation is one of the most important types of agent interaction. Thus far, most negotiation engines rely on analytic techniques to maximize the social welfare of agent communities. Such engines are developed with total disregard of the possibility of enabling agents to analyze offers made by their negotiation opponents. The analysis of the offers leads to making tradeoffs that result into agreeing on (selecting) solution options that maximize the social welfare of the negotiation agents. Therefore, this paper presents an agent negotiation engine that supports the following: evaluation of solution options, analysis of tradeoffs, analysis of offers, and management of negotiation deadlocks. Moreover, the paper presents a simulation experiment that illustrates the capabilities of the negotiation engine.

1 Introduction

Negotiation is one of the most important forms of agent interactions in multi-agent systems, because it provides the basis for managing the expectations of the individual negotiating agents, and it enables selecting solutions that satisfy all the agents as much as possible. Consequently, a variety of agent negotiation models have been proposed [2, 3]. Most of these models can be classified into two categories, namely: analytic models and knowledge based models. The analytic models are based on analytic techniques such as, *Game Theory*, to determine the solution that maximizes the social welfare of the negotiating agents. These models minimize communication among the negotiating agents; however, they have the following drawbacks:

- The agents have no control over the tradeoffs made during the negotiation process. That is, the models consider only the quantity of the tradeoffs, disregarding their quality. In other words, analytic based negotiation models are used with an implicit assumption that the negotiating agents accept any tradeoffs so long as they are the smallest in quantity. However, this is not always true, since agents may sometimes be more willing to give larger concessions on some decision variables, than to give small concessions on other variables.
- The analytic based models do not follow the natural process of negotiation, where, in between offers and counter offers, multiple negotiation decision

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 46 – 57, 2006.

variables are traded-off against one another in order to identify the solution that maximizes the social welfare.

On the other hand, knowledge based models such as the strategic negotiation model for multi-agent systems proposed in Kraus [7]; implicitly depend on tradeoffs made by negotiating agents to determine agreement solutions. Knowledge based agent negotiation models have the following major shortfalls:

- They normally do not give any guarantees that the agreement solution maximizes the social welfare of the negotiating agents.
- They normally do not support learning from the offers made by the agent negotiation opponents in order to enable the agents to make offers that are more socially acceptable, as the negotiation progresses.
- The agents have no way of knowing whether the negotiation is converging or not.

To circumvent the shortfalls of the analytic models, as well as the shortfalls of the knowledge based models, Faratin et al [4] have proposed an agent negotiation model which depends on utility, similar to the analytic models. Moreover, the model enables the agents to tradeoff during negotiation, like the knowledge based models. Given two negotiating agents, *Agent a* and *Agent b,* the model works as follows:

1. *Agent a* identifies the solutions it prefers using utility.
2. *Agent a* identifies the solution that *Agent-b* prefers perhaps through previous offers made by *Agent b*.
3. *Agent a* determines the similarity between each of the solutions it prefers and the solution that *Agent b* prefers.
4. *Agent a* selects the solution that is most similar to the solution *Agent b* prefers as the offer.

The negotiating agents can utilize the model proposed by Faratin et al even if they have partial information about the solution, thus the model has the potential of enabling the agents to search a larger solution space. However, in the context of Group-Choice Decision Making (GCDM) problems, the model has the following shortfall: It is viable for only two negotiating agents such as in buyer-seller negotiation problems. Since GCDM problems are normally characterized by a multiplicity of (more than two) agents, the approach of Faratin et al may not be applicable to GCDM problems in its current form.

This paper presents a negotiation engine that is based on a *Qualitative Reasoning* (QR) model as well as a *Game Theory* (GT) model. We call this engine the *A*gent *N*egotiation *E*ngine for *Co*llaborative *De*cision *M*aking (ANE-CODEM). The QR model is used in the Reasoning Component of the negotiation engine to estimate the preference models of the negotiation opponents of agents. On the other hand, the Game Theory model is used in the Social Welfare Component of the engine to integrate the solution rankings according to the individual negotiating agents, in order to generate the social (combined) ranking of the solutions. Moreover, our automatic negotiation engine allows the users to influence the negotiation behavior of their agents through the strategic settings, and it has capabilities for managing negotiation deadlocks. Although ANE-CODEM has many capabilities, the focus of this paper is how it assists agents while making and responding to negotiation offers.

The rest of this paper is arranged as follows, Section 2 presents the context in which ANE-CODEM operates, and the ANE-CODEM model is presented in Section 3. Section 4 deals with a simulation experiment that illustrates the capabilities of ANE-CODEM. Finally, conclusions are given in Section 5.

2 Context in Which ANE-CODEM Operates

We developed ANE-CODEM to use it in the negotiation component of an integrated design system which we designed. The system integrates decisions of the various domains (e.g. marketing, design, manufacturing, and sales) of Product Configuration Manufacturing, into the design process. This way, design flaws would be identified and fixed early in the manufacturing process. By the nature of this target problem, ANE-CODEM turned out to be an appropriate negotiation engine for most Group-Choice Decision Making (GCDM) problems, where a number of agents are involved in choosing a single solution from a set of available solution options. Therefore, we extended the use of ANE-CODEM to the Decision Support System (DSS) for the selection of Commercial-Off-The-Shelf (COTS) software products, which we have developed into a fully functional GCDM system for COTS selection. In Group-Choice problems, it is common for the agents to have different preference models. This is caused by the differing preferences, experience, background and constraints of the clients of the agents. Consequently, ANE-CODEM enables each agent to evaluate the solution options independently, and to carryout tradeoff analysis in order to determine what the agent loses or gains if a solution other than its optimum solution is selected. Finally, ANE-CODEM enables the agents to identify the mutually acceptable solution (i.e. to reach agreement on a single solution).

2.1 Evaluation of Solution Options

Selection of a solution from a set of options is facilitated by evaluating each solution option on a set of decision variables, using a Multi-Criteria Decision Making MCDM technique. The simplest MCDM model for estimating the performance of solution option k against the set of decision variables is given by Equation (1). We used this model as a starting point for developing and testing our automatic negotiation engine, as well as its constituent components; in the future we shall replace this model with more flexible models such as Ordered weighted Averaging [5, 10] and the Logic Scoring Preference model [1]. This shall require modification of algorithms in the various components of our negotiation engine, since different MCDM techniques are based on different concepts and assumptions.

$$Score_k = \sum_{i=1}^{n} a_{ki} w_i \cdot \tag{1}$$

Where, a_{ki} is the strength of solution option k in decision variable (criterion) i.

w_i is the preference value (weight) of decision variable i.

n is the total number of decision variables.

2.2 Tradeoff

Generally, a tradeoff is where at least one of the negotiating agents lowers the preference values of some of the decision variables, and at the same time increases the values of other variables [4]. Tradeoff is carried out after comparing and contrasting the solution options on each of the decision variables, which leads to identifying solution options that have the same value, in spite of the differences in their characteristics. This increases the prospects of identifying a mutually acceptable solution. For example, a customer of a COTS-based software development project may find a high-priced COTS product which has all the required security features to have the same value with a low-priced COTS product which necessitates implementing some of the required security features in house, leading to an extra implementation cost. However, from the software developer's perspective, the former COTS product may be much more acceptable than the later. In short, tradeoff is performed to identify offers (solution options) that have the same value with the most preferred solution for the propose, but have much greater benefit for the negotiation opponents [4]. This requires the agents to have some information about the preferences of the negotiation opponents. Consequently, tradeoff is carried out in between the negotiation rounds, after the agents have estimated the preferences of their negotiation opponents.

The process of comparing and contrasting the solution options on the decision variables is referred to as tradeoff analysis. This process identifies the decision variable that can be traded between pairs of the solution options. The tradeoff analysis model used in ANE-CODEM is based on a *Qualitative Reasoning* (QR) model presented in Wanyama and Far [9]. The representation of the model uses the performance of one of the solution options to provide the baseline performance, and then identifies what the agent gains and/or loses if a solution option other than the one providing the baseline performance is selected. Furthermore, ANE-CODEM ensures that the limits of the preference values are not violated during tradeoff.

2.3 Negotiation

For negotiation to take place between two or more agents there is need for a negotiation protocol that defines the rules of the game. Within the negotiation game, each agent adopts a strategy which determines the actions it takes in response to the actions of other agents. The following subsections present the negotiation protocol associated with ANE-CODEM, and the negotiation strategies of the agents that utilize ANE-CODEM.

Negotiation Protocol: The *negotiation protocol* associated with ANE-CODEM, which we used in our COTS selection DSS works as follows:

- One of the agents makes an offer, then each of the other agents respond independently accepting the offer, rejecting it, or opting out of the negotiation.
- The negotiation stops if all agents accept the offer. In some domains, the negotiation may stop if one of the agents opts out, but if the negotiation continues then a different agent makes another offer.

- To be fair, the offers are made in negotiation rounds, and in every round, each agent makes an offer once, unless an agreement is reached before the turns of some of the agents to make their offers.

It should be noted that ANE-CODEM can work with other agent negotiation protocols. In fact, we tested it in systems where offers are made randomly, and it worked well. However, we believe that the above protocol gives all the agents a fair chance to participate in the negotiation process.

Negotiation Strategies: Negotiation strategies for GCDM have two components, the first component being the coalition which the agent decides to belong. Coalition formation increases the bargaining power of individual agents, thus the agents that join coalitions trade their autonomy for increased utility [6]. But coalition formation can lead to increase utility only if the agents have the capability to analyze the offers of their coalition partners in order to identify the socially acceptable solution options; this capability is what ANE-CODEM offers agents in GCDM systems. It should be noted; however, that the coalition formation process is not addressed in this paper.

Ideally, each member of the coalition acts to maximize the welfare of the coalition, instead of its own welfare. But in practice, the agents normally retain some autonomy, which results in the maximizing of the coalition welfare to be subjected, to some extent, to maximizing the welfare of the individual agents. Consequently, agents negotiate with varying levels of commitment to their coalitions. The *Commitment Level* (ω) of an agent to its coalition is the second component of the negotiation strategy of the agent. At the extremes, a totally committed agent ($\omega = 0$) acts to maximize the coalition welfare without taking into account the tradeoffs it makes. On the other hand, a totally uncommitted agent ($\omega = 1$) acts to maximize its own welfare without taking into account the welfare of its coalition. Effectively, such an agent retains all its autonomy, and acts on it own (does not belong to the coalition). Since in practice agents negotiate with commitment levels that lay between 0 and 1, we define an *Acceptance Factor* (A_{f-i}) for each solution option i, given by Equation (2).

$$A_{f-i} = \omega \ . \ T_{f-i} + (1-\omega) . \ S_{f-i} \ . \tag{2}$$

Where, T_{f-i} is the *Tradeoff Factor* and S_{f-i} the *Social Fitness Factor*.

The *Acceptance Factors* of solution options take into account the coalition of the agents, the commitment levels of the agents to their coalitions, and the tradeoffs associated with the solution option with respect to particular agents. The *Acceptance Factor* (A_{f-i}) of a solution option is a measure of the social acceptance of the option. In the context of ANE-CODEM, agents use the *Acceptance Factors* of solution options as a basis for accepting and/or rejecting offers made by their negotiation opponents,. This enables agents in distributed multi-agent systems to make independent, tactical (tradeoff) decisions that maximize their individual, as well as coalition welfare. The *Tradeoff Factor* (T_{f-i}) of a solution option is a measure of tradeoff associated with the option with respect to the concerned agent. *Tradeoff Factors* of the solution options are generated using a tradeoff algorithm that is based on the QR model presented in Wanyama and Far [9]. The *Social Fitness Factor* (S_{f-i}) of a solution option is a

measure of how well the option satisfies the combined preferences of a group of negotiating agents. *Social Fitness Factor*s are generated using an algorithm that is based on a *Game Theory* model of an *n-person- general sum game with complete information*. The inputs to the model are the scores of the solution options with respect to each of the negotiating agents, and the outputs are the *Social Fitness Factors* of the solution options. The *Game Theory* model is presented in detail in Wanyama and Far [8].

We believe that assuming a linear relationship between *Acceptance Factors*, *Tradeoff Factors* and *Social Fitness Factors* may limit the applicability of our rationale for making or responding to offers. However, we note that the assumption only affects the model and not the general procedure for responding to offers. Moreover, we would like to note that the negotiation approaches that are based on analytic techniques implicitly assume that all the negotiating agents have the same strategy corresponding to $\omega = 0$ and a single grand coalition (coalition of all negotiating agents); in which case all the agents are totally committed, and are willing to accept any solution that is believed to optimize social welfare no matter the amount and/or quality of the tradeoffs some of them would have to make in order to utilize the agreement solution. On the other hand, purely knowledge based techniques assume $\omega = 1$; implying that all the agents have the same negotiation strategy of acting to maximize the individual welfare without taking into account the preference of the negotiation opponents. These two extreme approaches are not practical in the context of GCDM, because in such problems the agents seek to optimize their individual gains, as well as the social welfare, thus the issue here is modeling how much an agent puts the needs of the coalition before its own needs.

3 The ANE-CODEM

In this section we briefly describe the eight major components of the Agent Negotiation Engine for Collaborative Decision Making (ANE-CODEM), and Figure 1 illustrates how the components interact to provide the functionalities of ANE-CODEM.

- *The Reasoning Component*: This component has a *Qualitative Reasoning* algorithm that is used for two purposes. The first purpose being to estimate the preference models of the agent negotiation opponents based on their offers, and the second purpose being to determine the *Tradeoff Factors* associated with each of the solution options.
- *The Evaluation Component*: The evaluation component has a MCDM model that integrates preference models of agents with the strength of the solutions in each of the decision variables to generate scores that represent the ability of each of the solution options to satisfy the agent preferences.
- *The Social Welfare Component*: This component has a *Game Theory* model which is used to determine the *Social Fitness Factors* of the solution options. The input to the Game Theory model are the estimated scores of the solution options for the coalition mates of the concerned agent, as well as the actual solution scores for the agent.

- *The Preferences Component*: This component stores the preference models of the agents. In other words, the component stores the decision variables (evaluation criteria) the agent is interested in, and the preference values (criteria weights) of the decision variables.
- *Auto-Tradeoff Component*: This component stores an algorithm that is used to modify the agent preference models automatically. The input to this component is the *Acceptance Factors* of the solution options and the preference model to be modified. The automatic tradeoff algorithm is based on the QR model that is presented in Wanyama and far [9].

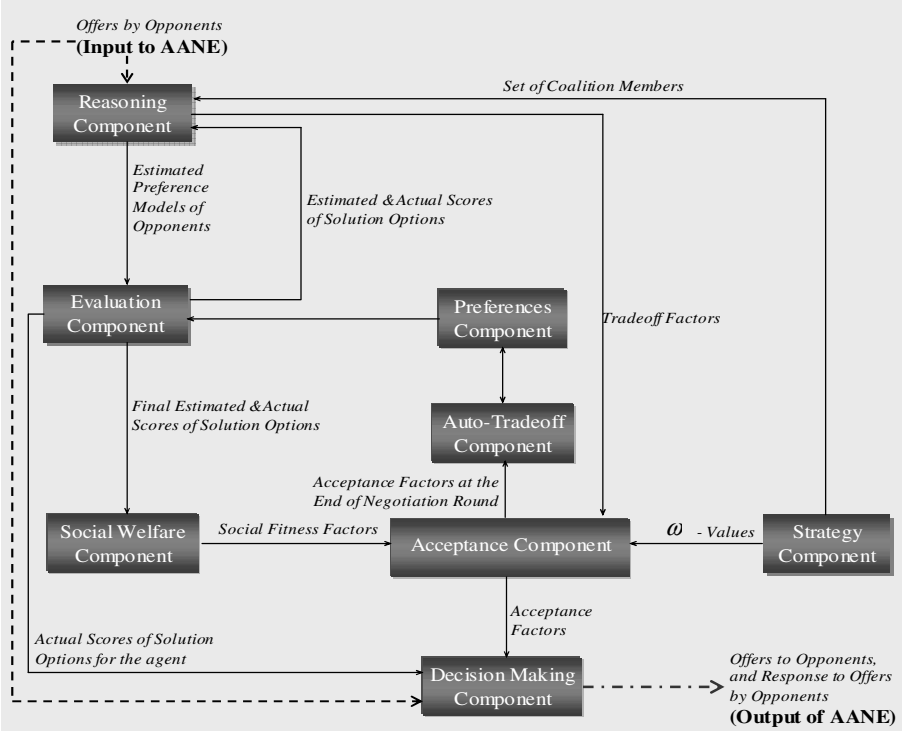


Fig. 1. The ANE-CODEM

- *The Strategy Component*: This component stores the strategies of the agents. This to say, the component stores the coalition to which the agent belongs, as well as the *commitment level* (ω) of the agent to the coalition.
- *Acceptance Component*: This component has an algorithm for combining the *Social Fitness Factors,* the *Tradeoff Factors*, and the parameters of the agent strategies, to determine the *Acceptance factors* of the solution options (see Equation 2).

- *The Decision Making Component*: This component makes decision of the offers to make based on the ranking of solution options generated by the evaluation component. In addition, the component decides on how to respond to offers made by the opponents of the agent, based on the *Acceptance Factor*s of the offers.

3.1 Making Offers

In the context of ANE-CODEM, each of the agents evaluates the solution options by integrating its preference model and the capabilities of the solution options using the MCDM model presented in Section 2.1. This generates solution scores which are sent to the Decision Making Component of ANE-CODEM, from where the agent selects the solution option that have the highest scores as its offer, which it proposes when its turn to make an offer comes .

3.2 Responding to Offers

Responding to offers in much more complicated than making them, because the agent has to analyze the offers to determine whether or not they are acceptable. On receiving an offer, the agent checks it according to the following scenarios:

1. The offer is the same as the solution option that the agent prefers. In which case, the offer is accepted.
2. The offer is not the preferred solution option of the agent, and it is made by an agent that is not a member of the agent's coalition. Such a solution is sent to the Decision Making Component of ANE-CODEM to determine whether it satisfies the acceptance criteria before accepting or rejecting it.
3. The offer is not the preferred solution option of the agent, and it is made by a member of the agent's coalition. The offer is sent to the Reasoning Component of ANE-CODEM to finally estimate the *Acceptance Factors* of the solution options. The *Acceptance Factors* are thereafter sent to the Decision Making Component of ANE-CODEM to determine whether the offer satisfies the acceptance criteria.

The *Acceptance Factors* of the solution options are updated whenever a new offer is received from a coalition member. This is done by determining the new *Social Fitness Factors* of all the solution options, based on the preferences of all coalition members that have already made their offers. If an agreement is not reached by the end of a negotiation round, the concerned agent sends the final *Acceptance Factors* of the solution options to the Automatic Tradeoff Component of ANE-CODEM to be used in the process of modifying the preference model. The agent modifies its preference model by adjusting the preference values of some decision variable in such a way as to increase the score of the solution option with the 'best' *Acceptance Factor*; if that solution is not the agent's most preferred, then the modified preference model is used to evaluate the solution option at the beginning of the next negotiation round.

3.3 Deadlock Management

Deadlock management is a challenge to automated negotiation for Multi-Agent Systems (MAS). In the context of GCDM, a deadlock is said to have occurred if at least one of the negotiating agents cannot modify it preference model to attain agreement with its opponents. This is normally caused by one of the following:

- The preference values of all the tradeoff decision variables are at their limits, meaning that there is no solution in the current solution space that satisfies the combined needs of all the negotiating agents.
- Adopting a strategy that results into selfish behavior.

Deadlocks due to the former cause are broken by changing or modifying the solution space, or by changing the preference value limits of the concerned agents. On the other hand, deadlock due to the latter cause can be broken by changing the negotiation strategy of the concerned agents, and this is the type of deadlocks that are automatically managed by ANE-CODEM.

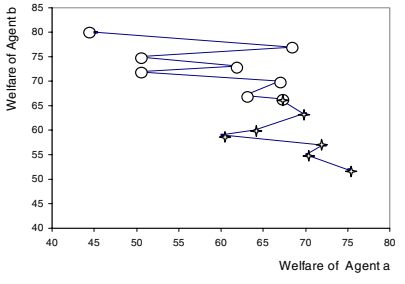
If the offers of all the agents are the same as the previous offers, ANE-CODEM recognizes that a deadlock has occurred, and triggers deadlock breaking procedures. There are two types of procedures for breaking deadlocks, namely: Procedures that modify the agent coalitions to include more members by lowering the membership requirements, and procedures that increase the commitment of the agents to their coalitions by reducing the value of ω . Some agents may employ both types of procedures to expedite the process of breaking the deadlock. However, the actual type of procedure adopted by a particular agent depends on the background knowledge of the agent.

4 Simulation Experiments

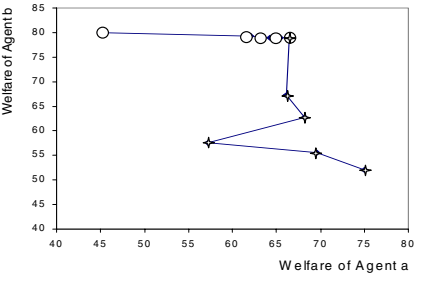
We performed a simulation experiment to investigate the capabilities of ANE-CODEM. In the experiment, agents were required to select a Commercial-Off-The-Shelf (COTS) product to be used in the development of a web-shop, from a set of four solution options (N.B. this is a typical group-choice decision making problem). A COTS expert ranked the products in each of the twelve predefined decision variables (evaluation criteria), and the initial preference value functions of the agents were generated using a truncated random number generator. The experiment had the following components:

- Negotiating without using ANE-CODEM (i.e. no tradeoff analysis and tracking of the negotiation process in order to identify socially acceptable solution options; agents adjust their preferences randomly in between negotiation rounds).
- *Agent a* and *Agent c* employing ANE-CODEM only if the offer is made by one of them, and using the existing decision variables to respond to the offers made by *Agent b*. Moreover, *Agent b* not employing ANE-CODEM. (i.e. *Agent a* and *Agent c* make tradeoffs that lead to making offers that are acceptable to both of them, and *Agent b* does not consider the preferences of its negotiation opponents, as revealed by their offers).

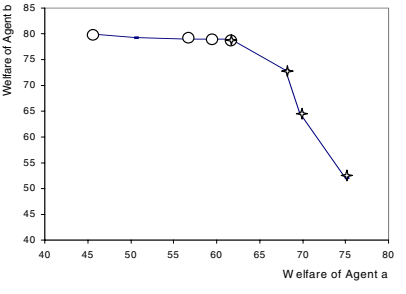
- *Agent a* and *Agent c* employing ANE-CODEM only if the offer is made by one of them, and *Agent b* employing ANE-CODEM regardless of the source of the offer being considered. (This means that *Agent a* and *Agent c* make tradeoffs that lead to making offers that are acceptable to each other, while *Agent c* makes tradeoffs that lead to making offers that are acceptable to all the three negotiating agents).
- All the three agents belonging to a single grand coalition and all of them employing ANE-CODEM regardless of the source of the offer being considered.



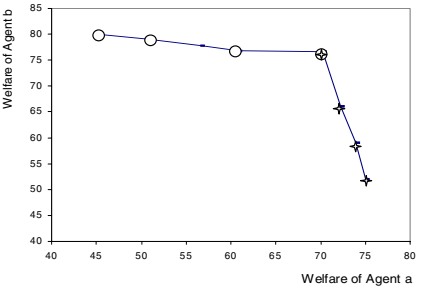
(a) No Coalition Formation



(b) Coalition between Agent b and another Agent, Agent a not Considering Preferences of Agent b and of the other Agent



(c) Coalition between Agent b and another Agent, Agent a Considering Preferences of Agent b and of the other Agent



(d) All the Three Agents Being Members of the Single Grand Coalition

Fig. 2. Effects of coalition formation on the negotiation process and on the welfare of the agents

In all the four components of the experiment, the ω-values of the agents were initially set at midpoints, (i.e. $\omega = 0.5$), and the utilities of the agents corresponding to the offers were noted. Figure 2 shows the utilities of two of the agents (*Agent a,* and *Agent b*) for the offer made during the negotiation processes for all the four components of the experiment. In the figure, the circles represent the offers made by *Agent a,* and the crosses represent the offers made by *Agent b*. A cross inside a circle represents the agreement offer.

4.1 Discussion of Results

Figure 2 illustrates that ANE-CODEM does not only increase the welfare of the individual negotiating agents, but also stabilizes the negotiation process (compare Figure 2(a) to Figures 2(c) and 2(d)). Moreover, Figure 2 demonstrates that ANE-CODEM enforces coalition formation and tradeoff analysis which in turn expedites the negotiation process (forming coalitions resulted in few negotiation rounds: compare Figure 2(a) to Figure 2(d)), and enables agents to negotiate in a smart manner, where they make offers that aim at identifying a mutually acceptable solution. Generally, Figure 2 shows that negotiating without deliberately making tradeoff that lead to the identification of socially acceptable solution options (i.e. without using a negotiation engine such as ANE-CODEM, that enforces coalition formation, as well as reasoning about offers made by negotiation opponents in order to identify mutually acceptable solution options), results in random behavior of agents, where agreements are reached almost by chance. Moreover, we have results (not presented) that reveal that when ANE-CODEM is armed with deadlock management capabilities, over 90% of negotiation deadlocks can be broken. However, it should be noted that ANE-CODEM results in increased utilization of computational resources.

5 Conclusion and Future Work

This paper presents an agent negotiation engine called ANE-CODEM, that is applicable to distributed multi-agent system for GCDM. The engine expedites the negotiation process and increases negotiation efficiency by enabling the estimation of the preferences of the negotiation opponents. Moreover, ANE-CODEM has capabilities for breaking negotiation deadlocks caused by the negotiation strategies, and it supports dynamic coalition formation, where each agent decides individually when to join or to leave a coalition. The simulation results presented in this paper show that ANE-CODEM is able to guide agents in the process of identifying mutually acceptable solution options, without having to share their preference models. Furthermore, the results illustrate that tradeoff analysis and indeed coalition formation enforced by ANE-CODEM has an effect of increasing the welfare of the negotiating agents.

For the future, this research will be extended in three major directions. Firstly, we would like to carry out a study to establish a practical meaning of ANE-CODEM parameters such as the *Acceptance Factor*, the *Social Fitness Factor*, and the *Tradeoff Factor*. This will help us to determine and set of acceptable ranges of values of such important decision parameters. Secondly, we would like to investigate the coalition formation process in order to establish an efficient algorithm for the formation of coalitions for the purpose of making tradeoffs that lead to making offers of socially acceptable solution options. Lastly, we believe that there is a need to reduce the complexity of ANE-CODEM before it can be customized for use in systems with limited computational, and/or memory facilities.

References

1. J. J. Dujmovic, W. Y. Fang, .Reliability of LSP Criteria., *LNAI 3131*, pp. 151-162, 2004.
2. N. R. Jennings, P. Faratin, A. R. Lomuscio, S. Parsons, C. Sierra, and M. Wooldridge, Automated Negotiation: Prospects Methods, and Challenges, *International Journal of Group Decision and Negotiation*, Vol. 10, No. 2, pp. 199-215, 2001.
3. N. R. Jennings, P. Faratin, M.J. Johnson, P. O'Brien, and M.E. Wiegand, Using Intelligent Agents to Manage Business Processes. *In the Proceedings of the Practical Application of Intelligent Agents and Multi-Agent Technology Conference*, London The United Kingdom. PAAM 1996, pp. 345-360, 1996
4. P. Faratin, C. Sierra, N.R. Jennings, Using similarity criteria to make issue trade-offs in automated negotiations. *Artificial Intelligence*, Vol. 142, pp. 205-237, 2002.
5. R. R. Yager, On ordered weighted averaging aggregation operators in multi-criteria decision making, *IEEE Trans. SMC*, vol.18, pp. 183-190, 1988.
6. S. Kraus 1997. Negotiation and Cooperation in Multi-Agent Environments, *Artificial Intelligence*, Vol. 94, No. 1-2, pp. 79-97, 1997.
7. S. Kraus, Strategic Negotiation in Multiagent Environments. Cambridge: Massachusetts Institute of Technology Press, pp. 17-29, 2001.
8. T. Wanyama and B. H. Far, Multi-Agent System for Group-Choice Negotiation and Decision Support. *In the proceedings of the 3rd Workshop on Agent Oriented Information Systems*, New York USA, 2004. AAMAS 2004, pp. 194-208, 2004
9. T. Wanyama and B. H. Far, Qualitative Reasoning Model for Tradeoff Analysis, *LNAI 3558*, pp. 99-109, 2005.
10. V. Torra, .The Weighted OWA operator., *International Journal of Intelligent Systems*, Vol. 12, pp. 153-166, 1997.

Learning Causal Bayesian Networks from Observations and Experiments: A Decision Theoretic Approach

Stijn Meganck[1], Philippe Leray[2], and Bernard Manderick[1]

[1] Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
smeganck, bmanderi@vub.ac.be
http://como.vub.ac.be
[2] INSA Rouen, Laboratoire PSI, BP 08 - Avenue de l'Université,
76801 St-Etienne du Rouvray Cedex, France
Philippe.Leray@insa-rouen.fr
http://psiserver.insa-rouen.fr/psi/

Abstract. We discuss a decision theoretic approach to learn causal Bayesian networks from observational data and experiments. We use the information of observational data to learn a completed partially directed acyclic graph using a structure learning technique and try to discover the directions of the remaining edges by means of experiment. We will show that our approach allows to learn a causal Bayesian network optimally with relation to a number of decision criteria. Our method allows the possibility to assign costs to each experiment and each measurement. We introduce an algorithm that allows to actively add results of experiments so that arcs can be directed during learning. A numerical example is given as demonstration of the techniques.

1 Introduction

Bayesian networks (BNs), introduced by Pearl [1], have become well known tools for working in domains with uncertainty. They allow performant probabilistic inference and give an intuitive representation of the domain.

BNs can also represent causal information when the edges represent causal relations between the corresponding variables [2]. The causal relation between two variables, in the form of a directed edge from a cause variable C to an effect variable E, is understood as the effect a manipulation of variable C (the cause) would have on variable E (the effect).

Learning BNs can be done from observations alone, by first learning the completed partially directed acyclic graph (CPDAG) and then choosing a possible complete instantiation in the space of equivalent graphs defined by this CPDAG.

This is impossible for causal Bayesian networks (CBNs), because there is only one true causal network that represents the underlying mechanisms, so the remaining edges have to be directed to represent the correct causal influence.

We discuss a decision theoretic approach for learning CBNs from a mixture of observational and experimental data. We assume we learn a CPDAG using

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 58–69, 2006.

a structure learning technique and then direct all the remaining arcs in the resulting CPDAG based on the results of experiments.

Algorithms exist to learn CBNs based on experiments [3, 4] and in [5] techniques have been developed to learn CBN from a mixture of experimental and observational data.

In [6, 7] it has been shown that in order to learn a complete structure at most $log_2(N)+1$, with N the number of variables, experiments are needed. This result is given as a theoretical bound for the worst case scenario.

The main difference with the Active Learning approaches in [3, 4] is that we assume that there is a number of observational data which we can use to form an initial CPDAG in which every directed edge is a representative of a causal mechanism. Next we introduce an experimentation phase in which we perform specific experiments in order to learn the completed CBN optimally based on some decision criterion. Our technique tries to find an optimal experimentation strategy in order to minimize the number of experiments, which should be lower than the bound derived in [7]. We allow the possibility to assign costs to an experiment, which might influence the decision of choice for performing an experiment. This type of setting is typical for medical applications where there is a lot of data from patients but it might be very costly to perform experiments.

The remainder of this paper is as follows, in the next section we introduce some notations and definitions. Then we will discuss several of the assumptions we make before introducing our decision theoretic approach and criteria used. We end with a conclusion and future work.

2 Notations and Definitions

In this work uppercase letters are used to represent variables or sets of variables, i.e. $V = \{X_1, \ldots, X_n\}$, while corresponding lowercase letters are used to represent their instantiations, i.e. x_1, x_2 and v is an instantiation of all x_i. $P(X_i)$ is used to denote the probability distribution over all possible values of variable X_i, while $P(X_i = x_i)$ is used to denote the probability distribution over the instantiation of variable X_i to value x_i. Usually, $P(x_i)$ is used as an abbreviation of $P(X_i = x_i)$.

$Ch(X_i)$, $Pa(X_i)$, $Ne(X_i)$ respectively denote the children, parents and neighbors of variable X_i in a graph. Furthermore, $Pa(x_i)$ represents the values of the parents of X_i.

A [illegible] (CBN) $\langle V, G, P(x_i|Pa(x_i))\rangle$, with:

- $V = \{X_1, \ldots, X_n\}$, a set of observable discrete random variables
- a directed acyclic graph (DAG) G, where each node represents a variable from V
- conditional probability distributions (CPD) $P(x_i|Pa(x_i))$ of each variable X_i from V conditional on its parents in the graph G.

is a Bayesian network in which the directed edges are viewed as representing [illegible] among the corresponding variables, while in a BN

the directed edges only represent a probabilistic dependency, and not necessarily a causal one.

3 Assumptions

In this section we will discuss some assumptions we make about the domain to apply our algorithm.

3.1 Faithful Distribution

We assume the observed samples come from a distribution faithful to a CBN, i.e. there are no hidden variables or confounding factors. This means that for each two variables X_i and X_j connected in the graph either $X_i \rightarrow X_j$ or $X_i \leftarrow X_j$ must hold [8]. During the remainder of this paper we will always accept this assumption.

3.2 Correct CPDAG

We assume that after an initial learning phase we are given the correct CPDAG as a consequence of applying a learning algorithm such as PC [9], BLCD [10], etc. For the technique presented in this paper we will accept this assumption.

In general it is possible that these algorithms do not retrieve the correct CPDAG, for a discussion on the properties of these algorithms see [9]. If this assumption is false, there are a lot of problems that might occur and we will have to adapt our algorithm. We are currently studying learning CBNs in this setting, and this is a part of our future research.

3.3 Modular Experiments

In order to find the causal relation between two variables X and Y we have to check whether randomizing X holding all other variables fixed at a certain value induces a variation in Y and/or vice versa. If we just randomized X and not put any constraints on the other variables then we can only detect which other variables covary with X but we cannot detect whether this relationship is direct or mediated by other variables.

If no variation is found by randomizing X this is possible due to the specific value assignment of all the other variables, and so it could be possible that we have to perform a randomization proces for all possible value assignments. Fortunately if we have the correct CPDAG, it is easier to find the causal relationships between all variables. The only possible unknown direct effects of a certain variable X are those that are connected to X by an undirected edge in the CPDAG. So performing an experiment at X, this is randomizing X, will give us direct information on the directionality of all undirected edges connected to X.

4 Decision Theoretic Approach

We use a decision theoretic approach to learn the CBN from a given CPDAG. In general a decision problem consists of three parts: values (symptoms, observables), actions and possible consequences. It is assumed that these are given in advance. It is possible to order the consequences by preference by using a utility function. Hence we can choose the action that will lead to our preferred result based on some decision criteria such as least risk or optimistic estimation.

Our decision problem is represented graphically in Figure 1, in which the possible actions are performing experiments, the values are the results of these, and the consequences are the relative utilities of the experiment. It is clear that we cannot construct the entire decision tree for this problem. Since the problem is iterative a decision can be dependent on the choice of a previous one, so we would have to construct a subtree for each possible sequence of actions and the size of the tree would explode.

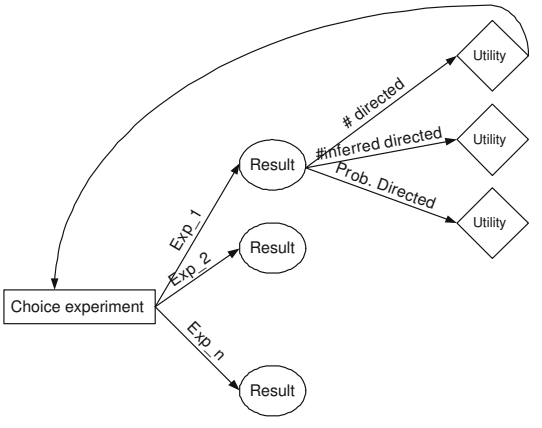


Fig. 1. Decision problem of learning a CBN from a given CPDAG

4.1 Utility Function

In general our utility function $U()$ will be a function of three variables: $gain(exp)$, $cost(exp)$, $cost(meas)$, respectively the gained information, the cost of performing an experiment and the cost of measuring other variables. If we denote performing an action (=experiment) at X_i by A_{X_i}, and measuring the neighboring variables by M_{X_i} then the utility function can be noted as:

$$U(A_{X_i}) = f(gain(A_{X_i}), cost(A_{X_i}), cost(M_{X_i})) \tag{1}$$

The only restriction that is placed on the utility function is that it is proportional to $gain(A_{X_i})$ and negative proportional to $cost(A_{X_i})$ and $cost(M_{X_i})$. In this paper we assume the following utility function:

$$U(A_{X_i}) = \frac{\alpha gain(A_{X_i})}{\beta cost(A_{X_i}) + \gamma cost(M_{X_i})} \tag{2}$$

where α, β and γ are measures of importance for every part. We will assume $\alpha = \beta = \gamma$ unless stated otherwise, this allows to simplify the notation.

Gain of an experiment. In this section we describe the information that can be retrieved after performing an experiment. Since it is our goal to direct all remaining undirected edges of our current CPDAG, the amount of edges we can direct after having results from an experiment is the gain of our experiment.

Lets assume that we perform an experiment on X_i and that we can measure all neigboring variables $Ne(X_i)$. In this case we can direct all links connecting X_i and $Ne(X_i)$ as a result of the experiment. So in this case the $gain(A_{X_i})$, with A_{X_i} = experiment on X_i, is based entirely on the number of variables that are connected to X_i by an undirected arc.

However it is possible that directing one arc can infer direction of other arcs, see the final phase of the PC-algorithm [9]. It is possible to take into account the possibility of inferred edges in $gain(A_{X_i})$. Note that the amount of edges of which the direction can be inferred after performing an experiment is entirely based on the instantiation of the undirected edges connected to the one being experimented on. An instantiation of an undireceted edge is assigning a direction to it, so for instance if we have an edge $X - Y$, then $X \rightarrow Y$ and $X \leftarrow Y$ are the two possible instantiations of that edge. We denote $inst(A_{X_i})$ as the set of instantiation of the undirected edges connected to X_i. The number of inferred edges based on $inst(A_{X_i})$ is noted as $\#inferred(inst(A_{X_i}))$.

Note that two parts of a graph that are not linked in any way by undirected edges can not be influenced by performing an experiment in the other part when the CPDAG is correct. Since we assume that all discovered arcs are correct no existing arcs can change based on inferration by experiments, and hence no new information can be inferred through a set of already directed arcs. So the calculation of the utilitiy of an experiment is only based on that part of the graph that is connected to the variable by undirected links. The problem can hence be separated in sub-problems, each concerning a part of the graph linked by undirected edges. In the remainder of this paper we will introduce solutions for a single substructure that is entirely constituted of undirected links. This result can then be mimicked for the other undirected substructures.

Cost of experiment and measurement. The cost of an experiment can be the time needed, the amount of space it takes or simply the amount of money it costs to perform an experiment. It is dependent on the situation in which every experiment takes place and will typically be given by experts.

It is important to note that there are certain experiments that can not be performed, either because of ethical reasons (e.g. infecting people with HIV) or simply because it is impossible to do so (e.g. changing the season). These types of experiments will be assigned a cost value of infinity (∞) and thus the gain of performing such an experiment will be 0, and therefore it will not add any new information.

In order to gain anything from an experiment we have to perform measurements on the variables of interest. It is however important to note that measuring itself can be costly and can diminish the usefulness of an experiment although it does not directly concern the variable that is being altered. For instance injecting someone with a certain fluid might not cost that much, but when the only way to check for changes is performing a CT-scan, measuring the results might add a huge cost factor.

5 Decision Criteria

In this section we will discuss a number of decision criteria for our learning problem. Our approach allows the possibility to maximize any of these criteria in order to converge optimally to the solution for these criteria. Depending on the type of situation in which to perform the experiments it might be advantageous to choose a specific criterion. We will show by example the working mechanism of the different criteria on the network in Figure 2, a comparative study is part of future research.

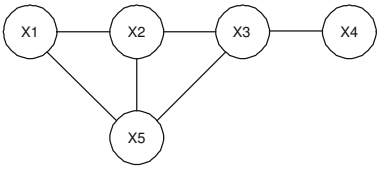


Fig. 2. Example CPDAG on which we show all optimization criteria

5.1 Maximax

The, , , decision criterion is an optimistic one, which means that we choose the action that could give the best result, i.e. the one that might direct the most arrows. In our case this means that we perform an experiment on X_{best} with:

$$X_{best} = \underset{X_i}{argmax} \left(\frac{Ne_U(X_i) + \underset{inst(A_{X_i})}{max}(\#inferred(inst(A_{X_i})))}{cost(A_{X_i}) + cost(M_{X_i})} \right) \quad (3)$$

This is the sum of the number of undirected edges connected to X_i and the maximum number of inferred edges by any of the instantiations of the directions of the undirected edges connected to X_i, divided by the cost.

In our example in Figure 2, if all costs are equal all variables except X_4 will have an equal maximal utility value, $U(X_i) = 6, i = 1, 2, 3, 5$.

5.2 Maximin

The, , , decision criterion is a pessimistic one, which means that we assume that for each experiment at a variable X_i the least number of possible inferred

edges can be found. This means the minimum amount of edges oriented by any instantiation of all edges connected to X_i. In our case this means that we perform an experiment on X_{best} with:

$$X_{best} = \underset{X_i}{argmax} \left(\frac{Ne_U(X_i) + \underset{inst(A_{X_i})}{min}(\#inferred(inst(A_{X_i})))}{cost(A_{X_i}) + cost(M_{X_i})} \right) \quad (4)$$

The instantiation of edges that would induce the least inferred edges would be the one where all arrows are pointing at X_i, but this might create new v-structures and thus is not always possible. So if two neighbors of X_i are not directly connected, one of the links has to be out of X_i and hence leads to inferred edges.

In the example Figure 2 the optimal choices are variable X_2, X_3 and X_5, since they all orient minimal 3 edges. For example, if we peform an experiment at X_2, a minimal instantiation would be: $X_1 \rightarrow X_2$, $X_3 \rightarrow X_2$, $X_5 \rightarrow X_2$.

5.3 Laplace

Using the [illegible] criterion means that we assume that all directionalities of edges are equally probable, for example for any two connected variables $P(X_i \rightarrow X_j) = P(X_i \leftarrow X_j) = 0.5$, for all non-directed edges $X_i - X_j$.

Every instantiation has thus a probability of $\frac{1}{\#inst(A_{X_i})}$ and for every possible instantiation we can calculate the number of inferred edges that will be directed.

In this case it would mean that we perform the experiment on X_{best} with:

$$\underset{X_i}{argmax} \left(\frac{Ne_U(X_i) + \frac{\sum_{inst(A_{X_i})} \#inferred(inst(A_{X_i}))}{\#inst(A_{X_i})}}{cost(A_{X_i}) + cost(M_{X_i})} \right) \quad (5)$$

In the example Figure 2 this is the variable X_3 with $U(X_3) = 3 + \frac{10}{5} = 5$. The derivation of this result is based entirely on the number of inferred edges for each instantiation. The results for X_3 are:

instantiation	#inferred	inferred
$\rightarrow X_2, \rightarrow X_4, \rightarrow X_5$	2	$X_2 \rightarrow X_1$, $X_5 \rightarrow X_1$
$\leftarrow X_2, \rightarrow X_4, \rightarrow X_5$	3	$X_2 \rightarrow X_1$, $X_5 \rightarrow X_1$, $X_2 \rightarrow X_5$
$\rightarrow X_2, \leftarrow X_4, \rightarrow X_5$	2	$X_2 \rightarrow X_1$, $X_5 \rightarrow X_1$
$\rightarrow X_2, \rightarrow X_4, \leftarrow X_5$	3	$X_2 \rightarrow X_1$, $X_5 \rightarrow X_1$, $X_2 \rightarrow X_5$
$\leftarrow X_2, \rightarrow X_4, \leftarrow X_5$	0	none

in which $\rightarrow X_i$ indicates an arrow from X_3 to X_i and vice versa.

5.4 Expected Utility

The expected utility is based on a distribution of the directions of the links. Based on this distribution it is possible to calculate the probability of any instantiation of directions that might occur. We will discuss several ways to give distributions for the directionalities.

Probabilities based on equivalence class for general graph structure. Instead of just assuming a uniform distribution of the edges we can look at all possible dags in the equivalence class of the discovered CPDAG and count for each pair $X_i - X_j$, the number of times $X_i \rightarrow X_j$ and $X_i \leftarrow X_j$ appears and hence we can assume that:

$$P_{eq}(X_i \rightarrow X_j) = \frac{\#(X_i \rightarrow X_j)}{\#\text{members of eq. class}} \tag{6}$$

$$P_{eq}(X_i \leftarrow X_j) = \frac{\#(X_i \leftarrow X_j)}{\#\text{members of eq. class}} \tag{7}$$

Note that in future steps in the learning phase we no longer have a CPDAG, because some arcs may be directed based on knowledge from experiments. We should then take into account all members of the original equivalence class that share the exact same directed edges, for convenience we will still refer to this set of dags as the members of the equivalence class of the current PDAG.

Using this approach it would mean that we perform the experiment on the variable X_{best} with:

$$\underset{X_i}{argmax} \left(\frac{Ne_U(X_i) + \sum\limits_{inst(A_{X_i})} \#inferred(inst(A_{X_i}))P_{eq}(inst(A_{X_i}))}{cost(A_{X_i}) + cost(M_{X_i})} \right) \tag{8}$$

with $P_{eq}(inst(A_{X_i}))$ meaning the number of times a certain instantiation is present in the equivalence class divided by the number of members in that class.

The problem with this approach is that we need to know the exact number of elements in the equivalence class. As far as we know there is no exact way of calculating the number of elements in the equivalence class of a certain DAG. In theory it is possible to construct all equivalent graphs, but this is very time consuming, certainly for large graphs. Hence, we can not calculate $P_{eq}(X_i \rightarrow X_j)$ in practice for a general graph.

However we can solve the problem for tree structures, since the number of elements in the equivalence class of a tree is equal to the number of nodes (= number of edges + 1). We will use this property to construct an approximation for general structures.

Approximate probabilities for general graph structure. Since the number of members in an equivalence class (of a CPDAG or PDAG as introduced earlier) is generally unknown and hard to compute we will need an approximation to solve the problem.

Checking the utility of performing an experiment at X_i means running over all possible instantiations of $X_i - Ne(X_i)$, so in our approximate technique we need to have information on all these directions. We have seen that for tree structures this problem can be solved, so instead of working with the original structure we will use a Minimum Weight Spanning Tree (MWST) algorithm in which we force that all edges $X_i - Ne(X_i)$ are present and no edges except those in the original structure are present to approximate the original structure. The weights on all the edges are given by the mutual information between the variables based on the observational data. We then use the technique for trees and use the results as approximations for the general structure.

For example if we want to check for the expected utility based on the equivalence class of performing an experiment at X_1 in Figure 2, a possible MWST is given in Figure 3.

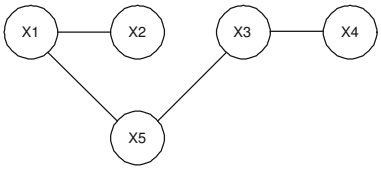


Fig. 3. Possible MWST for structure given in Figure 2

Keep in mind that we will look at all possible instantiations of $X_i - Ne(X_i)$ as they occur in the original graph. For example in Figure 3 we will also look at the instantiation $X_2 \rightarrow X_1 \leftarrow X_3$, although this is a v-structure in the tree.

Expert knowledge. It is also possible that an expert can give a certain probability to the direction of edges in the CPDAG. This is the least costly procedure since at the time of performing experiments an expert is present and it is no added cost to obtain the extra knowledge. So in this case the probabilities are based on the belief of an expert and is noted as:

$$P_{exp}(X_i \rightarrow X_j) = \text{belief of expert} \tag{9}$$

The utility function would be the same as given in equation 8 but with $P_{exp}()$ instead of $P_{eq}()$.

6 Learning Algorithm

In this section we propose our learning algorithm. We introduce an adaptive algorithm in which it is assumed that experiments are performed during learning. The first phase of the algorithm consists of applying a learning algorithm to obtain the correct CPDAG representing the equivalence class of all BNs faithful the distribution.

As stated, we allow the possibility to add newly discovered knowledge due to the experiments during the learning phase. Since experiments are performed we gain information on the direction of certain links, these may remove the need to perform certain other experiments. Remember that parts of the graph that are not connected by undirected links can be treated separately, so multiple instances of the algorithm can be applied in parallel to the substructures.

The complete algorithm is given in Algorithm 1.

Algorithm 1. Adaptive learning of CBN

Require: Observational data set.
Ensure: A CBN.

1. Apply a learning algorithm on the data-set to obtain CPDAG G.
2. Compute for each node X_i for which $\#Ne_U(X_i) > 0$ in G $U(A_{X_i})$ with equation 3, 4, 5 or 8.
3. Perform an experiment at the node with the optimal $U(A_{X_i})$ value in relation to the decision criterion, X_{best}.
4. For all $X_j \in Ne_U(X_{best})$
   If distribution of X_j changed because of experiment,
   then orient $X_{best} - X_j$ as $X_{best} \rightarrow X_j$
   else orient $X_{best} - X_j$ as $X_{best} \leftarrow X_j$
   end
5. repeat
   if $X_i \rightarrow X_j$ and X_j and X_k are adjacent, X_k and X_i are not and there is no arrow into X_j then orient $X_j - X_k$ as $X_j \rightarrow X_k$.
   if there is a directed path from X_i to X_j and an edge between X_i and X_j then orient $X_i - X_j$ as $X_i \rightarrow X_j$.
   until no more edges can be oriented.
6. Return to Step (2) until all links are directed.
7. Return CBN G.

7 Example

For example lets assume that the correct causal network of Figure 2 is as given in Figure 4.

Run with Maximin. We have seen in Section 5.2 that for the , , , criterion the optimal choices were X_2, X_3 and X_5. Suppose we choose to perform an experiment at X_2, this would result immediatly in the following instantiation: $X_2 \rightarrow X_1$, $X_2 \leftarrow X_3$ and $X_2 \rightarrow X_5$.

This instantiation then triggers the inferration of other directions in the next phase of the algorithm, $X_3 \rightarrow X_5$ and $X_5 \rightarrow X_1$.

The only remaining undirected arc is $X_3 - X_4$, so an experiment on either one of these will lead to the final result. In practice the experiment with highest utility will be chosen.

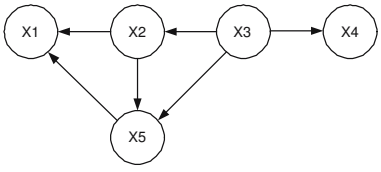


Fig. 4. The correct underlying CBN for the undirected structure in Figure 2

Run with tree approximation. In Figure 5 all MWSTs are given for each possible experiment, i.e. (a) is the tree for performing experiment at X_1, (b) for X_2, etc. Only arcs that where present in the original structure are used and all arcs connected to the variable being experimented on remain, as discussed previously.

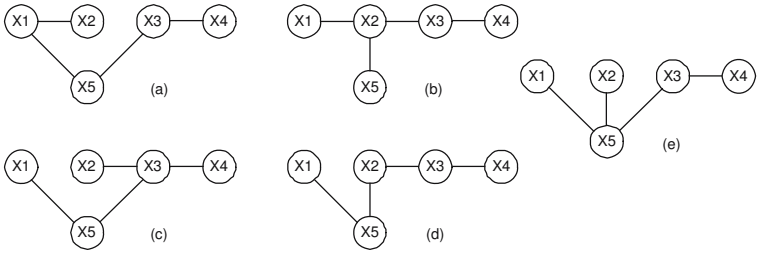


Fig. 5. All MWSTs for each possible experiment

The optimal choice (again assuming all costs equal) is variable X_3 (since not all arcs can be directed inwards in the original graph, there are less instantiations without inferred edges in the tree), and this would result immediatly in the following instantiation: $X_3 \rightarrow X_2$, $X_3 \rightarrow X_4$, $X_3 \rightarrow X_5$.

This instantiation then triggers the inferration of $X_2 \rightarrow X_1$ and $X_5 \rightarrow X_1$.

The only remaining undirected arc now is $X_2 - X_5$, thus again an experiment on any of these nodes will solve the problem.

8 Conclusion and Future Work

We discussed a decision theoretic approach to learn causal Bayesian networks from a mixture of experiments and observational data.

We used the information of observational data to learn a completed partially directed graph and tried to discover the directions of the remaining edges by means of experiment. Our method allows the possibility to assign costs to each experiment and each measurement.

We demonstrated that our approach allows to learn a causal Bayesian network optimaly with relation to a number of decision criteria. For the expected

utility for a general structure we gave an approximation based on the solution for tree structures.

We introduced an algorithm that is adaptive, since it allows to actively add results of experiments. The algorithm is a general description and can be used with any decision criteria or utility function.

A first part of future work is a comparative study between the different decision criteria for learning CBN structure. We would also like to extend this approach to take into account that the CPDAG learned in the first phase is not correct and allow hidden variables.

Acknowledgments. This work was partially funded by a IWT-scholarship. This work was partially supported by the IST Programme of the European Community, under the PASCAL network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

1. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann (1988)
2. Pearl, J.: Causality: Models, Reasoning and Inference. MIT Press (2000)
3. Murphy, K.P.: Active learning of causal bayes net structure. Technical report, Department of Computer Science, UC Berkeley (2001)
4. Tong, S., Koller, D.: Active learning for structure in bayesian networks. In: Seventeenth International Joint Conference on Artificial Intelligence. (2001)
5. Cooper, G.F., Yoo, C.: Causal discovery from a mixture of experimental and observational data. In: Proceedings of Uncertainty in Artificial Intelligence. (1999) 116–125
6. Eberhardt, F., Glymour, C., R.Scheines: N-1 experiments suffice to determine the causal relations among n variables. Technical report, Carnegie Mellon University (2005)
7. Eberhardt, F., Glymour, C., R.Scheines: On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In: Uncertainty in Artificial Intelligence Conference (UAI). (2005)
8. Shipley, B.: Cause and Correlation in Biology. Cambridge University Press (2000)
9. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction and Search. MIT Press (2000)
10. Mani, S., Cooper, G.F.: Causal discovery using a bayesian local causal discovery algorithm. In: MEDINFO 2004, IOS Press (2004) 731–735

The Pairwise Comparison Model: The Multiplicative and the Additive Approach

Antonio Jesús Herencia-Leva[1], M[a] Teresa Lamata[2], and Cristino Pérez-Meléndez[1]

[1] Depto de Psicología Social y Metodología de las Ciencias del Comportamiento Facultad de Psicología, Universidad de Granada,18071 Granada, Spain
herencia@ugr.es, cristino@ugr.es
[2] Depto de Ciencias de la Computación e Inteligencia Artificial, E.T.S de Ingeniería Informática, Universidad de Granada,18071 Granada, Spain
mtl@decsai.ugr.es

Abstract. The aim of this work is to study some of the differences between the additive and multiplicative representations associated with the Analytic Hierarchy Process. We present the Method of Pair Comparisons with the study of its properties from the point of view of representational measurement theory and scaling theory. From the first point of view it is impossible to differentiate two types of representations and therefore, the distinction has to be done in terms of the type of task that the subjects perform. The conclusion establishes some differences and relationships between the task of making judgments of proportion and judgments of distance.

1 Introduction

Within the tradition of psychological measurement there are a number of sub-schools. Perhaps the most notable is the division between The Scaling Theory and The Representational Theory of Measurement. Luce and Krumhansl [1] state that the scaling problem is one of finding how best to recast a set of data into a particular numerical representation within a chosen class of representations. Krantz et al [2] affirm that scaling literature involves mapping one numerical structure in another one.

The Representational Theory of Measurement means one or more specified operations and relations over a set of objects or events that are characterized by a number of empirically testable assumptions (Luce [3]). The measurement axioms offer a solution in that they can be considered expressions of potential empirical laws (Falmagne [4]).

Therefore, the principal difference between the Theory of Scaling and the Representational Theory of Measurement, is primarily a certain level of scale is assumed in the first moment (ordinal, interval or ratio, for example) and later by means of tests of adjustment it verifies the grade in which it is possible to support the above mentioned assumption. On the other hand, the Representational Theory of Measurement specified from the beginning a series of qualitative properties, that if they are satisfied, they guarantee the existence of a given representation and therefore, of a concrete level of scale.

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 70 – 80, 2006.

One of the most important tasks is the Method of Pairs Comparisons (MPC), this is well documented and widely employed by economists, psychometricians, decision-makers, etc…It provides a simple and practical way of eliciting the opinion of a referee from a set of $n(n\text{-}1)/2$ pairwise judgments that express the preferences among $n \geq 2$ alternatives which are compared with respect to an attribute/criterion. Hegelmaier introduced the MPC, published in German in 1852, and later, in 1860, Gustav Theodor Fechner [5], through two major contributions to this method, developed it as a basic tool for experimental psychology Siegler [6]. Fechner used the MPC in the framework of psychophysics, in which the attributes to be measured had to have a knowledge correspondence with an extensive attribute, i.e. an attribute for which the operation of addition is empirically defined.

In 1927 Thurstone [7] developed a model known as the "Law of Comparative Judgment" (LCJ), for the MPC. The principal importance of it was that:

a) It generalizes some of the psychophysical models to the case when the scaling attribute doesn't have a knowledge correspondence with an extensive attribute, i.e. the model was constructed without a mention of a reference (something like the referential set in Fuzzy Set Literature; see, for example, Norwich & Turksen, [8]).

b) It opens the door for measuring social attributes, allowing social psychology statistically defend its empirical studies.

The MPC was used later by Saaty [9-13] to build their model on decision-making introducing some changes in the response format. Some authors have reported some problems with this multiplicative response format and they have proposed alternative scales to make the decision makers judgments more consistent (see [14] for a review).

The questions related to consistency are one of the principal problems in the Saaty's multiplicative approach [9] [14] but they are not addressed in this paper.

The aim of this work is to study some of the differences between the additive and multiplicative approaches of the Saaty model. In section two, we present the classical preference structure related to the MPC. In section three we present the measurement theory approach: we describe various axioms systems which can represent the data obtained by the different versions of the task of the MPC. Finally, in section four we would like to remark that with the multiplicative and additive version of MPC we can construct a ratio scale, because, it has been shown that category ratio scaling has many advantages compared to magnitude matching, the various modifications introduced by the models mentioned earlier and how to compute the scale values within each model.

2 Classical Preference Structures

In a decision problem, a decision maker is usually confronted with a set of alternatives $A = (a_1, a_2, ..., a_n)$, among which, for instance, the best alternative has to be selected. In the following, we ask that the decision maker compares any two alternatives a_i and a_j in A. It is then acceptable to assume that the decision maker either

- prefers a_i to a_j
- prefers a_j to a_i
- is indifferent to a_i and a_j

Let the binary relation $\succ_C$ represent "more preferred than", $\sim_C$ "indifferent to" and $\prec_C$ "less preferred than" with respect to a criterion *C*. Hence, denote "a_i is preferred to a_j, with respect to *C*" by $a_i \succ_C a_j$. Likewise, if a_i and a_j are equally preferred, based on *C*, we write $a_i \sim_C a_j$, and $a_i \prec_C a_j$ represents the case where a_i is less preferred than a_j. Assuming that for every pair of alternatives a_i, a_j can be assigned a real number $P_C(a_i, a_j)$ (positive or negative) which represents their relative preference, such that:

- $a_i \succ_C a_j$ if and only if $e < P_c(a_i, a_j)$,
- $a_i \sim_C a_j$ iff $-e \le P_c(a_i, a_j) \le e$, and
- $a_i \prec_C a_j$ iff $P_c(a_i, a_j) < -e$, with respect to *C*.

e is the absolute threshold value in the structure and it determines the minimal absolute distance for which a stimulus is judged different to other stimulus.

Rating a set of *n* elements $(a_1, a_2, ..., a_n)$ can be a task even when a unique criterion is considered demanding. The comparison matrix (*CM*) represents all possible comparisons,

$$CM = \begin{pmatrix} P_C(a_1,a_1)\, P_C(a_1,a_2) \cdots P_C(a_1,a_j) \ldots P_C(a_1,a_n) \\ P_C(a_2,a_1)\, P_C(a_2,a_2) \cdots P_C(a_2,a_j) \ldots P_C(a_2,a_n) \\ \vdots \qquad\qquad\qquad P_C(a_i,a_j) \\ P_C(a_n,a_1)\, P_C(a_n,a_2) \ldots P_C(a_n,a_j) \ldots P_C(a_n,a_n) \end{pmatrix}$$

It provides a simple and practical way of eliciting the opinion of a referee from a set of *n(n-1)/2* pairwise judgments that express the preferences among $n \ge 2$ alternatives which are compared with respect to an attribute/criterion.

3 Measurement Theory Approach

In this section two possible types of numerical structures are studied. These can be used to represent the information obtained in the task of pair comparisons. Section 3.1 presents an algebraic difference structure as an example of additive representation. Section 3.2 outlines an algebraic ratio structure as an example of multiplicative representation.

3.1 Algebraic Difference Structure

The MPC is an algebraic difference structure, if and only if, we have a set $X=(x, y, z, ...)$ which is a nonempty set and a order relation "≥" on "X", i.e., a binary relation on "XxX" and the following axioms are satisfied [2] [15-17].

D.1. (XxX, ≥) is a strict weak order, if and only if on verify the properties.

D.1.1. Asymmetry
($\forall$ x, y, z, r $\in$ X), (if xy ≥ zr → not (zr ≥ xy))

D.1.2. Transitivity
($\forall$ x, y, z, r, v, w $\in$ X), (if xy ≥ zr & zr ≥ vw then xy ≥ vw)

D.2. Reciprocity
If xy ≥ zr then rz ≥ yx

D.3. Transitivity
If xy ≥ zr & yv ≥ rw then xv ≥ zw

D.4. Resolution
If xy ≥ zr ≥ xx, then there exist r', r'' $\in$ A, such that xr' ~ zr ~ r''y.

D.5. Arquimedian

If x_1, x_2, ..., x_i, ... is a strictly bounded standard sequence ($x_{i+1}x_i \sim x_2x_1$, for every a_i, a_{i+1} in the sequence; not $x_2x_1 \sim x_1x_1$; and there exist r', r''$\in$ X such that r'r''≥ x_ix_1 ≥ r''r', for all x_i in the sequence.

Proposition 1. If (XxX, ≥) is an algebraic-difference structure, then there exists a real-valued function ϕ_1 on X such that, for all x, y, z, w $\in$ X,

$$xy \geq zw \text{ iff } \phi_1(x) - \phi_1(y) \geq \phi_1(z) - \phi_1(w) \quad (1)$$

3.2 Algebraic Ratio Structure

The additive and multiplicative version of the MPC is an algebraic ratio structure if and only if we have a set "X" which is a nonempty set and a quaternary relation "≥" on "X", i.e., a binary relation on "XxX" and the later axioms are satisfied [2] [15-17].

Proposition 2. If (XxX, ≥) is an algebraic-difference structure, then there exists a real-valued function ϕ_2 on X such that, for all x, y, z, w $\in$ X,

$$xy \geq zw \text{ iff } \phi_2(x)/\phi_2(y) \geq \phi_2(z)/\phi_2(w) \quad (2)$$

An implication of propositions 1 and 2 is that we cannot differentiate between an additive and ratio representation because if equation (1) is true we can transform the function ϕ_1 into ϕ_2 and equation (2) is true too. So, in terms of mathematical models, we cannot differentiate between the two types of representations. In section 4, we propose to make the distinction in terms of the type of task that subjects do.

4 The Multiplicative and Additive Models

The Analytic Hierarchy Process (AHP) introduced by T.L. Saaty [9] is a well known and popular method of multicriteria decision making. Central to this method are the pairwise comparisons between criteria (and decision alternatives) made using a 9-unit scale. The appropriateness of Saaty's original one-to-nine (1–9) scale has been the subject of much debate and cause for concern.

4.1 The Multiplicative Approach

We have a set L of linguistic sentences L_i, where i often is $i = 1, 2, ..., 17$. These sentences express different values of the psychological distance between the stimuli of the pair. The task of the subject is to pick up the sentence that best express his/her subjective perception of the distance between the stimuli of the pair.

We can construct now a matrix similar to $CM=[P_C(a_i, a_j)]$, we denote it by $A=[a_{ij}]$. In this matrix a_{ij} denotes the evaluation about the absolute ratio of the row stimulus i versus the column stimulus j.

$$A=\left[a_{ij}\right]=\begin{pmatrix} 1 & a_{12} & \dots & a_{1j} & \dots & a_{1n} \\ 1/a_{12} & 1 & \dots & a_{2j} & \dots & a_{2n} \\ . & & & & & \\ 1/a_{1j} & 1/a_{2j} & \dots & a_{ij} & \dots & a_{in} \\ . & & & & & \\ 1/a_{1n} & 1/a_{2n} & \dots & 1/a_{in} & \dots & 1 \end{pmatrix}$$

We have that $a_{ij} = L_i \in L$.

In the pairwise comparisons method, stimuli (e.g. criteria or alternatives) are presented in pairs to one or more referees (e.g., experts or decision makers). It is necessary to evaluate individual alternatives, derive weights for the criteria, construct the overall rating the alternatives and identify the best alternative.

Let us denote the stimuli by $A=[A_1, A_2, \dots, A_n]$ (n is the number of compared stimuli), their actual weights by $w=[w_1, w_2, \dots, w_n]$ and the matrix of the ratios of all weights by $W = [w_i / w_j]$. The matrix of pairwise comparisons $A = [a_{ij}]$ represents the intensities of the expert's preference between individual pairs of alternatives (A_i versus A_j), for all $i,j = 1, 2,\dots, n$ chosen usually from the scale given by Saaty [9], it can be represented by

$$a_{ij} = \begin{cases} a & \text{if } i \text{ is preferred to } j \\ 1 & \text{if } i \text{ and } j \text{ are equally preferred} \\ 1/a & \text{if } j \text{ is preferred to } i \end{cases}$$

the values of a are obtained from the values $\langle 1, 2, ..., 9 \rangle$.

The elements a_{ij} are considered to be estimates of the ratios w_i / w_j, where w is the vector of actual weights of the stimuli, which is what we want to find. All the ratios are positive and satisfy the reciprocity property: $a_{ij} = 1/a_{ji}$ ($i,j = 1, 2, \dots, n$) Saaty´s

eigenvector solution of $Aw=\lambda w$ always exists if the consistency (or transitivity) condition $a_{ij} = a_{ik} \times a_{kj}$ $(i,j,k = 1, 2, \ldots, n)$ is satisfied.

In the eigenvector method, the vector of weights corresponds to the maximum eigenvalue λ_{max} of the matrix A. According to the Perron-Frobenius Theorem, the eigenvalue λ_{max} is positive and real. Furthermore, the vector w can be chosen with all positive coordinates. It is a normalized solution of the following equation:

$$Aw = \lambda_{max} w$$

4.2 The Additive Approach

For each pair of stimuli, the subjects treat with the function

$$b_{ij} = \begin{cases} b & \text{if } i \text{ is preferred to } j \\ 0.5 & \text{if } i \text{ and } j \text{ are equally preferred} \\ 1-b & \text{if } j \text{ is preferred to } i \end{cases}$$

in which, $b \in (0.5,1]$ that represents the additive model.

$$B = [b_{ij}] = \begin{pmatrix} 0.5 & b_{12} & \ldots & b_{1j} & \ldots & b_{1n} \\ 1-b_{12} & 0.5 & \ldots & b_{2j} & \ldots & b_{2n} \\ \cdot & & & & & \\ 1-b_{1j} & 1-b_{2j} & \ldots & b_{ij} & \cdots & b_{in} \\ 1-b_{1n} & 1-b_{2n} & \cdots & 1-b_{in} & \cdots & 0.5 \end{pmatrix}$$

We have now that

$$b_{ij} + b_{ji} = 1$$

and the diagonal cells of matrix B will are equal to 0.50.

By example, the assignment of 0.80 to one member of the pair and the remaining 0.20 to the other is same as to indicate that the former is four times as large as the latter. A split of 0.50-0.50 indicates that the stimuli are of the same magnitude [18] [19] .

For each subject we have a matrix similar to A, we denote it by $R=[r_{ij}]$. In this matrix r_{ij} denotes the evaluation about the absolute ratio of the row stimulus i versus the column stimulus j.

$$R = [r_{ij}] = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1j} & \ldots & r_{1n} \\ 1/r_{12} & 1 & \ldots & r_{2j} & \ldots & r_{2n} \\ \cdot & & & & & \\ 1/r_{1j} & 1/r_{2j} & \ldots & r_{ij} & \ldots & r_{in} \\ \cdot & & & & & \\ 1/r_{1n} & 1/r_{2n} & \ldots & 1/r_{in} & \ldots & 1 \end{pmatrix}$$

Now we can define r_{ij} as the ratio of the stimulus values indicated by the column to those indicated by the row:

$$r_{ij} = \frac{b_{ij}}{b_{ji}} = \frac{b_{ij}}{1 - b_{ij}}$$

so, we have that

$$r_{ij} = \frac{1}{r_{ji}}$$

but r_{ij} can be defined by

$$log\ r_{ij} = log\ b_{ij} - log\ b_{ji}$$

and if we make the sum in *j*, we can compute r_i by

$$\log r'_i = \sum_{j=1}^{n-1} (\log b_{ij} - \log b_{ji})$$

but $\log r'_i = \log b_i - \log 1$ and so, $\log r'_i = \log b_i$, $\log r'_i$ is the simulation of the geometric mean ratio of stimulus *i* all along the set *X* of stimuli as estimation of the weights, in place to use the eigenvector as in the multiplicative case. Now if we want to express the ratio value of *i* in function of another stimuli, we can compute:

$$r'_{ik} = e^{r'_i / r'_k}$$

that it already constitutes a ratio scale.

4.3 Relations Between the Two Approaches

It is possible to confirm that the two approaches represent the same models and it is possible to work with the two types of information, the multiplicative and the additive one. The major difference between them is on the scale; in the multiplicative case the scale is finite {1/9, 1/8,…, 1,…, 8, 9} while, in the additive case there is very much granularity, and the values of the matrix could be anyone in the (0, 1) interval. But, there is a very close relationship between these two approaches as we have shown previously. Now, we illustrate the relationships between these two approaches.

Definition 1. Let $A = \left[a_{ij} \right]$ be a multiplicative preference matrix, with values in {1/9, 1/8,…, 1,…, 8, 9} and let $B = \left[b_{ij} \right]$ be an additive preference matrix with values in [0,1] then, the relation between them is given by

$$f : \langle 1/9, ..., 1, ..., 9 \rangle \to [0,1]$$
$$a_{ij} \to b_{ij} = 1/2 \log_9 9 a_{ij}$$

such that

$$1/9 \to 1/2\log_9 1 = 0$$
$$\langle 1/9,...,1\rangle \to [0,1/2]$$
$$1 \to 1/2\log_9 9 = 1/2$$
$$\langle 1,...,9\rangle \to [1/2,1]$$
$$9 \to 1/2\log_9 81 = 1$$

The relations between a_{ij} and b_{ij} are given by

$$b_{ij} = 1/2\log_9 9a_{ij} \Rightarrow a_{ij} = 9^{2b_{ij}-1}$$

Definition 2. Let $A = \left[a_{ij}\right]$ be a multiplicative preference matrix, we call A a consistent multiplicative preference matrix, if $a_{ij} = a_{ik} \times a_{kj}$, $\forall$ $i,j,k \in N$.

Definition 3. Let $B = \left[b_{ij}\right]$ be an additive preference matrix, we call B a consistent additive preference matrix, if $b_{ij} = b_{ik} + b_{kj} - 1/2$, $\forall$ $i,j,k \in N$.

Theorem 1. Let $A = \left[a_{ij}\right]$ as in definition 1 with $a_{ij} = a_{ik} \times a_{kj}$ a consistent matrix, then the matrix $B = \left[b_{ij}\right]$ with $b_{ij} + b_{ji} = 1$ related to the additive case, also is consistent being this consistency given in [14] [20] by $b_{ij} = b_{ik} + b_{kj} - 1/2$.

Proof.
If

$$a_{ij} = a_{ik} \times a_{kj}$$

Then

$$81a_{ij} = 9a_{ik} \times 9a_{kj} \Rightarrow \log_9 9 \times 9a_{ij} = 1 + \log_9 9a_{ij} = \log_9 9a_{ik} + \log_9 9a_{kj}$$

if we multiply this equation by ½.

$$1/2(1 + \log_9 9a_{ij}) = 1/2\log_9 9a_{ik} + 1/2\log_9 9a_{kj} \Rightarrow 1/2 + b_{ij} = b_{ik} + b_{kj}$$

Theorem 2. Let $B = \left[b_{ij}\right]$ as in definition 1 with $b_{ij} + b_{ji} = 1$ and $b_{ij} = b_{ik} + b_{kj} - 1/2$ then the matrix $A = \left[a_{ij}\right]$ with $a_{ij} = a_{ik} \times a_{kj}$ also is consistent.

Proof.
If

$$b_{ij} = b_{ik} + b_{kj} - 1/2$$

Then

$$2b_{ij} = 2b_{ik} + 2b_{kj} - 1 = (2b_{ij} - 1) = (2b_{ik} - 1) + (2b_{kj} - 1) \Rightarrow$$
$$9^{(2b_{ij}-1)} = 9^{(2b_{ik}-1)+(2b_{kj}-1)}$$

and

$$a_{ij} = a_{ik} \times a_{kj}$$

Table 1. Table of equivalences between the multiplicative and the additive approaches

Multiplicative	Additive
(1:1)	[0.50, 0.50]
(1/2:2)	[0.33, 0.66]
(1/3:3)	[0.25, 0.75]
(1/4:4)	[0.20, 0.80]
(1/5:5)	[0.16, 0.83]
(1/6:6)	[0.19, 0.85]
(1/7:7)	[0.12, 0.87]
(1/8:8)	[0.11, 0.88]
(1/9:9)	[0.10, 0.90]

If a subject is able to judge stimuli with the values that are reflected in Table 1, the multiplicative and the additive models are equivalent.

In Figure 1 it can be consulted a comparison between the additive and multiplicative approaches. Assuming that in a comparison between two stimuli, a decision-maker chooses with the same probability the different linguistic labels available, the probability of choosing a label to which a numerical extreme value corresponds is higher in the multiplicative approach than in the additive approach, indicating that in this case the probability of an inconsistent judgment would be higher also.

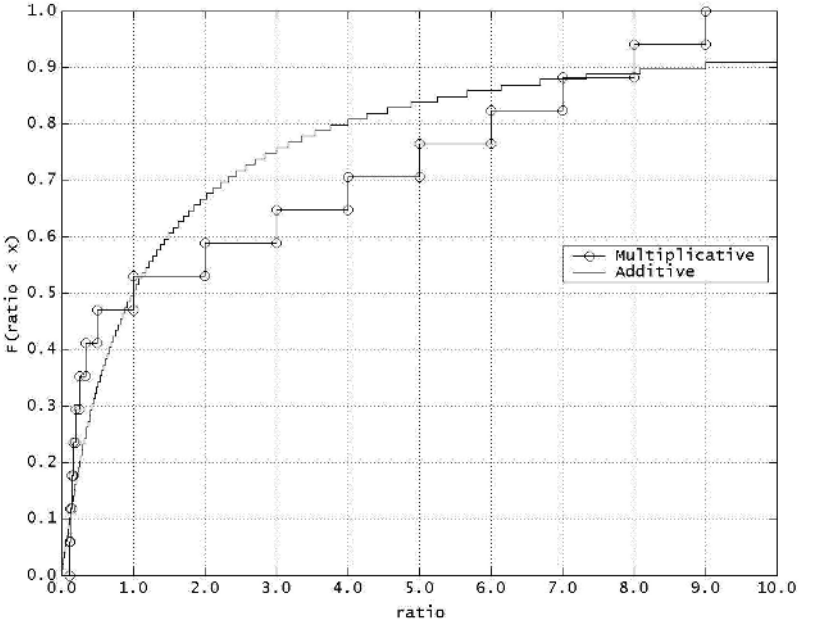


Fig. 1. Relationship between the multiplicative and the additive numerical scales in the pairwise comparison method

The Figure has been obtained assigning an equal probability mass to each numerical value that it expresses a possible ratio between a pair of stimuli. The horizontal axis shows the values that the ratio between two stimuli can reach (the maximum value being 10); the vertical axis shows the accumulated probability that a subject may randomly choose a label whose numerical value is less or equal to that ratio. The lines represent the distribution of probability for additive and multiplicative scales.

In the multiplicative approach the number of possible ratio values (x-axe in the figure) is finite and the cumulative probability function (y-axe in the figure) is staggered. Nevertheless, in the additive approach the number of possible values is infinite and the cumulative probability function is continuous.

In Figure 1, if P_M is the multiplicative cumulative probability function, P_A is the multiplicative cumulative probability function and x is any ratio value, then for x in the interval [1, 8], $P_A(x) \geq P_M(x)$ and for x in the interval [0,1) and x in the interval (8, 10] $P_A(x) \leq P_M(x)$. This indicates that for the multiplicative approach the probability of choosing a extreme ratio value is higher.

5 Conclusions

The AHP method is based on pairwise comparison between attributes, and can be used to evaluate the relative performance of decision alternatives with respect to one or several criteria. In the case study, the results from these comparison analyses "Saaty's multiplicative and additive measurement scales" show that the considered two methods for prioritization have different performance, but they are very closer.

The paper studied the differences between the additive and multiplicative approaches of the Saaty method for pairwise comparison. In this way, if the referees are consistent, the traditional method of Saaty is much closer to that in which the scale is not multiplicative. For all intents and purposes the two methods are the same, but they have one significant difference, this is the scale. So, while the scale of responses in the additive is any number between (*0, 1*), and therefore the granularity is large, in the case of Saaty this granularity is small. However, the Saaty scale is closer to the natural language.

The questions related to consistency are one of the principal problems in the multiplicative approach that are not addressed in this paper but it is possible to affirm that in the additive case this problem is reduced because it counts with much more granularity, i.e. theoretically, the decision maker has an infinity number of possible ratio values to choose. Nevertheless, in the Saaty´s multiplicative approach, if the decision maker has to make an infinity number of judgments, then it's obliged to choose a linguistic label that it makes inconsistent its judgments. So, theoretically the multiplicative approach makes inconsistent the decision maker judgments. This topic will be analyzed in depth by the authors in a future article.

Acknowledgements

This paper was partially financed by the Spanish research agency DGICYT within the framework of projects TIC2002-03411 and TIN2005-02418.

References

1. Luce, R.D., Krumhansl, C.: Measurement, scaling, and psychophysics. In R.C. Atkinson, R.J. Herrnstein, G. Lindzey, & R.D. Luce (Eds.) Stevens'Handbook of Experimental Psychology. New York: Wiley. Pp. (1998) 1-74
2. Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A..: Foundations of Measurement. Vol.: 1. Additive and Polynomial Representations. New York: Academic Press (1971)
3. Luce, R.D.: On the possible psychophysical laws. Psychological Review (1959) 66, 81-95. Reprinted in R.D. Luce, R.R. Bush, & E. Galanter (Eds.) Readings in Mathematical Psychology, Vol. I. New York: Wiley 1963. Pp. 69-83. Also in Bobbs-Merril Reprint Series.
4. Falmagne, J.C.: Psychophysical Measurement and Theory. J.-Cl. Falmagne, In: Handbook of Perception and Human Performance, Eds. K. Boff, L. Kaufman and J. Thomas. New York: John Wiley, (1986)
5. Fechner, G.T.: Elemente der Psychophysik. Leipzig: Breitkopf und Härtel, 2, p. 559. (1860) (Reprinted, Bristol: Thoemmes Press, 1999).
6. Brown, N. R., Siegler, R. S.: Metrics and mappings: A framework for understanding real-world quantitative estimation. Psychological Review, 100, (1993) 511-534
7. Thurstone, L.L. : A law of comparative judgment. Psychological Review, 34, (1927) 273-86.
8. Tüksen, I.B., Bilgiç, T.: Interval valued strict preference with Zadeh triples. Fuzzy sets and Systems (1996) 78, 183-195
9. Saaty, T.L.: The Analytic Hierarchy Process. New York: McGraw-Hill (1980)
10. Saaty, T.L.: The analytic hierarchy process: What it is and how it is used. Mathematical Modeling 9, (1987) 161–176.
11. Saaty, T.L.: Ratio scales are fundamental in decision making, Proceedings of the Fourth International Symposium on the Analytic Hierarchy Process, Simon Frasier University, Burnaby, B. C., Canada, July 12-15, (1996) 146-156.
12. Saaty, T.L.: A scaling method for priorities in hierarchical structures, Journal of Mathematical Psychology, 15, (1997) 234-281
13. Saaty, T.L., Ozdemir, M.S.: Why the Magic Number Seven Plus or Minus two. Mathematical and Computer Modelling, 38, (2003) 233-244.
14. Lamata, M.T., Peláez,J.I.: A New Consistency Measure for Reciprocal Matrices. Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems. Paris, Vol I, (1998) 211-215
15. Narens, L.: The measurement theory of dense threshold structures. Journal of Mathematical Psychology, 38, (1994) 301-321.
16. Narens, L.: A theory of ratio magnitude estimation. Journal of Mathematical Psychology, 40, (1996) 109-129.
17. Narens, L.: Symmetry, direct measurement and Torgenson's conjecture (Working paper nº 32). Irving, California: Institute of Behavioral Sciences (2005).
18. Torgerson, W.S.: Theory and methods of scaling. New York: John Wiley and Sons. Fourth Edition. (1958)
19. Blanco, M.J.: Psicofísica. Madrid: Editorial Universitas. (1996) 406-415.
20. Tanino, T. Fuzzy preference orderings in group decision making, Fuzzy Sets and Systems 12, (1984) 117–131.

Simultaneous Decision Networks with Multiple Objectives as Support for Strategic Planning

Ivan Blecic, Arnaldo Cecchini, and Giuseppe A. Trunfio

Laboratory of Analysis and Modelling for Planning,
Department of Architecture and Planning - University of Sassari,
Palazzo del Pou Salit, Piazza Duomo, 6 I07041 Alghero (SS), Italy
ivan@uniss.it, cecchini@uniss.it, trunfio@uniss.it

Abstract. Strategic planning can be schematised as a decision making process where, given a general outline of the desirable future, the decision makers need to choose a set of actions that should coherently lead a system (corporation, institution, city, region, etc.) toward that future. A more sophisticated case is when rather than only choosing actions, the decision maker also decides the allocation of available resources among different available actions. We show that in most cases the problem can be faced using a particular Decision Network with multiple objectives, in which actions are applied simultaneously and are modelled by variables representing the efforts spent on them. The main advantage of the proposed Simultaneous Decision Network is that it can be easily built by a panel of domain experts, under the assumption of the noisy-OR causal interaction. The problem of finding the best strategy in terms of resource allocation is formulated as a combinatorial optimisation, and solved through a multi-objective meta heuristic approach.

1 Introduction

While the exploration and the treatment of causal knowledge using Bayesian Network (BN) based approaches [1], has received great attention in the recent years, this has also highlighted a fundamental problem with the probabilistic reasoning, namely the large amount of a priori knowledge required for the specification of the Conditional Probability Table (CPT).

The problem becomes more evident when dealing with economical, political, and social issues, since most of the conditional probabilities required by the BN approach are difficult to estimate. In these fields often not enough data can be gathered to automatically determine the probabilities. In these cases the model construction turns out to be quite time consuming, since the input data must largely be derived from experts' opinions and value judgements sustained by knowledge, experiences, intuition and common sense.

An interesting issue, often belonging to the category of problems introduced, arises when a human decision maker needs to allocate an available amount of resources among various actions to be applied simultaneously, to reach a trade-off between many, often conflicting, objectives. This is usually the case with a

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 81–92, 2006.

Strategic Planning (SP) process in organisations, finalised toward the achievement of institutional goals and objectives.

Accounting for the uncertain knowledge usually available, the probabilistic effects of actions on the objectives can be modelled in terms of a Simultaneous Decision Network (SDN), that is, a BN augmented with actions applied 'in parallel' (i.e. all at the same instant), represented by variables.

Following the SDN approach, it should be take into consideration that the number of CPT entries to be estimated by human experts, for a node of the network, grows exponentially with the number of its direct causes (i.e. actions or events). Furthermore, when actions are variables belonging to domains with high cardinality, each variable representing the amount of [illegible] (i.e. money, energy, time, etc.) spent on the corresponding action, the problem becomes even more significant. Thus, the elicitation of such a model by human experts, as well as its solution, can be arduous in real applications.

On the other hand, in the context outlined above, causes are often expressions of quite independent sub-systems, in such a way that their effects can be modelled by the well known noisy-OR canonical model [2]. As will be shown below, this enables experts to build a probabilistic model in an easier way, estimating in a semi-qualitative manner the relationships which functionally link the amount of effort spent on actions and the probability of their effects.

Once the model is built, there is a feasible approach with SDN allowing to find the 'best' strategy in terms of effort allocation, with respect to the objectives. Such approach consists of solving all associated BN where an evidence is given (i.e. the values of action variables) and the probabilities of events influencing the objectives are to be computed [3]. In our case, the actions potentially belonging in sets with high cardinality, such an approach can bring about a need to perform a high number of BN inferences. This suggests the formulation of the problem as a multi-objective optimisation.

Some of the above outlined concepts were integrated in the Decision Support System (DSS) described in the rest of the paper. In particular, the subject of the next section is the underlying causal model, in section 3 the multi-objective search is described, which was developed in terms of a multi-objective genetic algorithm, and in section 4 an application example is presented. Section 5 concludes the paper with some final considerations.

2 The Causal Model

The entities of the model are represented as variables, which will be indicated by upper case letters (e.g. Y) or indexed upper-case letters (e.g. X_i). A specific value of a variable will be denoted by lower case letters (i.e. x). The set of all possible values of a variable Y will be indicated as $D(Y)$. Set of variables will be indicated by calligraphic style letters (e.g. $\mathcal{Y}$). The general scheme of the model, represented in Figure 1, includes:

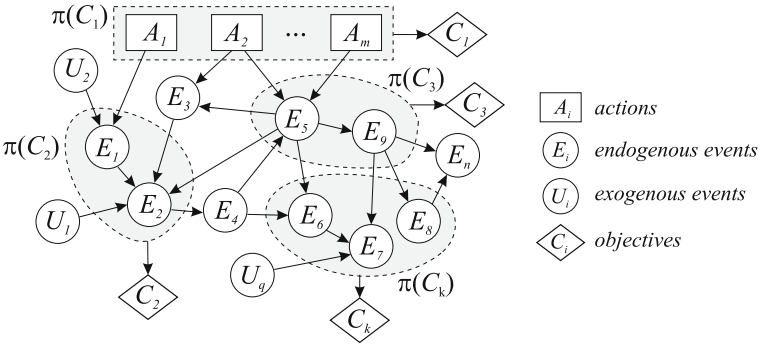


Fig. 1. A multi-objective SDN scheme and the involved entities

- a set $\mathcal{A}$ of m actions which are modelled as continuous real variables. The value a of an action $A \in \mathcal{A}$ represents the [illegible] spent on it. In general, the domain of definition of an action A is the discrete set $D(A) \subseteq [0, +\infty[$;
- a set $\mathcal{V}$ of events which are modelled as binary random variables. The two values of an event $Y \in \mathcal{V}$ will be denoted by the correspondent lower case letter y, with the meaning of 'occurrence', and negate lower case letter $\neg y$. Hence, the domain of definition of an event Y is $D(Y) = \{y, \neg y\}$. In order to facilitate the process of elicitation of the model by domain experts, the event set $\mathcal{V}$ is partitioned in the set $\mathcal{U}$ containing [illegible], which can not be influenced by any of the entities in the model, and the set $\mathcal{E}$ containing [illegible], which can be influenced by all other events in $\mathcal{V}$ and by actions in $\mathcal{A}$
- a set $\mathcal{C}$ of k objectives. Each $C \in \mathcal{C}$ is a variables defined in the set of real numbers $\Re$, reflecting the value of an objective function g_C.

A [illegible] α is defined as the m-uple $\alpha = (a_1, \ldots, a_m) \in \Pi_{i=1}^{m} D(A_i)$, representing an effort assignation for each action.

The employed model of interaction between variables is a BN, augmented with decision nodes and objective nodes, that is, a Decision Network (DN). It is worth to note that, differently from the more frequent utilisation of DNs, in this case the multiple decisions are assumed to be made in parallel and not sequentially. In other words, in the DNs used in this paper there is no precedence relationship among decision nodes, and they are not depending on any other model entity, so the model can be called Simultaneous Decision Network (SDN) (or Simultaneous Influence Diagram as in [3]).

As every BN, the causal model can be represented by a directed acyclic graph consisting of a set of nodes and the links among these nodes (see Figure 1). Each node is associated with a variable Y representing a model entity and, for simplicity, we will refer interchangeably to nodes and associated variables. Each variable Y has a parent set $\pi(Y)$ in the network. Actions and exogenous events have not ingoing edges, i.e., their parent sets are always empty. Endogenous events can have ingoing edges representing probabilistic dependence, whereas

objectives can have ingoing edges representing a functional dependence. In the DN, nodes representing objectives do not have children.

In particular, for each $E \in \mathcal{E}$, the distribution $P(E|\pi(E))$, where $\pi(E) \subseteq \mathcal{V} \cup \mathcal{A} \setminus \{E\}$, specifies the probability of each value of E given each possible assignment of $\pi(E)$. If the domains of definition of E and $\pi(E)$ are finite, such a distribution is called a Conditional Probability Table (CPT).

Each objective C takes a value given by an objective function $g_C : D(\pi(C)) \rightarrow \Re$, where $D(\pi(C)) = \Pi_{X \in \pi(C)} D(X)$. By hypothesis, as shown in Figure 1, the parent set of the objective C_1 is the set of actions, i.e., $\pi(C_1) = \mathcal{A}$, while for each objective C_i, $i = 2, \ldots, k$, is defined a specific parent set $\pi(C_i) \subseteq \mathcal{E}$ as part of the model.

In order to simplify the formalisation, the action variable A_i can be viewed, with respect to a strategy α, as a random variable whose probability is $P_\alpha(A_i) = 1$ if $A_i = a_i$ and $P_\alpha(A_i) = 0$ otherwise. Thus, temporarily neglecting the objective variables, the DN can be viewed as a BN representing the following joint probability:

$$P_\alpha(\mathcal{V}, \mathcal{A}) = \prod_{Y \in \mathcal{V}} P(Y|\pi(Y)) \prod_{i=1}^{m} P_\alpha(A_i) \tag{1}$$

The [illegible] of the objective variable $C \in \mathcal{C}$ in case of applying strategy α, is defined as

$$E_\alpha(C) = \sum_{\pi(C)} P_\alpha(\pi(C)) g_C(\pi(C)) \tag{2}$$

where $P_\alpha(\pi(C))$ denotes the marginal probability of the variables $\pi(C)$ in the BN, that is,

$$P_\alpha(\pi(C)) = \sum_{\mathcal{A} \cup \mathcal{V} \setminus \pi(C)} P_\alpha(\mathcal{V}, \mathcal{A}) \tag{3}$$

with the summation extended to all possible values in $\mathcal{A} \cup \mathcal{V} \setminus \pi(C)$.

It is worth to note that, given the above definitions, for the objective C_1, whose parent set $\pi(C_1)$ is the set of action variables $\mathcal{A}$, the expected value corresponding to the strategy α is simply $E_\alpha(C_1) = g_{C_1}(\{\alpha\})$.

The computation of the expected values of objectives C_i with $i > 1$, given the strategy α, involves the use of a proper inference procedure for the BN representing the joint probability (1). To this end, it is interesting to observe that approximate inference procedures, such as standard stochastic sampling methods, present quite fast convergence avoiding the problem of unlikely evidence, since the latter (i.e. the values of random variables representing actions) are concentrated in the roots of the BN graph.

2.1 Multi-attribute Utility Versus Multi-objective Approach

Following the [illegible], in most DN models the multiple objectives C_i are combined in a single terminal value node, through a multi-attribute utility function. In practice, the multi-objective problem is transformed into one with the single objective of maximising expected utility. The main

advantage of combining objectives lies in its simplicity: an [illegible] is the strategy that achieves the optimal value of the expected utility E_α. On the other hand, as pointed out in [4], the approach of combining objectives in a single utility function presents many drawbacks.

First, it is difficult to find a multi-attribute utility function which incorporate many conflicting objectives. When weights are used, there is a high subjectivity in their choice and, often, it is hard to use weights to combine objectives with different physical meaning. Second, it requires an explicit statement of preferences by decision maker prior to the solution process: if the preferences change, the entire solution process must be repeated. Another issue is related to the possible inaccuracy in the expected value of an objective: using a multi-attribute utility function, the entire output of the decision making process will be affected by that inaccuracy. Besides, while the solution process can potentially give valuable information about how the system represented by the DN works, this is precluded by combining all objectives in a single value.

A different approach exploited in the causal model described here (see Figure 1) consists in maintaining multiple objectives, emphasising the generation of a range of solutions and tradeoffs to be presented to the decision maker for consideration. The main purpose of this approach is to produce the entire set of non-dominated or effective solutions. The analyst then helps decision maker choose among possible solutions. The solution chosen is called the most preferred solution or the best-compromise solution and is attained by examining and exploring the various tradeoffs between objectives for the entire set of possible solutions.

2.2 Probabilistic Dependence of Events

Some types of conditional probability distributions can be approximated by canonical models that require fewer parameters. One type of model describing how an effect Y is probabilistically produced by n causes $\pi(Y) = \{X_1, \ldots, X_n\}$, is the noisy-OR gate [5, 2]. In the domain of application of SP, such canonical interaction often approximates the true distribution sufficiently well, reducing the model building effort significantly. On the other hand, as showed in [6], in case of models built by human experts, even if the distribution does not exactly respect the noisy-OR hypothesis, the elicitation error affecting the noisy-OR parameters might be smaller than the elicitation error of the complete CPTs. For these reasons the noisy-OR gate is used in our DSS.

With the purpose of briefly recalling the noisy-OR model, we say that the event represented by the variable X_i may be a cause of Y when it is [illegible] with respect to Y, that is, when it assumes a particular value $x_i^{\uparrow Y}$. It is worth to note that since in a large BN an event X_i may represent a cause for more than one effect, the activation value of X_i may be different for each effect. In general we can define the set:

$$\pi^*(Y) = \{X \in \pi(Y) : X = x^{\uparrow Y}\} \tag{4}$$

to which belong all the X_i's which are in the active state for Y. Formally, the noisy-OR assumptions are:

- each of the causes X_i is characterised with respect to its potential effect Y, by a parameter $p_i \in [0,1]$ called [illegible];
- the ability of each cause being sufficient is independent of the presence of other causes (i.e. when the effect Y has not been produced each cause X_i has failed independently).

The parameter p_i is defined as the probability that the effect Y will be true if the cause X_i is active and all other causes X_j, with $j \neq i$, are inactive:

$$p_i = P(y \mid x_i^{\uparrow Y} \wedge \neg x_j^{\uparrow Y} \forall j \in [1,n], j \neq i) \tag{5}$$

The above hypothesis allow us to specify the entire conditional probability distribution, given the non-empty activation set $\pi^*(Y)$ and only n parameters, i.e. a parameter p_i for every X_i. In particular the probability of y is given by the following formula:

$$P(y|\pi^*(Y)) = 1 - \prod_{i:X_i \in \pi^*(Y)} (1 - p_i) \tag{6}$$

which is sufficient to derive the complete CPT of Y conditional on its predecessor $X_1, X_2, \ldots, X_n$.

In practice, it is almost impossible to list all possible cause that can produce the effect Y. For this reason Henrion [5] proposed an extension of the binary Noisy-OR gate for situations where the effect can be produced even if all its explicit causes are inactive and called this extended model a [illegible] gate. This can be conceptualised by introducing an additional cause X_0 which is assumed to always be active. There is an additional parameter $p_0 \in [0,1]$, called the [illegible], associated to this cause, i.e. the probability that the effect Y is produced by the unmodelled causes when all the modelled causes are inactive. Hence:

$$p_0 = P(y \mid \neg x_j^{\uparrow Y} \forall j \in [1,n]) \tag{7}$$

Since X_0 is always active, the CPT is expressed by the equation:

$$P(y|\pi^*(Y)) = 1 - (1 - p_0) \prod_{i:X_i \in \pi^*(Y)} (1 - p_i) \tag{8}$$

where in this case p_i is the probability that the effect Y is produced by X_i if all modelled and unmodelled causes are inactive.

Causal strength shape functions. Equation (8) refers to the situation where an event Y has only other events X_i as its explicit causes in the model. It can be extended to the situation represented in Figure 2, where among additional causes of Y there is also the effort value of an action A. To this end, let us introduce an additional factor $1 - p(A)$ in Equation (8), which becomes:

$$P(y|\pi^*(Y), A) = 1 - (1 - p_0)\,(1 - p(A)) \prod_{i:X_i \in \pi^*(Y)} (1 - p_i) \tag{9}$$

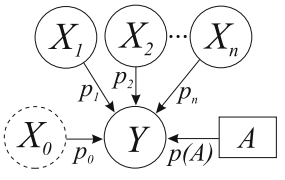


Fig. 2. Action causal strength

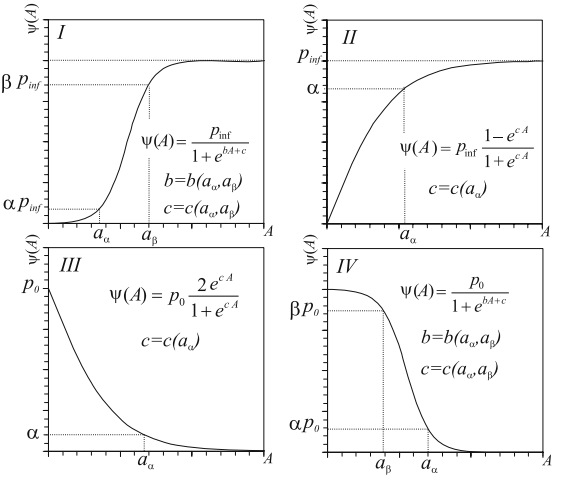


Fig. 3. Examples of causal strength functions

where $p(A)$ is the causal strength of the action A, with an obvious meaning in the context of the noisy-OR assumptions. In order to simplify the elicitation process, we assume that $p(A)$ can be expressed as:

$$p(A) = \psi(A;\, q_1, \ldots, q_n) \tag{10}$$

where ψ is a function, representing an [illegible]. These functions are chosen, during the modelling phase, by the domain experts on the basis of their knowledge and opinions about the interactions. In Figure 3, few examples of strength functions included in the model are represented. For example, the function *II* in Figure 3 allows to express the causal strength given the two parameters p_{inf} and a_α: the asymptotic value of ψ and the value of the action corresponding to strength $\alpha\, p_{inf}$, respectively.

2.3 Elicitation of Model Parameters

As explained in the section 1, in real-case applications it is frequent that not enough data (i.e. cases) are available to automatically learn the model. Hence, the BN must largely be derived from experts' opinions, both in terms of structure and parameters. For all these reasons, the developed DSS was designed in such a way the experts are involved in a collaborative interactive phase of modelling, considerably helped by the visual definition of the model interrelation as a directed graph (see Figure 4). In terms of parameters the model requires the estimation of:

- the occurrence probability $P(u_i)$ for each exogenous event U_i;
- the leak probability p_0 for each endogenous event E_i;

– the causal strength (i.e. a single scalar parameter) for each causal link from event to event;
– the causal strength in terms of the choice of a shape of function and associated scalar parameters, for each causal link from action to event.

Regarding the latter point, the DSS interactively represents the chosen shape of the function given its actual parameters.

3 Searching for Best Strategies

A way to find the trade-off [illegible] $\alpha = (a_1, \ldots, a_m) \in \Lambda = \Pi_{i=1}^{m} D(A_i)$, consists in the computation of the expected values of the k objectives C_i in correspondence to all possible strategies [3]. In our case such an approach might turn out to be too expensive from a computational point of view, being necessary to perform a complete inference of the BN representing the joint probability (1) for each possible strategy. In fact, the domain of definition of actions might present high cardinality, this potentially leading to a very large number of strategies in set Λ. On the other hand, the problem can be conveniently formulated as multi-objective search as follows:

$$\max_{\alpha \in \Lambda} E_\alpha(C_i), \quad i = 1 \ldots k \tag{11}$$

A more complete formulation may include a constraint on the total effort

$$\sum_{i=1}^{m} A_i \leq a_{max} \tag{12}$$

where $a_{\max}$ is a maximum allowed value.

The objective functions in problem (11) are not available in explicit form, this suggesting the use of techniques based on the local function knowledge such as Genetic Algorithms (GAs), which were already successfully exploited in conjunction with BN, although in a single objective form [7]. In our case, a GA can be used to evolve a randomly initialised population, whose generic element is a m-dimensional vector $\mathbf{s}$ (i.e. a chromosome) representing an element $\alpha \in \Lambda$. The i-th gene of the chromosome is obtained as the binary encoding of A_i, using a suitable bit numbers and its interval of definition $D(A_i)$. Each chromosome can be decoded in a strategy α and, performing the BN inference, the objective functions can be computed. In the present implementation of the program, the well known [illegible] was used as approximate BN inference algorithm.

To avoid the combination of multiple objectives, as illustrated in section 2.1, the GA is used in order to search for a Pareto-optimal set of solutions [8]. Thus, the comparison of two candidate solutions with respect to the different objectives is achieved through the introduction of the concepts of Pareto optimality and dominance, avoiding any a priori assumption about the relative importance of individual objectives. In particular, considering the optimisation problem (11), we say that a solution α (strongly) [illegible] the solution β if:

$$\forall\, i\; E_\alpha(C_i) \geq E_\beta(C_i) \wedge \exists\, j\; : E_\alpha(C_j) > E_\beta(C_j) \tag{13}$$

In other words, α dominates β if α is better or equivalent to β with respect to all objectives and better in at least one objective. A non-dominated solution is optimal in the Pareto sense (i.e. no criterion can be improved without worsening at least one other criterion), and a search based on such a definition of optimum almost always produces not a single, but a set of non-dominated solutions, from which the decision-maker should select one.

The employed multi-objective GA (MOGA) is based on the widely used Goldberg's 'non-dominated sorting' [9]. Briefly, the procedure proceeds as follows: (i) all non-dominated individuals in the current population are assigned to the highest possible rank; ,, these individuals are virtually removed from the population and the next set on non-dominated individuals are assigned to the next highest rank. The process is reiterated until the entire population is ranked.

The MOGA proceeds on the basis of such ranking: every individual belonging to the same rank class has the same probability to be chosen as a parent. The employed GA makes use of elitism as suggested by the recent research in the field [10], which means that from a generation to another the non-dominated individuals are preserved. This allows us to extract the Pareto-set from the last population. It is interesting to observe that some care is necessary using elitism in this case, because of the noise that may affect the fitness functions if an approximate stochastic method is used for the BN and the number of sampling was too low. In fact, noisy fitness functions coupled with elitism may produce outliers which remains forever in the population. Nevertheless, to cope with this problem it is sufficient to use a proper number of sampling in the stochastic evaluation of the fitness.

4 An Application Example

The application example illustrated in this section is relative to a policy-making case-study. The model graph is represented in Figure 4, including actions (i.e. rectangles) and events (i.e. ovals). In the graph a minus symbol as label of an arc from an event X to an event Y means $x^{\uparrow Y} = \neg x$, whereas a plus symbol means $x^{\uparrow Y} = x$ (see Section 2.2). A minus symbol as label of an arc from an action A to an event Y means that a decreasing causal strength function was used, whereas a plus symbol means that an increasing one was used (see also Figure 3). The thickness of the arcs represents the intensity of the causal strength (i.e. the maximum or the asymptotic value in case of arcs from actions to events). The model required the estimation of 15 leak probabilities, 16 constant causal strength and 16 shapes of functions. The analysis was based on three objectives, with parents defined as:, the set of actions $\mathcal{A}$ as $\pi(C_1)$;,, the set $\mathcal{G} \subseteq \mathcal{V}$, as $\pi(C_2)$, containing events reputed as positive facts by the panel of domain experts;,,, the set $\mathcal{B} \subseteq \mathcal{V}$, as $\pi(C_3)$ of events reputed as negative facts. In particular, the objective functions were defined as:

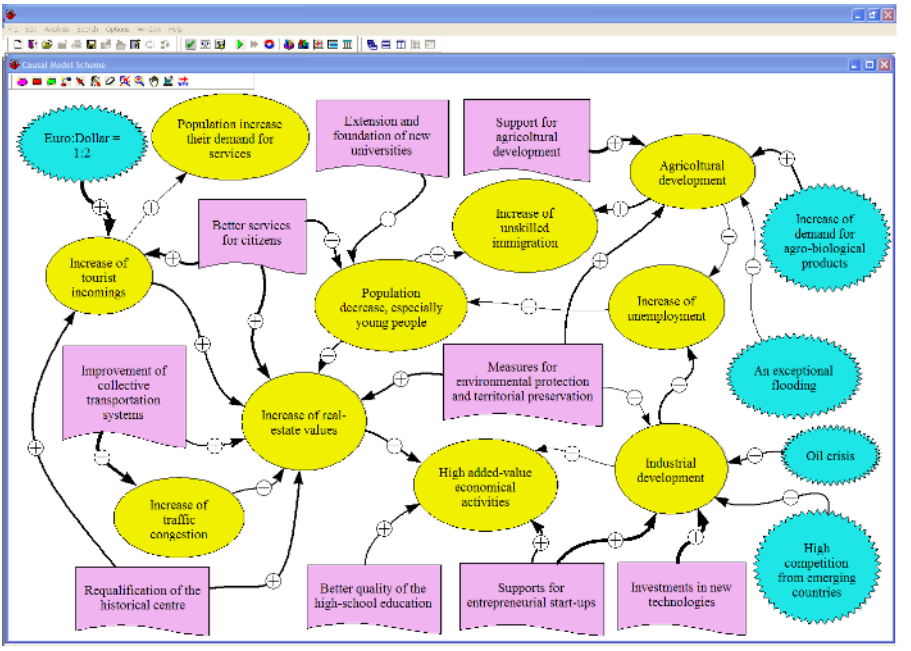


Fig. 4. The presented example graph from the developed DSS

$$g_{C_1} = \sum_{A_i \in \mathcal{A}}^{m} A_i, \qquad g_{C_2} = \frac{1}{\sharp \mathcal{G}} \sum_{X_i \in \mathcal{G}} x_i, \qquad g_{C_3} = \frac{1}{\sharp \mathcal{B}} \sum_{X_i \in \mathcal{G}} (1 - x_i) \tag{14}$$

where it is assumed that $x = 1$ and $\neg x = 0$. Thus, $E_\alpha(C_1)$ is simply the total effort spent on the strategy α, $E_\alpha(C_2)$ is the expected fraction of occurring events in $\mathcal{G}$, while $E_\alpha(C_3)$ is the expected fraction of not occurring events in $\mathcal{B}$. Table 1 reports all entities included in the model and their estimated characteristics, as well as the events' rating (i.e. positive and negative sets).

The randomly initialised GA population was composed of 200 chromosomes, each coding a strategy (i.e. the 9 effort values relative to the available actions). For each effort value a 12-bit string was employed.

The objective functions were evaluated performing a standard stochastic sampling procedure of the BN for each individual in the population. Given that the adopted GA was of elitist kind, the values of the objective function relative to the current Pareto set were conveniently stored from one generation into its successors (i.e. the BN inferences are not re-performed). In every generation, after the ranking, for each selected couple a one-site cross-over and subsequently a children mutation with probability $p_m = 0.003$ were applied. In this example, in order to explore the whole decision space, the effort constraints were not considered. The computation was simply terminated after 20 generations (the program allows a real-time monitoring of the Pareto-set evolution). Using a standard PC, less than ten minutes were required for the total computation.

Figure 5, representing the final non-dominated set, shows how the proposed

Table 1. The entity characteristics in the presented example (p_0 is the leak probability)

	Id	*Set*	*Description*	p_0
Endogenous Events	e_1	$\mathcal{G}$	Increase of tourist incomings	0.26
	e_2	$\mathcal{B}$	Population decrease, especially young people	0.35
	e_3		Increase of un-skilled immigration	0.12
	e_4	$\mathcal{G}$	High added-value economical activities	0.08
	e_5	$\mathcal{B}$	Increase of traffic congestion	0.42
	e_6	$\mathcal{G}$	Industrial development	0.41
	e_7	$\mathcal{G}$	Increase of real-estate values	0.40
	e_8		Population increase their demand for services	0.20
	e_9	$\mathcal{G}$	Agricultural development	0.50
	e_{10}	$\mathcal{B}$	Increase of unemployment	0.25
	Id		*Description*	*P*
Exogenous events	u_1		An exceptional flooding	0.06
	u_2		Increase of demand for agro-biological products	0.64
	u_3		High competition from emerging countries	0.39
	u_4		Euro:Dollar - 1:2	0.02
	u_5		Oil crisis	0.21
	Id		*Description*	
Actions	a_1		Better quality of the high-school education	
	a_2		Support for entrepreneurial start-ups	
	a_3		Better services for citizens	
	a_4		Investments in new technologies	
	a_5		Rehabilitation and re-qualification of the historical centre	
	a_6		Measures for environmental protection and territorial preservation	
	a_7		Extension and foundation of new universities	
	a_8		Improvement of collective transportation systems	
	a_9		Support for agricultural development	

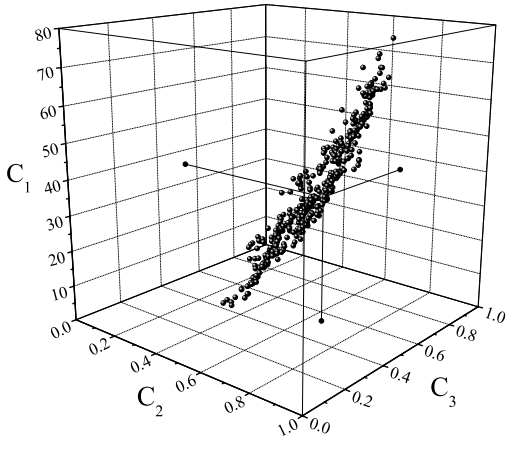


Fig. 5. The set of non-dominated solutions in the space of the objective functions

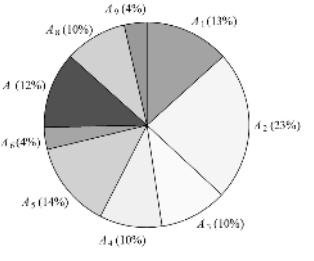


Fig. 6. Effort allocation corresponding to the solution highlighted in Figure 5

multi-objective approach allows the user to select a solution from a variety of possibilities. Clearly the final selection must be performed on the basis of some additional subjective decision. The selected strategy in our case, corresponding

to 272 effort units, is highlighted in the figure. In particular, Figure 6 reports the effort allocation suggested by the DSS analysis.

5 Conclusions and Future Works

We have presented a decision support system based on a simultaneous decision network coupled with a meta-heuristic search procedure. The latter, in the current version of the programme, is a multi-objective genetic algorithm, even if we are now trying different and more specific evolutionary algorithms that might lead to better overall behaviours. The DSS is purposeful for the solution of a problem often arising in strategic planning, and concerning the allocation of an available amount of resources among many actions. The formulation of the SDN is suitable when the model entities interact according to the noisy-OR model and it requires few parameters considering that actions are modelled as variables belonging to domains that might have high cardinality. The DSS can be used when the model can not be learnt from data and must be built by a panel of experts. In future we should explore better how the shape of strength functions can be obtained integrating data and experts' opinions. This work confirms that the multi-objective approach provides a number of advantages and, in particular, offers the decision-maker the choice of an adequate strategy, extending the knowledge about the variety of possibilities offered by the decision space. Besides, this approach allows the decision maker to be directly involved and in control of the process of evaluation of strategies.

References

1. Pearl, J.: Probabilistic reasoning in intelligent systems. Morgan Kaufman, San Mateo, CA (1988)
2. Díez, F.J.: Parameter adjustment in bayes networks. the generalized noisy or-gate. In: UAI. (1993) 99–105
3. Zhang, W., Ji, Q.: A factorization approach to evaluating simultaneous influence diagrams. (To appear in IEEE Trans. on Systems, Man and Cybernetics - A)
4. Diehl, M., Haimes, Y.Y.: Influence diagrams with multiple objectives and tradeoff analysis. IEEE Trans. on Systems, Man and Cybernetics - A **34** (2004) 293–304
5. Henrion, M.: Some practical issues in constructing belief networks. In: Uncertainty in Artificial Intelligence 3. Elsevier Science (1989) 161–173
6. Zagorecki, A., Druzdzel, M.J.: An empirical study of probability elicitation under noisy-or assumption. In: FLAIRS Conference. (2004)
7. de Campos, L.M., Gámez, J.A., Moral, S.: Partial abductive inference in bayesian belief networks using a genetic algorithm. Patt. Recogn. Lett. **20** (1999) 1211–1217
8. Sawaragi, Y., Nakayama, H., Tanino, T.: Theory of multiobjective optimization. Academic Press, Orlando, Florida (1985)
9. Goldberg, D.: Genetic Algorithms and Evolution Strategy in Engineering and Computer Science: Recent Advances and Industrial Applications. Wiley (1998)
10. Coello, C.C., Veldhuizen, D.V., Lamont, G.: Evolutionary Algorithms for Solving Multi-Objective Problems. Kluwer Academic Publishers (2002)

Evaluating Model Construction Methods with Objective Rule Evaluation Indices to Support Human Experts

Hidenao Abe[1], Shusaku Tsumoto[1], Miho Ohsaki[2], and Takahira Yamaguchi[3]

[1] Department of Medical Informatics, Shimane University, School of Medicine,
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
abe@med.shimane-u.ac.jp, tsumoto@computer.org
[2] Faculty of Engineering, Doshisha University,
1-3 Tataramiyakodani, Kyo-Tanabe, Kyoto 610-0321, Japan
mohsaki@mail.doshisha.ac.jp
[3] Faculty of Science and Technology, Keio University,
3-14-1 Hiyoshi, Kohoku Yokohama, Kanagawa 223-8522, Japan
yamaguti@ae.keio.ac.jp

Abstract. In this paper, we present a novel rule evaluation support method for post-processing of mined results with rule evaluation models based on objective indices. Post-processing of mined results is one of the key issues to make a data mining process successfully. However, it is difficult for human experts to evaluate many thousands of rules from a large dataset with noises completely. To reduce the costs of rule evaluation procedures, we have developed the rule evaluation support method with rule evaluation models, which are obtained with objective indices of mined classification rules and evaluations of a human expert for each rule. To evaluate performances of learning algorithms for constructing rule evaluation models, we have done a case study on the meningitis data mining as an actual problem. In addition, we have also evaluated our method on four rulesets from the four UCI datasets. Then we show the availability of our rule evaluation support method.

1 Introduction

In recent years, huge data are easily stored on information systems in natural science, social science and business domains, developing information technologies. With these huge data, people hope to utilize them for their purposes. Besides, data mining techniques have been widely known as a process for utilizing stored data on database systems, combining different kinds of technologies such as database technologies, statistical methods and machine learning methods. Especially, IF-THEN rules, which are produced by rule induction algorithms, are discussed as one of highly usable and readable output of data mining. However, to large dataset with hundreds attributes including noises, the process often obtains many thousands of rules. From such huge rule set, it is difficult for human experts to find out valuable knowledge which are rarely included in the rule set.

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 93–104, 2006.

To support such a rule selection, many efforts have done using objective rule evaluation indices such as recall, precision, and other interestingness measurements (we call them 'objective indices' later). However, it is also difficult to estimate a criterion of a human expert with single objective rule evaluation index, because his/her subjective criterion such as interestingness and importance for his/her purpose is influenced by the amount of his/her knowledge and/or a passage of time.

To above issues, we have been developed an adaptive rule evaluation support method for human experts with rule evaluation models, which predict experts' criteria based on objective indices, re-using results of evaluations of human experts. In Section 3, we describe the rule evaluation model construction method based on objective indices. Then we present a performance comparison of learning algorithms for constructing rule evaluation models in Section 4. With the results of the comparison, we discuss about the availability of our rule evaluation model construction approach.

2 Related Work

To avoid the confusion of real human interest, objective index, and subjective index, we clearly define them as follows: **Objective Index:** The feature such as the correctness, uniqueness, and strength of a rule, calculated by the mathematical analysis. It does not include any human evaluation criteria. **Subjective Index:** The similarity or difference between the information on interestingness given beforehand by a human expert and those obtained from a rule. Although it includes some human criterion in its initial state, the similarity or difference are mainly calculated with a mathematical analysis. **Real Human Interest:** The interest felt by a human expert for a rule in his/her mind.

Focusing on interesting rule selection with objective indexes, researchers have developed more than forty objective indexes based on number of instances, probability, statistics, information quantity, distance of rules or their attributes, and complexity of a rule [11, 22, 24]. Most of these indexes are used to remove meaningless rules rather than to discover really interesting ones for a human expert, because they can not include domain knowledge. In contrast, a dozen of subjective indexes estimate how a rule fits with a belief, a bias or a rule template formulated beforehand by a human expert. Although these subjective indexes are useful to discover really interesting rules to some extent due to their built-in domain knowledge, they depend on the precondition that a human expert is able to clearly formulate his/her interest. Although interestingness indexes were verified their availabilities on each suggested domain, nobody has validated their availabilities on the other domains or/and characteristics related to the background of a given dataset.

Ohsaki et. al [15] investigated the relation between objective indexes and real human interests, taking real data mining results and their human evaluations. In this work, the comparison shows that it is difficult to predict real human

interests with a single objective index. Based on the result, they indicated the possibility of logical combination of the objective indexes to predict real human interests more exactly.

3 Rule Evaluation Support with Rule Evaluation Model Based on Objective Indices

At practical data mining situations, costly rule evaluation procedures are repeatedly done by a human expert. In these situations, useful experiences of each evaluation such as focused attributes, interesting their combinations, and valuable facts are not explicitly used by any rule selection system, but tacitly stored in the human expert. To these problems, we suggest a method to construct rule evaluation models based on objective rule evaluation indices to describe a criterion of a human expert explicitly.

3.1 Constructing a Rule Evaluation Model

We considered the process for modeling rule evaluations of human experts as the process to clear up relationships between the human evaluations and features of input if-then rules. With this consideration, we decided that the process of rule evaluation model construction can be implemented as a learning task. Fig. 1. shows the process of rule evaluation model construction based on re-use of human evaluations and objective indices for each mined rule.

At the training phase, attributes of a meta-level training data set is obtained by objective indices such as recall, precision and other rule evaluation values. The human evaluations for each rule are joined as class of each instance. To obtain this data set, a human expert has to evaluate the whole or part of input rules at least once. After obtaining the training data set, its rule evaluation model is constructed by a learning algorithm. At the prediction phase, a human expert receives predictions for new rules based on their values of the objective indices. Since the task

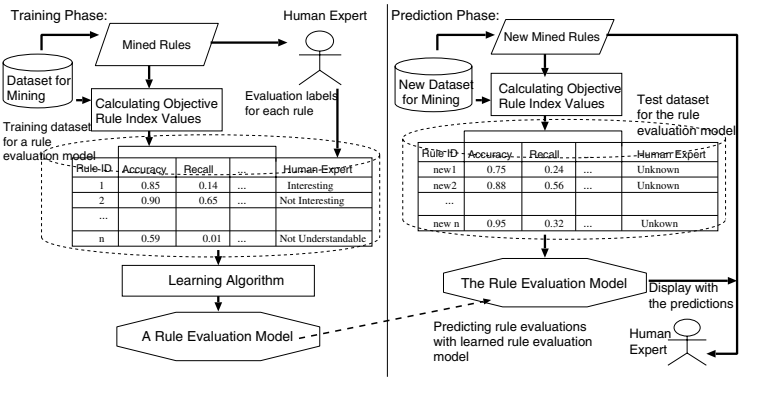


Fig. 1. Overview of the construction method of rule evaluation models

of rule evaluation models is a prediction, we need to choose a learning algorithm with higher accuracy as same as current classification problems.

4 Performance Comparison of Learning Algorithms for Rule Model Construction

To predict human evaluation labels of a new rule based on objective indices more exactly, we have to construct a rule evaluation model, which has higher predictive accuracy.

In this section, we firstly present the result of an empirical evaluation with the dataset from the result of a meningitis data mining[9]. Then to confirm the performance of our approach on the other datasets, we evaluated the five algorithms on four rule sets from four kinds of UCI benchmark datasets [10]. With the experimental results, we discuss about the following three view points: accuracies of rule evaluation models, analyzing learning curves of the learning algorithms, and contents of learned rule evaluation models.

As an evaluation of accuracies of rule evaluation models, we have compared predictive accuracies on the whole dataset and Leave-One-Out. The accuracy of a validation dataset D is calculated with correctly predicted instances $Correct(D)$ as $Acc(D) = (Correct(D)/|D|) \times 100$, where $|D|$ means the size of the dataset. Recalls of class i on a validation dataset is calculated with correctly predicted instances about the class $Correct(D_i)$ as $Recall(D_i) = (Correct(D_i)/|D_i|) \times 100$, where $|D_i|$ means the size of instances with class i. Also the precision of class i is calculated with the size of instances predicted i as $Precision(D_i) = (Correct(D_i)/Predicted(D_i)) \times 100$.

As for learning curves, we obtained learning curves about accuracies to the whole training dataset to evaluate whether each learning algorithm can perform in early stage of a process of rule evaluations. Accuracies from randomly sub-sampled training datasets are averaged with 10 times trials on each percentage of subset.

Looking at elements of the rule evaluation models to the whole dataset, we consider the characteristics of each learning algorithm on the attribute space consisted of the objective indices.

To construct a dataset to learn a rule evaluation model, values of objective indices have been calculated for each rule, taking 39 objective indices as shown in Table 1. The dataset for each rule set has the same number of instances as the rule set. Each instance consists of 40 attributes including the class attribute.

To these dataset, we applied five learning algorithms to compare their performance as a rule evaluation model construction method. We used the following learning algorithms from Weka[23]: C4.5 decision tree learner[19] called J4.8, neural network learner with back propagation (BPNN)[12], support vector machines (SVM)[1][18], classification via linear regressions (CLR)[2][3], and OneR[13].

[1] The kernel function was set up polynomial kernel.

[2] We set up the elimination of collinear attributes and the model selection with greedy search based on Akaike Information Metric.

Table 1. The objective rule evaluation indices for classification rules used in this research. **P:** Probability of the antecedent and/or consequent of a rule. **S:** Statistical variable based on P. **I:** Information of the antecedent and/or consequent of a rule. **N:** Number of instances included in the antecedent and/or consequent of a rule. **D:** Distance of a rule from the others based on rule attributes.

Theory	Index Name (**Abbreviation**) [Reference Number of Literature]
P	Coverrage(**Coverage**), Prevalence(**Prevalence**) Precision(**Precision**), Recall(**Recall**) Suppurt(**Support**), Specificity(**Specificity**) Accuracy(**Accuracy**), Lift(**Lift**) Leverage(**Leverage**), Added Value(**Added Value**)[22] Klösgen's Interestingness(**KI**)[14], Relative Risk(**RR**)[1] Brin's Interest(**BI**)[2], Brin's Conviction(**BC**)[2] Certainty Factor(**CF**)[22], Jaccard Coefficient(**Jaccard**)[22] F-Measure(**F-M**)[20], Odds Ratio(**OR**)[22] Yule's Q(**YuleQ**)[22], Yule's Y(**YuleY**)[22] Kappa(**Kappa**)[22], Collective Strength(**CST**)[22] Gray andOrlowska's Interestingness weighting Dependency(**GOI**)[7] Gini Gain(**Gini**)[22], Credibility(**Credibility**)[8]
S	χ^2 Measure for One Quadrant(χ^2**-M1**)[6] χ^2 Measure for Four Quadrant(χ^2**-M4**)[6]
I	J-Measure(**J-M**)[21], K-Measure(**K-M**)[15] Mutual Information(**MI**)[22] Yao and Liu's Interestingness 1 based on one-way support(**YLI1**)[24] Yao and Liu's Interestingness 2 based on two-way support(**YLI2**)[24] Yao and Zhong's Interestingness(**YZI**)[24]
N	Cosine Similarity(**CSI**)[22], Laplace Correction(**LC**)[22] ϕ Coefficient(ϕ)[22], Piatetsky-Shapiro's Interestingness(**PSI**)[17]
D	Gago and Bento's Interestingness(**GBI**)[5] Peculiarity(**Peculiarity**)[25]

4.1 Constructing Rule Evaluation Models on an Actual Datamining Result

In this case study, we have taken 244 rules, which are mined from six dataset about six kinds of diagnostic problems as shown in Table 2. These datasets are consisted of appearances of meningitis patients as attributes and diagnoses for each patient as class. Each rule set was mined with each proper rule induction algorithm composed by a constructive meta-learning system called CAMLET[9]. For each rule, we labeled three evaluations (I:Interesting, NI:Not-Interesting, NU:Not-Understandable), according to evaluation comments from a medical expert.

Table 2. Description of the meningitis datasets and their datamining results

Dataset	#Attributes	#Class	#Mined rules	#'I' rules	#'NI' rules	#'NU' rules
Diag	29	6	53	15	38	0
C_Cource	40	12	22	3	18	1
Culture+diag	31	12	57	7	48	2
Diag2	29	2	35	8	27	0
Course	40	2	53	12	38	3
Cult_find	29	2	24	3	18	3
TOTAL	—	—	244	48	187	9

Comparison on Classification Performance. In this section, we show the result of the comparisons of accuracies on the whole dataset, recall of each class label, and precisions of each class label. Since Leave-One-Out holds just one instance as the test data and remains as the training dataset repeatedly for each instance of a given dataset, we can evaluate the performance of a learning algorithm to a new dataset without any ambiguity.

The results of the performances of the five learning algorithms to the whole training dataset and the results of Leave-One-Out are also shown in Table 3 All of the accuracies, Recalls of I and NI, and Precisions of I and NI are higher than predicting default labels.

Table 3. Accuracies(%), Recalls(%) and Precisions(%) of the five learning algorithms

	On the whole training dataset							Leave-One-Out						
		Recall of			Precision of				Recall of			Precision of		
	Acc.	I	NI	NU	I	NI	NU	Acc.	I	NI	NU	I	NI	NU
J4.8	85.7	41.7	97.9	66.7	80.0	86.3	85.7	79.1	29.2	95.7	0.0	63.6	82.5	0.0
BPNN	86.9	81.3	89.8	55.6	65.0	94.9	71.4	77.5	39.6	90.9	0.0	50.0	85.9	0.0
SVM	81.6	35.4	97.3	0.0	68.0	83.5	0.0	81.6	35.4	97.3	0.0	68.0	83.5	0.0
CLR	82.8	41.7	97.3	0.0	71.4	84.3	0.0	80.3	35.4	95.7	0.0	60.7	82.9	0.0
OneR	82.0	56.3	92.5	0.0	57.4	87.8	0.0	75.8	27.1	92.0	0.0	37.1	82.3	0.0

Comparing with the accuracy of OneR, the other learning algorithms achieve equal or higher performance with combination of multiple objective indices than sorting with single objective index. Looking at Recall values on class I, BPNN have achieved the highest performance. As for the other algorithms, they show lower performance than OneR, because they have tended to be learned classification patterns for the major class NI.

Each value of Leave-One-Out estimation shows robustness of each learning algorithm to an unknown test dataset. On the accuracies, these learning algorithms have achieved from 75.8% to 81.9%. However, these learning algorithms have not been able to classify the instances with class NU, which is a minor class label in this dataset.

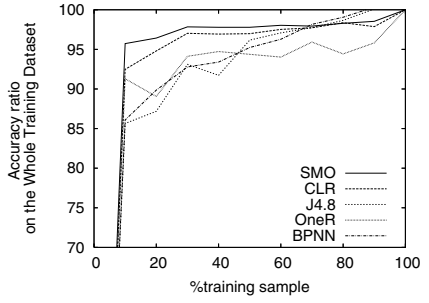
Looking at each learning algorithm, the values of BPNN show the trend of over fitting, comparing with its values of training dataset and its values of Leave-One-Out. Although OneR selects an adequate objective index to sort and classify 244 training datasets in the Leave-One-Out validation, the predictive performances to a new dataset have been limited because of the selection of just one objective index.

Learning Curves of the Learning Algorithms. Since the rule evaluation model construction method needs evaluations of mined rules by a human expert, we have investigated learning curves of each learning algorithm to estimate how many evaluations are needed to construct a valid rule evaluation model. The upper table in Fig. 2. shows accuracies to the whole training dataset with each subset of training dataset. The percentages of achievements for each learning

algorithm, comparing with the accuracy with the whole dataset, are shown in the lower chart of Fig. 2. As shown in these results, SVM and CLR, which learn hype-planes, achieves grater than 95% with only less than 10% of training subset. Although decision tree learner and BPNN could learn better classifier to the whole dataset than these hyper-plane learners, they need more training instances to learn accurate classifiers.

%training sample	10	20	30	40	50	60	70	80	90	100
J4.8	73.4	74.7	79.8	78.6	72.8	83.2	83.7	84.5	85.7	85.7
BPNN	74.8	78.1	80.6	81.1	82.7	83.7	85.3	86.1	87.2	86.9
SMO	78.1	78.6	79.8	79.8	79.8	80.0	79.9	80.2	80.4	81.6
CLR	76.6	78.5	80.3	80.2	80.3	80.7	80.9	81.4	81.0	82.8
OneR	75.2	73.4	77.5	78.0	77.7	77.5	79.0	77.8	78.9	82.4

%training sample	10	20	30	40	50	60	70	80	90	100
J4.8	82.7	85.3	92.8	88.7	93.2	92.7	93.2	92.9	94.0	97.9
BPNN	84.6	86.6	90.4	90.2	92.2	91.9	92.7	93.9	94.2	89.8
SMO	93.3	92.7	96.8	96.1	95.9	95.8	96.3	96.0	96.3	97.3
CLR	88.3	89.6	94.4	94.0	94.3	94.1	94.1	94.2	94.3	97.3
OneR	88.4	84.0	92.4	91.4	92.0	92.3	93.4	92.7	92.1	96.3



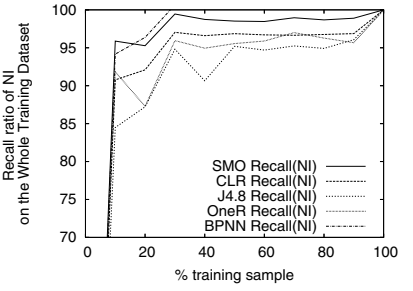


Fig. 2. Learning Curves of accuracies(%) on the learning algorithms with sub-sampled training dataset: The left table shows accuracies(%) on each training dataset to the whole dataset. The left graph shows their achievement ratio(%). Also the right table shows recalls(%), and the graph shows their achievement ratio(%).

To eliminate known ordinary knowledge from large rule set, it is needed to classify non-interesting rules correctly. The right upper table in Fig. 2. shows percentages of recalls on NI. The right lower chart in Fig. 2. also shows the percentages of achievements on recall of NI, comparing with the recall of NI on the whole training dataset. Looking at this result, we can eliminate NI rules with rule evaluation models from SVM and BPNN even if there is only 10% of rule evaluations by a human expert. This is guaranteed with no less than 80% precisions of all learning algorithms.

Rule Evaluation Models on the Actual Datamining Result Dataset. In this section, we present rule evaluation models to the whole dataset learned with OneR, J4.8 and CLR, because they are represented as explicit models such as a rule set, a decision tree, and a set of linear models.

The rule set of OneR is shown in Fig. 3(a). OneR has selected YLI1[24] to classify the evaluation labels. Although YLI1 corrects support to predict interestingness of a human expert, YLI1 estimates a correctness of each rule on a validation dataset.

As shown in Fig. 3(b), J4.8 leaned the decision tree. At the root node, this model takes Laplace Correction[22], which is a corrected Precision with constant

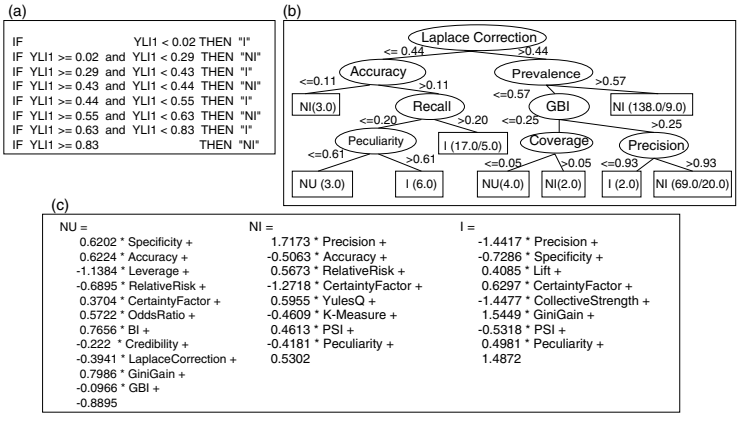


Fig. 3. Learned models to the meningitis data mining result dataset: (a) rule set learned from OneR, (b) decision tree learned from J4.8, (c) linear regression models learned from CLR

values. At the other levels, it takes indices evaluating a correctness of a rule such as Accuracy, Precision and Recall. Coverage and Prevalence are indices to evaluate a generality of the antecedent and the consequent of a rule. GBI[5] calculate index values with the classification result of a rule. Peculiarity[25] sums up differences of antecedents between one rule and the other rules in the same rule set.

Fig. 3(c) shows linear models to classify each class. The prediction has done with integrating the responses of these linear models. As for models to class NI and I, they have the same indices such as Precision, Certainty Factor, PSI, and Peculiarity with opposite coefficients. The strongest factors on these models are Precision and Gini Gain, which increase their values with the correctness of a rule. To class NU, the strongest factor is Leverage based on Precision with a correction using a generality of a rule.

4.2 Constructing Rule Evaluation Models on Artificial Evaluation Labels

We have also evaluated our rule evaluation model construction method with rule sets from three datasets of UCI Machine Learning Repository to confirm the lower limit performances on probabilistic class distributions.

We selected the following three datasets: Mushroom, Heart, Internet Advertisement Identification (called InternetAd later) and Letter. With these datasets, we obtained rule sets with bagged PART, which repeatedly executes PART[4] to bootstrapped training sub-sample datasets.

To these rule sets, we calculated the 39 objective indices as attributes of each rule. As for the class of these datasets, we set up three class distributions with multinomial distribution. The class distribution for 'Distribution I' is $P = (0.35, 0.3, 0.3)$ where p_i is the probability for class i. Thus the num-

ber of class i in each instance D_j become $p_i D_j$. As the same way, the probability vector of 'Distribution II' is $P = (0.3, 0.5, 0.2)$, and 'Distribution III' is $P = (0.3, 0.65, 0.05)$.

Table 4 shows us the datasets with three different class distributions.

Table 4. Datasets of the rule sets learned from the UCI benchmark datasets

	#Mined Rules	#Class labels			%Def. class
		L1	L2	L3	
Distribution I		(0.30)	(0.35)	(0.35)	
Mushroom	30	8	14	8	46.7
InternetAd	107	26	39	42	39.3
Heart	318	97	128	93	40.3
Letter	6340	1908	2163	2269	35.8
Distribution II		(0.30)	(0.50)	(0.20)	
Mushroom	30	11	16	3	53.3
InternetAd	107	30	53	24	49.5
Heart	318	99	140	79	44.0
Letter	6340	1890	3198	1252	50.4
Distribution III		(0.30)	(0.65)	(0.05)	
Mushroom	30	7	21	2	70.0
InternetAd	107	24	79	9	73.8
Heart	318	98	205	15	64.5
Letter	6340	1947	4062	331	64.1

Accuracy Comparison on Classification Performances. To above datasets, we have attempted the five learning algorithms to estimate whether their classification results can go to or beyond the accuracies of just predicting each default class. The left table of Table 5 shows the accuracies of the five learning algorithms to each class distribution of the three datasets. As shown in Table 5, J48 and BPNN always work better than just predicting a default class. However, their performances are suffered from probabilistic class distributions to larger datasets such as Heart and Letter.

Table 5. Accuracies(%) on whole training datasets labeled with three different distributions(The left table). Number of minimum training sub-samples to outperform %Def. class(The right table).

	J48	BPNN	SVM	CLR	OneR
Distribution I					
Mushroom	80.0	93.3	56.7	66.7	53.3
InternetAd	84.1	82.2	29.9	53.3	60.7
Heart	78.0	75.8	40.3	42.5	54.7
Letter	36.8	36.4	30.1	36.6	52.1
Distribution II					
Mushroom	93.3	93.3	80.0	80.0	76.7
InternetAd	73.8	79.4	49.5	59.8	60.7
Heart	72.3	69.2	35.9	47.8	55.7
Letter	51.0	51.0	50.4	50.4	57.0
Distribution III					
Mushroom	93.3	96.7	70.0	70.0	76.7
InternetAd	86.0	90.7	70.1	69.2	72.0
Heart	78.0	77.7	64.5	65.7	71.4
Letter	64.1	64.3	64.1	64.1	68.3

	J48	BPNN	SVM	CLR	OneR
Distribution I					
Mushroom	8	8	12	18	14
InternetAd	14	14	–	30	14
Heart	42	31	66	114	98
Letter	189	217	–	955	305
Distribution II					
Mushroom	6	4	4	6	12
InternetAd	24	24	52	42	70
Heart	52	40	–	104	92
Letter	897	>1000	451	–	>1000
Distribution III					
Mushroom	22	14	22	28	22
InternetAd	80	66	–	–	–
Heart	114	94	142	318	182
Letter	>1000	>1000	998	>1000	>1000

Evaluation on Learning Curves. As same as evaluations of learning curves on the meningitis rule set, we have estimated the minimum training subsets for a valid model, which works better than just predicting a default class. For each data point, we constructed rule evaluation models to each size of subsampled training datasets 10 times. Then the averaged accuracy of each set of rule evaluation models is calculated on each whole dataset.

The right table in Table 5 shows sizes of minimum training subsets, which can be constructed more accurate rule evaluation models than percentages of a default class by each learning algorithm. To smaller dataset, such as Mushroom and InternetAd, they can construct valid models with less than 20% of given training datasets. However, to larger dataset, they need more training subsets to construct valid models, because their performances with whole training dataset fall to the percentages of default class of each dataset as shown in the left table in Table 5.

4.3 Discussion

On the Classification Performances. As shown in Table 3 and the left table of Table 5, J4.8 decision tree learner and BPNN neural network learner work better than the other algorithms on both of the actual problem and a probabilistic problem. In section 4.1, the classification result about class 'I' indicates that these instances are difficult to separate with liner expressions in this attribute space based on the 39 objective indices. To predict such labels correctly, we should apply nonlinear classifier learned from nonlinear learners. Although these five learning algorithms have achieved 81.6% of the highest accuracy in the Leave-One-out estimation, we need to obtain more accurate rule evaluation models with meta-learning algorithms such as boosting, bagging and so forth.

On the Learning Curves. With this analysis of the learning curves about each amount of training samples, we consider the following guideline: At early stage of rule evaluation support, the system should select hyper-plane learners to construct better rule evaluation models rapidly. Then closing stage of evaluations, the system should select more accurate learning algorithm to predict minor but valuable rules. This guideline can be applied to a large rule set, considering the work done by Perlich et. al[16], which shows the result that regression learners can be learned faster on large datasets than decision tree learners.

On the Learned Rule Evaluation Models. Looking at indices used in learned rule evaluation models, they are not only the group of indices increasing with a correctness of a rule, but also they are used some different groups of indices on different models. This corresponds to the comment from the human expert. He said that he evaluated these rules not only correctness but also his interest based on his expertise. From the other viewpoint, this also indicates that the rule model construction method needs to select prior algorithms on data pre-processing algorithms, such as attribute construction and attribute selection, and a mining algorithm to construct an adequate rule evaluation model.

5 Conclusion

In this paper, we have described rule evaluation support method with rule evaluation models to predict evaluations for an IF-THEN rule based on objective indices, re-using evaluations of a human expert.

As the result of the performance comparison with the five learning algorithms, rule evaluation models have achieved higher accuracies than just predicting each default class. Considering the difference between the actual evaluation labeling and the artificial evaluation labeling, it is shown that the medical expert evaluated with noticing particular relations between an antecedent and a class/another antecedent in each rule. In the estimation of robustness to a new rule with Leave-One-Out, we have achieved more than 75.8% with these learning algorithms. On the evaluation with learning curves to the dataset of the actual datamining result, SVM and CLR have achieved more than 95% of achievement ratio compared to the accuracy of the whole training dataset with less than 10% of subset of the training dataset with certain human evaluations. These result related to performances of rule evaluation models indicate the availability of this rule evaluation support method for a human expert.

As future works, we will introduce a selection method of learning algorithms to construct a proper rule evaluation model according to each situation. We also apply this rule evaluation support method to estimate other data mining result such as decision tree, rule set, and committee of them with objective indices, which evaluate whole mining results.

References

1. Ali, K., Manganaris,S., and Srikant, R.: Partial Classification Using Association Rules. Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD-1997 (1997) 115–118
2. Brin, S., Motwani, R., Ullman, J., and Tsur, S.: Dynamic itemset counting and implication rules for market basket data. Proc. of ACM SIGMOD Int. Conf. on Management of Data (1997) 255–264
3. Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. H.: Using model trees for classification, Machine Learning, Vol.32, No.1 (1998) 63–76
4. Frank, E, Witten, I. H., Generating accurate rule sets without global optimization, in Proc. of the Fifteenth International Conference on Machine Learning, (1998) 144–151
5. Gago, P. and Bento, C.: A Metric for Selection of the Most Promising Rules. Proc. of Euro. Conf. on the Principles of Data Mining and Knowledge Discovery PKDD-1998 (1998) 19–27
6. Goodman, L. A. and Kruskal, W. H.: Measures of association for cross classifications. Springer Series in Statistics, 1, Springer-Verlag (1979)
7. Gray, B. and Orlowska, M. E.: CCAIIA: Clustering Categorical Attributes into Interesting Association Rules. Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD-1998 (1998) 132–143
8. Hamilton, H. J., Shan, N., and Ziarko, W.: Machine Learning of Credible Classifications. Proc. of Australian Conf. on Artificial Intelligence AI-1997 (1997) 330–339

9. Hatazawa, H., Negishi, N., Suyama, A, Tsumoto, S., and Yamaguchi, T.: Knowledge Discovery Support from a Meningoencephalitis Database Using an Automatic Composition Tool for Inductive Applications, in Proc. of KDD Challenge 2000 in conjunction with PAKDD2000 (2000) 28–33
10. Hettich, S., Blake, C. L., and Merz, C. J.: UCI Repository of machine learning databases [http://www.ics.uci.edu/˜mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science, (1998).
11. Hilderman, R. J. and Hamilton, H. J.: Knowledge Discovery and Measure of Interest, Kluwe Academic Publishers (2001)
12. Hinton, G. E.: "Learning distributed representations of concepts", Proc. of the Eighth Annual Conference of the Cognitive Science Society, Amherest, MA. REprinted in R.G.M.Morris (ed.) (1986)
13. Holte, R. C.: Very simple classification rules perform well on most commonly used datasets, Machine Learning, Vol. 11 (1993) 63–91
14. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. in Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy R. (Eds.): Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, California (1996) 249–271
15. Ohsaki, M., Kitaguchi, S., Kume, S., Yokoi, H., and Yamaguchi, T.: Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis, in Proc. of ECML/PKDD 2004, LNAI3202 (2004) 362–373
16. Perlich, C., Provost, F. J., and Simonoff, J. S.: Tree Induction vs. Logistic Regression: A Learning-Curve Analysis, Journal of Machine Learning Research, Vol. 4 (2003) 211–255
17. Piatetsky-Shapiro, G.: Discovery, Analysis and Presentation of Strong Rules. in Piatetsky-Shapiro, G., Frawley, W. J. (eds.): Knowledge Discovery in Databases. AAAI/MIT Press (1991) 229–248
18. Platt, J.: Fast Training of Support Vector Machines using Sequential Minimal Optimization, Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press (1999) 185–208
19. Quinlan, R.: C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, (1993)
20. Rijsbergen, C.: Information Retrieval, Chapter 7, Butterworths, London, (1979) http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html
21. Smyth, P. and Goodman, R. M.: Rule Induction using Information Theory. in Piatetsky-Shapiro, G., Frawley, W. J. (eds.): Knowledge Discovery in Databases. AAAI/MIT Press (1991) 159–176
22. Tan, P. N., Kumar V., and Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns. Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD-2002 (2002) 32–41
23. Witten, I. H and Frank, E.: DataMining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, (2000)
24. Yao, Y. Y. and Zhong, N.: An Analysis of Quantitative Measures Associated with Rules. Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD-1999 (1999) 479–488
25. Zhong, N., Yao, Y. Y., and Ohshima, M.: Peculiarity Oriented Multi-Database Mining. IEEE Trans. on Knowledge and Data Engineering, 15, 4, (2003) 952–960

A Multicriteria Fuzzy Decision System to Sort Contaminated Soils

M. García[1], E. López[1], V. Kumar[1], and A. Valls[2]

[1] Department of Chemical Engineering,
Escola Tècnica Superior d'Enginyeria Química, University Rovira i Virgili,
Campus Sescelades, Av. Paisos catalans, 26, 43007 Tarragona
miriam.garcia@urv.net
[2] Department of Computer Science and Mathematics,
Escola Tècnica Superior d'Enginyeria, University Rovira i Virgili,
Campus Sescelades, Av. Paisos catalans, 26, 43007 Tarragona
aida.valls@urv.net

Abstract. The poor characterisation of contaminated soils is likely to result in high costs, restricted choice in landfill disposal sites and future environmental impact. This makes that big quantities of soils are still waiting to be recovered. Until now the tools used to detect contaminated soil have been very generic, without any criteria of prioritization. Usually, simulation studies are used to classify contaminated soils, however these systems need a large quantity of data that is difficult to obtain and manage, which means that the results obtained are subject to large uncertainty. Recently, Artificial Intelligence techniques have been used to tackle this problem. In this work we propose the use of Fuzzy Expert Systems to classify the soils. Classical decision rules have shown to be interpretable, efficient, problem independent and able to treat large scale applications, but they are also recognised as highly unstable classifiers with respect to minor perturbations in the training data. In our problem, the data is subject to uncertainty, for this reason we propose the fuzzyfication of the variables. In the study many different variables have been taken into account: Physic and Chemical characteristics of the soil and pollutants, toxicological properties, and environment and social conditions. After applying Fuzzy Expert Systems at different levels, we obtain a ranking of the soils according to their risk of contamination. The results have been contrasted with another Multicriteria Decision Making system.

1 Introduction

Soil is a natural resource, not always renewable. Due to the deep aggressions carried out in the last years, the European Community considered to establish a Protecting Soil Directive that integrates the soil into the protected natural resources group, as water, atmosphere and biota (CEC 2002) [1]. The conditions of the soil have direct consequences in groundwater and water resources health [1]. Water quality is related to the quality of the soil, because soil performs storage, filtering, buffering and transformation functions. Between 1995 and 2005, around 4500 areas in Spain were classified as contaminated soils by the National Plan of Contaminated Soil Recovering.

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 105 – 116, 2006.

Before 1998, there was no legal rule to protect efficiently the soil from the contamination nor to detect and priorise the contaminated ones. The experience in other countries of E.U. helped to understand that huge areas of contaminated sites were still waiting to be characterized and recovered.

Until now, the classification methods used to evaluate contaminated soils only compare analytical results with the maximum concentration of contaminant established by the government. No recovery action is taken if the analytical results are under the established limit. However, many other factors influence the risk of a particular soil. Contaminated sites can vary widely in terms of their hydraulic conditions, the nature of the contamination, and the associated physical (e.g. mass transfer between different phases), chemical (e.g. oxidation and reduction), and biological (e.g. aerobic biodegradation) processes [2]. Therefore all the parameters involved in the processes described above should be also taken into account. Moreover, these data is sometimes very difficult to obtain and manage, leading to results subject to a great uncertainty. For this reason, there is a need of developing new computational technologies to optimise time and costs with respect to detection and recovery of contaminated sites.

Intelligent Decision Support Systems (DSSs) intend to rationally analyse complex sets of alternatives in order to find the solution to a decision problem. Some attempts to use DSSs for the characterization of contaminated soils have been done [2], but they do not use all the variables of the problem and they do not consider uncertainty in the data, leading to results that are not good enough. For this reason, a new approach is needed for developing a useful and user-friendly system which can be really adopted [3].

An expert system is an intelligent computer program using knowledge and inference procedures to emulate the decision-making ability of a human expert in a restricted domain [4]. They have been successfully applied in many areas such as business, medicine, science, and engineering. The basic idea of a knowledge-based expert system consists in supplying facts to the expert system, and receiving expert advice in response. The rules used in the inference stage can be automatically generated from decision trees [5]. Decision Trees have already shown to be easy to interpret, efficient, problem independent and able to treat large scale applications. But they are also recognised as highly unstable classifiers with attributes presenting high uncertainty in terms of variance. To cope with uncertainty, fuzzy decision trees have been developed [6]. Recently, different shells for implementing fuzzy expert systems have appeared (FuzzyCLIPS [7]. FuzzyShell [8], FuzzyJess [9]), so this kind of applications are growing [FS].

In our domain, we have lots of attributes to analyse in order to obtain a sorting of a set of soils into some predefined risk categories. In addition, many of these soil attributes are subjective measures, with some intrinsic uncertainty. For this reason, we propose, in this paper, the use of fuzzy sets to represent uncertainty and of fuzzy expert systems to solve the problem of soils classification and their priorisation for later recovery.

This paper proceeds as follows. Section 2 presents the characteristics of the problem we are trying to solve. Section 3 shows the architecture of the decision system we have designed and implemented, which is based on using fuzzy expert systems. Section 4 gives details about the modeling of the information with decision trees

made by a committee of experts in soil contamination, and explains the exploitation of the system. Then, section 5 is devoted to explain the tests of the system and the evaluation of the results in comparison with other approaches. Finally, section 6 presents some conclusions and new research directions.

2 Problem Domain

Many factors are relevant in the characterization of the risk of contamination in a soil, such as hazard assessment, toxicities and exposure assessment. To reduce the amount fo information, we performed a statistical analysis of the data to find the most relevant variables. Finally, 26 attributes were selected to build the knowledge base of a fuzzy expert system. This system will be used as a sorting method to determine the contaminated soil risk.

The information has been obtained from an expert commission, formed by a chemical engineer and an expert in contamination of soils, and bibliography (basically the Catalan Risk Evaluation Guide). For some months, a knowledge engineer had some interviews with the experts. In the first interview introduced and explained the concepts and tasks of the proposed methodology to tackle this problem. In the later interviews, when the knowledge engineer had gained some familiarity with the domain, they structured the information and modeled it in the form of decision trees, as will be explained in the following sections.

The experts decided that the rules should not consider together some variables, because they have different meaning and a different relation to the risk of contamination. For this reason, they put the variables into 3 types (see Table 1):

Table 1. Variables considered to build the expert system

Risk characterization variables		
Group nº 1: Source	***Group nº 2: Transfer vector***	***Group nº 3: Local properties***
Land dimension	*Porosity*	*Population*
Confinement	*Hydraulic conductivity*	*Population distance*
State of the contaminant	*Unsaturated zone deep*	*Population quantity*
Toxicity	*Recharge*	*Accessibility*
Contaminant Quantity	*Rainfall*	*Sensible activities*
Half life	*Groundwater proximity*	*Sensible ecosystem*
Solubility	*Surface water proximity*	*Species in danger*
Vapour pressure	*Soil used*	*Visual impact*
Installation age	*Water used*	

1. *Source attributes*: give details about the source of contamination.
2. *Transfer vector attributes*: refer to the possibility of dissemination of the contamination.
3. *Local properties*: are related to the possibly affect the health of human being and ecosystems.

For our environmental team of experts, one of the goals of this work is to study if we can determine the risk of contamination of a soil with less variables than it is usually done. So, we made a careful selection of the significant variables, and the structure of the expert system was widely discussed, before deciding to separate the variables. For each of this three groups we have build a separate expert system, that concludes which is the degree of risk of contamination in relation to these different points of view of the domain. Then, another expert system has rules that combine these three different measures of the risk. This approach with separate sets is a new proposal with respect to other previous approximations to this problem.

2.1 Fuzzyfying the Attributes

Environment decision-making involves complex and often ill-defined variables with a high degree of uncertainty due to incomplete understanding of the underlying issues [10]. These uncertainties can arise due to probability of an event, an expert's belief in an event, imprecision in real word measurements an experimenter's unclear notion of a value.

Fuzzy logic is useful for representing imprecise knowledge. Fuzzy expert systems show exceptional performance for working with processes which are adequately defined in qualitative terms and for which no precise mathematical model of the process exists [11]. So, we have taken this approach to design our system.

The group of experts in soils contamination analysis helped us to determine which attributes should use linguistic values rather than numerical ones, and also to design the membership functions associated to these linguistic values.

In this application, 16 attributes out of the 26 have been fuzzyfied. For almost all, we use three terms: low, medium and high risk. We have used trapezoidal sets to model their membership functions. The membership functions for each linguistic variable have been set up independently, to model the semantics of each attribute.

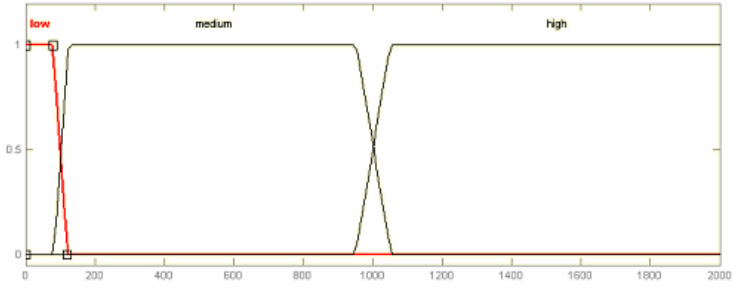


Fig. 1. Land dimension: surface in contact with contaminant (m^2)

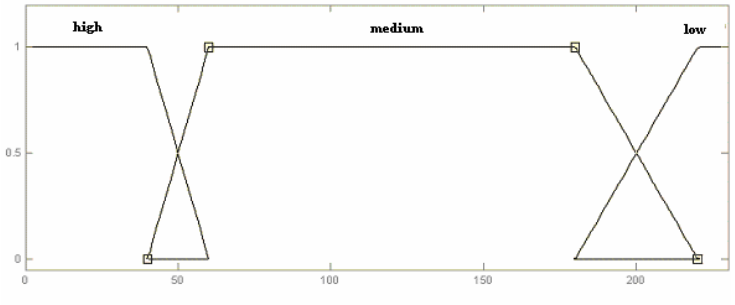


Fig. 2. Groundwater proximity: proximity of groundwater to contaminants (m)

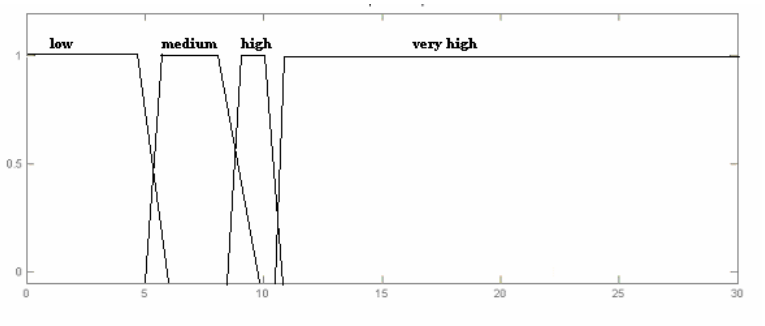


Fig. 3. Fuzzy values of the risk of contamination of a soil

Figures 1and 2 show the membership functions of the linguistic terms of the risk given the land dimension and groundwater proximity to the source of contaminant.

In the same way, for the decision variable (the risk of soil contamination) four linguistic values were chosen and their corresponding trapezoidal fuzzy sets fixed. This information is given in Figure 3.

3 The Expert System Architecture

In [2] a fuzzy expert system is designed to solve a problem quite similar to the one we are presenting here. The main differences that our proposal introduces are the following: (1) the set of attributes considered are not the same, (2) we organise the attributes in 3 sets and build an expert system for each one, due to that they refer to different aspects of the problem and seem to contribute differently to the final level of risk contamination. The results are used in the rules of another expert system to find the final risk of contamination. So, we are proposing to use a hierarchic fuzzy expert systems approach [12].

In Figure 4, the architecture of the hierarchic expert system is represented. It has two levels..Expert system SOURCE determines the risk of contamination having into consideration the attributes related to the source. Expert system TRANSFER uses the transfer vectors to know the risk of contamination. Expert system LOCAL-PROPERTIES

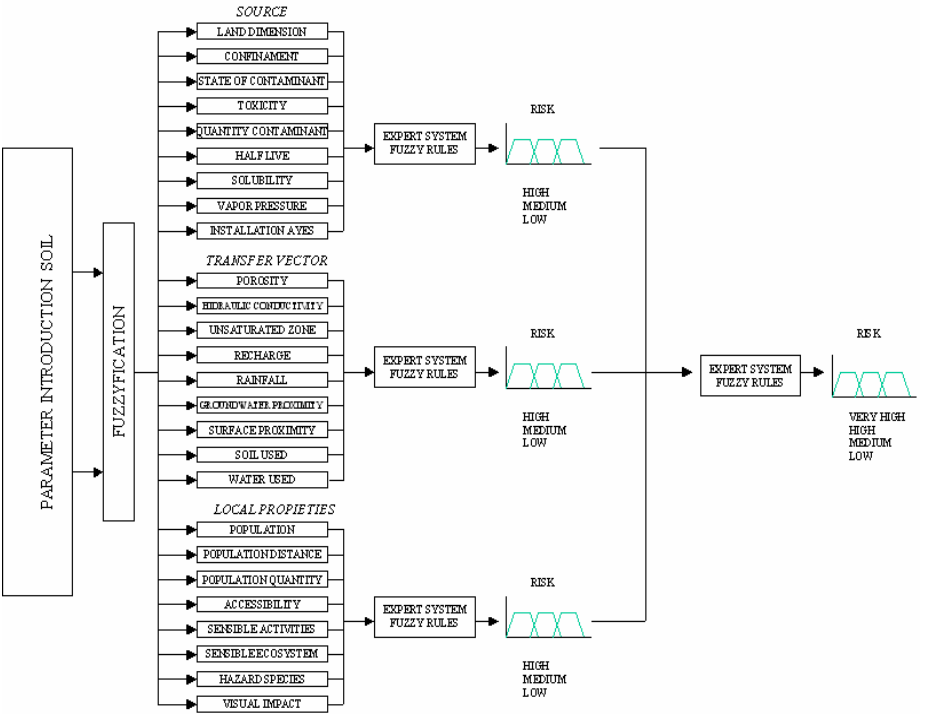


Fig. 4. Architecture of the decision making system

characterizes the risk in function of the local properties. We obtain a fuzzy degree of risk for each of these three aspects. Then, another expert system is used to integrate that values and determine the global risk of contamination.

4 Building the Expert Systems

This section is gives details about the design and development of the system we propose for this particular problem.

4.1 Modeling the Information with Decision Trees

Decision trees have been widely used in Artificial Intelligence applications. The knowledge is represented by nodes (corresponding to a particular attribute), branches (corresponding to the possible values of the attribute) and leaves (corresponding to decision values). So, decisions take part in the nodes, branches represent different ways to follow according to the value of the decision attribute, and leaves give the result of the problem.

Decision trees can be induced from a set of supervised examples using, for instance, the well-known ID3 algorithm [5]. Another way of obtaining a decision tree is building it with the help of an expert. We have used the latter approach because of the lack of a sufficient and significant set of supervised contaminated soils.

To build each of the four decision trees involved in our system, the knowledge engineer used a data-driven analysis, instead of the goal-driven approach proposed in [3]. The experts committee analyzed separately each of the three sets of variables to build a decision tree for each of them. The result of each tree is the risk of contamination associated to the source contaminant, the transfer vector and local properties, respectively. In addition, each attribute was given a weight according to its estimated contribution to the risk of contamination. Then, the new three fuzzy variables obtained from the three decision trees, were used to build the decision tree that integrates the partial results in order to find the global risk associated to a particular soil contaminated. In this case, the risk value of the three sets also contributes differently to the final level of risk contamination. The values of these weights were also given by the experts, who set that source is more important for risk assessment, so it takes a weight of 0.4, and transfer vector and local properties takes 0.3 each one.

4.2 Obtaining the Fuzzy Rules

To build the rule base of the expert systems, rules have been derived from the decision trees. Each path in the decision tree (form the root to the leave) corresponds to a rule. In Figure 6 we can see an example with a tree with two decision levels before determining the risk. From this tree we obtain the following rules:

R1: If toxicity is low then risk is low.
R2: If the toxicity is medium and the quantity of contaminant is low then the risk is low.
R3: If the toxicity is medium and the quantity of contaminant is medium then the risk is low.
R4: If the toxicity is medium and the quantity of contaminant is high then the risk is medium.
R5: If the toxicity is high and the quantity of contaminant is low then the risk is low.
R6: If the toxicity is high and the quantity of contaminant is medium then the risk is medium.
R7: If the toxicity is high and the quantity of contaminant is high then the risk is high.

To execute the expert system, the decision-maker provides the value of each of the attributes. Then the system selects the rules whose preconditions are fulfilled at some degree, and all these rules are fired to give a degree of risk of contamination.

To calculate the level of fulfillment of the conditions of a rule, we take the membership degree of thee values of each variable, and we use a weighted average operator to calculate the global degree of achievement, as it is done in [3]. This value is used to infer the degree of membership of the conclusion given by the rules (following [9]). In case, that the rules fired give different linguistic values in the conclusions, the final output of the system is the maximum value of the risk given by the rules. We take the maximum because the risk is directly associated with hazard to the health of human beings and ecosystems. This is a pessimistic approach. In fact, we want to study other methods.

In the example of Figure 5, we have supposed that Toxicity has a value of 4 (in a range from 1, best, to 10, worst), and the Quantity of Contaminant is 100 m^3 (from 0, best, to 2000, worst). In Figure 5, we can see the rules that are fulfilled in this example. Each of the conditions of these rules is satisfied to a different degree. In the nodes of the tree in Figure 5 we can find the membership degree for the corresponding variable. These degrees modify the final membership degree of the risk value in the consequent of the rule, together with the weights associated to each attribute. For

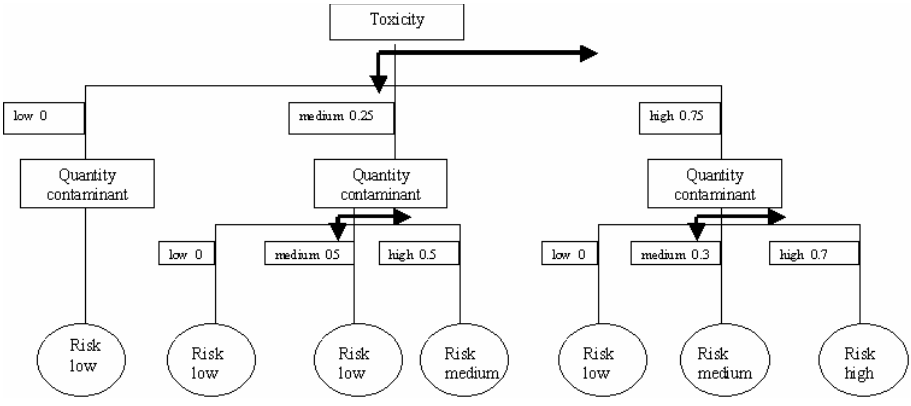


Fig. 5. Decision tree for a particular example

instance, for rule R3, we have that Toxicity is satisfied with a degree of 0.25 and Quantity of Contaminant with a degree of 0.5, so the risk is low with a degree of 0.3.

In Table 2 we can see the number of rules obtained from the decision trees for each expert system. We can see, depending on group of variables considered, the final number of rules obtained from the trees is different. The more complex part of the domain (the one that has more rules and with more complex conditions) is the group corresponding to the characterisation of the source of contamination.

Table 2 also shows some examples of rules of the four expert systems. Notice that each expert system has its own decision variable referring to RiskSource, RiskTransfer and RiskLocal respectively. Notice that the rules have different number of conditions; they use fuzzy linguistic variables but also non-fuzzy variables, like the second rule of the Local Properties system.

Table 2. Rules

Expert System	Number of rules	Examples
SOURCE	70	If Toxicity is low and ContaminantQuantity is high and HalfLife is low then RiskSource is low If VapourPressure is low then RiskSource is medium
TRANSFER VECTOR	33	If Rainfall is high and Recharge is medium then Climatology is high If Porosity is medium and Climatology is high then RiskTransfer is low
LOCAL PROPERTIES	25	If Population is adult and PopulationDistance is far and PopulationQuantity is low then RiskLocal is low If SensibleActivity=yes and SensibleEcosystems=yes and SpeciesInDanger=no then RiskLocal is medium
RISK	27	If RiskSource is medium and RiskTransfer is medium and RilkLocal is high then RiskTotal is high

5 Evaluation of the Results

To check whether the results provided by the fuzzy expert system are correct or not, we have done two types of studies. On one hand, we have tested some case studies of soils provided by the Risk Evaluation Guide published by the Catalan Government. Three different soils were considered, each one was tested with two different contaminants: DDT and Atrazine. The results were quite similar and, according to the experts' opinion, the results obtained with the fuzzy expert system were more adequate than the ones given by the guide. Although we must carry on a more exhaustive study, we can see that this system gives good results using fewer variables than the ones considered in the guide. We can also see that fuzzy values can capture the knowledge needed in this problem, so it seems better than dealing with precise numbers. This is very interesting because we can see that with less information, we can make good predictions, which results in a reduction of time and cost in the information elicitation stage.

On the other hand, we have compared our results with the ones given by another method that follows the Multi-Criteria Decision Aid (MCDA) philosophy. MCDA approaches attempt to consider a large set of criteria describing a set of different alternatives using consensus techniques [13]. In particular, we have focused on "sorting" methods, which are the ones classify the alternatives into a pre-defined set of ordered categories [13]. We have used the ClusDM system to perform this multicriteria sorting analysis, because ClusDM is able to deal with both numerical and qualitative data simultaneously [14].

ClusDM is based on the Multi-Attribute Utility Theory [15], so it has two initial stages consisting of aggregating all the information of the alternatives into a single criterion, and then ranking the values obtained. In this system, the aggregation is performed by means of unsupervised clustering techniques, which build a set of non-overlapping clusters (one for each desired final preference category). Then, the ranking is done using a similarity-based approach, comparing the prototypes of the clusters with the definition of the most risky contaminated soil and the least risky one. Finally, a linguistic preference value is assigned to each of the clusters according to its position in the ranking. The set of possible values has been provided by the user: {detest, can-not-stand, hate, do-not-like, do-not-mind, enjoy, like, love, absolutely-adore}. Then, the ClusDM algorithm selects the most appropriate term to describe the preference of each of the clusters.

Before presenting the results of the comparison between these two different decision making methods, we have to explain that ClusDM uses a different approach to uncertainty handling. Instead of defining the semantics of the linguistic values with fuzzy sets, ClusDM uses negation functions as defined in [16], where the negation of a term is the value that expresses the opposite meaning. We translated all the fuzzy set variables into the negation function representation, but the application is not exact. Consequently, there are some differences between the interpretations of the linguistic values done by each of the two systems. In Table 3 we can see the negation function of the *Land Dimension* criterion, with its associated numerical intervals. This negation maintains as much as possible the meaning of the terms given by its fuzzy variables, which were shown in Figure 1.

Table 3. Representation of the semantics of the Dimension variable with a negation function

Values of the Land Dimension	Negation of each value	Interval
Low	High	[0..0,25)
Medium	High	[0,25..0,5)
High	Low and Medium	[0,5..1]

Table 4. Results comparison

Id	FES		ClusDM		Dif^2
1	very-high	1	detest	0,04	0,001
13	very-high	0,91	detest	0,04	0,002
6	very-high	0,89	can-not-stand	0,12	6E-06
14	very-high	0,88	can-not-stand	0,12	1E-04
15	high	0,72	detest	0,04	0,058
8	high	0,69	do-not-mind	0,5	0,035
12	high	0,6	do-not-mind	0,5	0,01
10	high	0,6	hate	0,23	0,031
2	high	0,57	hate	0,23	0,041
16	high	0,56	do-not-like	0,39	0,003
7	medium	0,51	like	0,58	0,007
5	medium	0,48	hate	0,23	0,086
9	medium	0,25	love	0,69	0,003
11	medium	0,2	love	0,69	0,012
17	medium	0,19	love	0,69	0,013
18	low	0,18	love	0,69	0,016
3	low	0,13	like	0,58	0,089
4	low	0,09	love	0,69	0,048

Table 4 shows the results of a test with 18 different possible contaminated soils. FES stands for Fuzzy Expert System, and gives 4 terms to indicate the degree of risk of each of the soils. The following column gives the numerical value of risk in the range [0..1]. After this, we can find the preference values given by the ClusDM system. We can see that when FES says that site number 1 has a very high risk of contamination, ClusDM says that this site has a very low preference (value 0,04), it is detested, which means that the risk of contamination is high which is not the desirable state. The last column compares the two results giving the squared difference of their numerical values. The mean squared error of this test is 0,025.

Notice that both systems detect the worst and best situations. However, the ClusDM approach is not able to distinguish correctly the soils that have criteria indicating that it is risky and other criteria indicating that there is not a huge risk. With decision trees it is possible priorise the analysis of those criteria at different levels, so Fuzzy Expert System works better in this case.

6 Conclusions and Future Work

In this paper we propose the use of a hierarchic fuzzy expert system to characterise the risk of contamination of soils. The tests carried out indicate that the performance of the system is good, compared with manual evaluations and with results given by other techniques, such as a utility-based multicriteria decision aid approach.

The experts in this field that have participated in this study are quite confident in the use of this approach for the correct determination of the degree of risk of contamination of soils.

Remediation techniques are the following step to take into account in the management of contaminated soils. We would like to integrate the fuzzy expert system presented in this paper with the selection of the best remediation technique to be applied to recover the soil. This would be of great help for the people who have to decide how to manage the treatment of contaminated soils.

Other research line is the integration of the two approaches considered in this paper, in order to take advantage of the features of each of them. The advantage of using techniques as ClusDM is that these tools work directly with the set of criteria, not requiring any expert to build the rules of the system. Thus, they can easily change the set of criteria (including new variables or removing some of them). On the contrary, expert systems need the help of the expert committee each time that we have to modify the decision trees to incorporate new knowledge.

Acknowledgements

This work has been supported by a Spanish Ministry project (CTM2004-04720/ TECNO) and the Spanish Research network (REDEMAP: TIN2004-21700-E).

References

1. Commission of the European Community: Communication from the commission to the council, the European parliament, the economic and social committee and the committee of the regions "Towards a Thematic Strategy for Soil Protection". Brussels (2002)
2. Zhiying H., Gordon, C.W.C., Huanga, H.: A fuzzy expert system for site characterization. Expert Systems with Applications, 24 (2003) 123–131.
3. Geng L., C.Z., Chan C. W., Huang G.H.: An intelligent decision support system for management of petroleum-contaminated site. Expert systems with applications, 20 (2001) 251-260.
4. Garrouch A.: Development of an expert system for under balanced crilling using fuzzy logic. Petroleum Science & Engineering 31 (2001) 23-39.
5. J. R. Quinlan.: Induction of decision trees. *Machine Learning* 1 (1986) 81-106
6. Olaru, C. , Wehenkel, L.: A complete fuzzy decision tree technique. Fuzzy sets and systems 138 (2003) 221-254
7. Orchard, R.: FuzzyCLIPS Version 6.04 User's Guide, National Research Council Canada NRC/ERB-1054 (1996)
8. Pan, J., DeSouza, G.N, Kak, A.C.: FuzzyShell: a large-scale expert system shell using fuzzy logic for uncertainty reasoning, IEEE Transactions on Fuzzy Systems, 6 (1998) 563-581

9. Orchard, R.: Fuzzy reasoning in Jess: The Fuzzy J Toolkit and Fuzzy Jess, Proceedings of the Third International Conference on Enterprise Information Systems (ICEIS 2001), Setubal, Portugal (2001) 533-542
10. Andriantiatsaholinianina, L.A., Kouikoglou, V.S., Phillis, Y.A.: Evaluating strategies for sustainable development: fuzzy logic reasoning and sensitivity analysis. Ecologicaql economics 48 (2004) 149-172
11. Zimmerman, HJ: (1985) *Fuzzy Set Theory and its Applications*, Kluwer, Dordrecht
12. Vincke, P.: Multicriteria Decision-Aid, Ed. John Wiley & Sons (1992).
13. Valls, A.: ClusDM: a Multiple Criteria Decision Making Method for Heterogeneous Data Sets, Research Institute of Artificial Intelligence Series.17 (2003).
14. R.L. Keeney and H. Raiffa, Decisions with Multiple Objectives: Preferences and Value Tradeoffs, Wiley and Sons (1976).
15. Torra, V.: Negation Functions based Semantics for Ordered Linguistic Labels, International Journal of Intelligent Systems.11 (1996) 975-988.

A Comparing Method of Two Team Behaviours in the Simulation Coach Competition

José Antonio Iglesias, Agapito Ledezma, and Araceli Sanchis

Universidad Carlos III de Madrid,
Avda. de la Universidad, 30, 28911,
Leganés (Madrid), Spain
{jiglesia, ledezma, masm}@inf.uc3m.es

Abstract. The main goal of agent modelling is to extract and represent the knowledge about the behaviour of other agents. Nowadays, modelling an agent in multi-agent systems is increasingly becoming more complex and significant. Also, robotic soccer domain is an interesting environment where agent modelling can be used. In this paper, we present an approach to classify and compare the behaviour of a multi-agent system using a coach in the soccer simulation domain of the RoboCup.

1 Introduction

Modelling agent techniques have been used from Artificial Intelligence (IA) beginnings. The idea of extract knowledge from other agents' behaviour was used, originally, in the field of the game theory [1]. After not considering information about opponents in game mechanics, it was realized that knowledge about the opponent's strategy can enhance the game results. After this first research, many others researches have been working to develop different ways of modelling other agents. Carmel and Markovitch [2] proposed a method to infer a model of the opponent's strategy which is represented as a Deterministic Finite Automaton (DFA).

Perhaps, one of the most interesting environments where agent modelling has been used is the robotic soccer domain. The Robot World Cup Initiative (RoboCup) [3] is an international joint project to encourage AI, robotics and related fields. It provides a standard problem where many intelligent techniques must be integrated and examined. RoboCup chooses to use soccer game as a central topic of research, aiming at innovations to be applied for socially significant problems and industries [4]. One of the challenges for the near future is opponent modelling, a research which can be used both RoboCup and general multi-agent system.

One of the leagues competitions in RoboCup is *Simulation league*. In this league there are three competitions; *2D, Coach* and *3D* and each of them has different goals.

In the *2D Competition,* in order to perform a soccer game, technologies like multi-agent collaboration or strategy acquisition must be incorporated. There have been many papers and studies related to opponent modelling in this robotic soccer domain:

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 117 – 128, 2006.

Stone et al. [5] introduce "ideal-model-based behaviour outcome prediction" (IMBBOP) which models the result of the other agents' future actions in relation to their optimal actions based on an ideal world model. But, this first work into automated opponent-modelling, assumes that the opponent plays optimally. Ledezma et al. [6] present an approach to modelling low-level behaviour of individual opponent agents, OMBO (Opponent Modelling Based on Observation). Druecker et al [7] develop a neural network which is fed with the observed player positions and tries to classify them into a predefined set of formations. Riley et al. [8] propose a classification of the current adversary into predefined adversary classes. Based on this work, Steffens [9] presents a feature-based declarative opponent-modelling (FBDOM) which identify tactical moves of the opponent.

The *RoboCup Coach Competition* mainly deals with an agent (coach) providing advices to another agents about how to act. About the Coaching problem raised in *Coach Competition*, Riley et al. [11] present a general description of the coaching problem. This description is the first step in understanding advice-based relationships between automated agents. In this research, one of the reasons to consider a coach role in a team of autonomous agents is that a coach role provides a method of oversight for the agents and can aid in the creation of agents with adjustable autonomy [12]. Also, Riley et al. [13] justify that coaching can help teams to improve in simulated robotic soccer domain.

RoboCup Coach Competition changed recently in order to emphasize opponent-modelling approaches. The main goal of this competition is to compare two team behaviours, but in one of them a play pattern[1] (way of playing soccer) has been activated but not in the other one. With regards to this competition, our work on opponent modelling is driven by the goal of learning the behaviour of agents by observing them. Also, this behaviour must be stored in a way that it can be compared to another one.

This paper introduces a comparing process of agent behaviours as a general framework which can be used in different multi-agent systems. Our proposal is implemented and empirically evaluated in the RoboCup domain, specifically in the *Coach RoboCup Competition*.

The goal of this research has several aspects:

- Abstracting useful features from the multi-agent system log files.
- Analyzing these features in order to recognize different events.
- Proposing the storage of the events in a trie data structure.
- Comparing different tries to get the useful information.
- Recording this information in an easy way.

This paper is organized as follows: Section 2 presents a summary on our approach to agent modelling. Section 3 provides the trie data structure used to store events. Section 4 describes the proposed comparing method. The experimental results are shown in section 5. Finally, conclusions and future works are drawn in section 6.

[1] In this paper we use the term pattern as a contraction of play pattern.

2 The Modelling Environment

2.1 The RoboCup Simulation League

The robotic soccer domain has been chosen because is one of the most used and useful domains in this subject. The *RoboCup Simulation League* uses the Soccer Server System [10] to simulate the field and the objects. Each player consists of a unique process that connects via a standard network protocol to the server. This server keeps track of the current state of the world, executes the actions requested by the clients, and periodically sends each agent noisy information about the world. There are 11 players (10 fielders and 1 goalie) that can only perceive objects that are in their field of vision and both the visual information and the execution of the actions are noisy.

2.2 The RoboCup Online Coach Competition

The *RoboCup Online Coach Competition* is a sub-league for automated coaches which are able to advise a team of autonomous agents to perform better its behaviour in front of an opponent. In this case, in addition to the players, each team may connect a privileged client to the server, the coach agent. It gets global and noiseless information from the Soccer Server about the position and speed of all players and the ball. This coach can only support its team by giving messages to its player in a standard coach language called *CLang*, which was developed by members of the simulated soccer community [10].

This competition structure changed recently in order to emphasize opponent-modelling approaches. According to the RoboCup 2005 official rules for the *Coach Competition* [20], instead of having a coachable team playing against an opponent to score, this coachable team plays against a fixed opponent in which several playing patterns have been activated. The term pattern is used to describe a simple behaviour that a team performs which is predictable and exploitable for the coaches [14]. For example, a possible pattern could be that the player number 2 marks the attacker advancing from the lower part of the field. The coach is given a number of game logs and it must model them in order to detect the used patterns and report them. In other words, the coaches should be looking for the qualitative differences among the pattern log file and the corresponding no-pattern log file to recognize the pattern correctly.

Therefore, this competition requires two phases:

- **Offline analysis:** Is the first phase of each round and the coaches analyzes the log files of the patterns which will be used during the round.
- **Online Detection:** The only task of coach in this phase is to detect the pattern(s) activated in the opponent team.

The performance of a given coach is based only on its ability to detect and report patterns. The research focus is on team/opponent modelling and online adaptation. The coaches can work both by analyzing logs of previous games and by observing and adapting while a game is proceeding.

3 Classifying Behaviour

As we describe in the above section, *RoboCup Online Coach Competition* has two phases, but in this paper we consider only the first one (offline analysis) because this is a preliminary work on this domain. The goal of this first process is to extract and store in a useful way the important data from log files. This data must be relevant to classify the multi-agent/team behaviour.

3.1 Feature Extraction

Kuhlmann et al. [14] describe a procedure to identify high-level events in a play game. An event represents a recognized atomic behaviour. The goal of their work is to get a coach that learns to predict agent behaviour from past observations and generates advices to improve its team's performance. Based on that work, we carried out two processes: feature extraction and event recognition. Unlike Kuhlmann procedure, we use the result of these processes to describe a team's behaviour.

When running the *Soccer Server*, certain options can be used to store all the data for a given match. This information is recorded in a log file, so this file is a special stream of consecutive information data. As we are considering an offline mode, the coach receives data from the log files instead of getting it from the server. Hence, from these log files, we extract important features over all the information of the field.

At every cycle of the server, each agent updates its world model with data such as positions and velocities, as well as cumulative data such total travel distances and average positions. In this first phase, the most relevant data of these log files must be extracted, so we need to get the following information:

- *Cycle:* A number that enables arrange the events.
- *Ball Position:* The ball's position is stored like *xy* axis in a coordinate system.
- *Teammate Positions:* Each teammate´s position is stored as *xy* axis coordinates.
- *Opponent Position:* Each opponent´s position.
- *Ball Possessor:* A data that indicates who is the owner of the ball.

3. 2 Event Recognition

After extracting data from the logfiles, we have to infer what events have occurred. There is some uncertainty inherent in the results because there are events very hardly to identify, even if it is done by a soccer expert. Kuhlmann et al. work [14] propose nine different events (dribble, hold, goal, pass, foul, steal, missed shot, intercepted pass and clear) to create advices in RoboCup Simulation Soccer. In our work, we only identify seven events and the way we use them is different.

One of the most important data to extract to identify high-level events is the owner of the ball every cycle. When a ball possession change occurs, it means that an event is taking place. For our research, we classify the next events:

- **PassXtoY (T):** If a player (*X*) of the team (*T*) kicks the ball and a teammate (*Y*) gains possession, then the ball owner made a pass. (Perhaps the ball owner did not want to do this pass, but we can not consider this assumption). This event stores both the player who makes the pass and the player who gains possession.

- **DribbleX (T):** If the ball moves a significant distance since the player (*X*) gains possession until he kicks it.
- **InterceptedPassXtoY (T):** If the player (*X*) kicks the ball and the opponent (*Y*) gains possession within a reasonable distance of the ball owner, the event is assumed to be an intercepted pass.
- **StealXfromY (T):** If a player (*X*) kicks the ball and an opponent (*Y*) gains possession, then the opponent stole the ball from the ball owner.
- **GoalX (T):** If a player (*X*) kicks the ball and at the end of the interval, the ball is in the goal, the event is classified as a goal by the player *X*.
- **MissedShotX (T):** If a player (*X*) kicks the ball and at the end of the interval, the ball is out of bounds, the kick is considered as a missed shot.
- **Clear (T):** If the event cannot be classified as any of the above categories.

The result of this phase is a set of events ordered by time. Each stream of events is labelled by a team (T = Left or Right) and the result may look as follows:

{Pass1to2 (R) → Dribble2 (R) → Pass2to10 (R) → Goal10 (R)}, {Pass7to10(L) → MissedShot10 (L)}

3.3 Building a Trie

Once we have extracted the features from the log files and every event have been recognized, the next step is to analyze the behaviour sequences and choose the most appropriate sequences for our goal. In this and next sections, we try to choose, as well as possible, the behaviour sequences that describe the pattern followed by the team.

As we described in section 2.2, a pattern describes a simple behaviour that a team performs. As this pattern must be predictable for the coaches, we consider that the repeating events or behaviour sequences could be related to the activated pattern. Because of this supposition, in this paper we propose the use of a trie structure data [17], [18] to store the results of the event recognition.

A trie (which is abbreviated from "retrieval") is a kind of search tree similar to the data structure commonly used for page tables in virtual memory systems. This special search tree is used for storing strings in which there is one node for every common prefix and the strings are stored in extra leaf nodes. In this research we propose to use this data structure to store events in an effective way for our goal. The advantage of this kind of data structure is that every event is stored in the trie just once, but in a way that the event has a number that indicates how many times it has appeared.

In this research, every node represents an event so a path from the root to a node represents a sequence of events in the order they were played. Works by Kaminka [15] and Huang et al. [16] use the same data structure to analyze the behaviour sequences. The goals of these two researches are: to learn the coordinated sequential behaviour of teams [15] and to create frequent patterns in dynamic scenes [16].

An example of this trie data structure is shown as follow. If we get the next events sequences from the previous phase:

{Pass1to2 (R) → Dribble2 (R) → Pass2to10 (R) → Pass10to11 (R)} and *{Dribble2 (R) → Pass2to10(R) → Goal10 (R)}* and the trie is empty, the first event sequence to insert is *{Pass1to2 (R) → Dribble2 (R) → Pass2to10 (R) → Pass10to11 (R)}* and it is added the first branch of the tree. Each event is labelled with the number 1 that indicates how many times the event has been inserted in that sequence (In Figure 1, this number is

enclosed in brackets). Then, it is inserted the two remaining suffixes of this sequence: *{Dribble2 [1] → Pass2to10 [1] → Pass10to11 [1]}* and *{Pass2to10 [2] → Pass10to11 [2]}*. Next, we insert the second sequence *{Dribble2 (R) → Pass2to10(R) → Goal10 (R)}* and the remaining suffix: *{Pass2to10(R) → Goal10 (R)}* but in this case, there exist sequences like these inserted in the trie and we just add up one to the counter number of the events.

The trie built previously for the Right team is shown in the Figure 1.

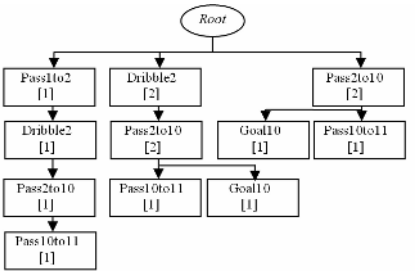


Fig. 1. Example of a trie representation

As we will see in section 5, the obtained trie from a log file game is quite big. So, in order to get the most important information and because of the pattern event sequences must be repeated to be able to extract the pattern, the tries are pruned. For pruning a trie, we eliminate the branches which have been introduced in the trie only once (it means that the node of level 1 of the branch has been introduced once).

3.4 Evaluating Dependence

Although there are few methods for discovering significance of sequences and subsequences, in this paper, we have used a statistical dependency test [16].

The main idea of this procedure is the proposal that the appearance of repetitive sequences may indicate a pattern. To evaluate the relation between the previous events sequence to a specific event (what we call prefix) and that event, we use one of the most popular statistics: Chi-square test [19]. Chi-square test is a statistical test that we propose to determine whether a prefix is dependent on the following events. This test enables us to compare observed and expected sequences objectively and evaluate whether a deviation appears. Hence, if we modify the trie structure that we have obtained in the previous section 3.3 for storing this value in every node (except the nodes of level 1 and the root), we can determine whether an event is or not relevant with its prefix.

To compute this test, a 2x2 contingency table (also known as a cross-tabulation table) has to be made. This table is filled with four frequency counts, as we can see in the Table 1. The counts are calculated as follows: The first number O_{11} indicates how many times the current node/event is following the prefix. The number O_{12} indicates how many times the prefix is followed by a different prefix. The number O_{21} indicates how many times a different prefix of the same length, is followed by the same event. The number O_{22} indicates how many times a different prefix of the same size, is followed by a different event.

Table 1. The contingency table

	Event	**Different event**	**Total**
Prefix	O_{11}	O_{12}	$O_{11} + O_{12}$
Different prefix	O_{21}	O_{22}	$O_{21} + O_{22}$
Total	$O_{11} + O_{21}$	$O_{12} + O_{22}$	$O_{11}+O_{12}+O_{21}+O_{22}$

The expected values are calculated as in *Equation 1*.

$$Expected\ (E_{ij}) = (Row_i\ Total\ x\ Column_j\ Total)\ /\ Grand\ Total \tag{1}$$

The formula to calculate chi-squared value, is giving in *Equation 2*.

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

where: O_{ij} is the observed frequency and E_{ij} is the expected frequency.

4 Comparing Process

The most interesting part in our research is to provide a method to compare two agent systems' behaviours. In this research, we only compare two agent systems' behaviours, the process to compare more than two behaviours is not considered in this work.

The input of this procedure is the result of the previous process. In this procedure, we compare two different tries which represent the behaviour of two agents systems. But, one of the agent system follows an unknown pattern. The result of the comparison is a pattern description as similar as possible to the pattern followed by the pattern-trie. The term *pattern-trie* is used to refer to the trie obtained from the agents system which followed a pattern, and *no-pattern-trie* terms the trie for the agents system that did not follow the pattern.

Before describing the comparing method, we will present our proposal to store the result of the comparison: our pattern description.

4.1 Our Pattern Description

In this work, a pattern defines recurring events to a recurring prefixes. Also, a pattern could consist of a set of simple behaviours. Because of this reason, our pattern description consists of a set of sub-patterns. Our sub-pattern is defined as the possibility, measured by chi-square test value (*chi-sq*), that an event (*ev*) occurs after a prefix (*pr*). Let OurPDescription = {[p1], [p2], ...} be the set of all sub-patterns $sub\text{-}p_i$. A sub-pattern is defined as follow:

$$sub\text{-}p_i = (ev, pr, chi\text{-}sq)$$

Because of this is a preliminary work on the *RoboCup Online Coach* domain, we do not treat the pattern description after it is obtained. However, it will be used in future works in the online detection of possible patterns in an opponent team.

4.2 Comparing Algorithm

In this section, we describe the proposed algorithm to compare two tries (*pattern-trie* and *no-pattern-trie*) in order to get the pattern description that the *pattern-trie* follows. Before describing the algorithm, we have to consider the following concepts about the trie data structure:

- As we describe in section 3.4, every node is represented by:
  - Event: A word that indicates a specific event and the agent/s that takes part in it (*passXtoY, InterceptedPassXtoY,…*).
  - Prefix: A set of the previous events in the trie.
  - Chi-Sq: A number that indicates the chi-square test value for the node.
- The depth of a trie is the maximum depth of any of its leaves.
- The level 0 in trie is the *root*.

Also, the main features of the algorithm are:

1. A threshold value (*ThresholdChiSqValue*) has to be established for accept or reject the chi-square test hypothesis in the events.
2. If the event of a node is represented in both tries at the same level and their prefix is the same:
   - Chi-square value of both nodes is compared and only if the difference is bigger than the threshold, the information of the node of the *pattern-trie* is stored as a sub-pattern in the pattern description. It means that a different behaviour in the trie has been found (and it is classified as a sub-pattern).
3. If the event and prefix of a node are represented only in the *pattern_trie:*
   - If the chi-square value of the event is bigger than a threshold value (*Threshold-ChiSqValue*), the information of the node is stored as a sub-pattern in the pattern description.

In order to make efficient and more comprehensible the proposed algorithm, we describe, first, a few used functions:

- *depthTrie (Trie T):* This function returns the maximum depth of any of its leaves.
- *getSetOfNodes* (*Level L, Trie T):* This function retrieves a result set that contains every node of the trie *T* in the level *L.*
- *getNode* (*Event E, Prefix P, SetOfNodes S):* This function returns a node consisting of the event *E* and which prefix is *P*, and is obtained from the set of nodes *S*. (If a node with these parameters does not exist in the trie *T*, the function returns *null*).
- *chi-Sq* (*node N*): This function returns the chi-square value of the node N.
- *Prefix* (*node N*): This function returns the prefix (set of events) of the node N
- *AddToOurPatternDesc (Event E, Prefix P, Chi-Sq Chi-SqValue*): This function adds to our pattern description the new sub-pattern consisting of the event *E*, the prefix *P* and the chi-square value *Chi-SqValue.*

The basic structure of the algorithm is shown in Figure 2.

```
CompareTries (pattern_trie, no_pattern_trie)
begin
   patternDesc ← null
   maxDepth ← depthTrie (pattern_trie)

   for level_i ← 2 to maxDepth do
      {Root is considered in level 0 and the events in level 1 have no prefix}

      set_P ← getSetOfNodes (level_i, pattern_trie)
      set_NP ← getSetOfNodes (level_i, no_pattern_trie)

      for all node_p in Set_P  do
         node_np ← getNode (event(node_p); prefix(node_p); set_NP)
         if (node_np = null)
         {The node is only in the pattern_trie}

            if (chi-Sq(node_p) > ThresholdChiSqValue)
            patternDesc←AddToOurPatternDesc(event(node_p),prefix(node_p),chi_sq(node_p))
            Fi

         else  (if node_np ≠ null)

            if ( (chi_sq(node_p) - chi_sq(node_np)) > ThresholdChiSqValue)
            {The node is in both tries, so the difference between them must be considered}
               DifChiSqValue = chi_sq(node_p)- chi_sq(node_np)
               patternDesc←AddToOurPatternDesc(event(node_p),prefix(node_p), DifChiSqValue)
            fi

         fi
      od
   od
   return (patternDesc)
end
```

Fig. 2. Basic structure of the algorithm

5 Experimental Results

We carried out a set of experiments in order to test whether our method is able to compare a game played by a RoboCup soccer team which follows a pattern to a game in which this pattern is not followed. Also, the result must be a description of the detected pattern.

Robocup 2005 Coach Competition has created a set of simple strategies to be used as the base strategies of the patterns [20]. The patterns are added to theses bases strategies. A no-pattern logfile describes the base strategy and the pattern file describes a simple behaviour that the opponent team performs.

The first step in our experiment is to obtain the pattern and its corresponding no-pattern logfiles. These files, that contain the same data that the online coach receives, have been obtained from the *"Patterns"* section of *"RoboCup 2005 Coach Competition*" web page [21]. The two logfiles last around 3500 times steps (a complete game lasts 6000 times steps). The pattern and no-pattern descriptions (defined by CLang rules) are described as follows:

- Pattern Description: The attackers (i.e. players 10 and 11) pass to each other right in front of the goal, before shooting.
- No-pattern Description: The attackers (i.e. players 10 and 11) dribble to the goal and shoot (i.e. they do not pass each other).

In order to extract the main features from the log files and analyze the behaviour of the two games, we have fully implemented a program. This program is able to recognize and report the events of each game. The result is a file for each game (pattern game and no-pattern game) where is defined every event of the game. This event representation is consisting of the event realized, the player/s who made this event and the team of the player.

Then, we build a trie for each logfile with the event sequences of the opponent team (this is the only team analyzed for our purpose). The *pattern trie* represents the game in which the opponent follows a pattern and *no-pattern trie* represents the game in which the pattern is not followed. As we have seen, a trie node consists of (1) the description of an event (2) a number that indicates how many times the event has been inserted in that sequence (3) Chi-square value for this node (except for the nodes of level 1 and the root).

In this experiment, the total node number of the tries is: 176 (pattern trie) and 121 (non-pattern trie). But, as we explained in section 3.3, in order to reduce the amount of no significant nodes, the trie is pruned and the branches which have been introduced only once are eliminated.

The main characteristics of the pruned tries are:

- Pruned *pattern trie*: Total node number: 49 nodes. Depth: 6.
  - Chi-square Value: Maximum (35,0); minimum (1,3) and average (11,85).
- Pruned *no-pattern trie*: Total node number: 56 nodes. Depth: 5.
  - Chi-square Value: Maximum (25,0); minimum (0,2) and average (11,91).

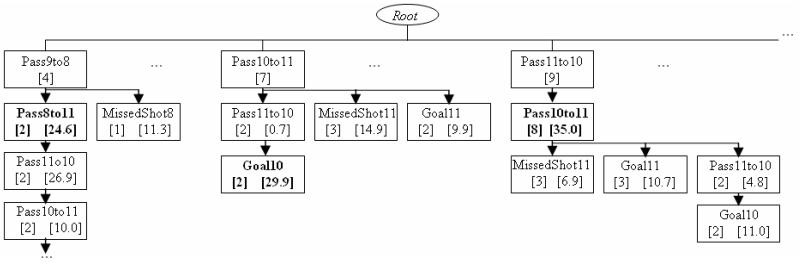


Fig. 3. Most important branches of the pattern trie

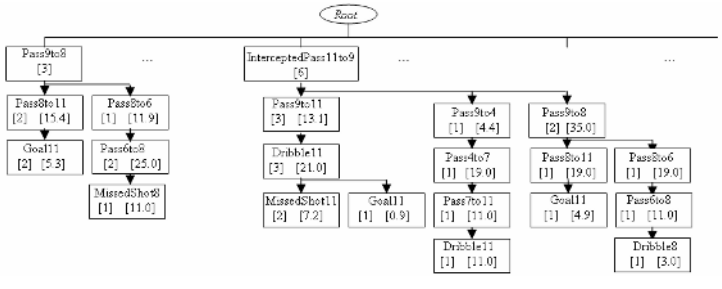


Fig. 4. Most important branches of the no-pattern trie

The most important branches of each trie are shown in Figures 3 and 4.

After obtaining the two tries that represent the two games, we can apply the algorithm described in section 4.2:

At level 2 of the pattern trie, we observe that the event “pass10to11” (and prefix “pass11to10”) does not appear in the no-pattern trie. Also, this event has the biggest chi-square value (35.0), so we insert this event and its prefix in the pattern

description. Another interesting node in this level, is the event "Pass8to11" (and prefix "Pass9to8"), but, in this case, this event appears in the no-pattern trie with a similar chi-square value, so we do not consider this event as a possible sub-pattern.

At level 3 of the pattern trie, the node "Goal10" is considered as a sub-pattern because of its chi-square value and because this node is not repeated in the no-pattern trie. So, we insert this sub-pattern in the pattern description. After analyzing every node in the pattern trie, Our Pattern Description is as follows:

OurPDescription = {[(pass10to11), (pass11to10), (35.0)],[(Goal10), (pass10to11 → pass11to10), (29.9)]}

As we can see, our pattern description represents a team behaviour very similar to the team behaviour represented by the original pattern description. So, in this case, our method is able to look for the qualitative difference between the pattern trie and its corresponding no-pattern trie. The only difference between our pattern description and the original pattern, is that the event "pass11to10" has not been identified as a sub-pattern. It is because of this event is in level 1 and at this level the nodes have no chi-square value. This improvement could be made in future works.

6 Conclusions and Future Work

In this paper we presented a comparing method of two different multi-agent systems behaviours, which one of them follows a specific pattern.

This method is applicable and useful in Coach RoboCup Soccer domain. In order to compare two different games, in which one of them is following a pattern, we use the trie data structure and it is demonstrated that a trie can be very usefulness to show the behaviour of a team in this domain.

The comparing method that we have applied works successfully when the pattern followed by a team is related to the players' actions. But, in our research, the different field regions in which the action occurs, has not been represented, so if the pattern followed by the team is related to this aspect, our proposed method would not be viable. Also, if the pattern is related to actions that occur when the player is not the ball owner, this method, as well, would not be viable.

This is a preliminary work on this domain and we consider that the result of the proposed method is very adequate in great amount of cases.

References

1. T. L. Turocy & B. Von Stengel. Game Theory. CDAM Research Report LSE-CDAM., 2001
2. D. Carmel & S. Markovitch. Opponent modelling in a multi-agent system. In G. Weiss and S. Sen, editors, Lecture note in AI, 1042: Adaptation and Learning in Multi-agent Systems, Lecutre Notes in Artifiical Intelligence. Springer-Verlag, 1996.
3. H.Kitano, M. Tambe, P. Stone, M. Veloso, S. Coradeschi, E. Osawa, H. Matsubara, I. Noda & M. Asada. The robocup synthetic agent challenge. In Proceedings of the Fifteenth Internacional Joint Conference on artificial Intelligence (IJCAI97), pages 24-49, San Francisco, CA, 1997.

4. H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda & E. Osawa. Robocup: The robot world cup initiative. In W. Lewis Johnson and Barbara Hayes-Roth, editors, Proceedings of the First International Conference on Autonomous Agents (Agents'97), pages 340-347, New York, 5-8, 1997. ACM Press.
5. P. Stone, P. Riley, and M. Veloso. Defining and using ideal teammate and opponent agent models. In Proceedings of the Twelfth Innovative Applications of Artificial Intelligence Conference (IAAI-2000), 2000.
6. A. Ledezma, R. Aler, A. Sanchis & D. Borrajo. *RoboCup 2004: Robot Soccer World Cup VIII*, volume 3276 of *Lecture Notes in Computer Science*, chapter Predicting Opponent Actions by Observation, pages 286-296. Springer Verlag, Lisbon (Portugal), 2005.
7. C. Druecker, C. Duddeck, S. Huebner, H. Neumann, E. Schmidt, U. Visser, & H.-G. Weland. Virtualweder: Using the online-coach to change team formations. Technical report, TZI-Center for Computing Technologies, University of Bremen, 2000.
8. P. Riley & M. Veloso, Distributed Autonomous Robotic Systems, volume 4, chapter On Behavior Classification in Adversarial Environments, pages 371-380. Springer-Verlag, 2000.
9. T. Steffens. Feature-based declarative opponent-modelling in multi-agent systems. Master´s thesis, Institute of Cognitive Science Osnabrück, 2002.
10. M. Chen, E. Foroughi, F. Heintz, Z. Huang, S. Kapetanakis, K. Kostiadis, J. Kummeneje, I. Noda, O. Obst, P. Riley, T. Steffens, Y. Wang & X. Yin. Soccerserver Manual v7. The RoboCup Federation, 2001.
11. P. Riley, M. Veloso & G. Kaminka. Towards any-team coaching in adversarial domains. In Onn Shehory, Thomas R. Ioerger, Julita Vassileva, and John Yen, editors, Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Sytemsi (AAMAS), Bologna, Italy, 2002.
12. P. Scerri, D. Pynadath, & M. Tambe. Adjustable autonomy in real-world multi-agent environments. In Agents-2001, 2001.
13. P. Riley, M. Veloso & G. Kaminka. An Empirical Study of Coaching. In H. Asama, T. Arai, T.Fukada, and T. Hasegawa, editors. Distributed Autonomous Robotic Systems 5, pp 215-224, Springer-Verlag, 2002.
14. G. Kuhlmann, P. Stone & J. Lallinger. The Champion UT Austin Villa 2003 Simulator Online Coach Team. In Daniel Polani, Brett Browning, Andrea Bonarini, and Kazuo Yoshida, editors. *RoboCup-2003: Robot Soccer World Cup VII*, Springer Verlag, Berlin, 2004.
15. G. A. Kaminka, M. Fidanboylu, A. Chang & M. Veloso. Learning the Sequential Coordinated Behavior of Teams from Observations. In Proceedings of the RoboCup-2002 Symposium, Fukuoka, Japan, June 2002.
16. Z. Huang, Y.Yang and X.Chen. An approach to plan recognition and retrieval for multi-agent systems. *Proc of AORC*, Sydney, Australia. Jan 2003.
17. D. E. Knuth. The Art of Computer Programming Vol. 3, Sorting and Searching. Addison-Wesley. 1972.
18. E. Fredkin. *Trie Memory.* Communication of the ACM**.** Vol. 3:9 (Sep 1960). pp. 490-499.
19. F.J. Rohlf, & R. R. Sokal. 1995. Statistical Tables, 3rd ed. W. H. Freeman and Company, New York, NY.
20. The RoboCup 2005 Simulation League Organizing Committee. RoboCup 2005 Official Rules for the Coach Competition. January, 2005.
http://staff.science.uva.nl/~jellekok/robocup/rc05/Coach-Rules-Revision0.3.pdf
21. http://staff.science.uva.nl/%7Ejellekok/robocup/rc05/

On the Use of Tools Based on Fuzzy Set Theories in Parametric Software Cost Estimation

F. Javier Crespo[1] and Óscar Marbán[2]

[1] SIP, Universidad Complutense de Madrid, Madrid, Spain
javier.crespo@fdi.ucm.es
[2] DLSIS, Universidad Politécnica de Madrid, Madrid, Spain
omarban@fi.upm.es

Abstract. The whole software industry has an awful footpath for delivering software on-time and on-budget. Probably, one reason is coming from not deal with the imperfection of information when they use a lot of human process. In this paper, we propose the use of fuzzy measures in contrast with crisp measures of traditional models and, therefore, apply of appropriate aggregators. Traditional models of software cost estimation are constructed from project databases and they describe cost drivers in terms of linguistic estimations using vague terms like "low" or "high", and such expressions are also used in obtaining actual predictions. But cost drivers are in many cases abstract concepts that are better estimated by breaking them down in a number of second-level aspects. The method proposed is based both, on a concrete study of the use of linguistic variable human categorizations and, on level aspects that are defined by layer and are easy to raise using appropriate aggregators. Moreover, the proposed scheme can have different planes according to the model morphology.

Keywords: Soft Computing, indiscernibility, Software Engineering, Fuzzy set theory, Software Development Cost, Aggregation Operators, Cost Drivers.

1 Introduction

1.1 Software Cost Estimation

The whole software industry has an awful footpath for delivering software on-time and on-budget[16].

No more than ten percent of software projects in large corporations actually finish on budget and on time. Over fifty percent of them cost in excess of twice the original budget. More than half of software projects are deployed without proposed functionalities in the original project because the time to complete them was widely surpassed, even over twice as long as planned and the budget go beyond of a painful overruns.

Effective management of any project requires quantification, measurement and modeling. Software measures provide or should be provided a quantitative basis for the development and corroboration of process of the software development. More specifically, they would have to give it in the process of Software Cost Estimation.

There are several ways to obtain estimation. The most basic approach is to record effort, duration or size estimates as well as estimating processes and assumptions, and

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 129–137, 2006.

then record the actual results from each estimated cost driver or activity. Comparing actual outcomes to the estimates helps generate more accurate estimates in the future. But, if you have not a good estimation, almost estimating procedures and templates that itemize tasks help avoid the common problem of overlooking necessary work.

1.2 The Use of Rough and Fuzzy Set Theories on Software Cost Estimation

Scattered previous research has addressed diverse aspects of fuzziness in software estimation models, including estimation by analogy [11], fuzzy function points [19], generalizing estimation formulas [13], and using fuzzy regression methods to adjust the models [9, 10]. But none of the approaches provide an experimental account regarding the role of natural language and human categorization [17] in the process of modeling and working with the inputs. At the same time, a study of a tool by evaluating, controlling and predicting some important attributes of software projects such as Usability.

The rest of this paper is structured as follows. Section 2 we report a concrete experience in investigating the fuzziness associated to the natural language expression of input values of one concrete COCOMO cost drivers, as representative case study of the modeling challenges that underlie many of the software and project attributes that is often used in parametric models. In Section 3, describes a concrete cost estimation setting under fuzziness, and a discussion of the resulting estimates. Finally, conclusions and future research directions are sketched in Section 4.

2 Fuzzines in Cost Drivers When They Are Described by Linguistic Variables

A simple on-line cost model for estimating the number of person-months required to develop software can be employed as educational examples of technology developed and used by cost engineers (Figure 1). This model COCOMO, described originally by Boehm in [4], estimates cost using factors to account for differences in constraints. These factors are calling cost drivers. Figure 1 show cost drivers and the linguistic variables associate by easy use of COCOMO.

Here is what a experiment says about the use of one cost drivers with linguistic values and the participants mental model.

The AEXP cost driver is dependent on the level of applications experience of the project team developing the software system or subsystem. The ratings are defined in terms of the project team's equivalent level of experience with this type of application. According to COCOMO indications, a very low rating is for application experience of less than 2 months, a very high rating is for experience of 6 years or more, and the intermediate labels are approximated by 6 months, one year and three years.

These assumptions can be contrasted by a membership elicitation process. Figure 2 provides the result of a membership exemplification process. The apparent "anomaly" in high values of rating scale are the consequence of a belief that when a developer has reached a certain degree of experience with a given technological context, he/she stops improving his/her level of knowledge, due to the relative degree of self-satisfaction.

Product Attributes

VL L N H VH XH : Required Reliability
VL L N H VH XH : Database Size
VL L N H VH XH : Product Complexity

Computer Attributes

VL L N H VH XH : Execution Time Constraint
VL L N H VH XH : Main Storage Constraint
VL L N H VH XH : Virtual Machine Volatility
VL L N H VH XH : Computer Turnaround Time

Personnel Attributes

VL L N H VH XH : Analyst Capability
VL L N H VH XH : Applications Experience
VL L N H VH XH : Programmer Capability
VL L N H VH XH : Virtual Machine Experience
VL L N H VH XH : Programming Language Experience

Fig. 1. COCOMO-81 Web tool input form

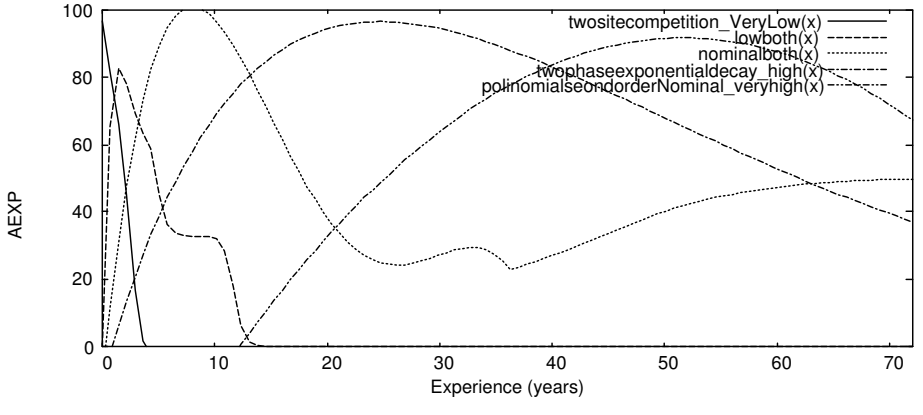


Fig. 2. Fuzzy sets for the AEXP cost driver

Another apparent “anomaly” appear by overlap. The large degree of overlapping becomes much more complex the work out, since categories overlap to larger extents. In addition, the separation between the centroids of the fuzzy numbers obtained by the proposed method and the values of the traditional model present a non-equitable

disparity. This can be due to cost drivers normally use crisp values to utilize algebra operator, even when they are described like intervals. The assignment of linguistic variables to these cost drivers, in order to make effortless their use, bring the confusion on theirs use.

3 The Use of Fuzzy Variable Entries and Aggregations Operator

Several existing software cost estimation models consider a number of factors often called cost drivers[5]. These cost drivers are of a diverse nature, encompassing development team's abilities, complexity and reliability of the software and schedule constraints, among others. But these cost drivers are in many cases of a somewhat abstract nature, so that estimating their values directly for a given project becomes a difficult task. For example, the influence of the overall documentation required is better expressed through the requirement for certain types of more specific documentation artifacts. In consequence, these high-level cost drivers can be broken up in more concrete and easily measurable "second-level" aspects that contribute to them to some extent. This raises the necessity for the design of aggregation operators from second-level (and possibly in some cases, from third and further levels) aspects to first-level ones.

Aggregation operators can be considered as mathematical objects that have the function of reducing a collection of numbers (or more generally, of values) to a unique representative or meaningful one. Perhaps the most commonly used aggregation operator is the weighted mean, but this does not mean that it's necessarily always the better choice. Since a considerable amount of research has been carried out in the last decade regarding the design of aggregation operators [6], a rich array of aggregation operator families with specific characteristics are available for the practitioner to better solve the problem at hand. These aggregation models should ideally be considered in each concrete aggregation process, or at least some well-known classes of Software Engineering problems should be investigated regarding the adequacy of such models for their underlying aggregation processes.

3.1 Case Study Description

Usability is a multifaceted concept [20] that encompasses several attributes regarding the interaction of humans with software systems, and its attainment is recognized as an important cost factor [2], since it must be considered at several stages in the development process and the costs of evaluation are usually high.

In many cases, these different attributes are summarized in several generic interdependent aspects like efficiency, learnability, memorability and satisfaction [15]. In consequence, estimation will typically include an aggregation stage in which partial estimations of importance regarding different attributes would need to be summarized in an overall number.

Let us consider a given project that account with three main functionalities f1, f2, f3 for which the required usability (USAB) is stated in term of three second-level aspects: efficiency (EFF), control (CTRL) and learnability (LRN).

Each of the functionalities have different requisites regarding each of these aspects. For example, f1 is the most commonly used function, a data entry form, that require high

Func.	Imp.	e1	e2	e3	c1	c2	c3	l1	l2	l3
f1	0.7	0.9	0.95	1	0.9	0.95	1	0.3	0.4	0.5
f2	0.2	0.4	0.5	0.7	0.6	0.7	0.8	0.2	0.3	0.4
f3	0.1	0.3	0.4	0.5	0.5	0.65	0.75	0.2	0.3	0.4

Fig. 3. Estimated requirements about each of the second-level usability aspects

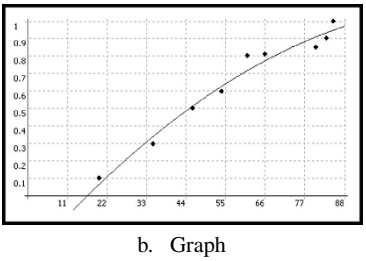
efficiency (i.e. speed in data entry) and control (usually measured as the absence of user errors). Functionalities f2 (a reservation form) and f3 (a devolution handling form) are considered to have less stringent requirements in terms of efficiency and control, since they occur less often. In addition, the three aspects have fairly similar requirements in terms of learnability (which impacts in the cost of training for the users of the application). In addition, the three functionalities have different importances with regard to the quality of their interfaces. For these departure assumptions, the first required step is that of estimating numerically the degrees of requirement for each usability factor.

A streamlined variant of the fuzzy Delphi method described by Ishikawa et al. [12], can be used for that purpose, estimating triangular fuzzy numbers. The results of a pilot study with three respondents and two rounds are showed in Figure 3.

In this Figure 3 the requirements for each aspect are expressed as triples – e.g.(e1, e2, e3) for efficiency – of numbers in [0,1] indicating the minimum acceptable level, the more reasonable level and the maximum level, respectively. In addition, the importance of each functionality is expressed as a weight. Now the problem is that of finding an appropriate overall usability requirement from the described sub-aspects. To do so, let us consider a model of the cost estimates based on the assumption that the costs of usability come from carrying out usability testing with users. According to the generic model described by Nielsen ([14], Chapter 6) the cost of usability tests can be estimated by using the formula $upf(i) = N(1 - (1 - \lambda)i)$, where i is the number of test users, N the (estimated) total number of usability problems in the interface, and the probability finding any single problem with any single user. For the above described cost estimating situation, we'll assume that the cost for EFF follows that formula with $N = 41$ and $\lambda = 0.31$, CTRL represents a 10% of that estimates (since the procedure for testing is the same than for EFF to a large extent), and LRN follows the formula with $N = 41$

i	upf_{EFF}	upf_{CTRL}	upf_{LRN}	μ_{USAB}
1	12.71	1.27	6.15	0.1
2	21.48	2.15	11.38	0.3
3	27.53	2.75	15.82	0.5
4	31.71	3.17	19.6	0.6
5	34.59	3.46	22.81	0.8
6	36.58	3.66	25.54	0.81
12	40.52	4.05	35.17	0.85
17	40.93	4.09	38.41	0.9
22	40.99	4.1	39.85	1

a. Table



b. Graph

Fig. 4. A possible mapping for usability given the number of test users and estimated overall usability with respect to the number of errors detected in the evaluation

and $\lambda = 0.15$ (of course the rationale for such approach may be arguable, but it's based on Nielsen's estimates and an inquiry about the type of user testing required). In consequence, the number of test users is related to the number of errors (probably) found, so that we can assess the degree of "usability" according to them. Figura 4.a shows some representative values for one of such possible characterizations.

The membership function $\mu USAB$ exemplified in Figure 4 provides the mean to decide on the required number of test users, depending on the required degree of usability expressed in the [0,1] interval. In consequence, the aggregation of the partial requirements in Table 1 determines the final decision regarding the cost estimate for the project. In the following section, a number of aggregation methods are discussed.

3.2 Using the OWA to Model Cost Driver Aggregation

Partial usability requirements can be aggregated following diverse schemes, from the more pessimistic1 (using the maximum) to the more optimistic (using the minimum), and also including the arithmetic mean. Each aggregation opera- tor will eventually yield a different decision that affects the cost estimates of the project. Table 3 shows the results obtained using several of them (global numbers are obtained by weighting each task according to its importance I).

The significant divergences between minimum and maximum operators can be reconciled by using the mean, but also by using more general operators. Among the latter, we have selected the OWA operator family. [21]. An OWA operator of dimension n is a mapping:

$$F(x_1, \dots x_n) = \sum_{j=1}^{n} x_{(j)} \cdot w_j \tag{1}$$

where $x_{(i)}$ are an increasing permutation of the input variables, and the vector of weighs $W = (w_1, \dots w_n)^T$ satisfies $w_i \in [0, 1]$ and $\sum_i w_i = 1$.

The selection of the weights of the OWA operator determines its behavior, and it can be adjusted from a set of experimental data. The values in Figure 3 have been obtained this way by using Beliakov's tool[1][1], from a data set consisting on estimation of aggregations for the values in Figure 3

$(.9, .9, .3) \approx .9$, $(.4, .6, .2) \approx .4$, $(.3, .5, .2) \approx .3$, $(.95, .95, .4) \approx .9$, $(.5, .7, .3) \approx .6$, $(.4, .65, .3) \approx .5$, $(1, 1, .5) \approx .9$, $(.7, .8, .4) \approx .75$, $(.5, .75, .4) \approx .6$.

The inputs for those values are examples of aggregation obtained from Figure 3 , and the outcomes have been obtained from the participants in the study using again a streamlined Delphi technique with two rounds, so that they reflect the consensus reached from the initial divergences in overall required usability given the required usability for each of the attributes EFF, CTRL and LRN respectively.

The resulting weights are $(0.348, 0.575, 0.077)$, and the degree of *orness* of the operator amounts to 0.636. Such degree of orness indicates that our operator is slightly closer to the maximum than to the minimum.

[1] http://www3.cm.deakin.edu.au/ gleb/aotool.html

F	I	n1	n2	n3	x1	x2	x3	m1	m2	m3	o1	o2	o3
f1	0.7	0.3	0.4	0.5	0.9	0.95	1	0.7	0.76	0.83	0.85	0.9	0.96
f2	0.2	0.2	0.3	0.4	0.6	0.7	0.8	0.4	0.5	0.63	0.45	0.55	0.71
f3	0.1	0.2	0.3	0.4	0.5	0.65	0.75	0.33	0.45	0.55	0.36	0.48	0.58
	0.1	0.27	0.37	0.47	0.5	0.65	0.75	0.6	0.68	0.76	0.72	0.79	0.87

Fig. 5. μ_{USAB} required values using the minimum (n),the maximum (x) operator, the median (m) and the OWA (o)

The implication of using one of the described operators are evidenced when returning to the mapping $\mu USAB$ illustrated in Figure 4.

If we use the minimum, our choices are constrained to using one or two test users (since the global $USAB$ values are lower than 0.5), while using the maximum lead to the decision of using six or more test users (being the better much around ten users). The mean and OWA operators indicate that the number of testers must be above four and five respectively, being closer to estimated cost-benefit trade-offs described by Nielsen [15]. As user testing is a resource consuming activity, these divergences may entail significant costs, especially for small projects subject to budget constraints.

Another example of such studies is described in [18]. In this paper, we approach such view through a concrete case study concerned with usability as a software cost driver. The case study illustrates how an OWA operator can be used as an alternative to more straightforward aggregation means to come up with a more realistic solution. Of course, this case study represents only a partial illustration of the rich variety of Software Engineering situations that would eventually benefit from a consideration of the research results on the area of mathematical aggregation operators, but it's intended to motivate further research in the area.

4 Conclusions

The contributions of this paper are two-fold: (i) a study of the expressed mental model mathematically of participants about cost drivers, and its relation with the values of traditional model cost drivers, such as COCOMO-81 cost drivers, which are constructed from project databases.

As the studies described in the previous section point out, the scales used in software parametric estimation models may include varying degrees of imprecision coming from human categorizations. Moreover, estimation cost model users make choices using a mental model that typically includes both explicit and implicit perceptions of concepts and uncertainties about real nature of cost drivers. The linguistic value choice is inferred because most of them cannot articulate that mental model with any precision. These choices can provide critical inputs based on theirs cost divers criteria and forecasts as well. Choice-based models have to change cost drivers selection and measures in order to a better and easier management and to get better results.

And, (ii) a concrete study of a proposed method base on levels scheme that is defined by layer and are easy to raise using appropriate aggregators. Furthermore, the proposed one allows developing cost drivers from others of second level. These second level ones should have more reliability of magnitude-estimation scaling as a measure and easier to be measure. These second level cost drivers let build up cost drivers much more descriptive as variable entries of model formula.

This raises the need for inquiry regarding the most appropriate aggregation processes. Recent research on aggregation operator design has resulted in a number of families of operators that should ideally be considered for each concrete situation. In this paper, we illustrate this view through a concrete case study in which the required usability of a given system is modelled in terms of three second level aspects, and the OWA operator is introduced as a realistic summarization tool. But already exists another as Choquet integral[8] or Double aggregation operators [7].

The results of these case studies point out the necessity to carefully examine the aspects that affect each given cost driver, and their relative influence and, eventually, their interactions in the overall result. Future work will investigate the composition of commonly used cost drivers and the aggregation mechanisms that may better serve each concrete situation. We propose that it is not pointless make changes using rough set and fuzzy set theories in order to cope with inherent imprecision or uncertainty that pervades estimation cost software process and they must be incorporated in the overall steps of the process.

References

1. Beliakov, G.: How to build aggregation operators from data? *International Journal of Intelligent Systems*, Vol. 18, (2003) 903–923
2. Bias, R.G., Mayhew, D.J. (eds.): Cost-Justifying Usability. *Academic Press*. (1994)
3. Bilgiç, T., Türksen, T.: Measurement of Membership Functions: Theoretical and Empirical Work. In: D. Dubois and H. Prade (eds.) Handbook of Fuzzy Sets and Systems Vol. 1, Chapter 3, *Fundamentals of Fuzzy Sets, Kluwer* (1999) 195–232
4. Boehm, B.W.: Software Engineering Economics. *Prentice-Hall, Englewood Cliffs*, NJ (1981)
5. Boehm, B., Abts, C. and Chulani, S.: Software Development Cost Estimation Approaches – A Survey. *Center for Software Engineering, University of California*, Technical Report USC-CSE-2000-505
6. Calvo,T. , Kolesárová, A. , Komornikóvá, M., Mesiar, R.: Aggregation operators: Basic concepts, issues and properties. In Calvo, T., Mayor, G. and Mesiar, R. (eds). *Aggregation Operators: New Trends and Applications. Studies in Fuzziness and Soft Computing*, Vol. 97, (2002) 3–106
7. Calvo, T., Pradera, A.: Double aggregation operators. Fuzzy Sets and Systems 142(1): (2004) 15–33
8. Choquet, G.: Theory of Capacities, Ann. Inst. Fourier 5, (1954) 131–296
9. Crespo, F.J., Sicilia, M.A., Cuadrado, J.J.: On Fuzzy Regression in Software Cost Estimation Models. In: *Proceedings of the 2003 ACM-IEEE International Symposium on Empirical Software Engineering* (ISESE'03). (2003)
10. Crespo, J.J., Sicilia, M.A., Cuadrado, J.J.: On the Use of Fuzzy Regression in Parametric Software Estimation Models: Integrating Imprecision in COCOMO Cost Drivers. *WSEAS Transactions on Systems* 1(3), (2004) 96–101

11. Idri, A., Abran, A., Khoshgoftaar, T. M.: Fuzzy Analogy: A New Approach for Software Cost Estimation. In: Dumke/Abran (eds.) *Current Trends in Software Measurement*, Shaker Publ., Aachen, Germany, (2001) 127–142
12. Ishikawa, A., Amagasa, M., Shiga, T., Tomizawa, G., Tatsuta, R., Mieno, H.: The max-min Delphi method and fuzzy Delphi method via fuzzy integration. *Fuzzy Sets and Systems* Vol. 55, (1993) 241–253
13. Musilek, P., Pedrycz, W., Succi, G., Reformat, M.: Software Cost Estimation with Fuzzy Models. *Applied Computing Review*, 8(2), (2000) 24–29
14. Nielsen, J.: Usability Engineering. *Morgan Kaufmann*. (1993)
15. Nielsen, J.: Why You Only Need to Test With 5 Users. *Jakob Nielsen's Alertbox*, March 19, 2000, available at http://www.useit.com/alertbox/20000319.html.
16. Rahman, M.: *Requirement Enginnering*. Oxford University Press. (2004)
17. Rosch, E.: *Principles of Categorization. Readings in Cognitive Science*, Erlbaum, (1988) 312–322
18. Sicilia, M.A., García, E., Calvo, T.: On the use of the Choquet integral for the aggregation of usability interface related scores. In *Proceedings of the 2003 International Summer School on Aggregation Operators and their Applications*, (2003) 159–164
19. Souza Lima Jr., O., Farias, P.M., Belchior, A.D.: Fuzzy function point analysis. In: *Proceedings of the 4th European Conference on Software Measurement and ICT Control*, Heidelberg, Germany, (2001) 161–172
20. Van Welie, M., van der Veer, G.C., Eliëns, A.: Breaking down usability. *Proc. of Interact'99*, (1999) 613–620
21. Yager, R.R.: On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making, *IEEE Transaction on Systems, Man and Cybernetics* **18** (1988) 183–190

Using Fuzzy Set Theory to Assess Country-of-Origin Effects on the Formation of Product Attitude

Kris Brijs[1], Koen Vanhoof[1], Tom Brijs[1], and Dimitris Karlis[2]

[1] Hasselt University, Department of Applied Economics, Agoralaan, gebouw D, 3590 Diepenbeek, Belgium
[2] Athens University of Economics and Business, Department of Statistics, 76 Patision Street, 10434 Athens, Greece
{kris.brijs, koen.vanhoof, tom.brijs}@uhasselt.be, karlis@aueb.gr

Abstract. Several researchers on country-of-origin (coo) have expressed their interest in knowing how consumers' emotional reactions toward coo-cues affect product attitude formation. This paper shows how Fuzzy Set Theory might serve as a useful approach to that problem. Data was gathered by means of self-administered questionnaires. Technically, orness of OWA-operators enabled us to distinguish consumers expressing highly positive versus less positive emotions toward coo. It appeared that this variance in emotional estate goes together with a difference in aggregating product-attribute beliefs.

1 Introduction to Product Attitude Formation

We start this paper by providing a short overview of the literature on product attitude formation. Without going into all the details, it provides the larger (marketing) context in which the technical contribution of our paper should be seen. The motivation for this is that the contribution of this paper does not only lay in technical aspects of the OWA operator (see section 5), but it also provides superior consumer information with managerial relevance that can not be offered by conventional statistical techniques that have been used for this kind of marketing research.

Attitude Theory states that consumers' behaviour toward products is determined to a large extent by their attitude toward them. In line with Peter et al. [19], we define 'attitude' as a person's overall evaluation of a concept. According to Eagly and Chaiken [6], this overall evaluative judgement can be seen as a psychological tendency that expresses some degree of (dis)favour toward the attitude object. The Expectancy Value Model developed by Fishbein and Ajzen [10] posits that this overall evaluative judgement of the product is mediated by the evaluation of salient beliefs. In other words, people combine or integrate product knowledge to form an overall evaluation of products. Thus, consumers' beliefs about product attributes are considered as crucial determinants of their attitude toward the product. The literature on advertising and emotions has challenged some of the basic principles behind this so-called multi-attribute theory. Its key-proposition was that advertisements can generate several affective reactions which also influence the formation of consumers' attitudes toward products. Peterson et al. [20] for instance, stated that ad evoked affects can

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 138 – 149, 2006.

play various roles in consumer decision making, ranging from influencing the ways in which information is processed and stored in memory to determining product choices.

Past research indicates that advertising cues indeed produce different types of affective reactions. According to Derbaix and Pham [5], these range from emotions, feelings, moods and temperaments to preference, attitude and appreciation. As recommended by Verlegh [26], our attention more specifically goes to 'feelings'.

As put by Burke and Edell [2], ad evoked feelings influence consumers' brand attitude through their attitude toward the ad or via their brand attribute beliefs. Holbrook and Batra [13] also found that ad attitude mediates the effect of ad feelings on brand attitude. In addition, they tested whether ad evoked feelings had a direct impact on brand attitude and found that it was rather limited. Results reported by Stayman and Aaker [22] went in the same direction, although they established that the effects of ad feelings on brand attitude were not always necessarily mediated by ad attitude. Yet, on the average, we can state that ad evoked feelings rather exert an indirect influence on the consumer's attitude toward the product. Ad attitude and product attribute beliefs both seem to function as important mediators. Since the concept of ad attitude is not within our scope of interest, we will further concentrate on the other path where feelings evoked by advertising cues like country-of-origin are posited to influence the consumer's overall evaluative judgement of the product indirectly, that is, via the formation and subsequent processing of attribute beliefs.

2 Coo-Effects: The Affective Approach

Although coo-effects have been traditionally approached from an information theoretic perspective, several scholars working within the field already argued attention should be paid to the coo-cue's capacity to evoke all kinds of symbolic and emotional connotations which might interfere with the consumer's intent to evaluate a product. Some interesting examples in support of this assumption have been cited by Obermiller and Spangenberg [17]. For instance, they mention the negative reaction toward high quality Israeli-made precision instruments expressed by Americans of Arab origin. Friedman [11] in turn, explains the American Jews' boycott of German-made products during the first decades after the Second World War by the fact that 'Made in Germany' labels elicited all kinds of traumatic feelings. Klein et al. [15] established that previous or ongoing military, political and economic events between Japan and the People's Republic of China generate feelings of 'animosity' affecting Chinese consumers' buying decisions. Still recently, Verlegh [26] demonstrated how Dutch consumers' tendency to identify with their home country is accompanied by less positive feelings toward a foreign coo. These in turn, appeared to influence the formation of beliefs about the attributes of products coming from abroad in a negative way.

Thus, we might conclude that the emotional reactions triggered by the coo-cue act as potential determinants of the consumer's attitude toward foreign sourced products. However, it still remains unclear how they affect this process of product attitude formation. Our attention will be focused on that problem. Throughout the following sections, we will first elaborate on our conception of attitudes. In our effort to explain how coo-related emotions might affect the consumer's product attitude, we will base ourselves on insights from the literature on ad emotions.

3 Coo Emotions and Information Processing

According to Isen [14], affective reactions evoked by ads influence our cognitive activities in many ways. As she puts it, "[…] the evidence suggests that rather than causing people not to think, affect (at least some affects) can influence thought by influencing what people think about, how they relate things to one another, what they try to accomplish, and how they go about solving problems. Thus feelings can have a substantial influence on thought processes and resultant behaviour." [14]. Before turning to our vision on the functioning of emotions evoked by advertising cues like the product's coo, we will briefly review some insights coming from the literature on coo-emotions. To begin with, it is striking to establish how the majority of these studies are concentrated on the functioning of rather extreme negative feelings like animosity [15], ethnocentrism [21], or patriotism [12]. Overall, it is found that these are directly transferred to the product. This leads to situations where people decide not to buy, purely based on their aversive feelings toward the product's coo.

Verlegh [26] wondered whether such powerful effects would also be triggered by milder affective reactions toward coo. As he argued, extremely negative feelings toward foreign nations only manifest themselves in very particular occasions and cannot always be generalized to the context of daily life. In his opinion, the average consumer will rather be characterized by the expression of less intensive feelings toward other countries. Therefore, he focused more on the role of softer feelings. More in detail, he proposed a framework where such weaker coo-related feelings are modelled as determinants of the consumer's product attribute beliefs. Although partial and inconsistent, he found significant supportive evidence in both cases of positive and negative feelings toward coo. Thus, it appears that milder coo-feelings bias our perception of a product's attributes.

However, Han [12] thinks people's perception of a product's quality attributes is not fundamentally determined by the way we feel about the place where it was made. Obermiller and Spangenberg [17] subscribe to this reasoning in positing that consumers who experience extremely negative feelings toward certain countries still acknowledge that products from those nations are of superior quality. Thus, coo-specific feelings apparently do not alter our perception of a product's quality attributes per se, even if we vividly experience them. Han [12] designed a study to examine this problem and found that consumers expressing positive feelings toward the product' s coo only tentatively rated that product's attributes more favourably.

Thus, in general, it seems that for weaker feelings toward coo, some doubts remain on how they precisely affect our product attitudes. Verlegh [26] thinks they determine our perception of a product's quality attributes albeit that his results and those obtained by others are not very consistent. The question of knowing how such milder affective reactions toward coo influence product attitude formation thus remains open.

Our key proposition will be that softer coo-specific feelings will influence the way in which consumers process these attribute beliefs. More in detail, we argue that less intensive coo-specific feelings will affect the way in which consumers cognitively combine or integrate their product attribute beliefs.

Our reasoning is based on the principles behind the encoding-specificity mechanism developed by Tulving and Thomson [25]. The underlying idea is that affects experienced by individuals can activate thoughts which have been stored in memory as relevant and related to those affects. As put by Cacioppo and Petty, affects indeed

can "bias issue-relevant thinking by making affectively consonant thoughts and ideas more accessible in memory." [3]. Isen [14] continues that several studies have shown how people being happy show better recall of positive material. Thus, it appears that affective reactions elicited by ads can lead to greater receptiveness of positive or persuasive communication. In line with this reasoning, we assume that individuals will be more inclined to process those particular attribute beliefs which correspond best with their actual emotional state. Therefore, we formulate the following hypothesis:

H: Consumers expressing more positive feelings toward the product's country-of-origin will process the stronger valued attribute beliefs while consumers expressing less positive feelings toward the product's country-of-origin will process the weaker valued attribute beliefs.

4 Methodology

A study was designed to determine how feelings evoked by coo-cues influence the respondents' cognitive attribute processing. More specifically, the products selected for our study were DVD-players (utilitarian) and beer (hedonic). The decision to opt for two distinct types of products was taken in order to increment the external validity of our study. Additional motivation for the selection of these two products can be found in the frequent use that is made of them by other coo-researchers. The countries-of-origin selected for our study were Spain and Denmark. Both countries were sufficiently familiar to our respondents and mutually different on a number of country-specific aspects. This made participants feel confident enough in filling out the questionnaire. Also, we obtained two samples of which the overall level or intensity of country-specific feelings aroused substantially varied.

As evaluation function we have chosen the ordered averaging operator (OWA). This operator was originally introduced by Yager [29] to provide a means for aggregating scores with the satisfaction of multiple criteria, which unifies in one operator both conjunctive and disjunctive behaviour. Examples of alternative aggregation operators include the Weighted Mean and the Weighted OWA [23]. However, we have chosen the OWA because the orness-measure can be directly learned from the data.

More formally, an OWA operator [30] of dimension n is a mapping :

$$f : R^n \rightarrow R \tag{1}$$

that has an associated weighting vector W

$$W = \begin{bmatrix} W_1 & W_2 & \dots & W_n \end{bmatrix}^T \tag{2}$$

such that

$$\sum_i W_i = 1 \quad , \quad W_i \in [0,1] \tag{3}$$

and where

$$f(a_1, \dots, a_n) = \sum_{j=1}^{n} W_j b_j \tag{4}$$

where b_j is the j-th largest element of the collection of the aggregated objects a_1, a_2, …, a_n. The function value $f(a_1, …, a_n)$ determines the aggregated value of arguments a_1, a_2, …, a_n.

A fundamental aspect of the OWA operator is the re-ordering step, in particular an argument a_1 is not associated with a particular weight w_i but rather a weight w_i is associated with a particular ordered position i of the arguments. A known property of the OWA operators is that they include the Max, Min and arithmetic mean operators for the appropriate selection of the vector W.

The operator has proven to be very useful because of its versatility and its measure that can quantify or express the nature of the behaviour of the evaluator like pessimistic or optimistic. This measure, called the 'orness measure' of the aggregation, is defined as

$$orness(W) = \frac{1}{n-1}\sum_{i=1}^{n}(n-i)W_i \tag{5}$$

As suggested by Yager [30] this measure, which lies in the unit interval, characterizes the degree to which the aggregation is like an or (Max) operation. It can be shown that:

$$\begin{aligned} &\text{orness}\left(\left[1\ 0\ ...\ 0\right]^{\mathrm{T}}\right) = 1, \\ &\text{orness}\left(\left[0\ 0\ ...\ 1\right]^{\mathrm{T}}\right) = 0, \\ &\text{orness}\left(\left[1/n\ 1/n\ ...\ 1/n\right]^{\mathrm{T}}\right) = 0.5 \end{aligned} \tag{6}$$

Therefore the Max, Min and arithmetic mean operators can be regarded as OWA operators with degree of orness, respectively, 1, 0 and 0.5. The orness measure can be seen as the optimistic degree of the evaluator. The interested reader can find more information on the orness of an aggregation in [8].

Data was gathered by means of two surveys (one for Spain/Spanish products and one for Denmark/Danish products). These were distributed to respectively 616 and 609 graduate students of Belgian nationality. The questionnaire was always administered at the beginning of a regular classroom session. The use of student samples for studying coo-effects is encouraged by Baughn and Yaprak [1] because of their homogeneous composition. In addition, several meta-analyses [16, 27] report that there are no significant differences in the estimates of coo-effects sizes between student and non-student samples. The questionnaire consisted of 4 sections. First, subjects indicated sex and age. The second section included a multi-item measure of subjects' feelings toward coo. The PANAS scale [28] served as a basis for operationalization. More in detail, it consists of 20 items that describe different emotional states. We limited ourselves to the 10 items referring to positive emotions. For each of these,

subjects had to indicate on a 7-point semantic differential scale how intensively they felt the item in question. The decision to limit ourselves to the use of items standing for positive feelings is based on our motivation to concentrate explicitly on the role of milder positive feelings toward coo. However, as will be pointed out later, after filling out the questionnaire, both samples were subdivided into a group of subjects expressing high positive feelings (137 cases for Spain vs. 194 cases for Denmark) toward coo and into another group of individuals showing less positive coo-specific feelings (134 cases for Spain vs. 74 cases for Denmark), based on the emotion scores.

The third section contained two 4-item scales measuring subjects' beliefs about DVD-player- and beer attributes. For each item, respondents had to indicate on a 7-point Likert scale ranging from 1 [definitely not agree] to 7 [fully agree] whether they believed the product to possess the attribute in question. For both products, the items (i.e., reliability, durability, performance and easiness of use for DVD-players and taste, naturalness, aroma and prestige for beer) were extracted from the coo-literature. Finally, subjects' evaluative judgement of DVD-players and beer was measured by means of a single-item 7-point semantic differential scale probing for the quality of the product.

Given are a collection of m respondents (observations) each comprised of an n-tuple of belief values (a_{k1}, a_{k2}, …, a_{kn}) called the arguments (i.e., reliability, durability, performance and easiness of use for DVD-players and taste, naturalness, aroma and prestige for beer), and an associated single value called the aggregated value (i.e., the quality of the product), which we shall denote as d_k.

Our goal will be to obtain an OWA operator, a weighting vector W that models the process of aggregation and its associated orness measure. We need a OWA operator, W, such that for a given group of respondents the following condition is satisfied as much as possible for any k:

$$f\left(a_{k1}, a_{k2}, ..., a_{kn}\right) = d_k \tag{7}$$

We shall relax this formulation by looking for a vector of OWA weights $W = [w_1 \; w_2 \ldots w_n]^{\mathrm{T}}$ that approximates the aggregation operator by minimizing the instantaneous errors e_k where

$$e_k = \tfrac{1}{2}\left(b_{k1}w_1 + b_{k2}w_2 + ... + b_{kn}w_n - d_k\right)^2 \tag{8}$$

The situation is complicated by the fact that the above minimization problem is a constrained optimization problem, since the OWA weights w_i have to satisfy the following two properties:

$$\sum_{i=1}^{n} w_i = 1; \quad \text{and} \quad w_i \in [0,1], \quad i = (1,...,n). \tag{9}$$

Therefore, the following transformation is introduced:

$$W_i = \frac{e^{\lambda_i}}{\sum_{j=1}^{n} e^{\lambda_j}} \tag{10}$$

From the above transformation it becomes clear that for any values of the parameters λ_i the weights w_i will be positive and will sum to 1. Therefore, the constrained minimization problem is transformed to the following unconstrained nonlinear programming problem:

Minimize the instantaneous errors e_k:

$$e_k = \frac{1}{2}\left(b_{k1} \frac{e^{\lambda_1}}{\sum_{j=1}^{n} e^{\lambda_j}} + b_{k2} \frac{e^{\lambda_2}}{\sum_{j=1}^{n} e^{\lambda_j}} + \ldots + b_{kn} \frac{e^{\lambda_n}}{\sum_{j=1}^{n} e^{\lambda_n}} - d_k \right)^2 \tag{11}$$

with respect to the parameters λ_i .

The gradient descent method was used to learn the weights [9]. See Torra [24] for an alternative estimation method.

5 Statistical Inference

The methodology described above measures the orness from a sample rather than a population and hence it is susceptible to random error. It would be interesting to infer statistically about the results based on a sample. Such inference could answer questions whether the orness (or any other similar measure) differs between different groups, to construct confidence intervals for the quantities under investigation and to test hypotheses for the population values. To our knowledge, however, there is no such technique for statistical inference available. Derivation of theoretical results is not simple because of the complicated nature of the measurements. For this reason, we base our statistical inference on resampling methods, namely non-parametric bootstrap. We construct confidence intervals for the orness measure based on non-parametric bootstrap. Bootstrap is a recently fashionable way for statistical inference for quantities for which theoretical and/or even asymptotic results are hard to derive. In these cases simulated inference based on bootstrap [7] is a key tool for inference. Each resample is analyzed exactly as if it were for the real data. To implement the non-parametric bootstrap, observations are sampled with replacement from the original data set until sample size is equal to that for the real data. These observations comprise the first bootstrap resample, denoted as X_1^*. The process is repeated a number of B times, and we end up with B resamples, denoted by $X_1^*, X_2^*, \ldots, X_B^*$. The key idea is that all these resamples can be considered as samples from the unknown population (or at least they look like the unknown population).

Now, denote the orness measure based on sample X_i^* as O_i^*. Hence if we calculate the orness (or any other measure) for all the B resamples, we have B realizations

of the quantity of interest $O_1^*, O_2^*, ..., O_B^*$, and in fact we have a random sample from the sampling distribution of this quantity. Hence, as the sample mean estimates the unknown population mean, we can estimate every quantity of interest based on those B values. By this approach, we can estimate variances, biases or any other quantity of interest including the construction of confidence intervals. The standard deviation for the orness will be simply the standard deviation of the values $O_1^*, O_2^*, ..., O_B^*$, i.e.

$$s(\hat{O}) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(O_i^* - \overline{O}\right)^2} \tag{12}$$

where $\overline{O} = \frac{1}{B}\sum_{i=1}^{B} O_i^*$.

There are several different ways to construct confidence intervals based on bootstrap values. We adapt the simple quantile based confidence intervals and hence a 95% confidence interval is constructed as $[k_{0.025}, k_{0.975}]$,where k_a is the a% sample quantile of the bootstrap values $O_1^*, O_2^*, ..., O_B^*$.

In a similar fashion, one can construct confidence intervals for any quantity of interest as for example for the w_i's. We emphasize that for the latter the standard approach to treat them as merely proportions is not correct as they are correlated proportions since they have to sum to one. Our bootstrap approach creates correct intervals in the sense that it takes into account the correlation structure that exists.

6 Results

Table 1-4 below present the results of the orness measure and the OWA weights (with standard errors between brackets) for Spanish/Danish DVD players and beer based on the outcome of the questionnaire. More specifically, these tables show the results for three groups of respondents. The first group is always the entire sample (616 cases for Spanish survey vs. 609 cases for Danish survey), whilst the second and third group are those respondents expressing respectively high positive feelings toward coo (137 cases for Spain vs. 194 cases for Denmark) and rather low positive feelings toward coo (134 cases for Spain vs. 74 cases for Denmark). Standard errors are based on B=1000 bootstrap replications using the procedure described above.

When comparing the group of respondents with high positive feelings toward coo (say group A) versus those expressing less positive feelings toward coo (say group B), table 1 (i.e., results for Spanish DVD-players) shows that the orness measure for group A is higher than for group B. When constructing 95% confidence intervals we found that for group A the interval is (0.439, 0.672), while for group B (0.347, 0.521), which implies a certain overlap. Statistically speaking, the differences between group A and B are not significant on a 5% level. According to our bootstrap results, it is however significant on the 10% although this decision depends on the bootstrap experiment used. Qualitatively, however, it is clear that group A has a larger orness, which somehow confirms our hypothesis that people expressing high positive feelings

Table 1. Results for Spanish DVD players

Data set	*Orness*	w_1	w_2	w_3	w_4
All cases (616)	0.4742	0.1338	0.2758	0.4695	0.1207
	(0.0259)	(0.0326)	(0.0858)	(0.0870)	(0.0382)
Group A: (137)	0.5499	0.1931	0.3866	0.2971	0.1230
	(0.0619)	(0.0682)	(0.2188)	(0.2191)	(0.0734)
Group B: (134)	0.4061	0.1495	0.2065	0.3567	0.2871
	(0.0556)	(0.0560)	(0.1350)	(0.1734)	(0.1124)
Significance	*S*	*NS*	*NS*	*NS*	*NS*

Table 2. Results for Spanish beer

Data set	*Orness*	w_1	w_2	w_3	w_4
All cases (616)	0.4290	0.1554	0.3171	0.1866	0.3407
	(0.0220)	(0.0299)	(0.0706)	(0.0798)	(0.0471)
Group A: (137)	0.4489	0.2015	0.1483	0.4452	0.2047
	(0.0521)	(0.0620)	(0.1568)	(0.1983)	(0.1038)
Group B: (134)	0.3438	0.1824	0.1907	0.1023	0.5243
	(0.0443)	(0.0666)	(0.1070)	(0.1217)	(0.0938)
Significance	*NS*	*NS*	*NS*	*S*	*S*

Table 3. Results for Danish DVD players

Data set	*Orness*	w_1	w_2	w_3	w_4
All cases (609)	0.5265	0.1901	0.3774	0.2544	0.1780
	(0.0215)	(0.0473)	(0.07370	(0.0651)	(0.0310)
Group A: (194)	0.5336	0.2491	0.2712	0.3107	0.1688
	(0.0504)	(0.0919)	(0.1652)	(0.1670)	(0.0595)
Group B: (74)	0.5133	0.1727	0.3005	0.4207	0.1060
	(0.0582)	(0.1030)	(0.1934)	(0.1867)	(0.0902)
Significance	*NS*	*NS*	*NS*	*NS*	*NS*

Table 4. Results for Danish beer

Data set	*Orness*	w_1	w_2	w_3	w_4
All cases (609)	0.4216	0.2099	0.1683	0.2983	0.3233
	(0.0204)	(0.0372)	(0.0732)	(0.0759)	(0.0399)
Group A: (194)	0.4166	0.2399	0.1223	0.2855	0.3522
	(0.0357)	(0.0699)	(0.1116)	(0.1205)	(0.0780)
Group B: (74)	0.3733	0.2266	0.0775	0.2849	0.4109
	(0.0548)	(0.07800	(0.1238)	(0.1406)	(0.0867)
Significance	*NS*	*NS*	*NS*	*NS*	*NS*

toward coo tend to use a more optimistic evaluation function toward evaluating the quality of Spanish DVD-players. In other words, they tend to base their quality evaluation more on the more positively evaluated attributes.

Confirmation of the encoding-specificity principle should, however, also be reflected by the individual OWA weights (w_1 to w_4) such that for group A versus group B, the ordered weights w_1 and w_2 should show higher values and the ordered weights w_3 and w_4 should show lower values. Based on results depicted in table 1 it can be observed that indeed w_1 and w_2 are higher in group A compared to group B. However, their individual differences are not statistically significant. Similarly, it can be seen from the values for w_3 and w_4 that they are higher in group B compared to group A, although their individual differences are again not statistically significant.

Table 2 presents the results obtained for Spanish beer. Also in this case, the orness measure for group A is surpassing that for group B, although in this case the difference is not statistically significant. The 95% confidence interval for group A is (0.351, 0.555) while for group B (0.262, 0.435). Yet, there is a clear indication that group A has a larger orness. This can again be seen as supportive evidence for our hypothesis. Thus, one could conclude that respondents with high positive coo-feelings tend to base their quality evaluation of Spanish beer rather on the more favourably evaluated attributes. However, in this case the results for the weight values are less convincing since the value of *w2* is larger in group B than in group A, and the value of w_3 is larger in group A than in group B.

Table 3 and 4 show the results for Danish DVD-players and beer. Although there is a tendency that the orness is again slightly higher for group A than for group B, the differences are much smaller compared to the results for Spain and not statistically significant. For example, for Danish DVD-players, the 95% confidence interval for group A is (0.435, 0.626) and for group B (0.402, 0.641), showing a large overlap. With respect to the values of w_1 to w_4 the results are not consistent.

Overall, it is interesting to observe that we can find much more evidence for our hypothesis in the case of Spanish products compared to Danish products.

7 Conclusion

From a practical point of view, our study shows how milder coo-specific feelings serve as a useful device for advertisers to direct consumers' processing of attribute beliefs. More in detail, their functioning can be understood as some kind of encoding-specificity mechanism. That is, consumers during their product evaluation ascribe most importance to those attribute beliefs which are closer in line with their internal affective state. Interestingly, support for our hypothesis was somewhat more substantial for the Spanish than for the Danish survey. Thus, the type of country seems to play a role in determining to what extent the encoding-specificity mechanism manifests itself.

From a technical point of view, we opted for an alternative methodology in using the OWA-operator. In our opinion, this is a useful approach while the interpretation of the OWA-weights is more straightforward compared to the more complex LISREL-models as they have been traditionally used for instance by Han [12]. An additional advantage lies in the fact that the 'orness measure' gives us the needed quantification of the optimistic degree of an evaluation. This aspect alone is already a huge advantage of the fuzzy set approach compared to the more traditional LISREL approaches where this degree of optimism cannot be extracted from the data. Finally,

we introduced a bootstrap procedure to estimate the orness and the level of uncertainty around it. This enables us to construct confidence intervals and conduct hypothesis tests. As far as we know, estimating this degree of uncertainty of the orness measure has never been introduced in the literature before.

References

1. Baughn, C.C., Yaprak, A.: Mapping Country-of-Origin Research: Recent Developments and Emerging Avenues. In: Papadopoulos, N., Heslop, L.A. (eds.): Product-Country Images: Importance and Role in International Marketing. International Business Press, New York (1993)
2. Burke, M.C., Edell, J.A.:The Impact of Feelings on Ad-Based Affect and Cognition. Journal of Marketing Research, 26 (February) (1989) 69-83
3. Cacioppo, J.T., Petty, R.E.: The Elaboration Likelihood Model: The Role of Affect and Affect-Laden Information Processing in Persuasion. In: Cafferata, P., Tybout, A.M. (eds.): Cognitive and Affective Responses to Advertising. Lexington Books, Toronto (1989)
4. Cafferata, P., Tybout, A.M.: Cognitive and Affective Responses to Advertising. Lexington Books, Toronto (1989)
5. Derbaix, C., Pham, M.T.: For the development of measures of emotion in marketing: Summary of prerequisites. In: Lambkin, M., Foxall, G., Van Raaij, F., Heilbrunn, B. (eds.): European Perspectives on Consumer Behaviour, Prentice Hall, London (1998)
6. Eagly, A.H., Chaiken, S.: The Psychology of Attitudes. Forth Worth: Harcourt Brace Jovanovich (1993)
7. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall. (1993)
8. Fernández Salido J.M., Murakami S.: Extending Yager's orness concept for the OWA aggregators to other mean operators. Fuzzy Sets and Systems, 139(3) (2003) 515-542
9. Filev D. and Yager R.R.: On the issue of obtaining OWA operator weights. Fuzzy Sets and Systems, 94 (1998) 157-169
10. Fishbein, M., Ajzen, I.: Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Reading, MA: Addison-Wesley (1975)
11. Friedman, M.: Consumer Boycotts: Effecting Change Through the Marketplace and the Media. Routledge, London (1999)
12. Han, C.M.: The Role of Consumer Patriotism in the Choice of Domestic versus Foreign Products. Journal of Advertising Research, 28 (3) (1988) 25-32
13. Holbrook, M.B., Batra, R.: Assessing the Role of Emotions as Mediators of Consumer Responses to Advertising. Journal of Consumer Research, 14 (December) (1987) 404-420
14. Isen, A.M.: Some Ways in Which Affect Influences Cognitive Processes: Implications for Advertising and Consumer Behavior. In Cafferata, P., Tybout, A.M. (eds.): Cognitive and Affective Responses to Advertising, Lexington Books, Toronto (1989)
15. Klein, J.G., Ettenson, R., Morris, M.D.: The Animosity Model of Foreign Product Purchase: An Empirical Test in the People's Republic of China. Journal of Marketing, 62 (1998) 89-100
16. Liefeld, J.P.: Experiments on Country-of-Origin Effects: Review and Meta-Analysis of Effect Size. In: Papadopoulos, N., Heslop, L.A. (eds.): Product-Country Images: Importance and Role in International Marketing, International Business Press, New York (1993)
17. Obermiller, C., Spangenberg, E.: Exploring the Effects of Country of Origin Labels: An Information Processing Framework. Advances in Consumer Research, 16 (1989) 454-459

18. Papadopoulos, N., Heslop, L.A.: Product-Country Images: Importance and Role in International Marketing, International Business Press, New York (1993)
19. Peter, J.P., Olson, J.C., Grunert, K.G.: Consumer Behaviour and marketing Strategy: European Edition, McGraw Hill, London (1999)
20. Peterson, R.A., Hoyer, W.D., Wilson, R.W.: The Role of Affect in Consumer Behavior: Emerging Theories and Applications. Lexington Books, Toronto (1986)
21. Shimp, T.A., Sharma, S.: Consumer Ethnocentrism: Construction and Validation of the CETSCALE. Journal of Marketing Research, 24 (August) (1987) 280-289
22. Stayman, D.M., Aaker, D.A.: Are All the Effects of Ad-Induced Feelings Mediated by Aad?. Journal of Consumer Research, 15 (December) (1988) 368-373
23. Torra V.: The weighted OWA operator. International Journal of Intelligent Systems, 12 (1997) 153-166
24. Torra V.: Learning weights for the quasi-weighted means. IEEE Transactions on Fuzzy Systems, 10(5) (2002) 653-666
25. Tulving, E., Thomson, D.M.: Encoding specificity and retrieval processes in episodic memory. Psychological Review, 80 (1973) 352-373
26. Verlegh, P.W.J.: Country-of-Origin Effects on Consumer Product Evaluations. Doctoral Dissertation, University of Wageningen (2001)
27. Verlegh, P.W.J., Steenkamp, J.-B.E.M.: A review and meta-analysis of country-of-origin research. Journal of Economic Psychology, 20 (5) (1999) 521-546
28. Watson, D., Clark, L.A., Tellegen, A.: Development and Validation of Brief Measures of positive and Negative Affect: The PANAS Scales. Journal of Personality and Social Psychology, 54 (6) (1988) 1063-1070
29. Yager, R.R.: An ordered weighted averaging operators in multicriteria decision making. IEEE Transactions on Systems, Man and Cybernetics, 18 (1988) 183-190
30. Yager, R.R. and Kacprzyck J.: The ordered weighted averaging operator, Theory and Applications, Kluwer Academic Publisher: Boston Dordecht London (1997)

Non-monotonic Fuzzy Measures and Intuitionistic Fuzzy Sets

Yasuo Narukawa[1] and Vicenç Torra[2]

[1] Toho Gakuen, 3-1-10 Naka,
Kunitachi, Tokyo, 186-0004 Japan
narukawa@d4.dion.ne.jp
[2] Institut d'Investigació en Intel·ligència Artificial,
Campus de Bellaterra, 08193 Bellaterra, Catalonia, Spain
vtorra@iiia.csic.es

Abstract. Non-monotonic fuzzy measures induced by an intuitinistic fuzzy set are introduced. Then, using the Choquet integral with respect to the non-monotonic fuzzy measure, the weighted distance between two intuitionistic fuzzy sets is defined. As it will be shown here, under some conditions, the weighted distance coincides with the Hamming distance.

Keywords: Fuzzy measure, Non-monotonic fuzzy measure, Choquet integral, Intuitionistic fuzzy sets,Hamming distance.

1 Introduction

The so-called [illegible] were proposed by Atanassov [1, 2, 3] to have additional degrees of freedom when defining the membership values in a fuzzy set. Since then, the theory has been developed. Several new concepts and methods have been introduced and studied.

Fuzzy measures and fuzzy integrals are basic tools for decision modeling. Fuzzy integrals can be used to combine the information supplied by different information sources or to integrate the evaluation of different criteria. In this setting, fuzzy measures are used to represent the basic information about the sources (, their importance).

Although fuzzy measures are, typically, monotonic set functions on the unit interval, non-monotonic fuzzy measures have been also considered in the literature. See [8, 10, 11, 14]. In this paper we establish some relationships between non-monotonic fuzzy measures and intuitionistic fuzzy sets.

We show that non-monotonic fuzzy measures can be defined from intuitionistic fuzzy sets. Thus, given an intuitionistic fuzzy set, we will consider the fuzzy measure induced by it. Then, we will study some properties that establish relationships between intuitionistic fuzzy sets and (non-monotonic) fuzzy measures. The concept of bounded variation [4, 10, 11] (either positive or negative variation) plays a central role in such properties.

The structure of the paper is as follows. In Section 2, we present some preliminaries that are needed later on in this paper. In Section 3, we review the concepts

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 150–160, 2006.

of bounded variation and we present some results on the Choquet integrals of non-monotonic fuzzy measures. In Section 4, we introduce non-monotonic fuzzy measures induced by intuitionistic fuzzy sets. Using the Choquet integral with respect to the non-monotonic fuzzy measure, the weighted distance between two intuitionistic fuzzy sets is defined. We show that under some conditions, the weighted distance coincides with the Hamming distance. The paper finishes with some conclusions.

2 Preliminaries

In this section, we review some preliminary definitions and propositions on fuzzy measures and intuitionistic fuzzy sets. In the following, we will use the following notation. Let X be an universal set and let $\mathcal{X}$ be σ-algebra of X. That is, $(X, \mathcal{X})$ is a measurable space.

Definition 1. [12] Let $(X, \mathcal{X})$ be a measurable space. A [illegible] m is a real valued set function, $m : \mathcal{X} \longrightarrow R^+$ with the following properties;

(1) $m(\emptyset) = 0$
(2) $m(A) \leq m(B)$ whenever $A \subset B$, $A, B \in \mathcal{X}$.

We say that the triplet $(X, \mathcal{X}, m)$ is a fuzzy measure space if m is a fuzzy measure.

We will use $\mathcal{F}(X)$ to denote the class of non-negative measurable functions. That is,

$$\mathcal{F}(X) := \{f | f : X \to R^+, f : \text{measurable}\}$$

Definition 2. [5, 9] Let $(X, \mathcal{X}, m)$ be a fuzzy measure space. [illegible] of $f \in \mathcal{F}(X)$ with respect to m is defined by

$$(C)\int f dm = \int_0^\infty m_f(r)dr,$$

where $m_f(r) = m(\{x | f(x) \geq r\})$.

Definition 3. [6] Let $f, g \in \mathcal{F}(X)$. Then, we say that f and g are [illegible] if

$$f(x) < f(x') \Rightarrow g(x) \leq g(x')$$

for $x, x' \in X$.

Proposition 4. [illegible] $(X, \mathcal{X}, m)$ [illegible] $f, g \in \mathcal{F}(X)$ [illegible]

$$(C)\int (f + g)dm = (C)\int f dm + (C)\int g dm.$$

We say that the additivity of the Choquet integral, according to this property, is comonotonic additivity.

Next, we define an intuitionistic fuzzy set by Attanassov (for conciseness, we denote them by A-IFS).

Definition 5. A X

$$A := \{< x, \mu_A(x), \nu_A(x) > | x \in X\}$$

$\mu_A : X \to [0,1]$ $\nu_A : X \to [0,1]$

$$0 \leq \mu_A(x) + \nu_A(x) \leq 1.$$

x $\mu_A(x)$ $\nu_A(x)$ $x \in X$ $A \subset X$

$$\pi_A(x) := 1 - \mu_A(x) - \nu_A(x).$$

Suppose that X is a finite set, that is, $X := \{x_1, x_2, \ldots, x_n\}$. The Hamming distance between two A-IFS are proposed by Szmidt and Kacprzyk.

Definition 6. $X := \{x_1, x_2, \ldots, x_n\}$ $A := \{< x, \mu_A(x), \nu_A(x) > | x \in X\}$ $B := \{< x, \mu_B(x), \nu_B(x) > | x \in X\}$

(1) $d_{IFS}(A,B)$ A B

$$d_{IFS}(A,B) := \sum_{i=1}^{n} (|\mu_A(x_i) - \mu_B(x_i)| + |\nu_A(x_i) - \nu_B(x_i)| + |\pi_A(x_i) - \pi_B(x_i)|)$$

(2) $l_{IFS}(A,B)$ A B

$$l_{IFS}(A,B) := \sum_{i=1}^{n} \frac{1}{2n} (|\mu_A(x_i) - \mu_B(x_i)| + |\nu_A(x_i) - \nu_B(x_i)| + |\pi_A(x_i) - \pi_B(x_i)|)$$

3 Non-monotonic Fuzzy Measure and Integral

Now, we turn into non-monotonic fuzzy measures and we show that the Choquet integral with respect to a non-monotonic fuzzy measure is comonotonically additive.

Definition 7. $(X, \mathcal{X})$ $\mathcal{X}$ $m(\emptyset) = 0$ $(X, \mathcal{X}, m)$ m

Definition 8. $(X, \mathcal{X}, m)$ $m^+(A)$ m $A \in \mathcal{X}$

$$m^+(A) = \sup\{\sum_{i=1}^{n} \max\{m(A_i) - m(A_{i-1}), 0\}\}$$

$$\emptyset = A_0 \subset A_1 \subset \cdots \subset A_n = A, A_i \in \mathcal{X}, i = 1, 2, \cdots n,$$

$m^-(A)$ m $A \in \mathcal{X}$

$$m^-(A) = \sup\{\sum_{i=1}^{n} \max\{m(A_{i-1}) - m(A_i), 0\}\}$$

$$\emptyset = A_0 \subset A_1 \subset \cdots \subset A_n = A, A_i \in \mathcal{X}, i = 1, 2, \cdots n$$

$|m|(A)$ m $A \in \mathcal{X}$

$$|m|(A) = m^+(A) + m^-(A).$$

$$m(A) = m^+(A) - m^-(A)$$

$A \in \mathcal{X}$ $|m|(X)$ $\|m\|$ m $\|m\| < \infty$.

Definition 9. $f \in \mathcal{F}(X)$ m

$$(C)\int f dm = \int_0^\infty m^+(\{x|f(x) \geq a\})da - \int_0^\infty m^-(\{x|f(x) \geq a\})da.$$

Since $m = m^+ - m^-$, the Choquet integral $C_m(f)$ is written by

$$C_m(f) := (C)\int f dm = \int_0^\infty m(\{x|f(x) \geq a\})da.$$

Let $f, g \in \mathcal{F}(X)$ be comonotonic. Then, the Choquet integrals with respect to m^+ and m^- are comonotonically additive. Therefore, the next proposition holds.

Proposition 10. m

Let $(X, \mathcal{X}, m)$ be a non monotonic fuzzy measure space and let $f, g \in \mathcal{F}(X)$ be comonotonic. Then,

$$\begin{aligned}(C)\int (f+g)dm &= (C)\int (f+g)dm^+ - (C)\int (f+g)dm^- \\ &= (C)\int fdm^+ + (C)\int gdm^+ - ((C)\int fdm^- + (C)\int gdm^-) \\ &= (C)\int fdm^+ - (C)\int fdm^- + (C)\int gdm^+ - (C)\int gdm^- \\ &= (C)\int fdm + (C)\int gdm. \square\end{aligned}$$

Let $\mathcal{A}$ be a chain of subsets of X, that is,

$$\mathcal{A} := \{A_i | i = 1, 2, \ldots, n, A_i \subset X, \emptyset \subset A_1 \subset \cdots \subset A_n = X\}.$$

Since 1_{A_i} and 1_{A_j} are comonotonic for every $i, j = 1, 2, \ldots n$ where 1_A is a characteristic function of A, we have

$$C_m(\sum_{i=1}^{n} a_i 1_{A_i}) = \sum_{i=1}^{n} a_i m(A_i)$$

for $a_i \geq 0$.

4 Non-monotonic Fuzzy Measure Induced by Intuitionistic Fuzzy Set

Let m be a non-monotonic fuzzy measure on X satisfying

$$0 \leq m(\{x\}) + m(X \setminus \{x\}) \leq 1, m(\{x\}) \geq 0, m(X \setminus \{x\}) \geq 0.$$

We can define an A-IFS $A := \{< x, \mu_A(x), \nu_A(x) > | x \in X\}$ by $\mu_A(x) := m(\{x\})$ and $\nu_A(x) := m(X \setminus \{x\})$. Conversely we can define a non-monotonic fuzzy measure from an A-IFS.

Definition 11. Let $A := \{< x, \mu_A(x), \nu_A(x) > | x \in X\}$ be an A-IFS. We define a non-monotonic fuzzy measure $\overline{m}_A : 2^X \to [0, 1]$ by

$$\overline{m}_A(B) = \begin{cases} 0 & \text{if } B = \emptyset \\ \sup_{y \in B} \mu_A(y) & \text{if } B \subsetneqq X \setminus \{x\} \text{ for all } x \\ \nu(x) & \text{if for some } x, B = X \setminus \{x\} \\ \inf_{x \in X} \sup_{y \in X \setminus \{x\}} \mu_A(y) & \text{if } B = X, \end{cases}$$

and a non-monotonic fuzzy measure $\underline{m}_A : 2^X \to [0, 1]$ by

$$\underline{\mathrm{m}}_A(B) = \begin{cases} 0 & \text{if } B = \emptyset \\ \sup_{y \in B} \nu_A(y) & \text{if } B \subsetneqq X \setminus \{x\} \text{ for all } x \\ \mu(x) & \text{if for some } x, B = X \setminus \{x\} \\ \inf_{x \in X} \sup_{y \in X \setminus \{x\}} \nu_A(y) & \text{if } B = X. \end{cases}$$

$\overline{m}_A$ A A A

Let A and B be an A-IFS. Then, we define the following non-monotonic fuzzy measures for $C \subset X$:

$$(\overline{m}_A - \overline{m}_B)(C) := \overline{m}_A(C) - \overline{m}_B(C),$$

$$|\overline{m}_A - \overline{m}_B|(C) := |\overline{m}_A(C) - \overline{m}_B(C)|,$$

$$(\underline{m}_A - \underline{m}_B)(C) := \underline{m}_A(C) - \underline{m}_B(C),$$

$$|\underline{m}_A - \underline{m}_B|(C) := |\underline{m}_A(C) - \underline{m}_B(C)|.$$

The next lemma follows from the definition of a positive variation and a negative variation.

Lemma 12. $A := \{< x, \mu_A(x), \nu_A(x) > | x \in X\}$ $\overline{m}_A$ A

$$\overline{m}_A^+(B) = \begin{cases} 0 & \text{if } B = \emptyset \\ \sup_{y \in B} \mu_A(y) & \text{if for some } x \in X, B \subsetneq X \setminus \{x\} \\ \sup_{y \in C, C \subsetneq B} \mu_A(y) & \text{if for some } x \in X, \quad B = X \setminus \{x\} \\ & \text{and } \sup_{y \in C, C \subsetneq B} \mu_A(y) > \nu_A(x) \\ \nu_A(x) & \text{if for some } x \in X, \quad B = X \setminus \{x\} \\ & \text{and } \sup_{y \in C, C \subsetneq B} \mu_A(y) \leq \nu_A(x) \\ \nu_A(x) & \text{if } B = X \text{ and for some } x \in X \\ & \inf_{x \in X} \sup_{y \in X \setminus \{x\}} \mu_A(y) \leq \nu_A(x) \\ \inf_{x \in X} \sup_{y \in X \setminus \{x\}} \mu_A(y) - \nu_A(x) & \\ \quad + \sup_{y \in C, C \subsetneq X \setminus \{x\}} \mu_A(y) & \text{if } B = X \text{ and for some } x \in X, \\ & \sup_{y \in C, C \subsetneq X \setminus \{x\}} \mu_A(y) \geq \nu(x) \\ \inf_{x \in X} \sup_{y \in X \setminus \{x\}} \mu_A(y) & \text{if } B = X \text{ and for some } x \in X, \\ & \inf_{x \in X} \sup_{y \in X \setminus \{x\}} \mu_A(y) \geq \nu(x) \\ & \text{and } \nu(x) \geq \sup_{y \in C, C \subsetneq X \setminus \{x\}} \mu_A(y) \end{cases}$$

$$\overline{m}_A^-(B) = \begin{cases} 0 & \text{if } B \subsetneq X \setminus \{x\} \text{ for all } x \\ 0 & \text{if for some } x \in X, B = X \setminus \{x\} \\ & \text{and } \sup_{y \in C, C \subsetneq B} \mu_A(y) \le \nu_A(x) \\ \sup_{y \in C, C \subsetneq B} \mu_A(y) - \nu_A(x) & \text{if for some } x \in X, B = X \setminus \{x\} \\ & \text{and } \sup_{y \in C, C \subsetneq B} \mu_A(y) \ge \nu_A(x) \\ \nu_A(x) - \inf x \in X \sup_{y \in X \setminus \{x\}} \mu_A(y) & \text{if } B = X, \text{ and for some } x \in X, \\ & \inf_{x \in X} \sup_{y \in X \setminus \{x\}} \mu_A(y) \le \nu_A(x) \\ \sup_{y \in C, C \subsetneq X \setminus \{x\}} \mu_A(y) \ - \nu_A(x) & \text{if } B = X, \text{ and for some } x \in X, \\ & \sup_{y \in C, C \subsetneq X \setminus \{x\}} \mu_A(y) \ge \nu_A(x) \\ 0 & \text{if } B = X, \text{ and and and for some } x \in X, \\ & \inf_{x \in X} \sup_{y \in X \setminus \{x\}} \mu_A(y) \ge \nu(x) \\ & \text{and } \nu(x) \ge \sup_{y \in C, C \subsetneq X \setminus \{x\}} \mu_A(y) \end{cases}$$

Since $||\overline{m}_A|| = \overline{m}_A^+(X) + \overline{m}_A^-(X)$, we have the next proposition.

Proposition 13. *A* $\overline{m}_A$ *A*

Since

$$\begin{aligned} |\overline{m}_A - \overline{m}_B| &= |\overline{m}_A^+ - \overline{m}_A^- + \overline{m}_B^+ - \overline{m}_B^-| \\ &\le |\overline{m}_A^+| + |\overline{m}_A^-| + |\overline{m}_B^+| + |\overline{m}_B^-|, \end{aligned}$$

we have the next corollary.

Corollary 14. *A B* $\overline{m}_A - \overline{m}_B$ $|\overline{m}_A - \overline{m}_B|$ $\underline{m}_A - \underline{m}_B$ $|\underline{m}_A - \underline{m}_B|$

It follows from Proposition 13 that we can define the Choquet integral with respect to a non-monotonic fuzzy measure induced by an A-IFS.

Definition 15. *A B* $f\ g\ h \in \mathcal{F}(X)$ $wdist_{f,g,h}$ *A B*

$$wdist_{f,g,h}(A,B) := C_{|\overline{m}_A - \overline{m}_B|}(f) + C_{|\underline{m}_A - \underline{m}_B|}(g) + C_{|(\overline{m}_A + \underline{m}_A) - (\overline{m}_B + \underline{m}_B)|}(h)$$

The weighted distance can be defined not only when X is a finite set, but also when X is infinite.

The next proposition immediately follows from this definition.

Proposition 16. *Let A B be C $f\ g\ h \in \mathcal{F}(X)$*

(1) $wdist_{f,g,h}(A, A) = 0$
(2) $wdist_{f,g,h}(A, B) = wdist_{f,g,h}(B, A)$
(3) $wdist_{f,g,h}(A, B) + wdist_{f,g,h}(B, C) \leq wdist_{f,g,h}(A, C)$
(4) *If $f > 0, g > 0$ $h > 0$ $wdist_{f,g,h}(A, B) = 0$ then $A = B$*

In the following suppose that X is a finite set, that is, $X := \{x_1, x_2, \ldots, x_n\}$.

The next lemma follows from the definition of comonotonicity.

Lemma 17. *Let $X := \{x_1, x_2, \ldots, x_n\}$ $f : X \to R$ $g : X \to R$ be comonotonic functions. Then*

(1) *$f(x_k) = \max_{x \in C} f(x)$ $g(x_k) = \max_{x \in C} g(x)$ $k = argmax_{x \in C} g(x)$ $C \subset X$*
(2) *$f(x_k) = \min_{x \in X} \max_{y \in X \setminus \{x\}} f(y)$ $g(x_k) = \min_{x \in X} \max_{y \in X \setminus \{x\}} g(y)$*

Proof. (1) Let $k = argmax_{x \in C} f(x)$. Since $|\{f(x) | x \in C\}| = n$, if $x \neq x_k$ then $f(x) < f(x_k)$.
Since f and g are comonotonic, $g(x) \leq g(x_k)$ for all $x \in C$.
Therefore $g(x_k) = max_{x \in C} g(x)$.

(2) Since $f(x_k) = \min_{x \in X} \max_{y \in X \setminus \{x\}} f(y)$, there exists $x_i \in X$ such that $f(x_k) = \max_{y \in X \setminus \{x_i\}} f(y)$. Then applying (1) we have $g(x_k) = \max_{y \in X \setminus \{x_i\}} g(y)$. Therefore $g(x_k) \geq \min_{x \in X} \max_{y \in X \setminus \{x\}} g(y)$.
Since for all $x \in X$

$$f(x_k) \leq \max_{y \in X \setminus \{x\}} f(y),$$

there exists $y \in X \setminus \{x_k\}$ such that $f(x_k) < f(y)$ since $y \neq x_k$. Then we have $g(x_k) \leq g(y)$, that is $g(x_k) \leq \max_{y \in X \setminus \{x_k\}} g(y)$. Therefore $g(x_k) \leq \max_{y \in X \setminus \{x\}} g(y)$ for all $x \in X$. that is, $g(x_k) \leq \min_{x \in X} \max_{y \in X \setminus \{x\}} g(y)$.

□

Choosing the classes $\mathcal{C}, \mathcal{D}, \mathcal{E}$ of subsets of X suitably, using the previous lemma, we have the next proposition.

Proposition 18. *Let A B be*

$$A := \{< x, \mu_A(x), \nu_A(x) > | x \in X\},$$

$$B := \{< x, \mu_B(x), \nu_B(x) > | x \in X\},$$

μ_A μ_B ν_A ν_B $\mu_A + \nu_A$ $\mu_A + \nu_B$ be comonotonic functions. Then $\mu_A, \mu_B, \mu_A + \nu_A$ are comonotonic.

Let $\mathcal{C} := \{C_i\}$ $\mathcal{D} := \{D_i\}$ $\mathcal{E} := \{E_i\}$ be partitions of X such that

$$wdist_{f,g,h}(A, B) = \sum_{i=1}^{n} (a_i | \sup_{x \in C_i} \mu_A(x) - \sup_{x \in C_i} \mu_B(x) | + b_i | \sup_{x \in D_i} \nu_A(x) - \sup_{x \in D_i} \nu_B(x) |$$
$$+ c_i | \sup_{x \in E_i} \pi_A(x) - \sup_{x \in Ei} \pi_B(x) |)$$

where f, g, h are comonotonic functions such that $a_i \geq 0, b_i \geq 0, c_i \geq 0$ and $f := \sum_i a_i 1_{C_i}, g := \sum_i b_i 1_{D_i}, f := \sum_i c_i 1_{E_i}$

. Let $C'_i := \{x_1, x_2, \ldots, x_i\}$, $f := \sum_{i=1}^n a'_i 1_{C'_i}$, $g := \sum_{i=1}^n b'_i 1_{C'_i}$ and $h := \sum_{i=1}^n c'_i 1_{C'_i}$ with $a'_i \geq 0, b'_i \geq 0, c_i \geq 0$.

Since each $1_{C'_i}$ and $1_{C'_j}$ are comonotonic,

$$\begin{aligned}
C_{|\overline{m}_A - \overline{m}_B|}(f) &= \sum_{i=1}^n C_{|\overline{m}_A - \overline{m}_B|}(a'_i 1_{C'_i}) \\
&= \sum_{i=1}^n a'_i |\overline{m}_A - \overline{m}_B|(C'_i) \\
&= \sum_{i=1}^n a'_i |\overline{m}_A(C'_i) - \overline{m}_B(C'_i)| \\
&= \sum_{i=1}^{n-2} a'_i |\sup_{x \in C'_i} \mu_A(x) - \sup_{x \in C'_i} \mu_B(x)| \\
&+ a'_{n-1} |\nu_A(x_n) - \nu_B(x_n)| + a'_n |\min_{x \in X} \max_{y \in X \setminus \{x\}} \mu_A(y) - \min_{x \in X} \max_{y \in X \setminus \{x\}} \mu_B(y)| \\
&= \sum_{i=1}^{n-2} a'_i |\sup_{x \in C'_i} \mu_A(x) - \sup_{x \in C'_i} \mu_B(x)| \\
&+ a'_{n-1} |\nu_A(x_n) - \nu_B(x_n)| + a'_n |\mu_A(x_{n-1}) - \mu_B(x_{n-1})|,
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
C_{|\underline{m}_A - \underline{m}_B|}(g) &= \sum_{i=1}^n C_{|\underline{m}_A - \underline{m}_B|}(b'_i 1_{D_i}) \\
&= \sum_{i=1}^n b'_i |\underline{m}_A - \underline{m}_B|(D_i) \\
&= \sum_{i=1}^n b'_i |\underline{m}_A(D_i) - \underline{m}_B(D_i)| \\
&= \sum_{i=1}^{n-1} b'_i |\sup_{x \in D_i} \nu_A(x) - \sup_{x \in D_i} \nu_B(x)| \\
&+ b'_{n-1} |\mu_A(x'_n) - \mu_B(x'_n)| + b'_n |\nu_A(x'_{n-1}) - \nu_B(x'_{n-1})|,
\end{aligned}$$

$$\begin{aligned}
C_{|(\overline{m}_A + \underline{m}_A) - (\overline{m}_B + \underline{m}_B)|}(f) &= \sum_{i=1}^n c'_i C_{|(\overline{m}_A + \underline{m}_A) - (\overline{m}_B + \underline{m}_B)|}(1_{E_i}) \\
&= \sum_{i=1}^n c'_i |(\overline{m}_A + \underline{m}_A) - (\overline{m}_B + \underline{m}_B)|(E_i) \\
&= \sum_{i=1}^n c'_i |(\overline{m}_A(E_i) + \underline{m}_A(E_i)) - (\overline{m}_B(E_i) + \underline{m}_B(E_i))|.
\end{aligned}$$

Changing coefficients a_i and the menber of the class $\mathcal{C}, \mathcal{D}, \mathcal{E}$ suitably, we have the concluding equality. □

Using the previous Proposition, we have the next proposition.

Proposition 19. $A := \{< x, \mu_A(x), \nu_A(x) > | x \in X\}$ $B := \{< x, \mu_B(x), \nu_B(x) > | x \in X\}$ μ_A μ_B ν_A ν_B $\mu_A + \nu_A$ $\mu_A + \nu_B$ $\mu_A, \nu_A, \mu_A + \nu_A$ f g h X

$$wdist_{f,g,h}(A,B) = \sum_{i=1}^{n} a_i(|\mu_A(x_i) - \mu_B(x_i)| + b_i|\nu_A(x_i)| - \nu_B(x_i)| + c_i|\pi_A(x) - \pi_B(x_i)|)$$

$a_i \geq 0, b_i \geq 0, c_i \geq 0$

Define f, g, h such that $a_i = b_i = c_i = 1$ for all i, we have the next corollary.

Corollary 20. $A := \{< x, \mu_A(x), \nu_A(x) > | x \in X\}$ $B := \{< x, \mu_B(x), \nu_B(x) > | x \in X\}$ μ_A μ_B ν_A ν_B $\mu_A + \nu_A$ $\mu_A + \nu_B$ $\mu_A, \nu_A, \mu_A + \nu_A$ f g h X

$$wdist_{f,g,h}(A,B) = d_{IFS}(A,B),$$

$wdist$

Let $C_i := \{x_1, x_2, \ldots, x_i\}, i = 1, 2, \ldots, n$ and let $f := \sum_{i=1}^{n} 1_{C_i}$, $g := \sum_{i=1}^{n} 1_{C_i}$ and $h := \sum_{i=1}^{n} 1_{C_i}$.

Then, it follows from the proof of Proposition 12,

$$C_{|\overline{m}_A - \overline{m}_B|}(f) = \sum_{i=1}^{n-2} |\mu_A(x_i) - \mu_B(x_i)| + |\nu_A(x_n) - \nu_B(x_n)| + |\mu_A(x_{n-1}) - \mu_B(x_{n-1})|,$$

$$C_{|\underline{m}_A - \underline{m}_B|}(f) = \sum_{i=1}^{n-2} |\nu_A(x_i) - \nu_B(x_i)| + |\nu_A(x_n) - \nu_B(x_n)| + |\mu_A(x_{n-1}) - \mu_B(x_{n-1})|,$$

and

$$C_{|\overline{m}_A + \underline{m}_A - \overline{m}_B - \underline{m}_B|}(f) = \sum_{i=1}^{n} |\pi_A(x_i) - \pi_B(x_i)|.$$ □

Define f, g, h such that the form of f is $f := \sum_{i=1}^{n} (1/n) 1_{C_i}$, then we have the next corollary.

Corollary 21. $A := \{< x, \mu_A(x), \nu_A(x) > | x \in X\}$ $B := \{< x, \mu_B(x), \nu_B(x) > | x \in X\}$ $B := \{< x, \mu_B(x), \nu_B(x) > | x \in X\}$ μ_A μ_B ν_A ν_B $\mu_A + \nu_A$ $\mu_A + \nu_B$ $\mu_A, \nu_A, \mu_A + \nu_A$ f g h X

$$wdist_{f,g,h}(A, B) = l_{IFS}(A, B).$$

5 Conclusions

In this paper we have proposed the definition of non-monotonic fuzzy measures in terms of intuitionistic fuzzy sets. We have seen that the Choquet integral of non-monotonic fuzzy measures permits to define the weighted distance between two intuitionistic fuzzy sets. We have also shown that under some conditions the weighted distance can be made equal to the Hamming distance.

References

1. K.T. Atanassov, Intuitionistic fuzzy sets, VII ITKR's Session, Sofia, 1983 (deposed in Central Sci.-Technical Library of Bulg. Acad. of Sci., 1697/84) (in Bulgarian).
2. K.T. Atanassov, Intuitionistic fuzzy sets, Fuzzy Sets and Systems 20 (1986) 87-96.
3. K.T. Atanassov, Intuitionistic Fuzzy Sets, Physica-Verlag, Heidelberg, New York, 1999.
4. R.J.Aumann and L. S. Shapley, Values of Non-atomic Games, Princeton Univ. Press, 1974.
5. G. Choquet, Theory of capacities. Ann. Inst. Fourier, Grenoble, 5 (1954/1955) 131-295.
6. C. Dellacherie, Quelques commentaires sur les prolongements de capacités, Séminaire de Probabilités 1969/1970, Strasbourg, Lecture Notes in Mathematics, 191 (1971) 77-81.
7. D. Denneberg, Non additive measure and Integral, Kluwer Academic Publishers, Dordorecht, 1994.
8. Y. Rébillé, Sequentially continuous non-monotonic Choquet integrals, Fuzzy Sets and Systems, 153:1 (2005) 79-94.
9. T. Murofushi and M. Sugeno, An interpretation of fuzzy measures and the Choquet integral as an integral with respect to a fuzzy measure, Fuzzy Sets and Systems, 29 (1989) 201-227.
10. T. Murofushi, M. Sugeno and M. Machida, Non-monotonic fuzzy measure and the Choquet integral, Fuzzy sets and Systems, 64:1 (1994) 73-86.
11. Y. Narukawa, T. Murofushi and M. Sugeno, Space of fuzzy measures and convergence, Fuzzy Sets and Systems, 138:3 (2003) 497-506.
12. M. Sugeno, Theory of fuzzy integrals and its applications, Doctoral Thesis, Tokyo Institute of Technology, 1974.
13. E. Szmidt, J. Kacprzyk, On distances between intuitionistic fuzzy sets, Fuzzy Sets and Systems 114 (2000) 505-518.
14. A. D. Waegenaere, P. P. Wakker, Nonmonotonic Choquet integrals, Journal of Mathematical Economics, 36 (2001) 45-60.

A Defuzzification Method of Fuzzy Numbers Induced from Weighted Aggregation Operations

Yuji Yoshida

Faculty of Economics and Business Administration, University of Kitakyushu,
4-2-1 Kitagata, Kokuraminami, Kitakyushu 802-8577, Japan
yoshida@kitakyu-u.ac.jp

Abstract. An evaluation method of fuzzy numbers is presented from the viewpoint of aggregation operators in decision making modeling. The method is given by the quasi-arithmetic means induced from weighted aggregation operators with a decision maker's subjective utility. The properties of the weighted quasi-arithmetic mean and its translation invariance are investigated. For the mean induced from the dual aggregation operators, a formula for the calculation is also given. The movement of the weighted quasi-arithmetic means is studied in comparison between two decision maker's utilities, which are essentially related to their attitude in decision making. Several examples are examined to discuss the properties of this defuzzification method.

1 Introduction

The most popular methods to evaluate fuzzy numbers in decision making problems are the defuzzification and the ordering of fuzzy numbers/fuzzy quantities ([8],[15],[16],[18]), and many authors have examined defuzzification methods for fuzzy numbers in various applications ([2],[9]). The aim of this paper is to present a new evaluation method of fuzzy numbers for decision making modeling from the viewpoint of aggregation operators. We deal with quasi-arithmetic means induced from weighted aggregation operators with decision maker's subjective utility, and we estimate fuzzy numbers by the quasi-arithmetic means. It is well-known that the weighted aggregation operation can be represented with a continuous increasing function (Kolmogorov [11], Nagumo [13], Aczél [1]). Torra and Godo [14] discussed a defuzzification method with weighted aggregation operations and its applications. In this paper, taking the continuous increasing function as a utility function in decision making, we discuss the decision maker's judgment by weighted means based on the utility. We analyze the properties of the weighted quasi-arithmetic mean and we investigate its translation invariance. We also introduce the weighted quasi-arithmetic mean induced from the dual aggregation operators, and we give a formula for its calculation. The movement of the weighted quasi-arithmetic means is studied in comparison between two decision maker's utilities, which are essentially related to their attitude in decision making. In the next section, starting from the notion of weighted aggregation operations of several variables on $[0, 1]$, we construct a weighted quasi-arithmetic

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 161–171, 2006.

mean of intervals step by step. In Section 3, the weighted quasi-arithmetic mean is applied to fuzzy numbers as a defuzzification method under the decision maker's subjective utility. In Section 4, the weighted quasi-arithmetic mean induced from the dual aggregation operators is discussed. In Section 5, the movement of the weighted quasi-arithmetic means is studied in comparison between two decision maker's utilities. In Section 6, we examine several examples and give formulae for triangle-type fuzzy numbers and trapezoidal-type fuzzy numbers.

2 Weighted Aggregation Operators

Weighted aggregation operators are given as follows. Let n be a fixed natural number, and let $\xi^n : [0,1]^n \mapsto [0,1]$ be a function. We represent the function as $\xi^n(x_1, x_2, \cdots, x_n)$ for $(x_1, x_2, \cdots, x_n) \in [0,1]^n$.

Definition 1. (n-ary weighted aggregation operator [3, 4]). A function $\xi^n : [0,1]^n \mapsto [0,1]$ is called an n [illegible] if it satisfies the following conditions (A.i) – (A.v):

(A.i) $\xi^n(x_1, x_2, \cdots, x_n) \leq \xi^n(y_1, y_2, \cdots, y_n)$ whenever $x_i \leq y_i$ for all $i = 1, \cdots, n$.
(A.ii) $\xi^n(x_1, x_2, \cdots, x_n) < \xi^n(y_1, y_2, \cdots, y_n)$ whenever $x_i < y_i$ for some $i = 1, 2, \cdots, n$.
(A.iii) ξ^n is continuous on $[0,1]^n$.
(A.iv) $\xi^n(x, x, \cdots, x) = x$ for all $x \in [0,1]$.
(A.v) It holds that

$$\begin{aligned}&\xi^n(\xi^n(x_{11}, x_{12}, \cdots, x_{1n}), \xi^n(x_{21}, x_{22}, \cdots, x_{2n}), \cdots, \xi^n(x_{n1}, x_{n2}, \cdots, x_{nn}))\\ &= \xi^n(\xi^n(x_{11}, x_{21}, \cdots, x_{n1}), \xi^n(x_{12}, x_{22}, \cdots, x_{n2}), \cdots, \xi^n(x_{1n}, x_{2n}, \cdots, x_{nn})).\end{aligned}$$

The condition (A.ii) together with (A.i) is called [illegible], and the properties (A.iii), (A.iv) and (A.v) are called [illegible] and [illegible] respectively. The following well-known result regarding the weighted aggregation operations is given by Aczél [1].

Lemma 1. (Aczél [1]). [illegible] $\xi^n : [0,1]^n \mapsto [0,1]$ [illegible] $f : [0,1] \mapsto [0,1]$ [illegible] $\{w_i | i = 1, 2, \cdots, n\}$ [illegible] $w_i > 0$ $i = 1, 2, \cdots, n$ $\sum_{i=1}^n w_i = 1$ [illegible]

$$\xi^n(x_1, x_2, \cdots, x_n) = f^{-1}\left(\sum_{i=1}^n f(x_i) w_i\right) \tag{1}$$

[illegible] $(x_1, x_2, \cdots, x_n) \in [0,1]^n$

Definition 2. (weighted aggregation operator [3, 4]). A function $\xi : \bigcup_{n \geq 1} [0,1]^n \mapsto [0,1]$ is called a [illegible] if it is given by n-ary weighted aggregation operators ξ^n such as $\xi = \xi^n$ on $[0,1]^n$ for each $n = 1, 2, \cdots$.

From Lemma 1, when a continuous strictly increasing function $f : [0,1] \mapsto [0,1]$ and a continuous function $w : [0,1] \mapsto (0,\infty)$ are given, we can present n-ary weighted aggregation operators ξ^n in the following form:

$$\xi^n(x_1, x_2, \cdots, x_n) = f^{-1}\left(\sum_{i=1}^{n} f(x_i)w(x_i) \Big/ \sum_{i=1}^{n} w(x_i)\right) \tag{2}$$

for $(x_1, x_2, \cdots, x_n) \in [0,1]^n$ and all $n = 1, 2, \cdots$. Therefore, with a fixed continuous strictly increasing function f and a fixed continuous function $w : [0,1] \mapsto (0,\infty)$, we give a weighted aggregation operator $\xi : \bigcup_{n\geq 1}[0,1]^n \mapsto [0,1]$ such that

$$\xi(x_1, x_2, \cdots, x_n) = \xi^n(x_1, x_2, \cdots, x_n) = f^{-1}\left(\sum_{i=1}^{n} f(x_i)w(x_i) \Big/ \sum_{i=1}^{n} w(x_i)\right) \tag{3}$$

for $(x_1, x_2, \cdots, x_n) \in [0,1]^n$ and $n = 1, 2, \cdots$. First, we construct a weighted quasi-arithmetic mean under subjective decision-making from the viewpoint of aggregation of every point in an interval. Let $[a,b]$ be a closed interval satisfying $0 \leq a < b \leq 1$. Let $\{[c_{i-1}, c_i] | i = 1, 2, \cdots, n\}$ be a partition of the interval $[a,b]$ such that

$$c_i := a + \frac{i(b-a)}{n} \quad \text{for } i = 0, 1, 2, \cdots, n.$$

Take a temporary aggregated point x_i of the interval $[c_{i-1}, c_i]$ such that $x_i \in [c_{i-1}, c_i]$ for each $i = 1, 2, \cdots, n$. From (3), we define a weighted quasi-arithmetic mean as follows

$$M := \lim_{n\to\infty} \xi(x_1, x_2, \cdots, x_n) = \lim_{n\to\infty} f^{-1}\left(\sum_{i=1}^{n} f(x_i)w(x_i) \Big/ \sum_{i=1}^{n} w(x_i)\right), \tag{4}$$

where $x_i (\in [c_{i-1}, c_i])$ is a temporary aggregated point on $[c_{i-1}, c_i]$ $(i = 1, 2, \cdots, n)$. By the definition of Riemann integral, we obtain

$$M = f^{-1}\left(\int_a^b f(x)w(x)\,\mathrm{d}x \Big/ \int_a^b w(x)\,\mathrm{d}x\right) \tag{5}$$

for $[a,b] (\subset [0,1])$ such that $0 \leq a < b \leq 1$ since we have

$$\begin{aligned} M &= \lim_{n\to\infty} f^{-1}\left(\sum_{i=1}^{n} f(x_i)w(x_i) \Big/ \sum_{i=1}^{n} w(x_i)\right) \\ &= f^{-1}\left(\lim_{n\to\infty}\left(\sum_{i=1}^{n} f(x_i)w(x_i) \Big/ \sum_{i=1}^{n} w(x_i)\right)\right) \\ &= f^{-1}\left(\lim_{n\to\infty}\sum_{i=1}^{n} f(x_i)w(x_i)\frac{b-a}{n} \Big/ \lim_{n\to\infty}\sum_{i=1}^{n} w(x_i)\frac{b-a}{n}\right) \\ &= f^{-1}\left(\int_a^b f(x)w(x)\,\mathrm{d}x \Big/ \int_a^b w(x)\,\mathrm{d}x\right). \end{aligned}$$

Let $D(\subset(-\infty,\infty))$ be an interval. Extending the domain from the closed interval $[0,1]$ to D, in the next section we study a defuzzification of fuzzy numbers induced from the [illegible] of closed subintervals of D in the form

$$M = f^{-1}\left(\int_a^b f(x)w(x)\,\mathrm{d}x \Big/ \int_a^b w(x)\,\mathrm{d}x\right) \tag{6}$$

for $[a,b]\subset D$ $(a<b)$, where $f: D\mapsto(-\infty,\infty)$ is a continuous strictly increasing function for utility and $w: D\mapsto(0,\infty)$ is a continuous function for weighting.

3 A Defuzzification of Fuzzy Numbers Induced from Weighted Aggregation Operators

Let $\mathbb{R}$ denote the set of all real numbers. A fuzzy number is denoted by its membership function $\tilde{a}:\mathbb{R}\mapsto[0,1]$ which is normal, upper-semicontinuous, fuzzy convex and has a compact support. $\mathcal{R}$ denotes the set of all fuzzy numbers, and $\mathcal{R}_c$ also denotes the set of fuzzy numbers with continuous membership functions. Refer to Zadeh [21] regarding fuzzy set theory. In this paper, we identify fuzzy numbers with their corresponding membership functions. The α-cut of a fuzzy number $\tilde{a}(\in\mathcal{R})$ is given by $\tilde{a}_\alpha := \{x\in\mathbb{R}\mid \tilde{a}(x)\geq\alpha\}$ $(\alpha\in(0,1])$ and $\tilde{a}_0 := \mathrm{cl}\{x\in\mathbb{R}\mid\tilde{a}(x)>0\}$, where cl denotes the closure of an interval. The α-cut is also written by closed intervals $\tilde{a}_\alpha=[\tilde{a}_\alpha^-,\tilde{a}_\alpha^+]$ $(\alpha\in[0,1])$. Hence we introduce a partial order $\succeq$, so called the [illegible] on fuzzy numbers $\mathcal{R}$: Let $\tilde{a},\tilde{b}\in\mathcal{R}$ be fuzzy numbers. $\tilde{a}\succeq\tilde{b}$ means that $\tilde{a}_\alpha^-\geq\tilde{b}_\alpha^-$ and $\tilde{a}_\alpha^+\geq\tilde{b}_\alpha^+$ for all $\alpha\in[0,1]$. Then $(\mathcal{R},\succeq)$ becomes a lattice. An addition and a scalar multiplication for fuzzy numbers are defined as follows: For $\tilde{a},\tilde{b}\in\mathcal{R}$ and $\zeta\in\mathbb{R}$, the addition $\tilde{a}+\tilde{b}$ of $\tilde{a}$ and $\tilde{b}$ and the scalar multiplication $\zeta\tilde{a}$ of ζ and $\tilde{a}$ are fuzzy numbers given by their α-cuts $(\tilde{a}+\tilde{b})_\alpha := [\tilde{a}_\alpha^-+\tilde{b}_\alpha^-,\tilde{a}_\alpha^++\tilde{b}_\alpha^+]$ and $(\zeta\tilde{a})_\alpha := [\zeta\tilde{a}_\alpha^-,\zeta\tilde{a}_\alpha^+]$ if $\zeta\geq0$ and $(\zeta\tilde{a})_\alpha := [\zeta\tilde{a}_\alpha^+,\zeta\tilde{a}_\alpha^-]$ if $\zeta<0$, where $\tilde{a}_\alpha=[\tilde{a}_\alpha^-,\tilde{a}_\alpha^+]$ and $\tilde{b}_\alpha=[\tilde{b}_\alpha^-,\tilde{b}_\alpha^+]$ $(\alpha\in[0,1])$.

Let $f: D\mapsto\mathbb{R}$ be a continuous strictly increasing function for utility, and let $\tilde{a}\in\mathcal{R}_c$ be a fuzzy number whose membership function is continuous. Define the [illegible] $\tilde{E}:\mathcal{R}_c\mapsto\mathbb{R}$, which is weighted with the membership grades of the fuzzy number $\tilde{a}$, by

$$\tilde{E}(\tilde{a}) := f^{-1}\left(\int_{-\infty}^{\infty} f(x)\,\tilde{a}(x)\,\mathrm{d}x \Big/ \int_{-\infty}^{\infty}\tilde{a}(x)\,\mathrm{d}x\right), \tag{7}$$

which we obtain putting $w=\tilde{a}$ and taking a and b as $a<a_0^-$ and $a_0^+<b$ in (6) since a fuzzy number $\tilde{a}(\in\mathcal{R}_c)$ has a compact support. Next, for a fuzzy number $\tilde{a}\in\mathcal{R}$ whose membership function is upper-continuous, there exists a sequence of fuzzy numbers $\{\tilde{a}_n\}$ such that $\tilde{a}_n\downarrow\tilde{a}$ as $n\to\infty$. Therefore, through the monotone convergence theorem, we can obtain the quasi-arithmetic mean (7) for $\tilde{a}\in\mathcal{R}$. In the equation (7), we can understand that $f(x)$ is a utility value of each point x and the quotient of the integrands in (7) is an average weighted by the membership grades $\tilde{a}(x)$ of the fuzzy number $\tilde{a}$ at each point x. We discuss

a defuzzification of fuzzy numbers in the form of the quasi-arithmetic mean (7) under a decision maker's subjective utility f.

Theorem 1. $\tilde{E}(\cdot)$

(i) $\tilde{a} \in \mathcal{R}$ $\tilde{E}(\tilde{a}) \in \tilde{a}_0$ $\tilde{a}$ $\tilde{a} = 1_{\{\delta\}}$ $\delta \in \mathbb{R}$ $\tilde{E}(\tilde{a}) = \delta$

(ii) $\tilde{a}, \tilde{b} \in \mathcal{R}$ $\tilde{a} \succeq \tilde{b}$ $\tilde{E}(\tilde{a}) \geq \tilde{E}(\tilde{b})$ $\succeq$

(iii) $\tilde{E} : \mathcal{R} \mapsto \mathbb{R}$

$$\lim_{n\to\infty} \tilde{E}(\tilde{a}_n) = \tilde{E}(\tilde{a})$$

$\tilde{a} \in \mathcal{R}$ $\tilde{a}_n \in \mathcal{R}(n = 1, 2, \cdots)$ $\lim_{n\to\infty} \tilde{a}_n = \tilde{a}$

The quasi-arithmetic mean $\tilde{E}(\cdot)$ also has the following properties regarding translation invariance.

Lemma 2.

(i) $f(x) = x^\gamma$ D γ

$$\tilde{E}(r \cdot \tilde{a}) = r \cdot \tilde{E}(\tilde{a})$$

$\tilde{a} \in \mathcal{R}$, $r \in \mathbb{R}$ $r > 0$ f $f(xy) \gtreqless f(x)f(y)$ $x, y \in D$ $\tilde{E}(r \cdot \tilde{a}) \gtreqless r \cdot \tilde{E}(\tilde{a})$

(ii) $f(x) = \gamma x$ $\mathbb{R}$ γ

$$\tilde{E}(\tilde{a} + 1_{\{s\}}) = \tilde{E}(\tilde{a}) + s$$

$\tilde{a} \in \mathcal{R}$, $s \in \mathbb{R}$ $1_{\{s\}}$ s f $f(x + y) \gtreqless f(x) + f(y)$ $x, y \in D$ $\tilde{E}(\tilde{a} + 1_{\{s\}}) \gtreqless \tilde{E}(\tilde{a}) + s$

4 Dual Weighted Quasi-arithmetic Means

In this section, we introduce weighted quasi-arithmetic means induced from a dual weighted aggregation operator, and we discuss its corresponding defuzzification.

Definition 3. (the dual weighted aggregation operator [3, 4]). For a weighted aggregation operator $\xi : \bigcup_{n\geq 1}[0,1]^n \mapsto [0,1]$, $\xi^{dual} : \bigcup_{n\geq 1}[0,1]^n \mapsto [0,1]$ is given by

$$\xi^{dual}(x_1, x_2, \cdots, x_n) := 1 - \xi(1 - x_1, 1 - x_2, \cdots, 1 - x_n) \tag{8}$$

for $(x_1, x_2, \cdots, x_n) \in [0,1]^n$ and $n = 1, 2, \cdots$.

We can deal with a more general dual weighted aggregation operator ξ^{dual} with any fixed $\kappa \in \mathbb{R}$ instead of 1 in (8). Now we introduce a weighted quasi-arithmetic mean value induced from the dual weighted aggregation operator. Let $f : D \mapsto (-\infty, \infty)$ be a continuous strictly increasing function for utility, and let $\tilde{a} \in \mathcal{R}$ be a fuzzy number. Put a weighted quasi-arithmetic mean $\tilde{E} : \mathcal{R} \mapsto D$ by (7). Fix any $\kappa \in \mathbb{R}$. Let a semi-linear strictly decreasing function $\varphi : \mathbb{R} \mapsto \mathbb{R}$ by $\varphi(x) := \kappa - x$ $(x \in \mathbb{R})$. Let $\mathcal{R}_\kappa$ be the set of fuzzy numbers $\tilde{a} \in \mathcal{R}$ satisfying their support $\tilde{a}_0 \subset \{x | x \in D \text{ and } \kappa - x \in D\}$. Then the mean induced from the translation φ is called [illegible] $\tilde{E}^{dual} : \mathcal{R}_\kappa \mapsto D$, and it is given by

$$\tilde{E}^{dual}(\tilde{a}) := (f \circ \varphi)^{-1} \left(\int_{-\infty}^{\infty} (f \circ \varphi)(x) \tilde{a}(x)\, \mathrm{d}x \Big/ \int_{-\infty}^{\infty} \tilde{a}(x)\, \mathrm{d}x \right) \tag{9}$$

for $\tilde{a} \in \mathcal{R}_\kappa$, where the operation $\circ$ is the composition of maps. The following results imply that dual weighted quasi-arithmetic means have the same properties as Theorem 1 and they can be calculated by the following equation (10).

Lemma 3. [illegible] $\tilde{E}^{dual} : \mathcal{R}_\kappa \mapsto D$ [illegible]

(i) [illegible] $\tilde{a} \in \mathcal{R}_\kappa$ [illegible]

$$\begin{aligned} \tilde{E}^{dual}(\tilde{a}) =& \kappa - f^{-1} \left(\int_{-\infty}^{\infty} f(\kappa - x) \tilde{a}(x)\, \mathrm{d}x \Big/ \int_{-\infty}^{\infty} \tilde{a}(x)\, \mathrm{d}x \right) \\ =& \kappa - \tilde{E}(-\tilde{a} + 1_{\{\kappa\}}). \end{aligned} \tag{10}$$

[illegible] f [illegible] $f(x) = \gamma x$ $(x \in \mathbb{R})$ [illegible] γ [illegible]

$$\tilde{E}^{dual}(\tilde{a}) = -\tilde{E}(-\tilde{a}).$$

(ii) [illegible] $\tilde{a} \in \mathcal{R}_\kappa$ [illegible] $\tilde{E}^{dual}(\tilde{a}) \in \tilde{a}_0$

(iii) [illegible] $\tilde{a}, \tilde{b} \in \mathcal{R}_\kappa$ [illegible] $\tilde{a} \succeq \tilde{b}$ [illegible] $\tilde{E}^{dual}(\tilde{a}) \geq \tilde{E}^{dual}(\tilde{b})$

(iv) [illegible] $\tilde{E}^{dual} : \mathcal{R}_\kappa \mapsto D$ [illegible]

$$\lim_{n \to \infty} \tilde{E}^{dual}(\tilde{a}_n) = \tilde{E}^{dual}(\tilde{a})$$

[illegible] $\tilde{a} \in \mathcal{R}_\kappa$ [illegible] $\tilde{a}_n \in \mathcal{R}_\kappa (n = 1, 2, \cdots)$ [illegible] $\lim_{n \to \infty} \tilde{a}_n = \tilde{a}$ [illegible]

5 Quasi-arithmetic Means and Utilities

Let $\tilde{a} \in \mathcal{R}$ be a fuzzy number. Define an arithmetic weighted mean value $C(\tilde{a})$ by

$$C(\tilde{a}) := \int_{-\infty}^{\infty} x\, \tilde{a}(x)\, \mathrm{d}x \Big/ \int_{-\infty}^{\infty} \tilde{a}(x)\, \mathrm{d}x. \tag{11}$$

This defuzzification is well-known and is called the , which is derived from the center of gravity. The quasi-arithmetic mean (7) can be regarded as a generalized form of (11). Now we discuss how quasi-arithmetic means $\tilde{E}^f$ move corresponding to utilities f. Let $g : D \mapsto \mathbb{R}$ be another continuous strictly increasing function for utility. Let $\tilde{E}^g : \mathcal{R} \mapsto \mathbb{R}$ be the weighted mean value defined by g instead of f in the way of (7):

$$\tilde{E}^g(\tilde{a}) := g^{-1}\left(\int_{-\infty}^{\infty} g(x)\,\tilde{a}(x)\,\mathrm{d}x \Big/ \int_{-\infty}^{\infty} \tilde{a}(x)\,\mathrm{d}x\right). \tag{12}$$

The following results give relations among the quasi-arithmetic means $\tilde{E}^f(\tilde{a})$ and $\tilde{E}^g(\tilde{a})$ and the center of gravity method $C(a)$.

Theorem 2. $f : D \mapsto \mathbb{R}$ $g : D \mapsto \mathbb{R}$ f g C^2 D $\tilde{a} \in \mathcal{R}$
(i) f g $f''/f' < g''/g'$ D $\tilde{E}^f(\tilde{a}) < \tilde{E}^g(\tilde{a})$
(ii) f g $f''/f' \le g''/g'$ D $\tilde{E}^f(\tilde{a}) \le \tilde{E}^g(\tilde{a})$
(iii) f $f(x) = rx + s$ $r, s \in \mathbb{R}$ $r > 0$ $\tilde{E}^f(\tilde{a}) = C(\tilde{a})$

Corollary 1. f C^2 D $\tilde{a} \in \mathcal{R}$
(i) f $f'' < 0$ D $\tilde{E}^f(\tilde{a}) < C(\tilde{a})$
(ii) f $f'' \le 0$ D $\tilde{E}^f(\tilde{a}) \le C(\tilde{a})$
(iii) f $f'' > 0$ D $\tilde{E}^f(\tilde{a}) > C(\tilde{a})$
(iv) f $f'' \ge 0$ D $\tilde{E}^f(\tilde{a}) \ge C(\tilde{a})$

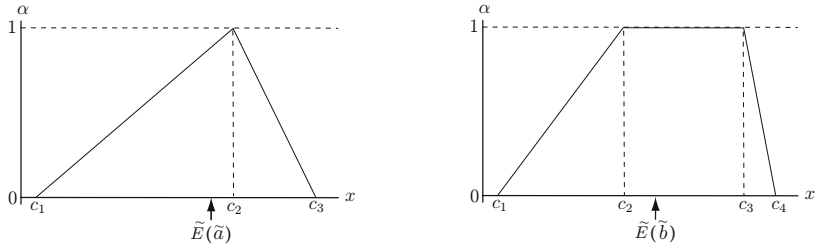
In Corollary 1, $f'' = 0$ means the decision maker's neutral attitude, $f'' < 0$ means the decision maker's risk averse attitude, and $f'' > 0$ means the decision maker's more risk tolerant attitude. Therefore, when we choose two functions f and g as decision maker's utilities, Theorem 2 implies that the utility f yields more risk averse results than g if $f''/f' \le g''/g'$ on D.

6 Examples

In this section, we examine the defuzzification methods in the previous sections for triangle-type fuzzy numbers and trapezoidal-type fuzzy numbers (Fig.1). Let $\tilde{a} \in \mathcal{R}_c$ be a triangle-type fuzzy number (13) and let $\tilde{b} \in \mathcal{R}_c$ be a trapezoidal-type fuzzy number (14):

$$\tilde{a}(x) = \begin{cases} 0 & \text{if } x < c_1 \\ (x - c_1)/(c_2 - c_1) & \text{if } c_1 \le x < c_2 \\ (x - c_3)/(c_2 - c_3) & \text{if } c_2 \le x < c_3 \\ 0 & \text{if } x \ge c_3, \end{cases} \tag{13}$$

where c_1, c_2, c_3 are real numbers satisfying $c_1 < c_2 < c_3$.



(a) A triangle-type fuzzy number $\tilde{a}$ (b) A trapezoidal-type fuzzy number $\tilde{b}$

Fig. 1. Fuzzy numbers and mean values

$$\tilde{b}(x) = \begin{cases} 0 & \text{if } x < c_1 \\ (x - c_1)/(c_2 - c_1) & \text{if } c_1 \le x < c_2 \\ 1 & \text{if } c_2 \le x < c_3 \\ (x - c_4)/(c_3 - c_4) & \text{if } c_3 \le x < c_4 \\ 0 & \text{if } x \ge c_4, \end{cases} \tag{14}$$

where c_1, c_2, c_3, c_4 are real numbers satisfying $c_1 < c_2 < c_3 < c_4$.

Example 1. Let κ be a positive real number. For a triangle-type fuzzy number $\tilde{a} \in \mathcal{R}_c$ given by (13) and a trapezoidal-type fuzzy number $\tilde{b} \in \mathcal{R}_c$ given by (14), we discuss examples of the mean values $\tilde{E}(\tilde{a}), \tilde{E}(\tilde{b}), \tilde{E}^{dual}(\tilde{a})$ and $\tilde{E}^{dual}(\tilde{b})$ corresponding to several kinds of utility functions f.

(i) Take a function $f(x) = cx + d$ on a domain $D = (-\infty, \infty)$ with constants c, d such that $c > 0$. Then, we have the centroid:

$$\tilde{E}(\tilde{a}) = \tilde{E}^{dual}(\tilde{a}) = \frac{c_1 + c_2 + c_3}{3}, \tag{15}$$

$$\tilde{E}(\tilde{b}) = \tilde{E}^{dual}(\tilde{b}) = \frac{c_1^2 + c_1c_2 + c_2^2 - (c_3^2 + c_3c_4 + c_4^2)}{3(c_1 + c_2 - c_3 - c_4)}. \tag{16}$$

(ii) Take a function $f(x) = x^\gamma$ on $D = (0, \infty)$ with a nonzero constant γ. Then, we can easily calculate

$$\tilde{E}(\tilde{a}) = \Big(\frac{2\big({c_1}^{2+\gamma}(c_2 - c_3) + {c_2}^{2+\gamma}(c_3 - c_1) + {c_3}^{2+\gamma}(c_1 - c_2)\big)}{-(c_1 - c_2)(c_2 - c_3)(c_3 - c_1)(1 + \gamma)(2 + \gamma)} \Big)^{\frac{1}{\gamma}}, \tag{17}$$

$$\tilde{E}(\tilde{b}) = \Big(\frac{2\big(({c_1}^{2+\gamma} - {c_2}^{2+\gamma})(c_3 - c_4) - ({c_3}^{2+\gamma} - {c_4}^{2+\gamma})(c_1 - c_2)\big)}{(c_1 - c_2)(c_1 + c_2 - c_3 - c_4)(c_3 - c_4)(1 + \gamma)(2 + \gamma)} \Big)^{\frac{1}{\gamma}}. \tag{18}$$

When $\gamma = 1$, we have the case (i) of the arithmetic mean. When $\gamma = -1$, we also obtain the case of the harmonic mean:

$$\tilde{E}(\tilde{a}) = \frac{-(c_1 - c_2)(c_2 - c_3)(c_3 - c_1)}{2\big((c_2 - c_3)c_1 \log c_1 + (c_3 - c_1)c_2 \log c_2 + (c_1 - c_2)c_3 \log c_3\big)}, \tag{19}$$

$$\tilde{E}(\tilde{b}) = \frac{(c_1 - c_2)(c_1 + c_2 - c_3 - c_4)(c_3 - c_4)}{2\big((c_3 - c_4)(c_1 \log c_1 - c_2 \log c_2) - (c_1 - c_2)(c_3 \log c_3 - c_4 \log c_4)\big)}. \tag{20}$$

As letting $\gamma \to 0$, we get the case of the geometric mean:

$$\tilde{E}(\tilde{a}) = \frac{(c_2 - c_3){c_1}^2 \log c_1 + (c_3 - c_1){c_2}^2 \log c_2 + (c_1 - c_2){c_3}^2 \log c_3}{-(c_1 - c_2)(c_2 - c_3)(c_3 - c_1)} - \frac{3}{2}, \tag{21}$$

$$\tilde{E}(\tilde{b}) = \frac{(c_3 - c_4)({c_1}^2 \log c_1 - {c_2}^2 \log c_2) - (c_1 - c_2)({c_3}^2 \log c_3 - {c_4}^2 \log c_4)}{(c_1 - c_2)(c_1 + c_2 - c_3 - c_4)(c_3 - c_4)} - \frac{3}{2}. \tag{22}$$

Fig.2 shows the the monotone movement of the quasi-arithmetic means $\tilde{E}(\tilde{a})$ and $\tilde{E}(\tilde{b})$ corresponding to the parameter γ. We can calculate the dual mean values $\tilde{E}^{dual}(\tilde{a})$ and $\tilde{E}^{dual}(\tilde{b})$ by (17), (18) and Lemma 3. If $\tilde{a}_0^+ < \kappa$ and $\tilde{b}_0^+ < \kappa$, then we can easily check

$$\tilde{E}^{dual}(\tilde{a}) = \kappa - \left(\frac{2d_1}{-(c_1 - c_2)(c_2 - c_3)(c_3 - c_1)(1+\gamma)(2+\gamma)}\right)^{\frac{1}{\gamma}}, \tag{23}$$

$$\tilde{E}^{dual}(\tilde{b}) = \kappa - \left(\frac{2d_2}{(c_1 - c_2)(c_1 + c_2 - c_3 - c_4)(c_3 - c_4)(1+\gamma)(2+\gamma)}\right)^{\frac{1}{\gamma}}. \tag{24}$$

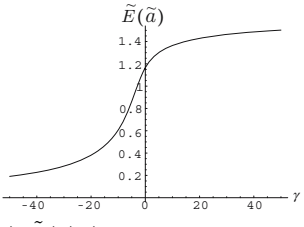
where $d_1 = (\kappa - c_1)^{2+\gamma}(c_2 - c_3) + (\kappa - c_2)^{2+\gamma}(c_3 - c_1) + (\kappa - c_3)^{2+\gamma}(c_1 - c_2)$, $d_2 = ((\kappa - c_1)^{2+\gamma} - (\kappa - c_2)^{2+\gamma})(c_3 - c_4) - ((\kappa - c_3)^{2+\gamma} - (\kappa - c_4)^{2+\gamma})(c_1 - c_2)$.

(iii) Take a concave function $f(x) = \gamma \log x$ on $D = (0, \infty)$ with a positive constant γ. Then, we obtain

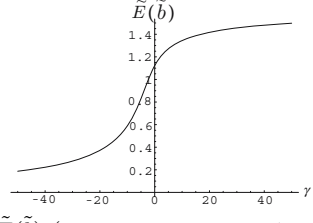
$$\tilde{E}(\tilde{a}) = e^{-3/2}\left({c_1}^{{c_1}^2(c_2 - c_3)} {c_2}^{{c_2}^2(c_3 - c_1)} {c_3}^{{c_3}^2(c_1 - c_2)}\right)^{1/d_1}, \tag{25}$$

$$\tilde{E}(\tilde{b}) = e^{-3/2}\left({c_1}^{{c_1}^2(c_3 - c_4)} {c_2}^{-{c_2}^2(c_3 - c_4)} {c_3}^{-{c_3}^2(c_1 - c_2)} {c_4}^{{c_4}^2(c_1 - c_2)}\right)^{1/d_2}, \tag{26}$$

where $d_1 := -(c_1 - c_2)(c_2 - c_3)(c_3 - c_1)$ and $d_2 := (c_1 - c_2)(c_1 + c_2 - c_3 - c_4)(c_3 - c_4)$.



(a) $\tilde{E}(\tilde{a})$ $(c_1 = 1, c_2 = 4c_3 = 5)$



(b) $\tilde{E}(\tilde{b})$ $(c_1 = 1, c_2 = 3, c_3 = 4, c_4 = 5)$

Fig. 2. The movement of quasi-arithmetic means corresponding to γ

(iv) Take a convex function $f(x) = e^{\gamma x}$ on $D = (-\infty, \infty)$ with a positive constant γ. Then, we also get

$$\tilde{E}(\tilde{a}) = \frac{1}{\gamma} \log \frac{2\big(e^{c_1\gamma}(c_2 - c_3) + e^{c_2\gamma}(c_3 - c_1) + e^{c_3\gamma}(c_1 - c_2)\big)}{-(c_1 - c_2)(c_2 - c_3)(c_3 - c_1)\gamma^2}, \tag{27}$$

$$\tilde{E}(\tilde{b}) = \frac{1}{\gamma} \log \frac{2\big((e^{c_1\gamma} - e^{c_2\gamma})(c_3 - c_4) - (e^{c_3\gamma} - e^{c_4\gamma})(c_1 - c_2)\big)}{(c_1 - c_2)(c_1 + c_2 - c_3 - c_4)(c_3 - c_4)\gamma^2}. \tag{28}$$

We give the following example as an application of Theorem 2, which shows a comparison of the means based on the decision makers' two different attitudes f and g.

Example 2. Take concave functions $f(x) = \log x$ and $g(x) = \sqrt{x}$ on $D = (0, \infty)$ (Example 1(ii),(iii)). Then we have

$$\frac{f''(x)}{f'(x)} = -\frac{1}{x} < -\frac{1}{2x} = \frac{g''(x)}{g'(x)} \tag{29}$$

for $x \in D$. From Theorem 2, we obtain $\tilde{E}^f(\tilde{a}) < \tilde{E}^g(\tilde{a})$, where $\tilde{E}^f(\tilde{a})$ is the quasi-arithmetic mean given by $f(x) = \log x$ and $\tilde{E}^g(\tilde{a})$ is the quasi-arithmetic mean given by $g(x) = \sqrt{x}$. This shows that $f(x) = \log x$ is more risk averse than $g(x) = \sqrt{x}$ as utilities.

7 Concluding Remarks

We constructed a weighted quasi-arithmetic mean of intervals from the notion of weighted aggregation operations of several variables on $[0, 1]$. The weighted quasi-arithmetic mean is applied to fuzzy numbers as a defuzzification method under the decision maker's subjective utility. This approach is constructed based on global properties of the utility. On the other hand, a local approach is found in Yager[17], and Torra and Godo [14] also developed a defuzzification method using ordered weighted aggregation operations. In this paper, we found several examples and gave formulae for triangle-type fuzzy numbers and trapezoidal-type fuzzy numbers. They will be applicable to many decision making problems in the field of artificial intelligence.

References

1. Aczél, J.: On weighted mean values. Bulletin of the American Math. Society **54** (1948) 392-400.
2. Campos, L., Munoz, A.: A subjective approach for ranking fuzzy numbers. Fuzzy Sets and Systems **29** (1989) 145-153.
3. Calvo, T., Kolesárová, A., Komorníková, M., Mesiar, R.: Aggregation operators: Basic concepts, issues and properties in: T.Calvo, Gmayor and R.Mesiar, eds., Aggregation Operators: New Trends and Applications, Phisica-Verlag, Springer (2002) 3-104.

4. Calvo, T., Pradera, A.: Double weighted aggregation operators. Fuzzy Sets and Systems **142** (2004) 15-33.
5. Carlsson, C., Fullér, R.: On possibilistic mean value and variance of fuzzy numbers. Fuzzy Sets and Systems **122** (2001) 315-326.
6. Fishburn, P.C.: Utility Theory for Decision Making. John Wiley and Sons, New York (1970).
7. Fodor, J., Roubens, M.: Fuzzy Preference Modelling and Multi-Criteria Decision Support. Kluwer Academic Publishers, Dordrecht Boston London (1994).
8. Fortemps, P., Roubens, M.: Ranking and defuzzification methods based on area compensation. Fuzzy Sets and Systems **82** (1996) 319-330.
9. Goetshel, R., Voxman, W.: Elementary fuzzy calculus. Fuzzy Sets and Systems **18** (1986) 31-43.
10. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice-Hall, London (1995).
11. Kolmogoroff, A.N.: Sur la notion de la moyenne. Acad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez. **12** (1930) 388-391.
12. López-Díaz, M., Gil, M.A.: The λ-average value and the fuzzy expectation of a fuzzy random variable. Fuzzy Sets and Systems **99** (1998) 347-352.
13. Nagumo, K.: Über eine Klasse der Mittelwerte. Japanese Journal of Mathematics **6** (1930) 71-79.
14. Torra, V., Godo, L.: Continuous WOWA operators with applications to defuzzification in: T.Calvo, Gmayor and R.Mesiar, eds., Aggregation Operators: New Trends and Applications, Phisica-Verlag, Springer (2002) 159-176.
15. Wang, X., Kerre, E.E.: Reasonable properties for the ordering of fuzzy quantities (I). Fuzzy Sets and Systems **118** (2001) 375-385.
16. Yager, R.R.: A procedure for ordering fuzzy subsets of the unit interval. Inform. Sciences **24** (1981) 143-161.
17. Yager, R.R.: OWA aggragation over a continuous interval argument with application to decision making. IEEE Trans. on Systems, Man, and Cybern. - Part B: Cybernetics **34** (2004) 1952-1963.
18. Yoshida, Y., Kerre, E.E.: A fuzzy ordering on multi-dimensional fuzzy sets induced from convex cones. Fuzzy Sets and Systems **130** (2002) 343-355.
19. Yoshida, Y.: A mean estimation of fuzzy numbers by evaluation measures in: M.Ch.Negoita, R.J.Howlett and L.C.Jain, eds., Knowledge-Based Intelligent Information and Engineering Systems, Part II. LNAI 3214. Springer (2004) 1222-1229.
20. Yoshida, Y.: Mean values on intervals induced from weighted aggregation operation. preprint.
21. Zadeh, L.A.: Fuzzy sets. Inform. and Control **8** (1965) 338-353.

Dependent OWA Operators

Zeshui Xu

Department of Management Science and Engineering,
School of Economics and Management,
Tsinghua University, Beijing 100084, China
Xu_zeshui@263.net

Abstract. Yager [1] introduced several families of ordered weighted averaging (OWA) operators, in which the associated weights depend on the aggregated arguments. In this paper, we develop a new dependent OWA operator, and study some of its desirable properties. The prominent characteristic of this dependent OWA operator is that it can relieve the influence of unfair arguments on the aggregated results. Finally, we give an example to illustrate the developed operator.

1 Introduction

The ordered weighted aggregation (OWA) operator as an aggregation technique has received more and more attention since it's appearance [2]. One important step of the OWA operator is to determine its associated weights. Many authors have focused on this issue, and developed some useful approaches to obtaining the OWA weights. For example, Yager [1] introduced some families of the OWA weights, including the ideal of aggregate dependent weights. Yager [2] introduced an approach to computing the weights of the OWA operator based on Zadeh's [3,4] concept of linguistic quantifiers. O'Hagan [5] established a mathematical programming model maximizing the entropy of the OWA weights for a predefined degree of orness. Xu and Da [6] extended O'Hagan's model to the situations where the weight information is available partially. Filev and Yager [7] developed two procedures to obtain the OWA weights, the first one learns the weights from a collection of samples with their aggregated value, and the second one calculates the weights for a given level of orness. Xu and Da [8] established a linear objective-programming model for obtaining the weights of the OWA operator by utilizing the given arguments under partial weight information. Xu [9] developed a normal distribution based method. We classify all these approaches into the following two categories: argument-independent approaches [1,2,5-7,9-12] and argument-dependent approaches [1,7,8,12-14]. The weights derived by the argument-independent approaches are associated with particular ordered positions of the aggregated arguments, and have no connection with the aggregated arguments, while the argument-dependent approaches determine the weights based on the input arguments. In this paper, we will pay attention on the second category, and develop a new argument-dependent approach to determining the OWA weights.

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 172 – 178, 2006.

2 Dependent OWA Operators

In [2], Yager defined the concept of an ordered weighted averaging (OWA) operator as follows:

An OWA operator of dimension n is a mapping, $OWA: R^n \to R$, that has an associated n vector $w = (w_1, w_2, ..., w_n)^T$ such that $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_j = 1$. Furthermore,

$$OWA(a_1, a_2, ..., a_n) = \sum_{j=1}^{n} w_j a_{\sigma(j)} \tag{1}$$

where $(\sigma(1),\sigma(2),...,\sigma(n))$ is a permutation of $(1,2,..., n)$ such that $a_{\sigma(j-1)} \geq a_{\sigma(j)}$ for all $j = 2,...,n$.

Clearly, the key point of the OWA operator is to determine its associated weights. Yager [1] introduced the ideal of aggregate dependent weights, which allows the weights to be a function of the aggregated arguments, in this case

$$OWA(a_1, a_2, ..., a_n) = \sum_{j=1}^{n} f_j(a_{\sigma(1)}, a_{\sigma(2)}, ..., a_{\sigma(n)}) a_{\sigma(j)} \tag{2}$$

The first family of the aggregate dependent weights that Yager [1] studied is as follows:

$$w_j = \frac{a_{\sigma(j)}^{\alpha}}{\sum_{j=1}^{n} a_{\sigma(j)}^{\alpha}}, \quad j = 1,2,..., n \tag{3}$$

where $\alpha \in (-\infty, +\infty)$. In this case, it leads to a neat OWA operator:

$$OWA(a_1, a_2, ..., a_n) = \frac{\sum_{j=1}^{n} a_j^{\alpha+1}}{\sum_{j=1}^{n} a_j^{\alpha}} \tag{4}$$

Note: An OWA operator is called neat if the aggregated value is independent of the ordering [1].

Another interesting case of the aggregate dependent weights is

$$w_j = \frac{(1 - a_{\sigma(j)})^{\alpha}}{\sum_{j=1}^{n} (1 - a_{\sigma(j)})^{\alpha}}, \quad j = 1,2,..., n \tag{5}$$

In this case, it follows that

$$OWA(a_1, a_2, ..., a_n) = \frac{\sum_{j=1}^{n} (1-a_j)^{\alpha} a_j}{\sum_{j=1}^{n} (1-a_j)^{\alpha}} \tag{6}$$

which is also a neat aggregation.

Yager [2] also considered a case where the aggregation is not neat, that is

$$w_i = \frac{a_{\sigma(n-i+1)}^{\alpha}}{\sum_{j=1}^{n} a_{\sigma(j)}^{\alpha}}, \quad j = 1,2,....,n \tag{7}$$

in this case, it yields

$$OWA(a_1, a_2, ..., a_n) = \frac{\sum_{j=1}^{n} a_{\sigma(n-j+1)}^{\alpha} a_{\sigma(j)}}{\sum_{j=1}^{n} a_{\sigma(j)}^{\alpha}} \tag{8}$$

In many actual situations, the arguments $a_1, a_2, ..., a_n$ are usually given by n different individuals. Some individuals may provide unduly high or unduly low preference arguments for their preferred or repugnant objects. In such a case, we shall assign very low weights to these "false" or "biased" opinions, that is to say, the closer a preference argument to the average value, the more the weight [9]. All the above argument-dependent approaches, however, should be unsuitable for dealing with this case. Therefore, it is worth paying attention to this issue, in the following we will develop a novel argument-dependent approach to determining the OWA weights.

Definition 1. Let $a_1, a_2, ..., a_n$ be a collection of arguments, and let μ be the average value of these arguments, i.e., $\mu = \frac{1}{n}\sum_{j=1}^{n} a_j$, $(\sigma(1), \sigma(2), ..., \sigma(n))$ is a permutation of $(1,2,...,n)$ such that $a_{\sigma(j-1)} \geq a_{\sigma(j)}$ for all $j = 2,...,n$, then we call

$$s(a_{\sigma(j)}, \mu) = 1 - \frac{|a_{\sigma(j)} - \mu|}{\sum_{j=1}^{n} |a_j - \mu|}, \quad j = 1,2,..., n \tag{9}$$

the similarity degree of the j-th largest argument $a_{\sigma(j)}$ and the average value μ.

Let the $w = (w_1, w_2, ..., w_n)^T$ be the weight vector of the OWA operator, then we define the following:

$$w_j = \frac{s(a_{\sigma(j)}, \mu)}{\sum_{j=1}^{n} s(a_{\sigma(j)}, \mu)}, \quad j = 1,2,..., n \tag{10}$$

where $s(a_{\sigma(j)}, \mu)$ is defined by Eq.(9). Clearly, we have $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_j = 1$. Since

$$\sum_{j=1}^{n} s(a_{\sigma(j)}, \mu) = \sum_{j=1}^{n} s(a_j, \mu) \tag{11}$$

then Eq.(10) can be rewritten as

$$w_j = \frac{s(a_{\sigma(j)}, \mu)}{\sum_{j=1}^{n} s(a_j, \mu)}, \quad j = 1,2,..., n \tag{12}$$

In this case, we have

$$OWA(a_1, a_2, ..., a_n) = \frac{\sum_{j=1}^{n} s(a_{\sigma(j)}, \mu)\, a_{\sigma(j)}}{\sum_{j=1}^{n} s(a_{\sigma(j)}, \mu)} = \frac{\sum_{j=1}^{n} s(a_j, \mu)\, a_j}{\sum_{j=1}^{n} s(a_j, \mu)} \tag{13}$$

It is easy to see that this is a neat and dependent OWA operator.

By Eq.(12), we can get the following results easily:

Theorem 1. Let $a_1, a_2, ..., a_n$ be a collection of arguments, and μ be the average value of these arguments, $(\sigma(1), \sigma(2), ..., \sigma(n))$ is a permutation of $(1,2,..., n)$ such that $a_{\sigma(j-1)} \geq a_{\sigma(j)}$, for all $j = 2,..., n$, and let $s(a_{\sigma(j)}, \mu)$ be the similarity degree of the j-th largest argument $a_{\sigma(j)}$ and the average value μ, if $s(a_{\sigma(i)}, \mu) \geq s(a_{\sigma(j)}, \mu)$, then $w_i \leq w_j$.

Corollary 1. Let $a_1, a_2, ..., a_n$ be a collection of arguments, if $a_i = a_j$, for all i, j, then $w_j = \frac{1}{n}$, for all j.

From Eq.(12) and Theorem 1, we know that a prominent characteristic of this dependent OWA operator are that it can relieve the influence of unfair arguments on the aggregated results by assigning low weights to those "false" or "biased" ones.

Yager [1,2] defined two important measures associated with an OWA operator. The first measure, called the dispersion of the weighting vector w of an OWA operator is defined as

$$disp(w) = -\sum_{j=1}^{n} w_j \ln w_j \tag{14}$$

which measures the degree to which w takes into account the information in the arguments during the aggregation.

The second measure, called orness measure, is defined as

$$orness(w) = \frac{1}{n-1}\sum_{j=1}^{n}(n-j)w_j \tag{15}$$

which lies in the unit interval [0,1], and characterizes the degree to which the aggregation is like an *or* operation. From Eqs.(12), (14) and (15), it follows that

$$disp(w) = -\frac{\sum_{j=1}^{n} s(a_j,\mu)\ln\dfrac{s(a_j,\mu)}{\sum_{j=1}^{n} s(a_j,\mu)}}{\sum_{j=1}^{n} s(a_j,\mu)} \tag{16}$$

$$orness(w) = \frac{1}{n-1}\frac{\sum_{j=1}^{n}(n-j)s(a_j,\mu)}{\sum_{j=1}^{n} s(a_j,\mu)} \tag{17}$$

Example 1. Suppose that there are seven decision makers $d_j\ (j=1,2,...,7)$, these decision makers provide their individual preferences for a university faculty with respect to the criterion *research.* Assume that the given preference arguments are as follows:

$$a_1 = 80,\ a_2 = 75,\ a_3 = 100,\ a_4 = 40,\ a_5 = 90,\ a_6 = 95,\ a_7 = 70$$

Therefore, the re-ordered arguments $a_j\ (j = 1,2,...,7)$ in descending order are

$$a_{\sigma(1)} = 100,\ a_{\sigma(2)} = 95,\ a_{\sigma(3)} = 90,\ a_{\sigma(4)} = 80,\ a_{\sigma(5)} = 75$$
$$a_{\sigma(6)} = 70,\ a_{\sigma(7)} = 40$$

then by Eqs.(9) and (12), we have

$$w_1 = 0.13145,\ w_2 = 0.13967,\ w_3 = 0.14789,\ w_4 = 0.16432$$
$$w_5 = 0.16080,\ w_6 = 0.15258,\ w_7 = 0.10329$$

which are shown in Fig. 1.

By Eqs. (16) and (17), we have

$$\begin{aligned} disp(w) &= -\sum_{j=1}^{7} w_j \ln w_j \\ &= -[0.13145\times\ln(0.13145) + 0.13967\times\ln(0.13967) + 0.14789\times\ln(0.14789) \\ &\quad + 0.16432\times\ln(0.16432) + 0.16080\times\ln(0.16080) + 0.15258\times\ln(0.15258) \\ &\quad + 0.10329\times\ln(0.10329)] \\ &= 1.9363 \end{aligned}$$

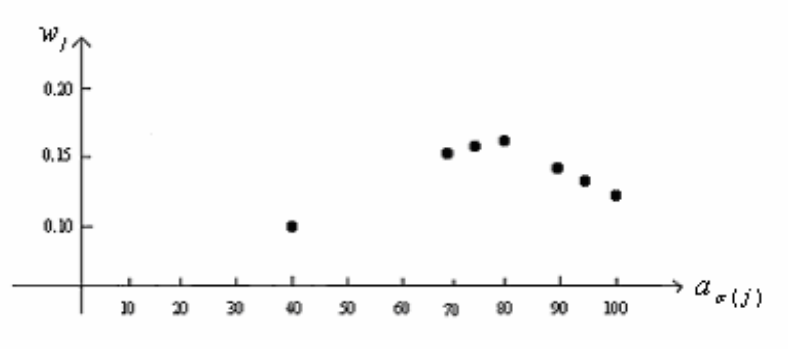


Fig. 1. The weights w_j of $a_{\sigma(j)}$ $(j = 1,2,...,7)$

and

$$orness(w) = \frac{1}{7-1}\sum_{j=1}^{7}(7-j)w_j$$

$$= \frac{1}{6}\times(6\times 0.13145 + 5\times 0.13967 + 4\times 0.14789 + 3\times 0.16432 + 2\times 0.16080 + 1\times 0.15258 + 0\times 0.10329)$$
$$= 0.5076$$

By Eq.(13), we have

$$OWA(a_1, a_2, ..., a_n) = 0.13145\times 100 + 0.13967\times 95 + 0.14789\times 90 + 0.16432\times 80 + 0.16080\times 75 + 0.15258\times 70 + 0.10329\times 40$$
$$= 79.74155$$

hence, the collective preference argument is 79.74155 .

To relieve the influence of unfair arguments on the aggregated results, in the above example, we assign low weights to those "false" or "biased" ones, that is to say, the closer a preference argument to the average value $\mu = 78.57$, the more the weight. For example, we assign the lowest weight $w_7 = 0.10329$ to the lowest preference value $a_4 = 40$, which has the biggest departure from the average value, and assign the second lowest weight $w_1 = 0.13145$ to the maximal preference value $a_3 = 100$, which has the second biggest departure from the average value. We assign the most weight $w_4 = 0.16432$ to the preference value $a_1 = 80$, which is closest to the average value, and assign the second most weight $w_5 = 0.16080$ to the preference value $a_{\sigma(5)} = 75$, which has the second least departure from the average value. We assign the value 0.5076 to the *orness* measure, and give the value 1.9363 to the dispersion measure.

3 Concluding Remarks

In this paper, we have investigated the dependent OWA operators, and developed a new argument-dependent approach to determining the OWA weights, which can relieve the influence of unfair arguments on the aggregated results. We have verified the practicality and effectiveness of the approach with a numerical example.

Acknowledgement

The work was supported by the National Natural Science Foundation of China under Grants 70571087.

References

1. Yager, R. R.: Families of OWA operators. Fuzzy Sets and Systems 59(1993) 125-148.
2. Yager, R. R.: On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics 18(1988) 183-190.
3. Zadeh, L.A.: A computational approach to fuzzy quantifiers in natural languages. Computers & Mathematics with Applications 9(1983) 149-184.
4. Zadeh, L.A.: A computational theory of dispositions. International Journal of Intelligent Systems 2(1987) 39-64.
5. O'Hagan, M.: Aggregating template rule antecedents in real-time expert systems with fuzzy set logic. Proc. 22nd Ann. IEEE Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, pp. 681-689, CA(1988).
6. Xu, Z.S., Da, Q.L.: Approaches to obtaining the weights of the ordered weighted aggregation operators. Journal of Southeast University 33(2003) 94-96.
7. Filev, D.P., Yager, R.R.: On the issue of obtaining OWA operator weights. Fuzzy Sets and Systems 94(1998) 157-169.
8. Xu, Z.S., Da, Q.L.: The uncertain OWA operator. International Journal of Intelligent Systems 17(2002) 569-575.
9. Xu, Z.S.: An overview of methods for determining OWA weights. International Journal of Intelligent Systems 20(8) (2005) 843-865.
10. Fullér, R., Majlender, P.: An analytic approach for obtaining maximal entropy OWA operator weights. Fuzzy Sets and Systems 124(2001) 53-57.
11. Yager, R.R., Filev, D.P.: Parameterized "andlike" and "orlike" OWA operators. International Journal of General Systems 22(1994) 297-316.
12. Xu, Z.S.: Uncertain Multiple Attribute Decision Making: Methods and Applications. Tsinghua University Press, Beijing (2004).
13. Filev, D.P., Yager, R.R.: Learning OWA operator weights from data. In: Proc. 3rd IEEE International Conference on Fuzzy Systems, pp. 468-473, Orlando, FL(1994).
14. Yager, R.R., Filev, D.P.: Induced ordered weighted averaging operators. IEEE Transactions on Systems, Man, and Cybernetics—Part B 29(1999) 141-150.

Decision Aggregation in an Agent-Based Financial Investment Planning System*

Zili Zhang

Intelligent Software and Software Engineering Laboratory,
Southwest University, Chongqing 400715, China
zhangzl@swu.edu.cn
School of Information Technology,
Deakin University, Geelong Victoria 3217, Australia
zzhang@deakin.edu.au

Abstract. Agent technology provides a new way to model many complex problems like financial investment planning. With this observation in mind, a financial investment planning system was developed from agent perspectives with 12 different agents integrated. Some of the agents have similar problem solving and decision making capabilities. The results from these agents require to be combined. Ordered Weighted Averaging (OWA) operator was chosen to aggregate different results. Details on how OWA was applied as well as appropriate evaluation are presented.

Keywords: Multi-Agent Systems, Intelligent Agents, Financial Investment Planning, Decision Aggregation, Ordered Weighted Averaging Operator.

1 Introduction

An agent here is a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives [1]. Multi-agent systems are systems composed of multiple interacting agents. These agents work together to solve problems that are beyond the capabilities or knowledge of individual agents. Agents offer a new and often more appropriate route to the development of complex systems, especially in open and dynamic environments.

Real-world applications such as financial investment planning are almost always made up of a large number of components that interact in varying and complex ways. This leads to complex behaviour that is difficult to understand, predict and manage. Take one sub-task of financial planning – financial portfolio management – as an example. The task environment has many interesting features, including [2]:

- the enormous amount of continually changing, and generally unorganized, information available;

* This paper is partially supported by the Ministry of Education of China Key Project 104160.

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 179–190, 2006.

- the variety of kind of information that can and should be brought to bear on the task (market data, financial report data, technical models, analysts' reports, breaking news, etc.); and
- the many sources of uncertainty and dynamic change in the environment.

It is obvious that financial planning is a typical complex problem. To design and develop systems that can solve such complex problems is not easy due to a large number of parts or components involved with many interactions. Fortunately, some researchers in agent research community have given a qualitative analysis to provide the intellectual justification of precisely why agent-based methodology is well suited to engineering complex software systems [3]. To this end, an agent-based system for financial investment planning was developed. In the system, the following models/techniques were integrated together: a client financial risk tolerance model and a client asset allocation model, both are based on fuzzy logic [4]; two interest rate prediction models, one based on neural networks, the other based on fuzzy logic and genetic algorithms [5]; three portfolio selection models–Markowitz's model [6], the fuzzy probability model, and the possibility distribution model [7]; and expert systems with explanation mechanisms. In addition to these models/techniques, an operations research software package called for solving quadratic programming ( ) and a matrix software package called for solving eigenvalues of matrices ( ) were also integrated. When developing the agent system, we used these models or software packages and wrapped them as agents.

In the multi-agent system for financial planning, there are agents with similar problem solving and decision making capabilities. When they are asked to accomplish the same task, the results may be different. In such situations, we need to aggregate the results to obtain a final one. Because there is much fuzzy information in our application, we need an aggregation method that is able to deal with fuzzy information. After we compared several kinds of fuzzy operators for aggregation such as fuzzy averaging, weighted MIN, and Ordered Weighted Averaging (OWA) (Of course, we did not compare all available averaging operators), we find out that OWA is more flexible. The OWA operator can take into account as much information as possible in the aggregation process. For example, the fuzzy averaging method can not take the decision maker's attitude into account, but the OWA operator can. This implies that the aggregated results using OWA are more reliable. This is very important for our financial investment application. That is why OWA was chosen for our application. This paper focuses on how OWA is applied and implemented in the system as well as how the aggregated results are evaluated.

The rest of the paper is structured as follows: Section 2 is a brief introduction to OWA operator. The architecture of the agent-based financial investment planning system is given in Section 3. Section 4 discusses the decision aggregation in the system. Experimental results and evaluation are presented in Section 5. Section 6 concludes the paper.

2 Introduction to OWA Operators

Yager introduced the ordered weighted averaging (OWA) operator to provide a family of aggregators having the properties of mean operators [8]. Here, we will briefly provide an introduction to the ordered weighted averaging (OWA) operator[8] [10] [11] [12] [15].

Definitions: A mapping $F : R^n \longrightarrow R$ is called an OWA operator of dimension n if it has an associated weighting vector W of dimension n such that its components satisfy

(1) $w_j \in [0, 1]$
(2) $\sum_{j=1}^{n} w_j = 1$
and

$$F_w(a_1, a_2, \ldots, a_n) = \sum_{j=1}^{n} w_j b_j$$

where b_j is the jth largest of the a_i.

A fundamental feature of this operator is the reordering process which associates the arguments with the weights. This process introduces a nonlinearity into the aggregation process. It should be observed that we can express this aggregation in a vector notation as

$$F_w(a_1, a_2, \ldots, a_n) = W^T B$$

In this expression, W is the OWA weighting vector associated with the aggregation, and B is the ordered argument vector; where the jth component in B, b_j is the jth largest of the a_i.

This operator can easily be seen to be a mean operator in that it is commutative, monotone, and is always bounded by the max and min of the arguments

$$Min_i[a_i] \leq F_w(a_1, a_2, \ldots, a_n) \leq Max_i[a_i]$$

It can be seen that this is idempotent, $F_w(a_1, a_2, \ldots, a_n) = a$.

Expressing the OWA operator $F_w(a_1, a_2, \ldots, a_n)$ in its vector notation form $W^T B$ makes very clear the distinct components involved in the performance of this operation. First, we have a weighting vector W; this is required to have components w_j which lie in the unit interval and sum to one. The second part of the OWA aggregation is the vector B, called the ordered argument vector. This vector is composed of the arguments of the aggregation.

To solve a specific problem using OWA operator, we need to find out the appropriate weighting vector W and the ordered argument vector B.

There are two characterising measures associated with the weighting vector W [8] [10] [19]. The first of these, the α value of an OWA operator, is defined as

$$\alpha(W) = \frac{1}{n-1} \sum_{j=1}^{n} w_j (n - j) \tag{1}$$

This measure, which takes its values in the unit interval, is determined by the weighting used in the aggregation. The actual semantics associated with α are

dependent upon the application in which it is being used. In our case, the α can be the degree that the aggregation prefers decisions with high confidence, or the attitude of the decision making agent.

The second measure is called the dispersion (or entropy) of W and is defined as

$$H(W) = -\sum_{j=1}^{n} w_j \ln(w_j) \tag{2}$$

Equation (2) helps measure the degree to which W takes into account all of the information in the aggregation.

There are a few alternative ways to determine the weights [18]. One of the methods requires the solution of the following mathematical programming problem:

Maximise $-\sum_{j=1}^{n} w_j \ln(w_j)$ subject to
(1) $\alpha(W) = \frac{1}{n-1} \sum_{j=1}^{n} w_j(n-j)$;
(2) $w_j \in [0,1]$;
(3) $\sum_{j=1}^{n} w_j = 1$.

Generally, the process of obtaining the vector B can be seen as an assigning operation[10]. It is an assigning operation in the sense that it assigns an argument element to a component value in W. There are three general approaches for obtaining W. The first is based upon a learning of the weights from data. An algorithm for learning the OWA weights from data is discussed in [13]. The second approach, which is based upon the close connection between OWA operators and linguistic quantifiers, uses linguistic elements to generate the weight, and is discussed in [9]. The third approach makes use of a single parameter, such as the α measure [14], to obtain the OWA weights.

3 Agent Behaviors in the Agent-Based Financial Planning System

The design and implementation of the agent-based financial planning system are detailed in [17]. This paper focuses on the decision aggregation mechanism used in the system. To facilitate the discussion of decision aggregation in the system, the behaviors of each kind of agent in the system are briefly described below:

Interface Agent. This agent interacts with the user (or user agent). It asks the user to provide his personal information and requirements, and provides the user with a final decision or advice that best meets the user's requirements.
Planning Agent. The planning agent is in charge of the activation and synchronization of different agents. It elaborates a work plan and is in charge of ensuring that such a work plan is fulfilled. It receives the assignments from the interface agent.
Decision Making Agent. It is application-specific, i.e., it has its own knowledge base; it must have some meta-knowledge about when it needs the help of

intelligent technique agents (e.g., pre or post processing some data); it can ask intelligent technique agents to accomplish some sub-tasks.

Serving Agent. The serving agent is a matchmaker – one kind of middle agent [16]. It keeps track of the names, ontologies, and capabilities of all registered intelligent technique agents in the system; it can reply to the query of a decision making agent with appropriate intelligent technique agent's name and ontology.

Service Provider Agent. Most of the service provider agents in the system are intelligent technique agents (the agents based on different [illegible] like fuzzy logic and neural networks are called [illegible]). Each intelligent technique agent can provide services for decision making agents with one or some kind of combined intelligent techniques; it can send back the processed results to decision making agents; it must advertise its capabilities to the serving agent.

Decision Aggregation Agent. When decision making agents finish the assigned tasks they return the results to the decision aggregation agent. The aggregation agent chooses one of the alternative decisions, or performs an aggregation of the different results into a final one. The behaviors of such agents and the algorithm used for aggregation in the system are detailed in Section 4. The empirical evaluation is given in Section 5.

4 OWA-Based Aggregation in the System

In the system, aggregation is required to determine the final investment policy of a user and to provide the final portfolio for a user. In this section, we take the investment policy determination as an example to discuss the aggregation process, which was implemented in the system.

Suppose a user wants to know whether his investment policy (IP) should be aggressive or conservative. First, the user gives his [illegible] (AI) and [illegible] (TNW) to the decision making agents through the user interface agents. The decision making agents use their own knowledge to evaluate user's [illegible] (RT) ability using rules such as [illegible] AI [illegible] L [illegible] TNW [illegible] L [illegible] RT [illegible] L. Here, AI, TNW, and RT are variables taking values from {low(L), medium(M), high(H), ...}; L, M, and H are fuzzy subsets of $[0, 1]$. Please note that different decision making agents may have different rules similar to this one.

Now suppose there are n decision making agents. Each agent has rules in its knowledge base such as

$$If\ RT\ is\ H\ and\ P_1\ is\ B_1\ and\ \ldots\ then\ IP\ is\ C_i \tag{3}$$

where $C_i (i = 1, 2, \ldots, n)$ is a fuzzy subset indicating the aggression or conservation of the investment policy.

Because the knowledge of the decision making agents and their decision attitudes may be different, the answers to the same question may also be different. Usually they are either close or conflicting to various degrees. They have to be combined or reconciled in order to produce one decision.

Assume that the agents' decisions are represented by trapezoidal numbers, which aare widely used to represent fuzzy information. If $C_i = (a_{i1}, b_{i1}, b_{i2}, a_{i2})$, $i = 1, 2, \ldots, n$, are trapezoidal numbers, then

$$C_{OWA} = (F_w(a_{11}, \ldots, a_{n1}), F_w(b_{11}, \ldots, b_{n1}), F_w(b_{12}, \ldots, b_{n2}), F_w(a_{12}, \ldots, a_{n2})) \tag{4}$$

where F_w is an OWA operator. To aggregate different results using formula (4), what we need to do is to decide the weighting vector W and the ordered argument vector B in different situations.

Assume that three decision making agents $DA_i (i = 1, 2, 3)$ use their rules like (3) and obtain their own decisions $C_i (i = 1, 2, 3)$ independently. These decisions are represented by trapezoidal numbers as follows:

$$C_1 = (-100, -100, -50, -30), C_2 = (-10, 10, 10, 30), C_3 = (60, 90, 100, 100); \tag{5}$$

C_1 indicates [illegible], C_2 [illegible], and C_3 [illegible]. If all the agents have equal importance in the decision making process, the weighting vector can be obtained: $W = [w_1, w_2, w_3] = [1/3, 1/3, 1/3]$. The arguments are ordered by their values. It is easy to calculate the aggregated result using Formula (4). The defuzzification value of the aggregated result according to the [illegible] [4] is 10. It suggests a policy on the aggressive side of the scale but a very cautious one.

Now consider the case when the opinions of the three conflicting agents have different importance on a scale from 0 to 10. The ranking of agents DA_1, DA_2, and DA_3 is assumed to be $r_1 = 4, r_2 = 6$, and $r_3 = 10$, correspondingly (agents actually ranking is application dependent). We map the degrees of importance into unit interval, and obtain $U = [u_1, u_2, u_3] = [0.2, 0.3, 0.5]$. As the ranking information reflects the importance of different agents opinions, they can be treated as the weights in OWA. That is, the weights w_i for DA_i: $w_1 = 0.2, w_2 = 0.3, w_3 = 0.5$. In this case, the arguments are ordered using the values of r_i, i.e., let b_j be the a_i value which has the jth largest of r_i, and $W = [w_1, w_2, w_3] = [u_3, u_2, u_1] = [0.5, 0.3, 0.2]$. Formula (4) is then used to aggregate. The defuzzification value of the aggregated result is 35.54. It indicates that the investment policy should be cautiously aggressive.

In both cases above, the same results can be obtained by using fuzzy averaging. The problem here is that the degrees of importance in aggregation were not used directly. Actually in this case, the arguments which need to be aggregated are pairs such as

$$(u_1, a_{11}), (u_2, a_{21}), \ldots, (u_n, a_{n1})$$

Here, the formula $G(u, a) = \bar{\alpha}\bar{u} + ua$ is used to transform the tuple into a single value [15] (pp. 41-49), where α is defined by (1). The following are the steps of the procedure:

1. Calculate the α value of the OWA operator:

$$\alpha = \sum_{j=1}^{3} \frac{3-j}{3-1} w_j = w_1 + w_2/2 = 0.5 + 0.3/2 = 0.65$$

2. Transform each of the argument tuples using $G(u_j, a_j) = \bar{\alpha}\bar{u_j} + u_j a_j$, hence

$$G(u_1, a_{11}) = -49.72, G(u_2, a_{21}) = -2.755, G(u_3, a_{31}) = 30.175$$

$$G(u_1, b_{11}) = -49.72, G(u_2, b_{21}) = 3.245, G(u_3, b_{31}) = 45.175$$
$$G(u_1, b_{12}) = -9.72, G(u_2, b_{22}) = 3.245, G(u_3, b_{32}) = 50.175$$
$$G(u_1, a_{12}) = -5.72, G(u_2, a_{22}) = 9.245, G(u_3, a_{32}) = 50.175$$

3. Calculate C_{OWA}

$$\begin{aligned} C_{OWA} = (&F_w(-49.72, -2.755, 30.175), F_w(-49.72, 3.245, 45.175), \\ &F_w(-9.72, 3.245, 50.175), F_w(-5.72, 9.245, 50.175)) \\ &= (4.32, 13.62, 24.12, 26.72) \end{aligned}$$

The defuzzification value is 18.87. This still indicates a very cautious investment policy – much more cautious than one not using the degrees of importance.

The concept of agents' decision making attitudes is also important. Because the agents usually have different knowledge, this results in different attitudes when making decisions. Some are aggressive, some conservative. Here, α_i ($\alpha_i \in [0, 1]$) is used to indicate the agents' attitudes. The bigger the value of α_i, the more aggressive the attitude of the decision making agent DA_i.

Suppose there are still three agents, and their attitudes are $\alpha_1 = 0.3, \alpha_2 = 0.5$ and $\alpha_3 = 0.8$. The decisions they make, and their degrees of importance, remain unchanged, as described above.

To aggregate, the first step is to decide the attitude α of all the agents (in this case three). The OWA operator is still used. Degrees of importance are mapped to unit interval as the weighting vector for combining α_i, called $W(\alpha)$, and

$$W(\alpha) = [w(\alpha)_1, w(\alpha)_2, w(\alpha)_3] = [0.5, 0.3, 0.2]$$

Then

$$\alpha = F_{W(\alpha)}(\alpha_1, \alpha_2, \alpha_3) = \alpha_3 \times w(\alpha)_1 + \alpha_2 \times w(\alpha)_2 + \alpha_1 \times w(\alpha)_3 = 0.61$$

By solving the mathematical programming problem with $\alpha = 0.61$, the weighting vector W is obtained for the final aggregation as follows:

$$W = [w_1, w_2, w_3] = [0.45, 0.32, 0.23]$$

The arguments are ordered according to the values of r_i. The final aggregation using (4) gives $C_{OWA} = (0.8, 20.7, 36.7, 47.3)$. The defuzzification value according to the [illegible] in fuzzy averaging is 28.7 [4]. This suggests a policy on the aggressive side of the scale, but a cautious one – more cautious than that using fuzzy averaging. This is because the decision attitude of DA_1 is slightly conservative, but its decision is very conservative. Taking all the information into account, the investment policy should be cautiously aggressive.

If the degrees of importance are used directly in the aggregation in this case, $C_{OWA} = (1.26, 9.93, 21.38, 24.22)$ is obtained. The defuzzification value is 15.66.

5 Experimental Results and Evaluation

The system can provide reasonable financial investment planning information based on the user provided data and some relevant models. Figure 1 shows the asset allocation results when the annual income is $\$100,000$, networth $\$700,000$, age 40, investment amount $\$50,000$, and investment attitude is aggressive (level 3). By clicking the "explanation" button, the corresponding explanation of how to get the results is displayed in the "result display" window (Figure 2).

If the growth part is invested in stock market, the system can provide a portfolio for the user (Figure 1). The portfolio is the aggregated result of three portfolios based on Markowitz's portfolio selection model, the fuzzy probability portfolio selection model, and the possibility distribution portfolio selection model, respectively. The four portfolios are marked as Powa, Pmar, Pfuz, and Ppos, respectively. The aggregation algorithm used is ordered weighted averaging (OWA) aggregation algorithm. By clicking the "evaluation" button, the system will provide the comparisons of the four portfolios (Figure 3). An empirical evaluation of the aggregated results is given in the following.

At this stage, one important problem is how to verify the aggregated portfolio. There is no systematic way available to answer this question. Instead, some experiments were conducted.

The first experiment conducted was to use the first 12 years return data described in [6] and produce three portfolios based on the three models. Based on the analysis in [7], it is known that the fuzzy model is a direct extension of Markowitz's model, while the possibility model is more reasonable than the fuzzy model. Thus the three portfolios are ordered as P_{POS}, P_{FUZ}, and P_{MAR}, and $\alpha = 0.7$ (the degree that the aggregation prefers decisions with high confidence) is chosen when using OWA operator to aggregate the three portfolios. The weight vector with $\alpha = 0.7$ is $W = [0.554, 0.292, 0.154]$. The selected portfolios as well as corresponding risks of investment are shown in Table 1. The portfolios in Table 1 are also selected with an expected average return $r_s = 17\%$.

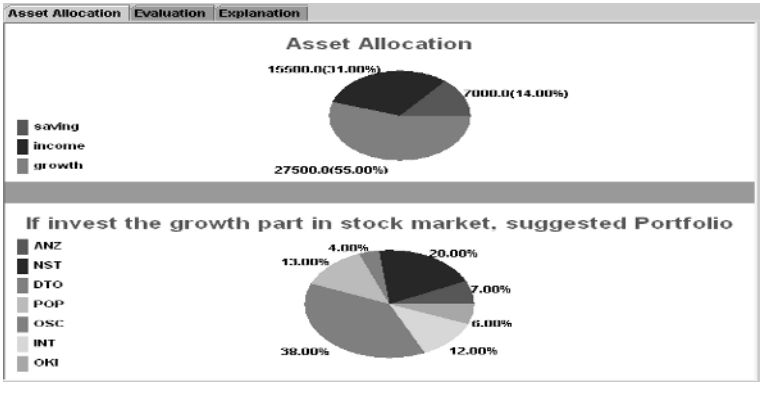


Fig. 1. Example Asset Allocation Results

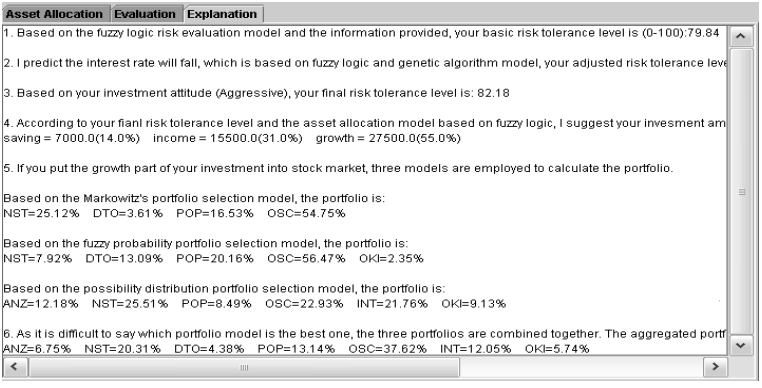


Fig. 2. Example Explanations

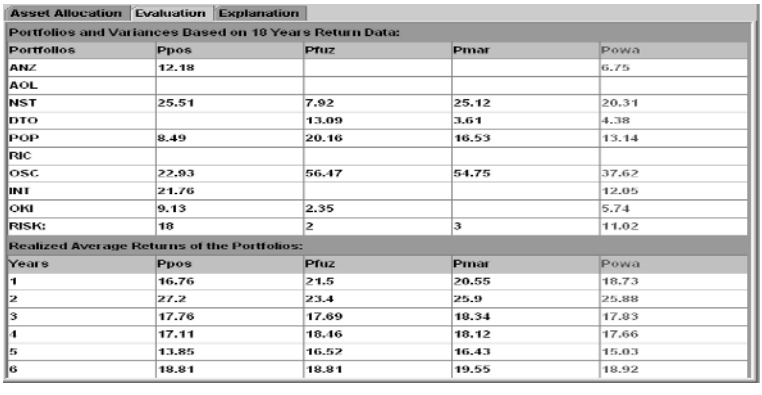


Fig. 3. Comparison of the Portfolios

Table 1. Portfolios and Variances Based on 12 Years Return Data

P	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	Variance
P_{POS}	5.15			13.95	19.53			39.75	21.62	0.30
P_{FUZ}					23.05		46.55		30.40	0.04
P_{MAR}					23.14		46.60		30.26	0.05
P_{OWA}	2.85			7.73	21.12		20.77	22.02	25.51	0.18

The last 6 years return data in [6] are used to verify the realized average returns of the four portfolios. The realized average returns of the four portfolios from one year to six years are listed in Table 2.

From Table 2, one can see that the average returns of P_{OWA} are better than those of P_{FUZ} and P_{MAR}, and slightly less than those of P_{POS}. The variance (risk or uncertainty degree of the investment) of P_{OWA} is greatly reduced (from 0.30 to 0.18) compared with that of P_{POS}.

Table 2. Realized Average Returns of the Portfolios (%)

Year(s)	P_{POS}	P_{FUZ}	P_{MAR}	P_{OWA}
1	11.421	7.501	7.554	9.681
2	27.694	15.665	15.700	22.334
3	18.748	14.457	14.467	16.836
4	19.684	14.990	15.006	17.593
5	14.005	13.318	13.334	13.701
6	19.703	14.590	14.607	17.425

Table 3. Returns on Twelve Securities from ASX

Yr	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}
86	.675	.33	.053	.946	.081	1.026	.42	.198	.405	.249	.75	.08
87	.648	.93	.864	1.097	.826	.897	.535	.207	1.023	.333	.994	.684
88	.248	.432	-.157	.656	.515	1.044	.623	.05	.132	.174	.338	.403
89	-.007	.405	.838	.698	.136	.544	.358	1.296	.468	.271	.355	.467
90	-.087	-.08	-.168	-.029	-.228	-.115	.243	-.233	-.203	-.251	-.208	.096
91	.314	.291	.787	.341	.37	.598	.577	.57	.037	.391	.382	.263
92	.246	.004	.025	-.161	.01	-.148	-.116	-.03	.018	-.127	-.177	-.138
93	.105	.126	.107	-.091	-.105	-.126	.133	-.013	.011	.005	-.164	-.03
94	.133	.135	.264	-.17	.058	.112	0	.292	.232	.129	.014	-.006
95	.385	.433	.543	.042	.167	.293	.675	.646	.55	.62	.42	.361
96	.053	.105	.319	-.947	.121	-.179	-.487	-.214	.226	-.046	-.299	-.265
97	-.2	.109	-.046	1.192	-.305	-.168	-.06	.259	-.213	-.088	-.111	0

Table 4. Portfolios and Variances Based on ASX 8 Years Return Data

P	S_1	S_2	S_3	S_4-S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	Variance
P_{POS}	22.2	26.45	12.06		9.95	0.07		28.65		0.62	0.26
P_{FUZ}	59.11				12.10	6.19				22.60	0.03
P_{MAR}	43.10				33.77	12.58				10.55	0.04
P_{OWA}	36.20	14.66	6.68		14.25	3.78		15.87		8.56	0.15

To further verify the aggregated portfolio, 12 securities listed in Australian Stock Exchange Limited (ASX) were selected and 12 years average returns (from 1986 to 1997) were collected (see Table 3). The ASX security codes of S_1 to S_{12} are AKC, AFI, AGL, BPC, CSR, EML, GUD, SMI, HAH, OPS, PDP, and WYL, respectively.

Similar to experiment one, the first 8 years (1986 to 1993) return data was used to generate the portfolios, while the last 4 years (1994 to 1997) data was used to verify. When the expected average return $r_s = 17\%$, the selected portfolios based on the three models and the aggregated portfolio based on OWA with $\alpha = 0.7$ are listed in Table 4. Based on the four portfolios, the realized average returns from one year to four years are shown in Table 5.

Table 5. Realized Average Returns of the Portfolios Based on ASX Data (%)

Year(s)	P_{POS}	P_{FUZ}	P_{MAR}	P_{OWA}
1	13.432	9.556	9.351	11.672
2	30.833	25.035	28.315	28.752
3	19.584	11.304	8.524	15.463
4	12.644	4.887	3.919	9.035

The results in Table 5 are consistency with those in Table 2. Thus the same conclusion can be reached. The average returns of P_{OWA} are better than those of P_{FUZ} and P_{MAR}, and slightly less than those of P_{POS}. The variance (risk) of P_{OWA} is greatly reduced (from 0.26 to 0.15) compared with that of P_{POS}.

Finally, different expected average return values (from 10% to 20%) were used to test the four portfolios based on the two sets of return data, the same conclusion was reached.

6 Concluding Remarks

Decision aggregation (including conflict resolution) is a considerable interest issue in distributed systems such as multi-agent systems. We presented the aggregation procedures by using OWA operator in the agent-based financial investment planning system. OWA aggregation is very flexible, and can take into account as much information as possible. Thus, the aggregation results using OWA are more accurate and reliable. Our work is another successful application of OWA aggregation and extends the application fields of OWA aggregation technique.

References

1. M. Wooldridge and N. R. Jennings, Intelligent Agents: Theory and Practice, *Knowledge Engineering Review*, Vol. 10, No. 2, 1995, 115-152.
2. K. Sycara, K. Decker and D. Zeng, Intelligent Agents in Portfolio Management, in: N. R. Jennings and M. J. Wooldridge (Eds.), *Agent Technology: foundations, Applications, and Markets,* Springer, Berlin, 1998, 267-281.
3. N. R. Jennings, On Agent-Based Software Engineering, *Artificial Intelligence*, 117(1):277-296, 2000.
4. G. Bojadziev and M. Bojadziev, *Fuzzy Logic For Business, Finance, and Management*, World Scientific, Singapore, 1997.
5. Stephen T. Welstead, *Neural Network and Fuzzy Logic Applications in C/C++*, Wiley, New York, 1994.
6. H. Markowitz, *Portfolio Selection: Efficient Diversification of Investments (2nd ed.)*, Blackwell, 1991.
7. H. Tanaka, P. Guo, and I. Turksen, Portfolio Selection Based on Fuzzy Probabilities and Possibility Distributions, *Fuzzy Sets and Systems*, Vol. 111, 2000, 387-397.
8. R. R. Yager, On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making, *IEEE Trans. Syst., Man, Cybern.*, Vol. 18, Jan/Feb, 1988, 183-190.

9. R. R. Yager, Interpreting Linguistically Quantified Propositions, *Int. J. of Intelligent Systems*, Vol. 9, 1994, 541-569.
10. R. R. Yager, New Models of OWA Information Fusion, *Int. J. of Intelligent Systems*, Vol. 13, 1998, 661-681.
11. P. A. Schaefer and H. B. Mitcell, A Generalized OWA Operator, *Int. J. of Intelligent Systems*, Vol. 14, 1999, 123-143.
12. H. B. Mitcell and D. D. Estrakh, An OWA Operator with Fuzzy Ranks, *Int. J. of Intelligent Systems*, Vol. 13, 1998, 69-81.
13. D. P. Filev and R. R. Yager, On the Issue of Obtaining OWA Operator Weights, *Fuzzy Sets and Systems*, Vol. 94, 1998, 157-169.
14. M. O'Hagan, Using Maximum Entropy-Ordered Weighted Averaging to Construct a Fuzzy Neuron, *Proc. 24nd Annual IEEE Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA, 1990, 618-623.
15. R. R. Yager and J. Kacprzyk (Eds.), *The Ordered Weighted Averaging Operators: Theory and Applications*, Kluwer Academic, Norwell, MA, 1997.
16. K. Decker, K. Sycara, and M. Williamson, Middle Agents for the Internet, *Proceedings of 15th International Joint Conference on Artificial Intelligence*, Nogoya, Japan, 1997, 578-583.
17. Z. Zhang and C. Zhang, An Agent-Based Hybrid Intelligent System for Financial Investment Planning, in *Proceedings of PRICAI 2002, LNAI-2417*, pp. 355-364, Springer, August 2002.
18. G. Beliakov, How to Build Aggregation Operators from Data, *International Journal of Intelligent Systems*, Vol. 18, No 8, 2003, 903-923.
19. J. M. Fernandez Salido, S. Murakami, Extending Yagers orness concept for the OWA aggregators to other mean operators, *Fuzzy Sets and Systems*, Vol. 139, 2003, 515-549.

Generated Universal Fuzzy Measures

Radko Mesiar[1,3], Andrea Mesiarová[2], and L'ubica Valášková[1]

[1] Slovak University of Technology,
Radlinského 11, 81368 Bratislava, Slovakia
[2] Mathematical Institute of SAS,
Štefánikova 49, 81473 Bratislava, Slovakia
[3] Institute for Research and Application of Fuzzy Modelling,
University of Ostrava, Czech Republic

Abstract. The concepts of generated universal fuzzy measures and of basic generated universal fuzzy measures are introduced. Special classes and properties of generated universal fuzzy measures are discussed, especially the additive, the symmetric and the maxitive case. Additive (symmetric) basic universal fuzzy measures are shown to correspond to the Yager quantifier-based approach to additive (symmetric) fuzzy measures. The corresponding Choquet integral-based aggregation operators are then generated weighted means (generated OWA operators).

Keywords: aggregation operator, Choquet integral, fuzzy measure, generator, universal fuzzy measure.

1 Introduction

For a fixed number $n \in \mathbb{N}$ of inputs (criteria score), commonly used aggregation techniques are based on some fuzzy integral (Choquet integral, Sugeno integral, etc. [2, 3, 5, 10, 13, 14, 21]) and a fuzzy measure m_n acting on the space $X_n = \{1, \ldots, n\}$ [3, 13, 21]. Recall that m_n is a non-decreasing mapping from $\mathcal{P}(X_n)$ to $[0, 1]$ preserving the bounds, i.e., $m_n(\emptyset) = 0$ and $m_n(X_n) = 1$. However, for an apriori unknown (but still finite) number of inputs, though the applied fuzzy integral is known, a system $(m_n)_{n\in\mathbb{N}}$ of fuzzy measures acting on spaces X_n, $n \in \mathbb{N}$, should be given.

As a measure counterpart of an extended aggregation function $\mathbf{A} : \bigcup_{n\in\mathbb{N}} [0, 1]^n \to [0, 1]$ acting on any finite number of inputs [2, 7], we have introduced in [11] the concept of universal fuzzy measures.

Definition 1. Denote $\mathcal{A} = \{(n, A) \mid n \in \mathbb{N}, A \subset X_n\}$, where $X_n = \{1, \ldots, n\}$, $n \in \mathbb{N}$. A mapping $M : \mathcal{A} \to [0, 1]$ is called a universal fuzzy measure whenever for each $n \in \mathbb{N}$, $M(n, \cdot)$ is non-decreasing, $M(n, \emptyset) = 0$ and $M(n, X_n) = 1$.

Recall that for any fuzzy measure μ acting on $\mathbb{N}$, $\mu : \mathcal{P}(\mathbb{N}) \to [0, 1]$, such that $\mu(\{1\}) > 0$, a universal fuzzy measure $M_\mu : \mathcal{A} \to [0, 1]$ can be introduced putting [11]

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 191–202, 2006.

$$M_\mu(n, A) = \frac{\mu(A)}{\mu(X_n)}. \tag{1}$$

Standard properties of fuzzy measures (additivity, symmetry, maxitivity, submodularity, subadditivity, belief, etc.) are extended to universal fuzzy measures in a natural way, requiring the relevant property to be satisfied by any fuzzy measure $m_n = M(n, \cdot)$, $n \in \mathbb{N}$. For more details we recommend [11] and [20].

Similarly as in the case of extended aggregation functions, formally the fuzzy measures $m_n = M(n, \cdot)$ and $m_k = M(k, \cdot)$ for $n \neq k$ and a given universal fuzzy measure M need not be related.

Two subsequent properties establish some connections between fuzzy measures $m_n = M(n, \cdot)$, $n \in \mathbb{N}$.

Definition 2. Let $M : \mathcal{A} \to [0, 1]$ be a universal fuzzy measure. Then M is called regular whenever for each $n, k \in \mathbb{N}$, $A \subset X_n$, it holds $M(n, A) \geq M(n + k, A)$. Moreover, M is called proportional whenever for each $n, k \in \mathbb{N}$, $A \subset X_n$, it holds $M(n, A) \cdot M(n + k, X_n) = M(n + k, A)$.

Evidently, each proportional universal fuzzy measure M is also regular, but not vice versa.

In decision making procedures based on fuzzy integrals determination of a fuzzy measure m fitting given real data is the crucial task. Several approaches trying to reduce the computational complexity were based on the idea of generated fuzzy measures [18, 19], but dealing with fixed number of inputs. The first approach dealing with possibly different number of inputs (at least known to authors) was presented recently in [1], including a free software tool available on http://www.it.deakin.edu.au/~gleb. However, in each of above mentioned references two types of generated universal fuzzy measures are discussed only and a general concept of generated (universal) fuzzy measures is still missing.

The aim of this paper is to fill this gap, introducing the concept of generated universal fuzzy measures in the next section. Moreover, the concept of basic generated universal fuzzy measures is also introduced. In Section 3, we discuss special classes of generated universal fuzzy measures and of basic generated universal fuzzy measures. Especially, additive basic and symmetric basic generated fuzzy measures are studied in detail (observe that exactly these two types of generated fuzzy measures were discussed and applied in [1, 18, 19]).

2 Generated Universal Fuzzy Measures

Definition 3. Let $g : [0, 1] \to [0, 1]$ be a non-decreasing mapping satisfying $g(0) = 0, g(1) = 1$, and let $h : \mathcal{A} \to \mathcal{P}(\mathbb{N})$ be a mapping non-decreasing in the second coordinate such that $h(n, \emptyset) = \emptyset$ and $h(n, X_n) = X_n$ for all $n \in \mathbb{N}$. Then the mapping g is called a generator, and a mapping $M_{h,g} : \mathcal{A} \to [0, 1]$ given by $M_{h,g}(n, A) = \sum_{i \in h(n,A)} \left(g(\frac{i}{n}) - g(\frac{i-1}{n})\right)$ (with convention $M_{h,g}(n, A) = 0$ if $h(n, A) = \emptyset$) is called a generated universal fuzzy measure. We denote by $\mathcal{G}$ the set of all generators and by $\mathcal{H}$ the set of all mappings h with above mentioned properties.

Mapping $M_{h,g}$ given in the above definition is well-defined universal fuzzy measure since for arbitrary $n \in \mathbb{N}$ it holds:

(i) $h(n, \emptyset) = \emptyset$ and thus $M_{h,g}(n, \emptyset) = 0$
(ii) $h(n, X_n) = X_n$ and thus $M_{h,g}(n, X_n) = 1$
(iii) for $A \subset B \subset X_n$ we have $h(n, A) \subset h(n, B)$ and thus $M(n, A) \leq M(n, B)$.

(i) Let $h_* : \mathcal{A} \to X_n$ be given by

$$h_*(n, A) = \begin{cases} X_n & \text{if } A = X_n, \\ \emptyset & \text{else.} \end{cases}$$

Then $M_{h_*,g}$ is the weakest universal fuzzy measure. On the other hand, let $h^* : \mathcal{A} \to X_n$ be given by

$$h^*(n, A) = \begin{cases} \emptyset & \text{if } A = \emptyset, \\ X_n & \text{else.} \end{cases}$$

Then $M_{h^*,g}$ is the strongest universal fuzzy measure. Independently on the generator g, $M_{h_*,g}(n, \cdot) = m_{*n}$ is the weakest fuzzy measure on X_n and $M_{h^*,g}(n, \cdot) = m_n^*$ is the strongest fuzzy measure on X_n. Evidently, for any $h \in \mathcal{H}$, $n \in \mathbb{N}$, $A \subset X_n$ we have $h_*(n, A) \subseteq h(n, A) \subseteq h^*(n, A)$.

(ii) Let $h : \mathcal{A} \to X_n$ be given by $h(n, A) = X_{|A|}$. Then $M_{h,g}$ is a symmetric universal fuzzy measure given by $M_{h,g}(n, A) = g(\frac{|A|}{n})$.

(iii) Let $h : \mathcal{A} \to X_n$ be given by $h(n, A) = \{\sigma_n(i) \mid i \in A\}$, where for each $n \in \mathbb{N}$, σ_n is a permutation of X_n. Then $M_{h,g}$ is an additive universal fuzzy measure for any generator g. As a special case we can take permutation $\sigma_n(i) = n - i + 1$ and then $M_{h,g}(n, A) = \sum_{i \in A} \left(g(\frac{n-i+1}{n}) - g(\frac{n-i}{n}) \right)$.

(iv) Let $h : \mathcal{A} \to X_n$ be given by $h(n, A) = X_{\max A}$. Then $M_{h,g}$ is a universal fuzzy measure given by $M_{h,g}(n, A) = g(\frac{\max A}{n})$ (with convention $\max \emptyset = 0$).

(v) Let $h : \mathcal{A} \to X_n$ be given by $h(n, A) = \{\min A, \ldots, n\}$. Then $M_{h,g}$ is a universal fuzzy measure given by $M_{h,g}(n, A) = 1 - g(\frac{\min A - 1}{n})$ (with convention $\min \emptyset = n + 1$ valid for each X_n, $n \in \mathbb{N}$).

(vi) Let $h : \mathcal{A} \to X_n$ be given by $h(n, A) = \{\min A, \ldots, \max A\}$. Then $M_{h,g}$ is a universal fuzzy measure given by $M_{h,g}(n, A) = g(\frac{\max A}{n}) - g(\frac{\min A - 1}{n})$ (with the above conventions).

In the following lemma we can see two duality relations for generated universal fuzzy measures. Note that for a given universal fuzzy measure M, its dual $M^d : \mathcal{A} \to [0, 1]$ is given by $M^d(n, A) = 1 - M(n, X_n \setminus A)$.

Lemma 4. $M_{h,g}^d$ [illegible] $M_{h,g}$. [illegible] $h^d(n, A) = X_n \setminus h(n, X_n \setminus A)$, $n \in \mathbb{N}$, $A \subset X_n$ [illegible] h, [illegible] $g^d(x) = 1 - g(1 - x)$, $x \in [0, 1]$ [illegible]

(i) $M^d_{h,g} = M_{h^d,g}$,
(ii) $M_{h,g^d} = M_{\widetilde{h},g}$, where $\widetilde{h}(n,A) = \{n-i+1 \mid i \in h(n,A)\}$, $n \in \mathbb{N}$, $A \subset X_n$.

The proof follows straightforwardly from Definition 3 and hence it is omitted.

Definition 5. Assume mapping $h \in \mathcal{H}$ which is independent of n (we will denote it by γ, i.e., $\gamma(A) = h(n,A)$ for all $n \geq \max A$). Then a generated universal fuzzy measure $M_{\gamma,g}(n,A) = \sum_{i \in \gamma(A)} \left(g(\frac{i}{n}) - g(\frac{i-1}{n})\right)$ will be called a basic generated universal fuzzy measure and the set of all set functions γ with the above property will be denoted by Γ. Moreover, $\gamma \in \Gamma$ will be called a basic set function.

Evidently, a set function $\gamma : \mathcal{F} \to \mathcal{F}$, where $\mathcal{F} = \{A \subset \mathbb{N} \mid A \text{ is finite }\}$, is a basic set function if and only if it is non-decreasing and $\gamma(X_n) = X_n$ for all $n \in \mathbb{N}$, $\gamma(\emptyset) = \emptyset$. The class Γ is a complete lattice induced by a partial ordering $\preceq$, $\gamma_1 \preceq \gamma_2$ if and only if $\gamma_1(A) \subseteq \gamma_2(A)$ for all $A \in \mathcal{F}$. The following result shows that Γ is a bounded lattice.

Proposition 6. For $\gamma_*(A) = X_{\max\{k, X_k \subseteq A\}} = X_{\min(\mathbb{N}\setminus A)-1}$ and $\gamma^*(A) = X_{\max A}$, where $X_0 = \emptyset$, $\gamma_*, \gamma^* \in \Gamma$ and for $\gamma \in \Gamma$ we have $\gamma_*(A) \subseteq \gamma(A) \subseteq \gamma^*(A)$, $A \subset \mathbb{N}$, i.e., γ_* is the smallest basic set function and γ^* is the greatest basic set function.

3 Special Classes of Generated Universal Fuzzy Measures

Recall that a universal fuzzy measure M is called additive if and only if $M(n, A \cup B) = M(n,A) + M(n,B)$ for all $n \in \mathbb{N}$, $A, B \subset X_n$, $A \cap B = \emptyset$. Similarly, M is called symmetric (maxitive) whenever $M(n,A) = M(n,B)$ for $A, B \subset X_n$, $|A| = |B|$ ($M(n, A \cup B) = \max(M(n,A), M(n,B))$ for all $n \in \mathbb{N}$, $A, B \subset X_n$).

For additive, symmetric and maxitive generated universal fuzzy measures we have the following result.

Proposition 7. Let $M_{h,g}$ be
(i) additive if and only if $g \in \mathcal{G}$ and $h(n,A) = h(n, X_{|A|})$ for all $n \in \mathbb{N}$, $A \subset X_n$.
(ii) symmetric if and only if $g \in \mathcal{G}$ and $h(n,A) = \bigcup_{i \in A} h(n, \{i\})$ and $h(n,\{i\}) \cap h(n,\{j\}) = \emptyset$ for all $n \in \mathbb{N}$, $A \subset X_n$, $i, j \in X_n$, $i \neq j$.
(iii) maxitive if and only if $g \in \mathcal{G}$ and $h(n, A \cup B) \in \{h(n,A), h(n,B)\}$ for all $n \in \mathbb{N}$, $A, B \subset X_n$.

Proof. In the necessity parts of the proof for each item (i)-(iii) we will deal with generators $g_n \in \mathcal{G}$, $n \in \mathbb{N}$, determined by $g_n(\frac{i}{n}) = \frac{2^i - 1}{2^n - 1}$ for $i \in \{1, \ldots, n-1\}$ (for example each g_n can be a continuous piecewise linear generator). Then $M_{h,g_n}(n,A) = M_{h,g_n}(n,B)$ imply $h(n,A) = h(n,B)$.

(i) Let $h(n,A) = h(n,X_{|A|})$ for all $n \in \mathbb{N}$, $A \subset X_n$ and let $A_1, A_2 \subset X_n$, $|A_1| = |A_2|$. Then $M_{h,g}(n,A_1) = M_{h,g}(n,X_{|A_1|}) = M_{h,g}(n,A_2)$.
Vice versa, let $M_{h,g}$ be a symmetric generated universal fuzzy measure for each generator g. Assume generator g_n. Then the equality $M_{h,g_n}(n,A) = M_{h,g_n}(n,X_{|A|})$ implies $h(n,A) = h(n,X_{|A|})$.

(ii) Let $h(n,A) = \bigcup\limits_{i\in A} h(n,\{i\})$ and $h(n,\{i\}) \cap h(n,\{j\}) = \emptyset$ for all $n \in \mathbb{N}$, $A \subset X_n$, $i,j \in X_n$, $i \neq j$. Then $M_{h,g}(n,A) = \sum\limits_{i\in h(n,A)} \left(g(\frac{i}{n}) - g(\frac{i-1}{n})\right) = \sum\limits_{i\in \bigcup\limits_{j\in A} h(n,\{j\})} \left(g(\frac{i}{n}) - g(\frac{i-1}{n})\right) = \sum\limits_{j\in A}\sum\limits_{i\in h(n,\{j\})} \left(g(\frac{i}{n}) - g(\frac{i-1}{n})\right) = \sum\limits_{j\in A} M_{h,g}(n,\{j\})$.
Vice versa, let $M_{h,g}$ be an additive generated universal fuzzy measure for each generator g. Then $M_{h,g}(n,A) = \sum\limits_{j\in A} M_{h,g}(n,\{j\})$ for each generator g.
Let us assume generator g_n and let $A \subset X_n$, $|A| = 2$, i.e., $A = \{k,j\}, k \neq j$. Then

$$M_{h,g_n}(n,A) = \sum_{i\in h(n,\{k,j\})} \left(g_n(\frac{i}{n}) - g_n(\frac{i-1}{n})\right) =$$

$$\sum_{i\in h(n,\{k\})} \left(g_n(\frac{i}{n}) - g_n(\frac{i-1}{n})\right) + \sum_{i\in h(n,\{j\})} \left(g_n(\frac{i}{n}) - g_n(\frac{i-1}{n})\right) \quad (2)$$

Denote the sets $h(n,\{k,j\}) = I$, $h(n,\{k\}) = K$, $h(n,\{j\}) = J$. They can be expressed as union of disjoint sets in the following manner:

$$\begin{aligned} I &= (K \setminus J) \cup (J \setminus K) \cup (K \cap J) \cup (I \setminus (K \cup J)),\\ K &= (K \setminus J) \cup (K \cap J),\\ J &= (J \setminus K) \cup (K \cap J). \end{aligned}$$

Then from (2) we have:

$$\sum_{i\in I\setminus(K\cup J)} \left(g_n(\frac{i}{n}) - g_n(\frac{i-1}{n})\right) = \sum_{i\in K\cap J} \left(g_n(\frac{i}{n}) - g_n(\frac{i-1}{n})\right)$$

and from the properties of g_n it follows $I \setminus (K \cup J) = K \cap J$. Since these two sets are disjoint we get $I \setminus (K \cup J) = K \cap J = \emptyset$ and thus also $I = K \cup J$. Hence $h(n,\{k\}) \cap h(n,\{j\}) = \emptyset$ and $h(n,\{k,j\}) = h(n,\{k\}) \cup h(n,\{j\})$ for all $k,j \in X_n$, $k \neq j$. Proof for $A \subset X_n$, $|A| > 2$ can be done analogically with the use of induction.

(iii) Let $h(n,A \cup B) \in \{h(n,A), h(n,B)\}$ for all $n \in \mathbb{N}$, $A,B \subset X_n$. Then $M_{h,g}(n,A \cup B) = M_{h,g}(n,A)$ or $M_{h,g}(n,A \cup B) = M_{h,g}(n,B)$, i.e.,

$$M_{h,g}(n,A \cup B) = \max(M_{h,g}(n,A), M_{h,g}(n,B)).$$

Vice versa, let $M_{h,g}$ be a maxitive generated universal fuzzy measure for each generator g. Then for the generator g_n, $A,B \subset X_n$ we have either

$M_{h,g_n}(n, A \cup B) = M_{h,g_n}(n, A)$ resulting to $h(n, A \cup B) = h(n, A)$ or we have $M_{h,g_n}(n, A \cup B) = M_{h,g_n}(n, B)$ resulting to $h(n, A \cup B) = h(n, B)$, i.e., $h(n, A \cup B) \in \{h(n, A), h(n, B)\}$.

In the following proposition we describe three special types of basic set functions γ.

Proposition 8. [illegible] $M_{\gamma,g}$

(i) [illegible] $g \in \mathcal{G}$ [illegible] $\gamma(A) = X_{|A|}$ [illegible] $A \subset \mathbb{N}$ [illegible] γ_s

(ii) [illegible] $g \in \mathcal{G}$ [illegible] $\gamma(A) = A$ [illegible] $A \subset \mathbb{N}$ γ [illegible] γ_a

(iii) [illegible] $g \in \mathcal{G}$ [illegible] $\gamma(A) = X_{\max A}$, [illegible] $\gamma = \gamma^*$.

The proof in all three cases is either a corollary of Proposition 7 or a modification of its proof.

[illegible] For any maxitive universal fuzzy measure $M_{h,g}$, and any $n \in \mathbb{N}$, $M_{h,g}(n, \cdot) = m_n$ is a possibility measure on X_n, compare [4, 22, 6]. Thus such universal fuzzy measure can be called a universal possibility measure. Evidently, its dual universal fuzzy measure $M^d_{h,g}$ yields a necessity measure $m^d_n = M^d_{h,g}(n, \cdot)$ for each $n \in \mathbb{N}$, see [4, 3, 13, 21], and thus $M^d_{h,g}$ can be called a universal necessity measure. Note that the basic generated universal possibility measure $M_{\gamma^*,g}$ is given by $M_{\gamma^*,g}(n, A) = g(\frac{\max A}{n})$. Similarly as in Proposition 8 (iii) we can show that the unique basic generated necessity measure is $M_{\gamma_*,g}$ and it is given by

$$M_{\gamma_*,g}(n, A) = g(\frac{\min A^c - 1}{n}),$$

where $A^c = \mathbb{N} \setminus A$.

[illegible] A symmetric basic universal fuzzy measure generated by a generator $g \in \mathcal{G}$ is given by $M_{\gamma_s,g}(n, A) = g(\frac{|A|}{n})$, and an additive basic universal fuzzy measure generated by a generator $g \in \mathcal{G}$ is given by $M_{\gamma_a,g}(n, A) = \sum_{i \in A} \left(g(\frac{i}{n}) - g(\frac{i-1}{n})\right)$. These two types of generated universal fuzzy measures were exploited and discussed in [1, 18]. In the following proposition we describe conditions for additive and symmetric basic universal fuzzy measures to be generated. First, let us note that for a universal fuzzy measure M_g generated by a generator $g \in \mathcal{G}$ it is enough to know the values of the generator g only on $[0, 1] \cap \mathbb{Q}$. Due to the required monotonicity of generator g, for irrational $x \in]0, 1[$ we can always assume $g(x) = \sup\{g(r) \mid r \in [0, x] \cap \mathbb{Q}\}$.

Proposition 9. (i) [illegible] M [illegible] $n, k, t \in \mathbb{N}$, $t \leq n$ [illegible] $M(n, X_t) = M(kn, X_{kt})$.

(ii) M [illegible] $n, k, t \in \mathbb{N}$, $t \leq n$ [illegible] $M(n, X_t) = M(kn, X_{kt})$ [illegible] $M(n, \{t\}) = M(kn, X_{kt} \setminus X_{k(t-1)})$

[illegible] (i) Let a symmetric universal fuzzy measure M be a basic generated universal fuzzy measure, i.e., $M(n, A) = g(\frac{|A|}{n})$. Then $M(n, X_t) = g(\frac{t}{n}) = g(\frac{kt}{kn}) = M(kn, X_{kt})$.

Vice versa, let a symmetric universal fuzzy measure M fulfils $M(n, X_t) = M(kn, X_{kt})$ for all $n, k, t \in \mathbb{N}$, $t \leq n$. Let us define a mapping $g : [0,1] \cap \mathbb{Q} \to [0,1]$ by putting $g(\frac{p}{q}) = M(q, X_p)$. We will show that g is a generator of universal fuzzy measure M. The mapping g is well-defined since for $\frac{p}{q} = \frac{r}{s}$ we have $g(\frac{p}{q}) = M(q, X_p) = M(sq, X_{sp}) = M(sq, X_{qr}) = M(s, X_r) = g(\frac{r}{s})$. Moreover, $g(\frac{0}{q}) = M(q, \emptyset) = 0$ and $g(\frac{1}{1}) = M(1, X_1) = 1$ and g is non-decreasing since for $\frac{p}{q} \leq \frac{r}{s}$ we have $g(\frac{p}{q}) = M(q, X_p) = M(sq, X_{sp}) \leq M(sq, X_{qr}) = M(s, X_r) = g(\frac{r}{s})$. Using the facts that $[0,1] \cap \mathbb{Q}$ is dense in $[0,1]$ and that g is non-decreasing we can extend the mapping g to the whole interval $[0,1]$. We get $M(n, A) = M(n, X_{|A|}) = g(\frac{|A|}{n})$, i.e., g is a generator of universal fuzzy measure M, i.e., $M = M_{\gamma_s, g}$.

(ii) First we will show that for $n, k, t \in \mathbb{N}$, $t \leq n$ conditions $M(n, X_t) = M(kn, X_{kt})$ and $M(n, \{t\}) = M(kn, X_{kt} \setminus X_{k(t-1)})$ are equivalent.

Let M be an additive universal fuzzy measure such that

$$M(n, X_t) = M(kn, X_{kt}).$$

Then because of additivity of M we have $\sum_{i=1}^{t} M(n, \{i\}) = M(n, X_t) = M(kn, X_{kt}) = \sum_{j=1}^{kt} M(kn, \{j\})$.

For $t = 1$ we have $M(n, \{1\}) = \sum_{j=1}^{k} M(kn, \{j\})$.

For $t = 2$ we have

$$M(n, \{1\}) + M(n, \{2\}) = \sum_{j=1}^{k} M(kn, \{j\}) + \sum_{j=k+1}^{2k} M(kn, \{j\}),$$

i.e., we get $M(n, \{2\}) = \sum_{j=k+1}^{2k} M(kn, \{j\})$ and analogically $M(n, \{t\}) = \sum_{j=(t-1)k+1}^{tk} M(kn, \{j\})$.

Vice versa, let M be an additive universal fuzzy measure such that

$$M(n, \{t\}) = M(kn, X_{kt} \setminus X_{k(t-1)}).$$

Then $M(n, X_t) = \sum_{i=1}^{t} M(n, \{i\}) = \sum_{i=1}^{t} \sum_{j=(i-1)k+1}^{ik} M(kn, \{j\}) = \sum_{j=1}^{kt} M(kn, \{j\}) = M(kn, X_{kt})$.

Let M be an additive basic universal fuzzy measure generated by a generator $g \in \mathcal{G}$, i.e., $M(n,A) = \sum_{i\in A}\left(g(\frac{i}{n}) - g(\frac{i-1}{n})\right)$. Then $M(n,X_t) = \sum_{i=1}^{t}\left(g(\frac{i}{n}) - g(\frac{i-1}{n})\right) = g(\frac{t}{n}) = g(\frac{kt}{kn}) = \sum_{i=1}^{kt}\left(g(\frac{i}{kn}) - g(\frac{i-1}{kn})\right) = M(kn, X_{kt})$.
Vice versa, let M be an additive basic universal fuzzy measure such that $M(n,X_t) = M(kn,X_{kt})$. Let us assume the mapping g from part (i) of this proof. We will show that g generates M. We have $M(n,A) = \sum_{i\in A} M(n,\{i\}) = \sum_{i\in A}(M(n,X_i) - M(n,X_{i-1})) = \sum_{i\in A}\left(g(\frac{i}{n}) - g(\frac{i-1}{n})\right)$, i.e., $M = M_{\gamma_a,g}$.

Similarly, we can show that a maxitive universal fuzzy measure M is a maxitive basic generated universal fuzzy measure if $M(n,A) = M(n,X_{\max A})$ and $M(n,X_t) = M(kn,X_{kt})$ for all $n,k,t \in \mathbb{N}$, $t \le n$, $A \subset X_n$. However, there are maxitive basic generated universal fuzzy measures not fulfilling the above conditions.

In the following propositions we describe the properties of a generator which generates an additive (symmetric) universal fuzzy measure which is regular (proportional).

Proposition 10. [illegible] $M_{\gamma_s,g}$ [illegible]

[illegible] Let $M_{\gamma_s,g}$ be a symmetric basic universal fuzzy measure generated by the generator g, $n_1 \le n_2$ and $A \subset X_{n_1}$. Then $\frac{|A|}{n_1} \ge \frac{|A|}{n_2}$ and since g is non-decreasing we get $g(\frac{|A|}{n_1}) \ge g(\frac{|A|}{n_2})$, i.e., $M_{\gamma_s,g}(n_1,A) \ge M_{\gamma_s,g}(n_2,A)$, i.e., $M_{\gamma_S,g}$ is regular.

Proposition 11. [illegible] $g \in \mathcal{G}$ [illegible] $]0,1[$. [illegible]

(i) $x \cdot g'(x)$ [illegible] $]0,1[$.
(ii) [illegible] $M_{\gamma_a,g}$ [illegible]
(iii) [illegible] $M_{\gamma,g}$ [illegible] $\gamma \in \Gamma$.

[illegible] (i) $\Rightarrow$ (ii) Let $x\cdot g'(x)$ be non-decreasing on $]0,1[$. Then for all $t \in [\frac{i}{n+1}, \frac{i}{n}]$ we have

$$t \cdot g'(t) \ge \frac{i}{n+1} g'(\frac{i}{n+1}),$$

$$g'(t) \ge \frac{i}{n+1} g'(\frac{i}{n+1}) \frac{1}{t},$$

$$\int_{\frac{i}{n+1}}^{\frac{i}{n}} g'(t)\mathrm{d}t \ge \frac{i}{n+1} g'(\frac{i}{n+1}) \int_{\frac{i}{n+1}}^{\frac{i}{n}} \frac{1}{t}\mathrm{d}t$$

$$\int_{\frac{i}{n+1}}^{\frac{i}{n}} g'(t)\mathrm{d}t \geq \frac{i}{n+1} g'(\frac{i}{n+1}) \ln \frac{n+1}{n}.$$

Similarly for all $t \in [\frac{i-1}{n+1}, \frac{i-1}{n}]$ we have

$$\frac{i-1}{n} g'(\frac{i-1}{n}) \geq t \cdot g'(t),$$

$$\frac{i-1}{n} g'(\frac{i-1}{n}) \ln \frac{n+1}{n} \geq \int_{\frac{i-1}{n+1}}^{\frac{i-1}{n}} g'(t)\mathrm{d}t.$$

Since $x \cdot g'(x)$ is non-decreasing we have:

$$\frac{i}{n+1} g'(\frac{i}{n+1}) \ln \frac{n+1}{n} \geq \frac{i-1}{n} g'(\frac{i-1}{n}) \ln \frac{n+1}{n},$$

and together we get

$$\int_{\frac{i}{n+1}}^{\frac{i}{n}} g'(t)\mathrm{d}t \geq \int_{\frac{i-1}{n+1}}^{\frac{i-1}{n}} g'(t)\mathrm{d}t,$$

i.e,

$$g(\frac{i}{n}) - g(\frac{i}{n+1}) \geq g(\frac{i-1}{n}) - g(\frac{i-1}{n+1})$$

for all $n \in \mathbb{N}$, $i = 1, \ldots, n$. Then $M_{\gamma_a,g}(n, \{i\}) \geq M_{\gamma_a,g}(n+1, \{i\})$ and from additivity of $M_{\gamma_a,g}$ it follows that $M_{\gamma_a,g}$ is regular.

(ii) $\Rightarrow$ (i) Let $M_{\gamma_a,g}$ be regular. Then $M_{\gamma_a,g}(n, A) \geq M_{\gamma_a,g}(n+1, A)$ for all $n \in \mathbb{N}$ and $A \subset X_n$. For any $x, y \in]0, 1[\cap \mathbb{Q}, x > y$ there are $n, i, j \in \mathbb{N}$, $j < i < n$ such that $x = \frac{i}{n}$ and $y = \frac{j}{n}$. Then also for all $k \in \mathbb{N}$ we have $x = \frac{ki}{kn}$ and $y = \frac{kj}{kn}$. From the regularity of $M_{\gamma_a,g}$ we have

$$M_{\gamma_a,g}(kn, X_{ki} \setminus X_{kj}) \geq M_{\gamma_a,g}(kn+1, X_{ki} \setminus X_{kj}),$$

i.e.,

$$g(\frac{ki}{kn}) - g(\frac{kj}{kn}) \geq g(\frac{ki}{kn+1}) - g(\frac{kj}{kn+1})$$

$$g(\frac{ki}{kn}) - g(\frac{ki}{kn+1}) \geq g(\frac{kj}{kn}) - g(\frac{kj}{kn+1})$$

$$g(x) - g(x - \frac{x}{kn+1}) \geq g(y) - g(y - \frac{y}{kn+1})$$

$$x \frac{g(x) - g(x - \frac{x}{kn+1})}{\frac{x}{kn+1}} \geq y \frac{g(y) - g(y - \frac{y}{kn+1})}{\frac{y}{kn+1}}.$$

And since $\lim_{k\to\infty} x\frac{g(x)-g(x-\frac{x}{kn+1})}{\frac{x}{kn+1}} = x\cdot g'(x)$ and $\lim_{k\to\infty} y\frac{g(y)-g(y-\frac{y}{kn+1})}{\frac{y}{kn+1}} = y\cdot g'(y)$ we get $x\cdot g'(x) \geq y\cdot g'(y)$. However, $]0,1[\cap\mathbb{Q}$ is dense in $]0,1[$ and g is continuously differentiable on $]0,1[$ and thus $x\cdot g'(x) \geq y\cdot g'(y)$ for all $x,y\in\,]0,1[\,, x>y$, i.e., $x\cdot g'(x)$ is non-decreasing on $]0,1[$.

(ii) $\Rightarrow$ (iii) Let $M_{\gamma_a,g}$ be regular and let $\gamma\in\Gamma$. Then $M_{\gamma_a,g}(n,\{i\}) \geq M_{\gamma_a,g}(n+1,\{i\})$ for all $n\in\mathbb{N}$ and $i\in X_n$, i.e.,

$$g(\frac{i}{n}) - g(\frac{i-1}{n}) \geq g(\frac{i}{n+1}) - g(\frac{i-1}{n+1})$$

and thus

$$\sum_{i\in\gamma(A)} \Big(g(\frac{i}{n}) - g(\frac{i-1}{n})\Big) \geq \sum_{i\in\gamma(A)} \Big(g(\frac{i}{n+1}) - g(\frac{i-1}{n+1})\Big)$$

for all $n\in\mathbb{N}$, $A\subset X_n$, i.e, $M_{\gamma,g}(n,A) \geq M_{\gamma,g}(n+1,A)$, i.e., $M_{\gamma,g}$ is regular.

(iii) $\Rightarrow$ (ii) This implication is obvious.

Proposition 12. (i) [illegible] $M_{\gamma_s,g}$ [illegible] $g\in\mathcal{G}$ [illegible] g [illegible] $g(x)=x^r$, $r>0$.

(ii) [illegible] $M_{\gamma_a,g}$ [illegible] $g\in\mathcal{G}$ [illegible] g [illegible] $g(x)=x^r$, $r>0$.

[illegible] (i) Let M be a symmetric basic universal fuzzy measure generated by a generator $g\in\mathcal{G}$, i.e., $M=M_{\gamma_s,g}$. Let $g(x)=x^r$, $r>0$. Then $M(n+k,A) = g(\frac{|A|}{n+k}) = (\frac{|A|}{n+k})^r = (\frac{|A|}{n})^r(\frac{|X_n|}{n+k})^r = g(\frac{|A|}{n})\cdot g(\frac{|X_n|}{n+k}) = M(n,A)\cdot M(n+k,X_n)$, i.e., $M_{\gamma_s,g}$ is proportional.

Vice versa, let $M_{\gamma_s,g}$ be proportional. Then $M(n+k,A) = g(\frac{|A|}{n+k}) = g(\frac{|A|}{n})\cdot g(\frac{|X_n|}{n+k}) = M(n,A)\cdot M(n+k,X_n)$ for all $k,n\in\mathbb{N}$, $A\subset X_n$. This means that the mapping g is a solution of the Cauchy functional equation $g(ab) = g(a)g(b)$ for all $a,b\in[0,1]\cap\mathbb{Q}$. Since $[0,1]\cap\mathbb{Q}$ is dense in $[0,1]$ and g is bounded and non-decreasing on $[0,1]$ we get $g(x)=x^r$, $r>0$.

(ii) Let M be an additive basic universal fuzzy measure generated by a generator $g\in\mathcal{G}$, i.e., $M=M_{\gamma_a,g}$. Let $g(x)=x^r$, $r>0$. Then $M(n+k,A) = \sum_{i\in A}\Big(g(\frac{i}{n+k}) - g(\frac{i-1}{n+k})\Big) = \sum_{i\in A}\Big((\frac{i}{n+k})^r - (\frac{i-1}{n+k})^r\Big) = \sum_{i\in A}\big((\frac{i}{n})^r - (\frac{i-1}{n})^r\big)\cdot(\frac{n}{n+k})^r = \sum_{i\in A}\big(g(\frac{i}{n}) - g(\frac{i-1}{n})\big)\cdot\sum_{i=1}^{n}\Big(g(\frac{i}{n+k}) - g(\frac{i-1}{n+k})\Big) = M(n,A)\cdot M(n+k,X_n)$, i.e., $M_{\gamma_a,g}$ is proportional.

Vice versa, let $M_{\gamma_a,g}$ be proportional. Then $M(n+k,A) = M(n,A)\cdot M(n+k,X_n)$, i.e.,

$$\sum_{i\in A}\left(g(\frac{i}{n+k})-g(\frac{i-1}{n+k})\right)=$$

$$\sum_{i\in A}\left(g(\frac{i}{n})-g(\frac{i-1}{n})\right)\cdot\sum_{i=1}^{n}\left(g(\frac{i}{n+k})-g(\frac{i-1}{n+k})\right),$$

$$\sum_{i\in A}\left(g(\frac{i}{n+k})-g(\frac{i-1}{n+k})\right)=\sum_{i\in A}\left(g(\frac{i}{n})-g(\frac{i-1}{n})\right)\cdot g(\frac{n}{n+k}).$$

Let $A=X_n\setminus X_{n-t-1}$, $t\in\{0,1,\ldots,n-1\}$. We get

$$g(\frac{n}{n+k})-g(\frac{n-t-1}{n+k})=\left(1-g(\frac{n-t-1}{n})\right)\cdot g(\frac{n}{n+k}),$$

i.e.,

$$g(\frac{n-t-1}{n+k})=g(\frac{n-t-1}{n})\cdot g(\frac{n}{n+k})$$

for all $k,n\in\mathbb{N}$. Hence, analogically as in part (i) g is a power function, $g(x)=x^r$, $r>0$.

4 Conclusion

We have introduced generated universal fuzzy measures. Moreover, we have discussed several properties of these universal fuzzy measures. As a promising field of applications of introduced concepts we can mention fuzzy integrals with respect to universal fuzzy measures yielding extended aggregation functions. For example, the Choquet integral [3, 13, 21] with respect to the additive basic generated universal fuzzy measure $M_{\gamma_a,g}$ is just the weighted mean W related to the weighting triangle $\triangle=(w_{in}\mid n\in\mathbb{N},i\in\{1,\ldots,n\})$, $w_{in}=g(\frac{i}{n})-g(\frac{i-1}{n})$. Moreover, the Choquet integral with respect to the symmetric basic generated universal fuzzy measure $M_{\gamma_s,g}$ yields the OWA operator [17, 19] related to already mentioned weighting triangle $\triangle$. Similarly, the Sugeno integral [13, 15, 21] with respect to the maxitive basic generated universal fuzzy measure $M_{\gamma^*,g}$ yields a weighted maximum operator $\mathbf{A}:\bigcup_{n\in\mathbb{N}}[0,1]^n\to[0,1]$ given by $A(x_1,\ldots,x_n)=\max_i(\min(x_i,g(\frac{i}{n})))$.

Another field of applications of (generated) universal fuzzy measures is the problem of densities of subsets of $\mathbb{N}$, see, e.g., [12, 16]. Indeed, for any subset $E\subset\mathbb{N}$, set functions $d^*,d_*:\mathcal{P}(\mathbb{N})\to[0,1]$ given by (for a fixed universal fuzzy measure M) $d^*(E)=\limsup_n M(n,E\cap X_n)$ and $d_*(E)=\liminf_n M(n,E\cap X_n)$ are densities in the sense of [12, 16].

Acknowledgement. This work was supported by grants MSM 619 8898 701, APVT 20-046402, APVT 20-003204 and VEGA 1/2032/05, VEGA 2/ 3163/23. We are grateful to the anonymous referees for their valuable comments.

References

1. Beljakov, G., Mesiar, R., Valášková, Ľ.: Fitting generated aggregation operators to empirical data. Int. J. Uncertainty, Fuzziness and Knowledge-Based System **12** (2004) 219–236
2. Calvo, T., Kolesárová, A., Komorníková, M., Mesiar, R.: Aggregation Operators: Properties, Classes and Construction Methods. In: T. Calvo, G. Mayor, R. Mesiar, eds. Aggregation Operators. Physica- Verlag, Heidelberg (2002) 3–107
3. Denneberg, D.: Non-additive Measure and Integral. Kluwer Acad. Publ., Dordrecht (1994)
4. Dubois, D., Prade, H.: Fuzzy Sets and Systems: Theory and Applications. Academic Press, New York (1980)
5. Grabisch, M., Murofushi, T., Sugeno, M.: Fuzzy Measures and Integrals. Physica-Verlag, Heidelberg (2000)
6. Klir, G., J., Folger, T., A.: Fuzzy Sets. Uncertainty, and Information. Prentice-Hall, Englewoods Cliffs, NJ (1988)
7. Kolesárová, A., Komorníková, M.: Triangular norm-based iterative compensatory operators. Fuzzy Sets and Systems **104** (1999) 109–120
8. Komorníková, M.: Generated aggregation operators. Proc. EUSFLAT'99, Palma de Mallorca, (1999) 355–358.
9. Komorníková, M.: Aggregation operators and additive generators. Int. J. Uncertainty, Fuzziness and Knowledge-Based System **9** (2001) 205–215
10. Mesiar, R., Mesiarová, A.: Fuzzy integrals. Modeling Decisions for Artificial Intelligence (V. Torra, and Y. Narukawa, eds.), LNAI 3131, Springer, Berlin, (2004) 7–14
11. Mesiar, R., Valášková, Ľ.: Universal fuzzy measures. Proc. 10th IFSA World Congress, Istanbul, Turkey (2003) 139-142
12. Mišík, L.: Sets of positive integers with prescribed values of densities. Math. Slovaca **52** (2002) 289–296
13. Pap, E.: Null-additive Set Functions. Kluwer Acad. Publ., Dordrecht (1995)
14. Struk, P.: Extremal fuzzy integrals. Soft Computing **9** (2005) in press.
15. Sugeno, M.: Theory of Fuzzy Integrals and Applications. PhD Thesis, Tokyo Institute of Technology (1974)
16. Štrauch, O., Tóth, J.: Asymptotic density and density of ratio set $R(A)$. Acte Arithmetica **LXXXVII** (1998) 67–78
17. Yager, R., R.: On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Trans. Syst., Man, Cybern. **18** (1988) 183–190
18. Yager, R., R., Filev, D., P.: Essentials of Fuzzy Modelling and Control. J. Wiley & Sons, NewYork (1994)
19. Yager, R., R., Kacprzyk, J.: The Ordered Weighted Averaging Operators, Theory and Applications. Kluwer Academic Publishers, Boston (1997)
20. Valášková, Ľ.: Non-additive Measures and Integrals. PhD Thesis, Slovak University of Technology (2005)
21. Wang, Z., Klir, G., J.: Fuzzy Measure Theory. Plenum Press, New York (1992)
22. Zadeh, L., A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems **1** (1978) 3–28

Aggregation of Valued Relations Applied to Association Rule Interestingness Measures

Jean-Pierre Barthélemy[1], Angélique Legrain[1], Philippe Lenca[1], and Benoît Vaillant[1,2,⋆]

[1] GET / ENST Bretagne – Cnrs Umr 2872 Tamcic, Technopôle de Brest Iroise, CS 83818, 29238 Brest Cedex 3, France
forename.name@enst-bretagne.fr
[2] UBS - IUT de Vannes - Département STID, 8, rue Montaigne, BP 561, 56017 Vannes, France

Abstract. One of the concerns of knowledge discovery in databases is the production of association rules. An association rule $\mathtt{A} \rightarrow \mathtt{B}$ defines a relationship between two sets of attributes A and B, caracterising the data studied. Such a rule means that objects sharing attributes of A will "likely" have those contained in B. Yet, this notion of "likeliness" depends on the datamining context.

Many interestingness measures have been introduced in order to quantify this likeliness. This panel of measures is heterogeneous and the ranking of extracted rules, according to measures, may differ largely.

This contribution explores a new approach for assessing the quality of rules: aggregating valued relations. For each measure, a valued relation is built out of the numerical values it takes on the rules, and represents the preference of a rule over another. The aim in using such tools is to take into account the intensity of preference expressed by various measures, and should reduce incomparability issues related to differences in their co-domains. It also has the advantage of relating the numerical nature of measures compared to pure binary approaches.

We studied several aggregation operators. In this contribution we discuss results obtained on a toy example using the simplest of them.

1 Basic Considerations

1.1 Association Rules

An aim of data mining concerns the extraction of informations (also called "knowledge discovery") from very large data bases (or data warehouses). In this paper we shall restrict to the case where data are objects (or recordings, according to the database terminology), described by q binary attributes (or properties): either the object i has the property x, or not. We shall denote by

⋆ Benoît Vaillant would like to thanks Brest Métropole Océane for financial support of his Ph.D. thesis.

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 203–214, 2006.

$N = \{1, \ldots, n\}$ the set of considered objects and by $Q = \{a, b, \ldots\}$ the set of properties. The searched information will be restricted to association rules [1]. The general idea when considering the association rule $\mathtt{A} \rightarrow \mathtt{B}$ is that if a subset of N significantly possesses the features of the set $\mathtt{A} \subseteq Q$, then it significantly possesses the features from the set $\mathtt{B} \subseteq Q$. It is important to point out that such rules, by construction, focus only on the expressed properties of the objects, that a rule will never express the fact that a set of objects may not posses a set of properties. Without assessing the notion of significance for the moment, these considerations lead to the following formal definition:

Definition. A B Q $\mathtt{A} \cap \mathtt{B} = \emptyset$

The association rule $(\mathtt{A}, \mathtt{B})$ is usually denoted $\mathtt{A} \rightarrow \mathtt{B}$. There are, $3^q - q - 1$ possible rules (each property either belonging to A, B or neither of them). Galois lattices constitute a natural framework for this approach [17]. Comming back to the notion of significance, not any $\mathtt{A} \rightarrow \mathtt{B}$ is eligible as a realistic association rule. An extreme case is pure implication, $\mathtt{A} \Rightarrow \mathtt{B}$, which means that if an object fulfils all the properties from A (head of the rule), then it definitely fulfils all the properties from B (tail of the rule). Generaly speaking, eligible association rules are filtered from their interestingness: a rule must be supported by enough examples and may be contradicted by a small proportion of counter examples. To make this precise we introduce the following quantities: for $\mathtt{A}, \mathtt{B} \subseteq Q$, $n_{\mathtt{A}}$ is the number of objects fulfilling all the properties from A and $n_{\mathtt{AB}}$ denotes the number of objects fulfilling all the properties from $\mathtt{A} \cup \mathtt{B}$. While $n_{\mathtt{AB}}$ involves the notion of examples, $n_{\mathtt{AB}}/n_{\mathtt{A}}$ relativizes it as a proportion of examples, and $n_{\mathtt{A}}$-$n_{\mathtt{AB}}$ involves the notion of counter-examples.

From these quantities, we deduce the probabilities $p_{\mathtt{A}} = n_{\mathtt{A}}/n$, $p_{\mathtt{AB}} = n_{\mathtt{AB}}/n$ and the conditional probability $p_{\mathtt{B|A}} = p_{\mathtt{AB}}/p_{\mathtt{A}}$. The quantities $p_{\mathtt{AB}}$ and $p_{\mathtt{AB}}/p_{\mathtt{A}}$ are respectively called the **support** and the **confidence** of the rule $\mathtt{A} \rightarrow \mathtt{B}$.

In order to construct a set $\mathcal{R}$ of "interesting rules" we used Apriori [7], an implementation of the Ais algorithm proposed by [1]. This implementation only extracts rules of the form $A \rightarrow \{b\}$ (simply written as $A \rightarrow b$). The arguments for such a (classical) restriction are that in the first place, association rule mining usually produces too many rules, even if one confines oneself to rules with only one item in the consequent. Considering more complex rules makes the situation even worse. This requirement decreases the maximal cardinality of $\mathcal{R}$: there are at most $q.2^{q-1}$ rules when the tail is restricted to a single property. Secondly, more complex rules add few to the insights about the data. Indeed, should a complex rule be valid, then its simpler forms will be valid to, and thus extracted from the database. Even if presenting to an end-user a set of simple rules or its more complex form is not the same, the final gain is usually considered as non-worthy. Apriori follows a two steps approach in order to generate all rules having support and confidence values above some user-defined thresholds:

(i) Select the sets of "frequent properties", i.e. the sets $X \subseteq Q$ such that $p_X \geq \alpha$, where α is a given threshold (and will be the minimum support of each rule).

(ii) For each frequent set X, generate all the rules of the form $A \rightarrow b$, where $A = X \backslash \{b\}$, and select those whose confidence is largest than some threshold β (which will hence be the minimum confidence of each rule).

1.2 A Toy Example

To illustrate the notions introduced above, we propose a small dataset (see table 1) of 10 objects described by 7 properties. Table 2 lists the set of 21 rules extracted by the Apriori algorithm [7], using a support threshold of .30 and a confidence threshold of .60.

Table 1. Toy example

	a	b	c	d	e	f	g
01	0	0	1	1	1	0	0
02	1	0	1	0	1	1	0
03	1	0	1	1	0	0	1
04	0	1	0	1	0	1	0
05	1	1	0	1	1	0	1
06	1	1	1	1	1	0	1
07	0	1	0	1	1	1	0
08	1	0	0	1	0	1	0
09	0	0	1	1	0	1	0
10	0	1	1	0	1	1	0

Table 2. Rules extracted by Apriori

id	head $\rightarrow$ tail	p_A	p_{AB}	p_{AB}/p_A
r_1	{g} $\rightarrow$ a	.30	.30	1
r_2	{a} $\rightarrow$ g	.50	.30	.60
r_3	{g} $\rightarrow$ d	.30	.30	1
r_4	{b} $\rightarrow$ f	.50	.30	.60
r_5	{f} $\rightarrow$ d	.60	.40	.67
r_6	{b} $\rightarrow$ e	.50	.40	.80
r_7	{e} $\rightarrow$ b	.60	.40	.67
r_8	{b} $\rightarrow$ d	.50	.40	.80
r_9	{a} $\rightarrow$ c	.50	.30	.60
r_{10}	{a} $\rightarrow$ e	.50	.30	.60
r_{11}	{a} $\rightarrow$ d	.50	.40	.80
r_{12}	{c} $\rightarrow$ e	.60	.40	.67
r_{13}	{e} $\rightarrow$ c	.60	.40	.67
r_{14}	{c} $\rightarrow$ d	.60	.40	.67
r_{15}	{e} $\rightarrow$ d	.60	.40	.67
r_{16}	{g, a} $\rightarrow$ d	.30	.30	1
r_{17}	{g, d} $\rightarrow$ a	.30	.30	1
r_{18}	{a, d} $\rightarrow$ g	.40	.30	.75
r_{19}	{b, e} $\rightarrow$ d	.40	.30	.75
r_{20}	{b, d} $\rightarrow$ e	.40	.30	.75
r_{21}	{e, d} $\rightarrow$ b	.40	.30	.75

1.3 Interestingness Measures

Support and confidence constitute the usual framework used to extract a set $\mathcal{R}$ of rules, but they are not the only way of assessing the quality of a rule. On the contrary, they are considered as having only very few good properties for ranking a set of rules from the best to the lesser good ones, when compared to other interestingness measures ([31, 34, 25]). Still, they remain frequently used for algorithmic reasons and as references. We have selected and studied 20 more or less usual measures from the literature (see table 3 and [23]). Properties of these 20 measures have been thoroughly analysed, as in [34, 20] for example. In [26], we propose a Multicriteria Decision Aid approach aiming at selecting an interestingness measure, according to data and user expectations.

1.4 An Heterogeneous Landscape

At first glance, table 3 shows important variations between the formulae. This is due to the fact that measures do not tell the same story. This is also due to the fact that co-domains are quite different ([0, 1], [0, $+\infty$[,]$-\infty$, 1] and others... with

Table 3. Association rule quality measures

		Definition	Co-domain	Reference
CONF	confidence	$\frac{n_{ab}}{n_a}$	$[0,\ 1]$	[1]
CENCONF	centred confidence	$\frac{nn_{ab}-n_an_b}{nn_a}$	$[-\frac{n_b}{n},\ \frac{n_{\bar{b}}}{n}]$	
CONV	conviction	$\frac{n_an_{\bar{b}}}{nn_{a\bar{b}}}$	$[\frac{n_{\bar{b}}}{n},\ +\infty[$	[9]
IG	information gain	$\log(\frac{nn_{ab}}{n_an_b})$	$]-\infty,\ \log\frac{n}{n_b}]$	[10]
-IMPIND	implication index	$-\frac{n_an_b-nn_{ab}}{\sqrt{nn_an_b}}$	$[-\frac{\sqrt{n_an_b}}{\sqrt{nn_b}},\ \sqrt{\frac{n_an_{\bar{b}}}{n}}]$	[15]
INTIMP	intensity of implication	$P\Big[\mathcal{N}(0,1) \geq ImpInd\Big]$	$[0,\ 1]$	[15]
EII	entropic intensity of implication	$\Big\{[(1-h_1(\frac{n_ab}{n})^2)(1-h_2(\frac{n_{a\bar{b}}}{n})^2)]^{1/4}IntImp\Big\}^{1/2}$	$[0,\ 1]$	[16]
KAPPA	Kappa	$2\frac{nn_{ab}-n_an_b}{nn_a+nn_b-2n_an_b}$	$[-2\frac{n_an_b}{n_an_{\bar{b}}+n_{\bar{a}}n_b},\ 2\frac{n_an_{\bar{b}}}{n_an_{\bar{b}}+n_{\bar{a}}n_b}]$	[11]
LAP	Laplace	$\frac{n_{ab}+1}{n_a+2}$	$[\frac{1}{n_a+2},\ \frac{n_a+1}{n_a+2}]$	[14]
LIFT	lift	$\frac{nn_{ab}}{n_an_b}$	$[0,\ \frac{n}{n_b}]$	[18]
LOE	Loevinger	$\frac{nn_{ab}-n_an_b}{n_an_{\bar{b}}}$	$[-\frac{n_b}{n_{\bar{b}}},\ 1]$	[28]
BF	Bayes factor	$\frac{n_{ab}n_{\bar{b}}}{n_bn_{a\bar{b}}}$	$[0,\ +\infty[$	[19]
PS	Piatetsky-Shapiro	$\frac{nn_{ab}-n_an_b}{n^2}$	$[-\frac{n_an_b}{n^2},\ \frac{n_an_{\bar{b}}}{n^2}]$	[31]
PDI	probabilistic discriminant index	$P\Big[\mathcal{N}(0,1) > ImpInd^{CR/B}\Big]$	$[0^+,\ 1^-]$	[27]
R	linear correlation coefficient	$\frac{nn_{ab}-n_an_b}{\sqrt{nn_an_bn_{\bar{a}}n_{\bar{b}}}}$	$[-\sqrt{\frac{n_an_b}{nn_{\bar{a}}n_{\bar{b}}}},\ \sqrt{\frac{n_an_{\bar{b}}}{nn_{\bar{a}}n_b}}]$	[29]
SEB	Sebag-Schoenauer	$\frac{n_{ab}}{n_{a\bar{b}}}$	$[0,\ +\infty[$	[33]
SUP	support	$\frac{n_{ab}}{n}$	$[0,\ \frac{n_a}{n}]$	[1]
LC	least contradiction	$\frac{n_{ab}-n_{a\bar{b}}}{nn_b}$	$[-\frac{n_{\bar{a}}}{nn_b},\ \frac{n_a}{nn_b}]$	[3]
ECR	example and counter example rate	$\frac{n_{a\bar{b}}-n_{a\bar{b}}}{n_{ab}} = 1-\frac{1}{\frac{n_a}{n_{ab}}-1}$	$]-\infty,\ 1]$	
ZHANG	Zhang	$\frac{nn_{ab}-n_an_b}{\max\{n_{ab}n_{\bar{b}},n_bn_{a\bar{b}}\}}$	$[-1,\ 1]$	[37]

$ImpInd^{CR/\mathcal{B}}$ corresponds to IMPIND, centred reduced (CR) for a rule set $\mathcal{B}$

$h_1(t) = -(1-\frac{n_\cdot t}{n_a})\log_2(1-\frac{n_\cdot t}{n_a}) - \frac{n_\cdot t}{n_a}\log_2(\frac{n_\cdot t}{n_a})$ if $t \in [0, n_a/(2\,n)[$; else $h_1(t) = 1$

$h_2(t) = -(1-\frac{n_\cdot t}{n_b})\log_2(1-\frac{n_\cdot t}{n_b}) - \frac{n_\cdot t}{n_b}\log_2(\frac{n_\cdot t}{n_b})$ if $t \in [0, n_b/(2\,n)[$; else $h_2(t) = 1$

$\mathcal{N}(0,1)$ stands for the centred and reduced normal repartition function

bounds depending on n_A, n_B and/or n_{AB}). This heterogeneity is confirmed when looking at the values (listed in table 4) taken by the measures for each of the 21 rules of table 2. Moreover, if we look at the rules ranked as the "best" for a given measure and as the worst by others we see important variations. Table 5 gives the different rankings (from 1 to at most 11) of the rules evaluated by our 20 measures. This can be highlighted by comparing complete preorders induced by the measures on rule sets. In order to carry out such experimental campaigns, we have developed HERBS, a tool designed for case-based studies of measures [36]. In [35] we produced an experimental classification of the measures using 10 datasets retrived from the UCI repository [4]. This experimental classification is quite similar to a formal one, also presented in [35].

To attest of the variety of the different rankings we sorted the rules according to the measures' values in table 5: the "best ranked" (ranks close to 1) correspond to rules for which the considered measure takes its largest values, the "worth ranked" (ranks closest to 11) correspond to rules for which the considered measure takes its lowest values, and "medium ranked" (ranks in between). Of course this kind of representation is quite arbitrary. For instance the measure SUP admits only two ranks: every rule becomes either a good or a bad one. However, it is not surprising as SUP has for a long time been pointed out as

Table 4. Evaluation of the rules by the quality measures

id	Lift	CenConf	Conf	Sup	EII	IntImp	Kappa	Conv	IG	-ImpInd	PDI	Lap	Loe	BF	PS	R	Seb	LC	ECR	Zhang
1	2.00	0.50	1.00	0.30	0.94	0.89	0.60	$+\infty$	0.30	1.22	3.87	0.80	1.00	$+\infty$	1.50	0.65	$+\infty$	0.60	1.00	1.00
2	2.00	0.30	0.60	0.30	0.52	0.79	0.60	1.75	0.30	0.80	2.54	0.57	0.43	3.50	1.50	0.65	1.50	0.33	0.33	0.71
3	1.25	0.20	1.00	0.30	0.88	0.78	0.19	$+\infty$	0.10	0.77	2.45	0.80	1.00	$+\infty$	0.60	0.33	$+\infty$	0.38	1.00	1.00
4	1.00	0.00	0.60	0.30	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.57	0.00	1.00	0.00	0.00	1.50	0.17	0.33	0.00
5	0.83	−0.13	0.67	0.40	0.00	0.23	−0.36	0.60	−0.08	−0.73	−2.31	0.62	−0.67	0.50	−0.80	−0.41	2.00	0.25	0.50	−0.50
6	1.33	0.20	0.80	0.40	0.69	0.76	0.40	2.00	0.12	0.71	2.24	0.71	0.50	2.67	1.00	0.41	4.00	0.50	0.75	0.62
7	1.33	0.17	0.67	0.40	0.47	0.72	0.40	1.50	0.12	0.58	1.83	0.62	0.33	2.00	1.00	0.41	2.00	0.40	0.50	0.50
8	1.00	0.00	0.80	0.40	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.71	0.00	1.00	0.00	0.00	4.00	0.38	0.75	0.00
9	1.00	0.00	0.60	0.30	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.57	0.00	1.00	0.00	0.00	1.50	0.17	0.33	0.00
10	1.00	0.00	0.60	0.30	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.57	0.00	1.00	0.00	0.00	1.50	0.17	0.33	0.00
11	1.00	0.00	0.80	0.40	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.71	0.00	1.00	0.00	0.00	4.00	0.38	0.75	0.00
12	1.11	0.07	0.67	0.40	0.00	0.60	0.17	1.20	0.05	0.26	0.82	0.62	0.17	1.33	0.40	0.17	2.00	0.33	0.50	0.25
13	1.11	0.07	0.67	0.40	0.00	0.60	0.17	1.20	0.05	0.26	0.82	0.62	0.17	1.33	0.40	0.17	2.00	0.33	0.50	0.25
14	0.83	−0.13	0.67	0.40	0.00	0.23	−0.36	0.60	−0.08	−0.73	−2.31	0.62	−0.67	0.50	−0.80	−0.41	2.00	0.25	0.50	−0.50
15	0.83	−0.13	0.67	0.40	0.00	0.23	−0.36	0.60	−0.08	−0.73	−2.31	0.62	−0.67	0.50	−0.80	−0.41	2.00	0.25	0.50	−0.50
16	1.25	0.20	1.00	0.30	0.88	0.78	0.19	$+\infty$	0.10	0.77	2.45	0.80	1.00	$+\infty$	0.60	0.33	$+\infty$	0.38	1.00	1.00
17	2.00	0.50	1.00	0.30	0.94	0.89	0.60	$+\infty$	0.30	1.22	3.87	0.80	1.00	$+\infty$	1.50	0.65	$+\infty$	0.60	1.00	1.00
18	2.50	0.45	0.75	0.30	0.77	0.86	0.78	2.80	0.40	1.08	3.40	0.67	0.64	7.00	1.80	0.80	3.00	0.67	0.67	0.86
19	0.94	−0.05	0.75	0.30	0.00	0.41	−0.07	0.80	−0.03	−0.22	−0.71	0.67	−0.25	0.75	−0.20	−0.10	3.00	0.25	0.67	−0.25
20	1.25	0.15	0.75	0.30	0.63	0.68	0.23	1.60	0.10	0.47	1.50	0.67	0.38	2.00	0.60	0.25	3.00	0.33	0.67	0.50
21	1.50	0.25	0.75	0.30	0.69	0.76	0.40	2.00	0.18	0.71	2.24	0.67	0.50	3.00	1.00	0.41	3.00	0.40	0.67	0.67

poor measure in an knowledge discovery context. What is more, as the rules are extracted from a toy example dataset, it was to be expected that such extreme situations should happen, which hopefully is usually not the case when using larger datasets, as in [35].

1.5 Aggregation as a Tool to Face the Heterogeneity of the Measures

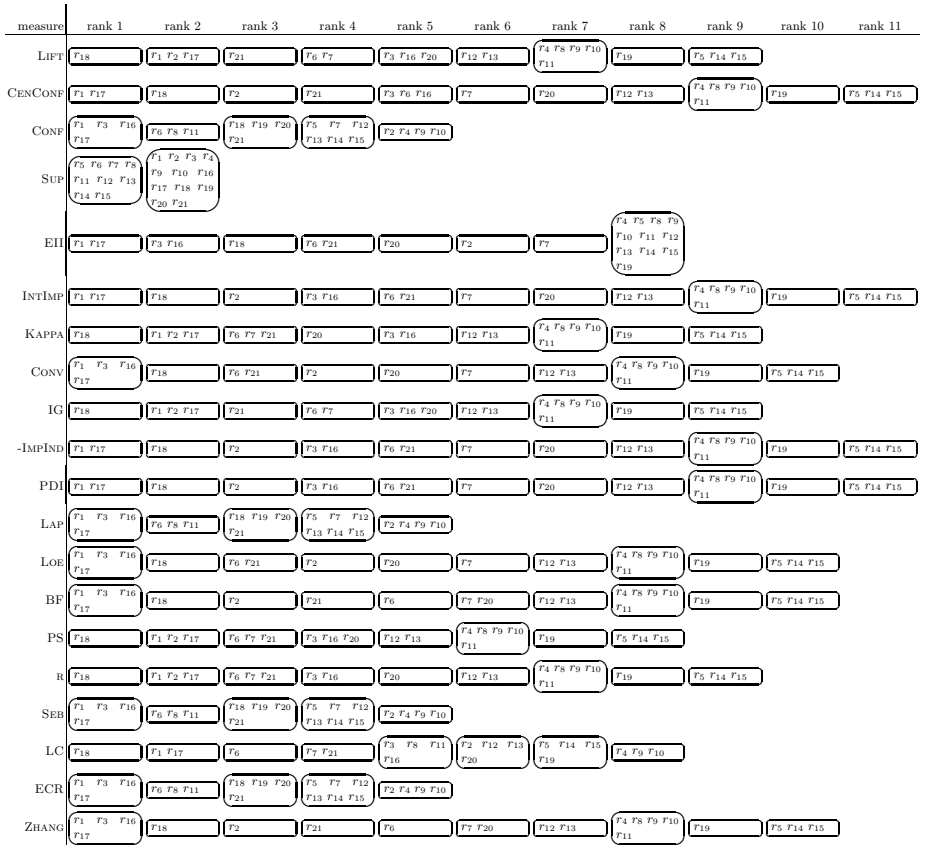
From the discussion of section 1.4 the following question arise: which rule(s) should be considered as the best one(s), according to our interestingness measures? In that framework different possibilities can be looked upon:

(i) dictatorship: use only one (preferred) measure, and forget the others;
(ii) try to find a consensus within a set of measures, judged as worth considering. In this paper we follow this second track. To achieve a consensus, two ways appear as "natural":
   a) direct aggregation of the measures into a single one, using some kind of generalised mean.
   b) aggregation of the rankings involved by the various measures into a single one.

A worry involved by the approach a) is the diversity of the induced scales. How to aggregate measures whose co-domains are $[0, 1]$, $[0, +\infty[$, or $]-\infty, 1]$? Approach b) has just the inverse drawback: rankings do not account for differences in evaluations. Moreover ordinal aggregation involves perilous "logical" problems (see, for example Arrow's 1951 theorem [2]). Thus we have chosen to follow a third, less straightforward, track: the aggregation of valued relations.

Rankings are complete preorders, i.e. special binary relations. We shall generalise them into valued relations. We recall that a valued relation (sometimes

Table 5. Rankings induced by measures on the rules



called a "fuzzy relation") on a set S is just a map R from $S \times S$ to the unit interval [0, 1]. Notice that an ordinary binary relation (sometimes "called crisp relation") is a particular instance of valued relation, whose values are 0 or 1. Valued relations allow to escape from scale effects, with a preservation of scale differences (because they are supposed to account for such differences). Aggregation of valued relation has been intensively studied in [13].

2 Valued Relations

2.1 Properties of Valued Relations

The various properties of usual binary relations extend mutatis mutandis to valued relations. For instance a valued relation R is said to be:

complete whenever $R(s,t) + R(t,s) \geq 1$, for $t \neq s$,
reflexive whenever $R(s,s) = 1$,
irreflexive whenever $R(s,s) = 0$,
symmetric whenever $R(s,t) = R(t,s)$,

reciproqual whenever $R(s,t) + R(t,s) = 1$,
min-antisymmetric whenever $\min\{R(s,t), R(t,s)\} = 0$, for $s \neq t$.

The notion of transitivity admits a number of extensions. They can be sorted in two categories:

(i) stochastic transitivities (see [12] for a review). In this case, $R(s,t)$ is an observed probability for s to be preferred to t, and R is reciprocal. Stochastic transitivities are often related to stochastic utility models.
(ii) fuzzy transitivities (see [13] for a review). In this case, $R(s,t)$ measures an intensity of preference (of s over t). The general expressions of fuzzy transitivities use the notions of "T-norm" and "T-co-norm", that generalize operators like min and max (and many others).

We shall use only three kinds of transitivities: the weak transitivity, the min-transitivity and the max-transitivity. The reason of these choices is that they have some properties of the relations are preservations under aggregation operators.

A valued preference R is said to be:

min-transitive whenever $R(s,u) \geq \min\{R(s,t), R(t,u)\}$
max-Δ-transitive whenever $R(s,u) \geq \max\{0, R(s,t) + R(t,u) - 1\}$
weakly transitive whenever $\{R(s,t) \geq \frac{1}{2}$ and $R(t,u) \geq \frac{1}{2}\} \Rightarrow R(s,u) \geq \frac{1}{2}$

It is easy to check that the min-transitivity implies both the max-Δ-transi- tivity and the weak-transitivity, and that there is no implication between the max-Δ-transitivity and the weak-transitivity. [32] assessed thoroughly the preservation of transitivities in an aggregation operators context.

2.2 Construction of a Valued Relation on a Set of Measures

From now on, we are concerned by valued relations on a set $\mathcal{R}$ of rules. In that purpose we denote by $r_1, \ldots, r_k$ the rules under consideration and by $\mu_1, \ldots, \mu_m$ the selected measures. Each measure μ_j will induce a valued relation R_j on $\mathcal{R}$. The general idea is that $R_j(r_i, r_{i'})$ correspond to a normalized difference between the rules r_i and $r_{i'}$, according to the measure μ_j.

The simplest idea is to consider these differences as linear:

$$R_j(r_i, r_{i'}) = \begin{cases} 0 & \text{if } \mu_j(r_i) - \mu_j(r_{i'}) < 0 \\ \frac{\mu_j(r_i) - \mu_j(r_{i'})}{\sigma_j} & \text{if } 0 \leq \mu_j(r_i) - \mu_j(r_{i'}) \leq \sigma_j \\ 1 & \text{if } \mu_j(r_i) - \mu_j(r_{i'}) > \sigma_j \end{cases} \quad (1)$$

In formula 1, σ_j is a threshold beyond which the intensity of preference becomes maximum (i.e. equal to 1). A drawback is that the transitions between "no preference" and "weak preference", and between "weak preference" and "strong preference" are rather "abrupt". That is the reason why we have preferred a smooth variant to formula 1, proposed by [8]:

$$R_j(r_i, r_{i'}) = \begin{cases} 1 - exp\left(-\frac{(\mu_j(r_i) - \mu_j(r_{i'}))^2}{2\sigma_j^2}\right) & \text{if } \mu_j(r_i) - \mu_j(r_{i'}) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

In formula 2, σ_j is the inflection point of the representative curve. It thus represents a threshold value between "weak preferences" and "strong preferences". The relation R_j of formula 2 is irreflexive (but can be made reflexive with an easy change). What is more, the relations R_j are min-transitive.

The two valued relation 1 and 2 are strongly related to strict fuzzy orderings [6, 5], since one of our goal is to build a global ranking on the rules.

3 Aggregation

3.1 Generalities

An aggregation operator is a map C from $I^* = \cup_{m\geq 1}[0,1]^m$ to [0,1]. If $R^* = (R_1, \ldots, R_m)$ is a m-uple of valued relations, an aggregation operator C will produce a consensus relation of R^*, abusively noted as $C(R^*)$. Following the notations of 2.2, the consensus of R^* is defined by:

$$C(R^*)(r_i, r_{i'}) = C(R_1(r_i, r_{i'}), \ldots, R_m(r_i, r_{i'}))$$

Many aggregation operators, fulfilling different properties, have been designed in the literature (see [13]): generalised means, OWA operators, Choquet and Sugeno integrals, weighted maximum and minimum...

Amongst the large number of possibilities, we have chosen to focus on generalised means. A generalised mean M is the aggregation operator defined by:

$$M(u_1, \ldots, u_m) = f^{-1}(\sum_{1\leq j\leq m} w_j f(u_j))$$

where f is a continuous monotonic function, f^{-1} its reciprocal function, and the w_j are non negative weights.

We shall look at some particular cases, in term of consensus relations (with obvious notations).

Weighted arithmetic mean(WMean): $f(u) = u$
Weighted geometric mean (WGeom): $f(u) = \log(u)$
Root-power mean (RPM): $f(u) = u^\alpha, \alpha \in \mathbb{R}$
Weighted harmonic mean (WHarm): $f(u) = 1/u$

3.2 Behavior Relatively to Transitivity and Other Properties

Table 6 (see [30]) summarizes the properties preserved by these four operators. We have retained only properties that apply to preference modelling (and, hence, eliminated various forms of dissimilarities). We focused mainly on transitivity as it is the one closest to some user's expectations (it guaranties that if a rule is better evaluated than a set of others, the aggregation procedure will maintain it above this set).

Table 6. Properties preserved by some aggregation operators

Operator	$R^* \Rightarrow R$
WMean	$t_2 + c \Rightarrow t_3 + c$
WGeom	$r + t_1 \Rightarrow r + t_2$
	$r + a + t_1 \Rightarrow r + a + t_2$
WHarm	$r + t_1 \Rightarrow r + t_2$
RPM	$c \Rightarrow c'$

r: reflexivity, c: reciprocity,
c': lower reciprocity, a: min-asymetry,
t_1: min-transitivity, t_2: max-Δ-transitivity,
t_3: weak transitivity

3.3 Experimental Results

Using the four proposed generalised means, we computed the aggregated preference for each pair of rules. The valued relation retained to express the individual preference of a rule over an other according to a measure requires that a σ_j parameter should be fixed (see formula 2). The strategy we choosed was to select the value taken at a quantile of the absolute difference between the evaluation of any two rules by a measure. For example, the value taken at a quantile of 0% of the differences corresponds to the smallest absolute difference (which obviously is 0, since the difference of evaluation between a rule and itself is null). The value taken at a quantile of 100% is the highest absolute difference of evaluation between any two rules, and the value at a quantile of 50% leads to the median value of all absolute differences.

In our example, this quantile is taken at 60% of all absolute differences, which is a high value. We will explain latter on why such a high value was needed. The weights choosen are all equal to 1/20. Other weights have been proposed in [21], through the analysis of what could be expected from some expert user's preferences [24].

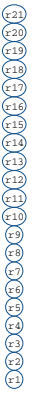


Fig. 1. WHarm

Once these parameters are fixed, the aggregation procedure produces a square matrix of values between 0 and 1, where each value is the aggregated preference of a rule over another. In order to produce an understandable (binary) output, a common way of transforming this matrix is to fix a threshold value, λ and compare the aggregated index to this threshold. Values below λ are considered as being equal to 0, and values above as equal to 1. Figures 1 to 4 are visual representations of this binary matrix, an arrow being drown between the rules (arrow pointing towards the preferred rule) when the preference of a rule over an other is 1. Of course, no rule is preferred to itself. The graph is automatically generated using Dot (`http://www.graphviz.org/`), and is sometimes hard to produce, especially when the number of edges increases.

Clearly, rules which are always "good" according to all measures (such as r_{18}) remain good ones, and rules which were poorly rated (such as r_{19}) end up on

the right side of the graph. More controversial rules (such as r_2) are less easy to follow. This could be attributed to the limited size of our toy example.

Figures 1 and 2 highlight a veto side-effect of some aggregation operators. Indeed, when a measure has the same value for two rules, the use of the log and inverse function produces "infinite" values, thus all other differences in evaluation by the measures are meaningless. As we already pointed out, the rules are grouped in only two ranks by SUP. Hence, a large pair of rules will always remain incomparable, in terms of aggregated preference.

The fact that SUP splits the rules in only two groups also accounts for the high value we selected in order to fix σ_j: if we had choosen a value lower than 60%, then the value for the quantile would have been null. In such a situation, formulae (1) and (2) are equivalent. A way of avoiding such a situation might be to add a slight level of noise in the values taken by the measures. This strategy -known as "jittering"- is commonly used in visualisation tools in order to plot a large number of possibly overlapping points. Note that the visualisation of the results is not a requirement, but a practical way of apprehending results on a toy example. Experiments were carried out on more large datasets in [21], still the automated plotting of preferences between rules needs more sophisticated methods. This possibility has been explored for example by [22] in visual data mining for association rules extraction.

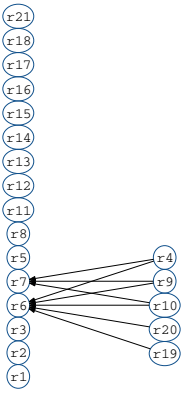


Fig. 2. WGeom

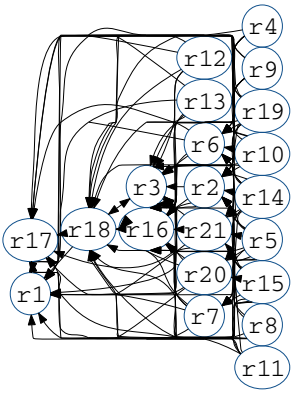


Fig. 3. RPM ($\alpha = 2$)

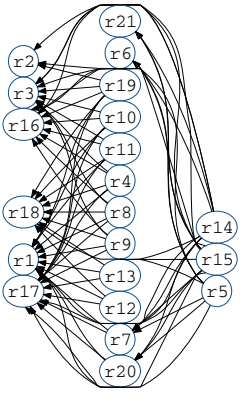


Fig. 4. WMean

4 Conclusion and Perspectives

In this study we were interested in the construction of aggregation operators on valued relations for association rule interestingness measures. The underlying objects and user's expectations constrained the available choices. Such constraints could be expressed mathematically, as properties that should be verified. Amongst the classical tools, few of them respect such constraints, still the generalised means comply with our requirements. A toy example study supported our approach. Experiments were also carried out on large datasets and look promising, still we pointed out some possible visualisation issues. Even if there are some conflictory points of view between the processing of very large rule sets and readable graphical exploration of them, we think that a promising way of solving such issues is to allow user-driven zooms on meaningful regions. Scalability on large datasets yet has to be explored.

Also, an alternative track to the quantile approach we used to fix σ_j thresholds would be to learn it automatically from tests in the presence of an expert. Such an approach should also be considered in order to learn the weights w_j of the aggregation operator.

References

1. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Int. Conf.*, pages 207–216, 1993.
2. K. J. Arrow. *Social Choice and Individual Values*. Cowles Foundations and Wiley, 1951.
3. J. Azé and Y. Kodratoff. Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'assocation. *EGC 2002*, 1(4):143–154, 2002.
4. C. Blake and C. Merz. UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.
5. U. Bodenhofer and M. Demirci. Strict fuzzy orderings in a similarity-based setting. In *Proc. Joint 4th Conf. EUSFLAT and* 11^e *LFA*, pages 297–302, 2005.
6. U. Bodenhofer and J. Küng. Fuzzy orderings in flexible query answering systems. *Soft Computing*, 8(7):512–522, 2004.
7. C. Borgelt and R. Kruse. Induction of association rules: Apriori implementation. In *15th Conf. on Computational Statistics*, 2002.
8. J. Brans, B. Mareschal, and P. Vincke. *PROMETHEE: A New Family of Outranking Methods in MCDM*. IFORS 84, 1984.
9. S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In J. Peckham, editor, *ACM SIGMOD 1997 Int. Conf. on Management of Data*, pages 255–264, 1997.
10. K. W. Church and P. Hanks. Word association norms, mutual information an lexicography. *Computational Linguistics*, 16(1):22–29, 1990.
11. J. Cohen. A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, 20:37–46, 1960.
12. P. Fishburn. *Handbook of Utility Theory*, volume 1, chapter Stochastic utility, pages 273–319. Kluwer Academic, Dordrecht, 1998.
13. J. Fodor and M. Roubens. *Fuzzy preference modelling and multicriteria decision support*. Kluwer Academic Publishers, 1994.

14. I. J. Good. The estimation of probabilities: An essay on modern bayesian methods. The MIT Press, Cambridge, MA, 1965.
15. R. Gras, S. Ag. Almouloud, M. Bailleuil, A. Larher, M. Polo, H. Ratsimba-Rajohn, and A. Totohasina. *L'implication Statistique, Nouvelle Méthode Exploratoire de Données. Application la Didactique, Travaux et Thèses*. La Pensée Sauvage, 1996.
16. R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *EGC 2001*, 1(1-2):69–80, 2001.
17. J. L. Guigues and V. Duquenne. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Math. Sci. hum.*, 25(95):5–18, 1986.
18. International Business Machines. *IBM Intelligent Miner User's Guide, Version 1 Release 1, SH12-6213-00*. 1996.
19. H. J. Jeffreys. Some tests of significance treated by the theory of probability. In *Proc. of the Cambridge Philosophical Society*, number 31, pages 203–222, 1935.
20. S. Lallich and O. Teytaud. Évaluation et validation de l'intérêt des règles d'association. *RNTI-E-1*, pages 193–217, 2004.
21. A. Legrain. Agrégation de mesures de qualité de règles d'association, rapport de DEA MIASH, 2004.
22. R. Lehn, F. Guillet, P. Kuntz, H. Briand, and J. Philippé. Felix: An interactive rule mining interface in a KDD process. In *HCP'99*, pages 169–174, 1999.
23. P. Lenca, P. Meyer, P. Picouet, B. Vaillant, and S. Lallich. Critères d'évaluation des mesures de qualité en ECD. *RNTI*, (1):123–134, 2003.
24. P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. A multicriteria decision aid for interestingness measure selection. Technical report, GET / ENST Bretagne, 2004.
25. P. Lenca, P. Meyer, B. Vaillant, and P. Picouet. Aide multicritère à la décision pour évaluer les indices de qualité des connaissances – modélisation des préférences de l'utilisateur. *EGC 2003*, 1(17):271–282, 2003.
26. P. Lenca, P. Meyer, B. Vaillant, P. Picouet, and S. Lallich. Évaluation et analyse multicritère des mesures de qualité des règles d'association. *RNTI-E-1*, pages 219–246, 2004.
27. I. Lerman and J. Azé. Une mesure probabiliste contextuelle discriminante de qualité des règles d'association. *EGC 2003*, 1(17):247–262, 2003.
28. J. Loevinger. A systemic approach to the construction and evaluation of tests of ability. *Psychological monographs*, 61(4), 1947.
29. K. Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia. *Philosophical Transactions of the Royal Society*, A, 1896.
30. V. Peneva and I. Popchev. Properties of the aggregation operators related with fuzzy relations. *Fuzzy Sets and Systems*, (139):615–633, 2003.
31. G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.
32. S. Saminger, R. Mesiar, and U. Bodenhofer. Domination of aggregation operators and preservation of transitivity. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 10(Suppl.):11–35, 2002.
33. M. Sebag and M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In *EKAW'88*, pages 28–1 – 28–20. 1988.
34. P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Eighth ACM SIGKDD Int. Conf.*, pages 32–41, 2002.
35. B. Vaillant, P. Lenca, and S. Lallich. A clustering of interestingness measures. In *Discovery Science*, volume 3245 of *LNAI*, pages 290–297. Springer-Verlag, 2004.
36. B. Vaillant, P. Picouet, and P. Lenca. An extensible platform for rule quality measure benchmarking. In *HCP'03*, pages 187–191, 2003.
37. T. Zhang. Association rules. In *PADKK 2000*, volume 1805 of *LNCS*, 2000.

On the Relationship Between the Quantifier Threshold and OWA Operators

Luigi Troiano[1] and Ronald R. Yager[2]

[1] RCOST — University of Sannio,
Viale Traiano, 82100 Benevento, Italy
troiano@unisannio.it
[2] Machine Intelligence Institute — IONA College,
New Rochelle, NY 10801, USA
yager@panix.com

Abstract. The OWA weighting vector and the fuzzy quantifiers are strictly related. An intuitive way for shaping a monotonic quantifier, is by means of the threshold that makes a separation between the regions of what is satisfactory and what is not. Therefore, the characteristics of a threshold can be directly related to the OWA weighting vector and to its metrics: the attitudinal character and the entropy (or dispersion). Generally these two metrics are supposed to be independent, although some limitations in their value come when they are considered jointly. In this paper we argue that these two metrics are strongly related by the definition of quantifier threshold, and we show how they can be used jointly to verify and validate a quantifier and its threshold.

1 Introduction

An Ordered Weighted Averaging (OWA) operator of dimension n is a mapping defined as [1]

$$M_{[w]}(a_1, \ldots, a_n) = \sum_{i=1}^{n} w_i \cdot a_{(i)} \tag{1}$$

where $(\cdot)$ means a decreasing ordered permutation of arguments $a_1, \ldots, a_n$, so that $a_{(i)} \geq a_{(j)} \forall i < j$. Weights are such that

$$\sum_{i=1}^{n} w_i = 1 \tag{2}$$

The OWA weighting vector is characterized by two metrics: the [illegible] and the [illegible] ([illegible]). The attitudinal character (a.k.a. [illegible]) is defined as [1]

$$\sigma = \frac{1}{n-1} \sum_{i=1}^{n} (n-i) w_i \tag{3}$$

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 215–226, 2006.

Noteworthy cases are:

- $w_1 = 1, w_{i>1} = 0 \Leftrightarrow M_{[w]}(x_1, \ldots, x_n) = \max\{x_1 \ldots, x_n\} \Leftrightarrow \sigma = 1$
- $w_i = \frac{1}{n} \Leftrightarrow M_{[w]}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} a_i \Rightarrow \sigma = \frac{1}{2}$
- $w_{i<n} = 0, w_n = 1 \Leftrightarrow M_{[w]}(x_1, \ldots, x_n) = \min\{x_1, \ldots, x_n\} \Leftrightarrow \sigma = 0$

The entropy [2] of OWA operators is defined as

$$H_y = -\sum_{i=1}^{n} w_i \ln w_i \tag{4}$$

Special cases are:

- $\exists i : w_i = 1 \Leftrightarrow H_y(w) = 0$ (minimum)
- $w_i = \frac{1}{n} \forall i \Leftrightarrow H_y(w) = \ln n$ (maximum)

Generally, it is useful to adopt a normalized measure of entropy

$$\overline{H}_y = -\frac{1}{\ln n} \sum_{i=1}^{n} w_i \ln w_i \tag{5}$$

Entropy has been generally adopted as a measure of weight dispersion for the OWA operators. O'Hagan [3], in his ground-breaking work, suggests to select the vector that maximizes the entropy of OWA weights (ME-OWA). Analytical solutions to this problem have been proposed by Filev and Yager [4], and Fullér and Majlender [5]. However, entropy is a strong characterization of OWA weights, because it satisfies several additional properties that do not enrich the notion of dispersion. More recently, Troiano and Yager [6] proposed a dispersion metric

$$\rho = \frac{1}{n-1} \frac{1 - w_{[1]}}{w_{[1]}} \tag{6}$$

where $w_{[1]} = \max_{i=1..n} w_i$. The main advantage of this metric consists in solving analytically Eq.(6), given a level of dispersion. Moreover, the maximal weight can be computed as

$$w_{[1]} = \frac{1}{(n-1)\rho + 1} \tag{7}$$

Generally the attitudinal character and the entropy (or dispersion) are considered as two independent characterizations of the weighting vector, although their values are not fully independent within the unit interval. For instance if $\sigma = 0$ or 1 then $\overline{H}_y = 0$; viceversa if $\overline{H}_y = 1$ then $\sigma = \frac{1}{2}$.

In this paper we argue that the two metrics can be related to the definition of fuzzy quantifiers, in particular to the definition of quantifier thresholds. Indeed, an OWA operator implements the quantification semantics of **at least** t, where t is a given threshold. The specification of t determines the OWA weighting vector and its metrics. In particular, the position and fuzziness of t affect the attitudinal character and the dispersion respectively. The remainder of this paper is

organized as follows: Section 2 introduces basic concepts of fuzzy quantifiers and how they are related to the semantics of OWA operators; Section 3 provides a characterization of monotonic quantifier thresholds; Section 4 discusses the relationship between the threshold and the OWA weighting vector metrics; Sections 5 outlines conclusions.

2 Fuzzy Quantifiers in OWA Operators

The OWA operators satisfy the compensatory property so that

$$\min(a_1, \ldots, a_n) \leq M_{[w]}(a_1, \ldots, a_n) \leq \max(a_1, \ldots, a_n) \tag{8}$$

When $w_n = 1$ ($w_{i \neq n} = 0$) we get

$$M_{[w]}(a_1, \ldots, a_n) = \min(a_1, \ldots, a_n) \tag{9}$$

Viceversa, when $w_1 = 1$ ($w_{i \neq 1} = 0$) we get

$$M_{[w]}(a_1, \ldots, a_n) = \max(a_1, \ldots, a_n) \tag{10}$$

More in general, the OWA operators build a continuum of aggregation operators between the minimum and the maximum by varying the weighting vector.

The minimum and the maximum operators have a specific semantics. Let $a_1, \ldots, a_n$ be degrees, such as the satisfaction of criteria in decision making problems, or the truth of predicates in logics. If they are boolean, so that $a_i \in \{0, 1\}$, the result of the aggregation is boolean. In particular,

$$\min(a_1, \ldots, a_n) = 1 \Leftrightarrow a_i = 1 \forall i \in 1..n \tag{11}$$

$$\max(a_1, \ldots, a_n) = 1 \Leftrightarrow \exists i \in 1..n | a_i = 1 \tag{12}$$

Therefore, to be $\min(a_1, \ldots, a_n) = 1$, it is required that **all** arguments are 1; for instance all criteria have to be satisfied or all predicates have to be true. Differently, to be $\max(a_1, \ldots, a_n) = 1$, the maximum requires that **at least one** argument is 1, that is at least one criterion is satisfied, or at least 1 predicate is true. **All** and **at least one** are two special cases of quantification. They are also respectively known as the [illegible] and the [illegible]. Other choices are possible, for instance **none** or **half**. Quantification provides the means for describing a collection of objects by counting the elements. Quantifiers can be rigorously formalized.

Definition 1 (Two-valued quantifier). [illegible]

$$Q_b : 2^U \to \{0, 1\} \tag{13}$$

[illegible]

$$\forall A_1, A_2 \in 2^U | \, card(A_1) = card(A_2) \Rightarrow Q_b(A_1) = Q_b(A_2) \tag{14}$$

[illegible] $U \neq \emptyset$ [illegible] domain 2^U [illegible] U [illegible] $card(A)$ [illegible] $A \in 2^U$

Definition 2 (Fuzzy quantifier).

$$Q_f : 2^U \rightarrow [0,1] \tag{15}$$

$$\forall A_1, A_2 \in 2^U | \, card(A_1) = card(A_2) \Rightarrow Q_f(A_1) = Q_f(A_2) \tag{16}$$

Since quantifiers depend only on the subset cardinality, we can describe a quantifier by its

$$Q : [0,1] \rightarrow [0,1] \tag{17}$$

that assumes as input the normalized cardinality $x = \frac{i}{n}$, with $i = card(A)$ and $n = card(U)$.

A quantifier is if $Q(\cdot)$ is a non-decreasing (non-increasing) function. Similarly, a quantifier is if $Q(\cdot)$ is unimodal; it is if $\max_{x \in [0,1]} Q(x) = 1$. Quantifiers are inherently defined on finite sets, whose cardinality is n. Therefore the function Q is defined on a finite subset of rational numbers. More specifically it is defined on the set of equidistant rational numbers, associated to the normalized cardinalities $x = \frac{i}{n}, i = 0..n$. However, since their semantics keeps valid for any value n, it is convenient to consider the extension of $Q(\cdot)$ to the unit interval of rational numbers. Thus, for the sake of simplicity, we can consider the extension of $Q(\cdot)$ to the unit interval or real numbers. Quantifiers are predicates associated to the cardinality of sets, whose semantics is **at least** t for non-decreasing quantifiers, and **no more than** t for non-increasing quantifiers. For instance, the quantifier depicted in Fig.1-(a) means **at least 40%**, and the quantifier depicted in Fig.1-(b) means **no more than 60%**. In Fig.1, the threshold has been precisely defined as 0.4 in (a) and 0.6 in (b). Therefore the degree of truth provided by $Q(x)$ sharply moves from 0 to 1 when the threshold is passed, and viceversa. If the threshold is imprecisely defined, then the quantifier becomes smoother, as depicted in Fig.2. A non-decreasing quantifier can be regarded as a cumulative probability distribution.

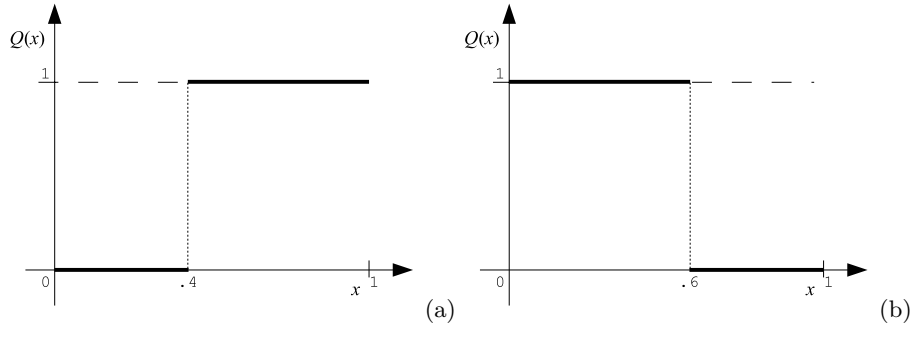


Fig. 1. Examples of non-decreasing (a) and non-increasing(b) quantifiers

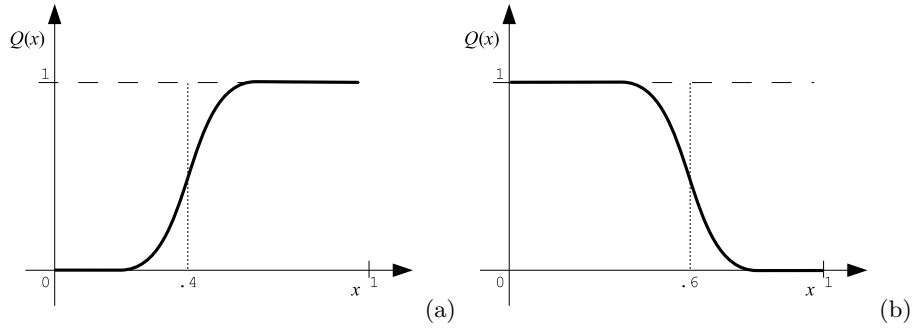


Fig. 2. Examples of non-decreasing (a) and non-increasing(b) quantifiers

Its derivative provides a probability density function

$$w(x) = \frac{d}{dx}Q(x) \tag{18}$$

For any finite n, the definition domain of the quantifier is finite, and the derivative defined by Eq.(18) becomes the 1-st finite backward difference

$$w_i = \nabla Q\left(\frac{i}{n}\right) = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right) \tag{19}$$

This equation is used by Yager [7, 8] to obtain the OWA weighting vectors from quantifiers. On the other hand, the OWA weighting vectors can be used to determine non-decreasing regular quantifiers.

In the continuous case, we have

$$\sigma = 1 - \int_0^1 \int_0^x w(y)dydx \tag{20}$$

and

$$H_y = 1 - \int_0^1 w(x)\ln w(x)dx \tag{21}$$

Dispersion can be related to $w(x)$, if we consider [6] that

$$\rho = \frac{\sum_{j=1}^{n} \frac{j-1}{n-1}\left(w_{[j]} - w_{[j+1]}\right)}{\sum_{j=1}^{n}\left(w_{[j]} - w_{[j+1]}\right)} \tag{22}$$

where $[\cdot]$ is a decreasing permutation of weights so that $w_{[j]} \geq w_{[l]}, \forall j < l$. The quantity $\frac{j-1}{n-1}$ is the number of weights but one whose values is at least $w_{[j]}$. In the continuous case, this quantity can be substituted by [1]

[1] It is possible to verify that the set $\{x \in [0,1] | w(x) \geq w\}$ is always a finite union of intervals.

$$v(w) = \text{length}\{x \in [0,1] | w(x) \geq w\} \tag{23}$$

and Eq.(22) becomes [2]

$$\rho = \frac{1}{w_{\max}} \int_0^1 v(w)dw = \frac{1}{w_{\max}} \int_0^1 w(x)dx = \frac{1}{w_{\max}} \tag{24}$$

where

$$w_{\max} = \max_{x \in [0,1]} w(x) \tag{25}$$

These measures can be directly linked to the quantifier characteristic function $Q(\cdot)$, by means of Eq.(18).

3 The Quantifier Threshold

In both the semantics **at least** t and **no more than** t, the threshold t is a key concept, and can be used to characterize or build the associated quantifier. From comparing Fig.1 and Fig.2, we notice that the smoothness of the quantifier is related to how precisely the threshold t is defined: less precise thresholds entail smoother quantifiers. The threshold can be described by a distribution of values $t : [0,1] \to [0,1]$, each x with a plausibility degree $t(x)$. Thus, in this context the fuzziness is associated to the plausibility of the threshold.

Describing a quantifier in terms of the threshold makes simpler to identify and choose properly the quantifier that better models the concept, then to determine the OWA weighting vector. Indeed, the decision maker can find more intuitive to choose a nominal value for the threshold, spread on an interval according to the value confidence, then to derive the related quantifier, instead of working directly with the quantifier. An example, of this process is presented in Fig.3. Similarly, given an OWA weighting vector, or a quantifier, it is possible to check the semantics against the related threshold.

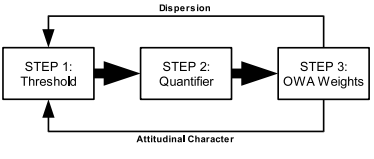


Fig. 3. Building the OWA weighting vector from the quantifier threshold

The first question to answer is about the relationship between the quantifier and the associated threshold. The first idea for describing the fuzzy threshold is to use the derivative of the quantifier, that it is a probability density function as

[2] This equation is valid only if $Q(\cdot)$ is continuous. $Q(\cdot)$ may have a finite number of discontinuities, entailing impulses in its derivative $w(\cdot)$. In this latter case, $w_{\max} = \infty$ and $\rho \to 0$.

noted in the previous section. In particular, if we assume that the more probable a threshold value is, the more plausible it is, we get

$$t_d(x) = \frac{1}{w_{\max}} \frac{d}{dx} Q(x) \tag{26}$$

However, although this solution looks appealing for its simplicity, it can be misleading. For example, let us consider the quantifier depicted in Fig.4-(a) whose characteristic function is

$$Q(x) = \begin{cases} 0.9x & 0 \leq x < 0.495 \\ 10.9(x - 0.495) + 0.4455 & 0.495 \leq x \leq 0.505 \\ 0.9x + 0.1 & 0.505 < x \leq 1 \end{cases} \tag{27}$$

The maximum value of derivation is $w_{\max} = 10.9$. Therefore, the threshold given by derivation is

$$t_d(x) = \begin{cases} 0.083 & 0 \leq x < 0.495 \\ 1 & 0.495 \leq x \leq 0.505 \\ 0.083 & 0.505 < x \leq 1 \end{cases} \tag{28}$$

Although the quantifier in Fig.4-(a) looks very similar to the linear quantifier in Fig.4-(b), the thresholds are very different, as depicted by Fig.5. Indeed, even for small deviation from linearity, the threshold could be dominated in truth by some few values, as depicted in Fig.5-(a). Therefore, although computing the threshold distribution by derivation is straightforward and would allow to link the weighting vector of an OWA operator only by a constant, the result is sometimes counter-intuitive. The intuitiveness of a threshold is important as the decision maker is asked to define the threshold assigning to each value x a plausibility degree $t(x)$; conversely it is also important as the threshold can be used to verify the consistency of a quantifier or OWA weighting vector.

Alternatively to derivation, it is possible to weigh the degree of plausibility in terms of the degree of truth assigned to intervals of values. The quantifier $Q(\cdot)$, defined by Eq.(27) and depicted by Fig.4-(a), is piecewise linear. In particular it has been obtained as a composition of two quasi-linear quantifiers $L_1(\cdot)$ and

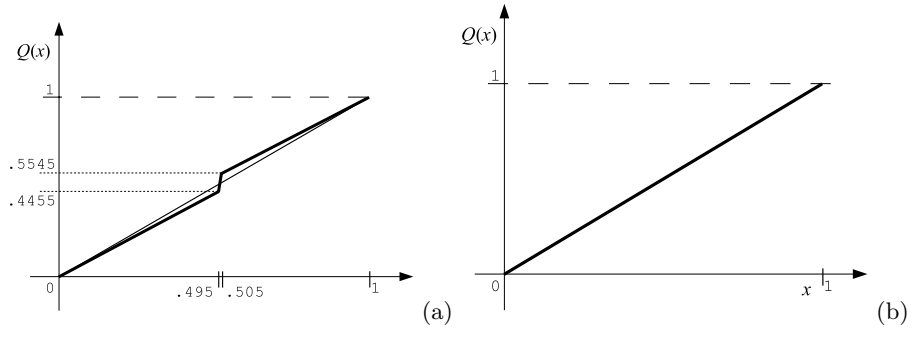


Fig. 4. Piecewise linear (a) and linear (b) quantifiers

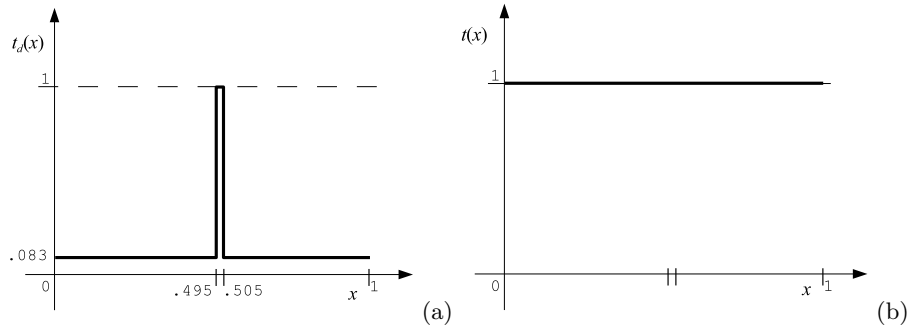


Fig. 5. The threshold of the piecewise linear (a) and linear (b) quantifiers

$L_2(\cdot)$. In particular,

$$Q(x) = 0.9L_1(x) + 0.1L_2(x) \tag{29}$$

A quasi-linear quantifier is characterized by a function as

$$L(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases} \tag{30}$$

as depicted in Fig.6.

When the [illegible] $[a, b] \equiv [0, 1]$, the quantifier is linear. A quasi-linear quantifier has a threshold uniformly distributed over the transition range

$$h(x) = \begin{cases} 1 & x \in [a, b] \\ 0 & otherwise \end{cases} \tag{31}$$

In the example above, we can express the threshold plausibility as

$$t(x) = 0.9h_1(x) + 0.1h_2(x) \tag{32}$$

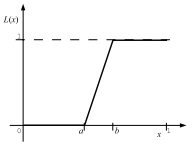


Fig. 6. A quasi-linear quantifier

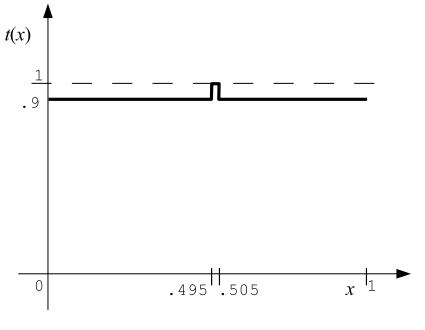


Fig. 7. A threshold plausibility function

where

$$h_1(x) = 1, \forall x \in [0,1] \tag{33}$$

$$h_2(x) = \begin{cases} 1 & x \in [0.495, 0.505] \\ 0 & otherwise \end{cases} \tag{34}$$

In this case, the threshold plausibility is almost everywhere equal to 0.9, with a small peak within $[0.495, 0.505]$, as depicted in Fig.7. This plausibility function appears more intuitive than the function drawn in Fig.5-(a).

We can apply this approach to a generic piecewise linear quantifier, as that one depicted in Fig.8. A piecewise linear quantifier can be always written as

$$Q(x) = \sum_{i=1}^{m} \eta_i L_i(x) \tag{35}$$

where L_i are quasi-linear quantifiers varying on $[a_i, b_i]$, and such that[3]

$$[a_i, b_i] \subseteq [a_j, b_j], \forall i > j \tag{36}$$

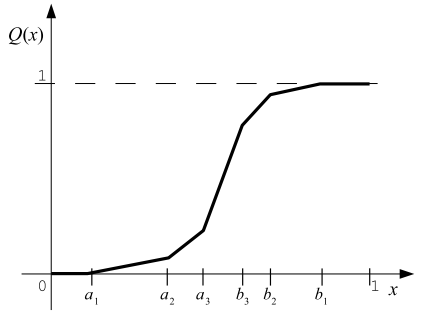


Fig. 8. A piecewise linear quantifier

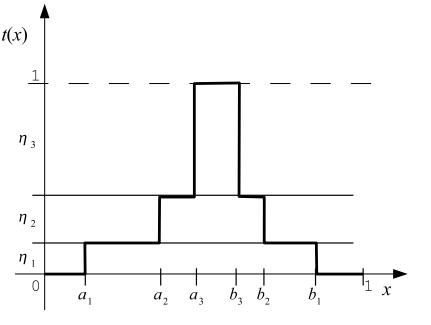


Fig. 9. A piecewise linear quantifier threshold

The coefficients η_i can be computed by solving a system of linear equations. Indeed,

$$\frac{d}{dx}Q(x) = \sum_{i=1}^{m} \eta_i \frac{d}{dx} L_i(x) \tag{37}$$

Within the interval $[a_i, a_{i+1}]$ $(i = 1..m-1)$, we have that

$$\frac{d}{dx}L_j(x) = \begin{cases} \frac{1}{b_j - a_j} & \forall j \leq i \\ 0 & \forall j > i \end{cases} \tag{38}$$

[3] This last condition is assured if the derivative $\frac{d}{dx}Q(x)$ is convex. Here convexity is meant in the sense of fuzzy sets. A fuzzy set is convex, if all α-cuts are convex. In the case of real numbers, convex α-cuts are intervals.

and

$$\frac{d}{dx}Q(x) = \frac{Q(a_{i+1}) - Q(a_i)}{a_{i+1} - a_i} \tag{39}$$

By varying $i = 1..m-1$, we get $m-1$ linear equations, each considering the unknown $\eta_1, \ldots, \eta_i$. The last equation can be computed considering the interval $[a_m, b_m]$. In this interval, we have

$$\frac{d}{dx}L_j(x) = \frac{1}{b_j - a_j} \ \forall j \tag{40}$$

and

$$\frac{d}{dx}Q(x) = \frac{Q(b_m) - Q(a_m)}{b_m - a_m} \tag{41}$$

The last equations considers all variables $\eta_1, \ldots, \eta_m$.

For instance, in the previous example, we get

$$\begin{aligned} \eta_1 &= 0.9 \\ \eta_1 + 100\eta_2 &= 10.9 \end{aligned} \tag{42}$$

whose solutions is

$$\begin{aligned} \eta_1 &= 0.9 \\ \eta_2 &= 0.1 \end{aligned} \tag{43}$$

as expected. After we determine coefficients $\eta_1, \ldots, \eta_m$, we can build the threshold plausibility function as

$$t(x) = \sum_{i=1}^{m} \eta_i h_i(x) \tag{44}$$

where h_i is associated to L_i, then it is a threshold uniformly distributed in $[a_i, b_i]$, according to Eq.33. Therefore, the resulting plausibility function $t(x)$ is a step-wise function as depicted in Fig.9.

The problem can be also inverted: given a step-wise threshold, we can build a piecewise quantifier. Indeed, $Q(x)$ can be still computed by Eq.(35), where

$$\eta_i = t(a_i) - t(a_{i-1}) \tag{45}$$

assuming $t(a_0) = 0$. Although, the idea of composing a quantifier as sum of quasi-linear quantifiers looks simple and intuitive, its application to the whole class of regular quantifiers is not easy, involving integrals complicate to solve analytically. Only in few simple cases, such as if the threshold is triangular, the integral can be still solved. More in general the integral regards ratio of non polynomial functions, and this can lead to equations hard to solve. However, continuous quantifiers can be approximated by piecewise linear quantifiers; thus, their threshold can be approximated by a stepwise threshold. The number m of intervals, can be chosen in order to minimize the approximation error, for both the quantifier and the threshold.

4 The Quantifier Threshold and the Weighting Vector Metrics

As we previously noticed, a fuzzy quantifier $Q(\cdot)$ is linked to the weighting vector w by Eq.(19), and in the previous section we dealt with the relationship between the threshold $t(\cdot)$ and the quantifier $Q(\cdot)$. Thus, it is convenient to establish a direct link between the quantifier threshold and the metrics characterizing an OWA weighting vector. The attitudinal character can be computed as

$$\sigma = 1 - \int_0^1 \int_0^x w(y)dydx = 1 - \int_0^1 Q(x)dx =$$
$$= 1 - \int_0^1 \sum_{i=1}^{m} \eta_i L_i(x)dx = \sum_{i=1}^{m} \eta_i \left(1 - \int_0^1 L_i(x)dx\right) = \sum_{i=1}^{m} \eta_i \sigma_i \quad (46)$$

where σ_i is the attitudinal character of the quasi-linear quantifier L_i. Therefore, the attitudinal character of a piecewise linear quantifier is the weighted mean the attitudinal character provided by the quasi-linear quantifiers that compose it. The attitudinal character of a quasi-linear quantifier is easy to compute. Indeed,

$$\sigma_i = 1 - \frac{a_i + b_i}{2} \quad (47)$$

Both information η_i and $[a_i, b_i]$, are related to the definition of the quantifier threshold. Indeed, from Eq.(45), we get η_i is the level increment of the α-cut $[a_i, b_i]$, as depicted in Fig.9.

According to Eq.(24), dispersion is

$$\rho = \min_{i=1..m} \rho_i = \min_{i=1..m} b_i - a_i = b_m - a_m \quad (48)$$

where ρ_i is the dispersion associated to the quasi-linear quantifier L_i, thus

$$\rho_i = b_i - a_i \quad (49)$$

If we consider that the piecewise quantifier $Q(\cdot)$ is composed by linear quantifier, it is useful to consider the average dispersion, defined as

$$\overline{\rho} = \sum_{i=1}^{n} \eta_i \rho_i = \sum_{i=1}^{n} \eta_i (b_i - a_i) \quad (50)$$

Therefore, the average dispersion is the area underlying $t(\cdot)$.

5 Conclusions

The OWA operators represent a way for implementing non-decreasing quantifiers, and quantifiers provide a semantic interpretation for the OWA operators.

Monotonic quantifiers can be fully characterized in terms of the the threshold that separate the region of truth from the region of falsity. The threshold can be imprecise, affecting the quantifier smoothness. Modeling a threshold is an intuitive way for building a quantifier, then an OWA weighting vector. The OWA weighting vector is characterized by metrics regarding the attitudinal character and the dispersion. Although, they have been generally considered as independent, they can jointly be used to verify and validate a quantifier threshold.

References

1. Yager, R.R.: On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Trans. on Systems, Man, and Cybernetics **18** (1988) 183–190
2. Yager, R.R.: Measures of entropy and fuzziness related to aggregation operators. Information Sciences **82** (1995) 147–166
3. O'Hagan, M.: Aggregating template rule antecedents in real-time expert systems with fuzzy set logic. In: 22th Annual IEEE Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA (1988)
4. Filev, D.P., Yager, R.R.: Analytic Properties of Maximum Entropy OWA Operators. Information Sciences (1995) 11–27
5. Fullér, R., Majlender, P.: An analytic approach for obtaining maximal entropy OWA operator weights. Fuzzy Sets and Systems **124** (2001) 53–57
6. Troiano, L., Yager, R.R.: A meaure of dispersion for OWA operators. In Y. Liu, G.C., Ying, M., eds.: Proceedings of the Eleventh International Fuzzy systems Association World Congress, Beijing, China, Tsinghua University Press and Springer (2005) 82–87 ISBN 7-302-11377-7.
7. Yager, R.R.: Quantifier guided aggregation using OWA operators. International Journal of Intelligent Systems **11** (1996) 49–73
8. Yager, R.R.: Families of OWA operators. Fuzzy Sets and Systems **57** (1993) 125–148 ISSN:0165-0114.

Integrated Fuzzy Approach for System Modeling and Risk Assessment

Vikas Kumar, Marta Schuhmacher, and Miriam García

Environmental Management and Engineering (AGA) Group,
ETSEQ, Universitat Rovira i Virgili
Avinguda dels Pasos Catalans, 26, 43007 Tarragona (Spain)
vikas.kumar@urv.net

Abstract. Subjective nature of system components makes the natural problems more complicated and harder to quantify. Thus, effective reflection of uncertainties, which is essential for generating reliable and realistic outcomes, has become a major concern for risk assessment. With the growing trend of fuzzy modeling and simulation of environmental problem, there is a need to develop a risk analysis approach which can use the fuzzy number output for characterization of risk. This study has been done to fulfil that need. Integration of system simulation and risk analysis using fuzzy approach allowed to incorporate system modelling uncertainty and subjective risk criteria. In this study, an integrated fuzzy relation analysis (IFRA) model is proposed for risk assessment involving multiple criteria. Model is demonstrated for a multi-components groundwater contamination problem. Results reflect uncertainties presented as fuzzy number for different modelling inputs obtained from fuzzy system simulation. Integrated risk can be calculated at different membership level which is useful for comprehensively evaluating risks within an uncertain system containing many factors with complicated relationship. It has been shown that a broad integration of fuzzy system simulation and fuzzy risk analysis is possible.

Keywords: Fuzzy modeling, risk analysis, uncertainty analysis.

1 Introduction

A common approach for estimating and assessing environmental risks is to reduce the complicated systems into mathematical models. Unfortunately, real-world problems are often uncertain with associated risks of pollution impacts. Commonly environmental models are calibrated to field data to demonstrate their ability to reproduce contaminant behaviour at site. However, system modeling presents a big uncertainty due to the lack of reliable field data. On the other hand, specific field situations can not be extrapolated over larger distances, even in the same site [12]. Also many system components may not be known with certainty but can be stated in linguistic terms. This makes the study problems more complicated and harder to quantify. Thus, effective reflection of uncertainties, which is essential for generating reliable and realistic outcomes, has become a major concern for risk assessment.

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 227 – 238, 2006.

Also, in the process of environmental risk assessment, a number of chemical, physical, biological and socio-economic factors may have to be considered due to their direct or indirect relations to risk impacts. For example, factors of water treatment technologies, site remediation practices, and related capital/operating costs may be closely related to the levels of risk impacts. Thus, an integrated approach for risk assessment that incorporates multiple system components within a general framework rather than examines them in isolation would be valuable for internalizing intricate interrelationships among a variety of system components.

Several risk assessment approaches in complicated environmental systems have been developed. Among them probabilistic approach is quite common and have been commonly used in the treatment and processing of uncertainty, environmental modeling and risk assessment. This technique requires knowledge of parameter values and their statistical distribution. However generally site investigation are not detailed to determine values for some of the parameters and their distribution pattern and sufficient data may not be collected for calibrating a model. These approaches suffer from obvious lack of precision and specific site-characterization which make it difficult to determine how much error is introduced into the result due to assumptions and prediction. Also Monte Carlo method for uncertainty propagation typically requires several model runs that use various combinations of input values, resulting in substantial computational demands.

Another major approach for risk assessment is through fuzzy set theory, which is better representative of uncertainty and imprecision and also suitable for situations when probabilistic information is not available. Fuzzy set is a mathematical theory for the representation of uncertainty [13][15]. Given a degree of uncertainty in the parameters, fuzzy set theory makes possible to evaluate the uncertainty in the results without any knowledge of probability distribution functions, thereby avoiding the difficulties associated with stochastic analysis. Fuzzy set approach has been applied recently in various fields, including decision making, control and modeling [1].

With the growing trend of fuzzy modeling and simulation of environmental problem, there is a need to develop a risk analysis approach which can use the fuzzy number output and integration of multiple risk criteria for characterization of risk. This study has been done to fulfil that need. Integration of system simulation and risk analysis using fuzzy approach allowed to incorporate system modeling uncertainty and subjective risk criteria. In this study, an integrated fuzzy relation analysis (IFRA) model is proposed for the environmental risk assessment involving multiple criteria. Fuzzy transformation method has been used for the simulation of fate and transport model of contaminants in the subsurface. The fuzzy outputs from system simulation is used for integrated risk assessment using IFRA approach. At the end, some conclusions are drawn and recommendations are made for the future research.

2 Fuzzy Alpha-Cut and Transformation Method

An alpha cut is the degree of sensitivity of the system to the behavior under observation. At some point, as the information value diminishes, one no longer want to be "bothered" by the data. In many systems, due to the inherent limitations of the mechanisms of observation, the information becomes suspect below a certain level of reliability.

Fuzzy alpha-cut (FAC) technique is based on the extension principle, which implies that functional relationships can be extended to involve fuzzy arguments and can be used to map the dependent variable as a fuzzy set. In simple arithmetic operations, this principle can be used analytically. However, in most practical modeling applications, relationships involve complex structures (e.g. partial differential equations) that make analytical application of the principle difficult. Therefore, interval arithmetic can be used to carry out the analysis [1].

Membership functions define the degree of participation of an observable element in the set, not the desirability or the value of the information. The membership function is cut horizontally at a finite number of α-levels between 0 and 1. For each α-level of the parameter, the model is run to determine the minimum and maximum possible values of the output. This information is then directly used to construct the corresponding membership function of the output which is used as a measure of uncertainty. If the output is monotonic with respect to the dependent fuzzy variable/s, the process is rather simple since only two simulations will be enough for each α-level (one for each boundary). Otherwise, optimization routines have to be carried out to determine the minimum and maximum values of the output for each α-level. The α-sublevel technique consists of subdividing the membership range of a fuzzy number into α-sublevels at membership levels $\mu_j = j/m$, for $j = 0, 1, ...m$ [8]. This allows to numerically represent the fuzzy number by a set of $m + 1$ intervals $[a(j), b(j)]$. Figure 1 shows a triangular fuzzy number, subdivided into intervals using $m = 5$.

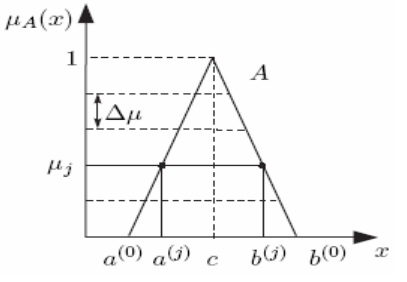


Fig. 1. A triangular fuzzy number with 5 α-level

2.1 Transformation Method (TM)

The TM presented by [9] uses a fuzzy alpha-cut approach based on interval arithmetic. The uncertain response reconstructed from a set of deterministic responses, combining the extrema of each interval in every possible way unlike the FAC technique where only a particular level of membership (α-level) values for uncertain parameters are used for simulation. The reduced TM used in the present study will be next explained.

Given an arithmetic function f that depends on n uncertain parameters $x_1, x_2, ..., x_n$, represented as fuzzy numbers, the function output $q = f(x_1, x_2, ..., x_n)$ is also a fuzzy number. Using the α-level technique, each input parameter is decomposed into a set P_i of $m + 1$ intervals $X_i^{(j)}$, $j = 0, 1, ...,m$ where

$$P_i = \{X_i^{(0)}, X_i^{(1)}, ..., X_i^{(m)}\} \tag{2.1}$$

$$\text{with } X_i^{(j)} = \left[a_i^{(j)}, b_i^{(j)}\right], \quad a_i^{(j)} \le b_i^{(j)}, \quad i=1,2,...,n, \qquad j=1,2,...,m. \tag{2.2}$$

where $a_i^{(j)}$ and $b_i^{(j)}$ denote the lower and upper bound of the interval at the membership level μ_j for the ith uncertain parameter. Instead of applying interval arithmetic like FAC method, intervals are now transformed into arrays $\hat{X}_i^{(j)}$ of the following form:

$$\hat{X}_i^{(j)} = \underbrace{\left(\alpha_i^{(j)}, \beta_i^{(j)}, \alpha_i^{(j)}, \beta_i^{(j)}, ..., \alpha_i^{(j)}, \beta_i^{(j)}\right)}_{2^{i-1}\, pairs} \tag{2.3}$$

with

$$\alpha_i^{(j)} = \underbrace{\left(a_i^{(j)}, ...a_i^{(j)}\right)}_{2^{n-1}\, elements}, \qquad \beta_i^{(j)} = \underbrace{\left(b_i^{(j)}, ...b_i^{(j)}\right)}_{2^{n-1}\, elements} \tag{2.4}$$

The evaluation of function f is now carried out by evaluating the expression separately at each of the positions of the arrays using the conventional arithmetic. Result obtained is deterministic in decomposed and transformed form which can be retransformed to get fuzzy valued result using recursive approximation.

3 Environmental Risk Assessment through Integrated Fuzzy Relation Analysis Method

For the purpose of quantifying uncertainty more effectively and integrating the risk assessment process with system modelling in a fuzzy environment, Integrated Fuzzy Relation Analysis (IFRA) has been proposed. The concept of fuzzy relation was first applied to medical diagnosis by [13]. In a very general setting, the process of fuzzy relation analysis can be conveniently described by pointing out relationships between a collection of pattern features and their class membership vectors. This analysis is useful for multifactorial evaluation and risk assessment under imprecision and uncertainty [11], [14]. The axiomatic framework of fuzzy set operation provides a natural setting for constructing multiattribute value functions in order to sort a set of potential actions and make an effective assessment. IFRA method is a generalization and refinement of the interval based methods such as Interval Parameter Fuzzy Relation Analysis (IPFRA) proposed by [10]. In IFRA the bounds vary according to the level of confidence one has in the estimation. One can think of a fuzzy number as a nested stack of intervals, each at a different level of presumption or possibility which ranges from zero to one and risk assessment can be performed at each level of possibility.

IFRA method for risk analysis will be explained in the context of multi-contaminants problem in the groundwater. Assuming that chronic daily intake and average human life expectancy are constant, the relationship between the risk and the pollutant concentration can be expressed as follows:

$$I = \sum_i C_i \times K_i \,, \tag{3.1}$$

Where:

I = Health Risk or Hazard Index;
C_i = Concentration of pollutant i the groundwater (mg/L);
K_i = Constant for the pollutant i $(\text{mg/L})^{-1}$.

Thus, the IFRA modeling computation can be initiated by first defining set U for pollutants and set V for risk levels as follows:

$$U = \{ u_i \mid \forall i \} \tag{3.2}$$
$$V = \{ v_j \mid \forall j \} \tag{3.3}$$

where u_i represent pollutant i, and v_j is for risk level j.
Fuzzy subsets of U and V can then be determined as follows:

$$\tilde{A} = a_1 / u_1 + a_2 / u_2 + ...a_m / u_m \,, \tag{3.4}$$

$$\tilde{B} = b_1 / v_1 + b_2 / v_2 + ... + b_n / v_n \,, \tag{3.5}$$

where a_i represents membership grade of u_i (for pollutant i) versus the multifactorial space, and b_j denotes an integrated membership grade for risk level j. The a_i value can be regarded as a weighting coefficient for pollutant i as follows:

$$\mathrm{a_i} = W_i \times \hat{C}_i \,, \tag{3.6}$$

where W_i is general weighting coefficient for pollutant i, which can be calculated using some multi-attribute decision-aiding model. Here a general weight for each pollutants has been decided according to the relative risk of the pollutants based on different health and ecological risk criteria. This weight was assigned with Analytic Hierarchy Process (AHP) using singular value decomposition (SVD). And $\hat{C}_i$ is normalized concentration of pollutant i.

A fuzzy subset of $U \times V$, which is a binary fuzzy relation from U to V, can be characterized through the following membership function:

$$\tilde{R} : U \times V \to [0,1] \tag{3.7}$$

Thus, we have fuzzy relation matrix:

$$\tilde{R} = \left\{ r_{ij} \middle| i = 0,1,...,m;\ j = 0,1,...,n \right\} \tag{3.8}$$

where r_{ij} is the membership grade of pollutant i versus risk level j, which is a function of pollutant concentration and risk level criteria.

With m pollutant under consideration, the pollutants concentration can be represented as follows:

$$C^{\pm} = \{ c_{ij}^{\pm} | i = 1,2,...,m;\ j = 1,2,...,k \} \tag{3.9}$$

where $c_{ij}^{\pm}$ denotes the lower and upper bound of the interval at the membership level μ_j for the i^{th} uncertain parameter.

For the problem of integrated risk assessment, the membership grade of fuzzy relation between given $c_{ij}^{\pm}$ at membership level μ_j for fuzzy number $C^{\pm}$ and risk level j can be calculated as follows.

Case 1: when $v_{i,j-1} \le c_{ij}^{\pm} \le v_{i,j}$:

$$r_{ij}^{+} = (c_{ij}^{+} - v_{i,j-1}) / (v_{i,j} - v_{i,j-1}), \forall i, j, \tag{3.10}$$

$$r_{ij}^{-} = (c_{ij}^{-} - v_{i,j-1}) / (v_{i,j} - v_{i,j-1}), \forall i, j, \tag{3.11}$$

Case 2: when $v_{i,j-1} \le c_{ij}^{+} \le v_{i,j}$ and $c_{ij}^{-} \le v_{ij-1}$

$$r_{ij}^{+} = (c_{ij}^{+} - v_{i,j-1}) / (v_{i,j} - v_{i,j-1}), \forall i, j, \tag{3.12}$$

$$r_{ij}^{-} = 0, \forall i, j, \tag{3.13}$$

Case 3: when $v_{i,j} \le c_{ij}^{\pm} \le v_{i,j+1}$:

$$r_{ij}^{+} = (v_{i,j+1} - c_{ij}^{-}) / (v_{i,j+1} - v_{i,j}), \forall i, j, \tag{3.14}$$

$$r_{ij}^{-} = (v_{i,j+1} - c_{ij}^{+}) / (v_{i,j+1} - v_{i,j}), \forall i, j, \tag{3.15}$$

Case 4: when $c_{ij}^{\pm} \le v_{i,j-1}$ or $c_{ij}^{\pm} \ge v_{i,j+1}$:

$$r_{ij}^{\pm} = 0, \forall i, j; \tag{3.16}$$

Case 5: when $c_{ij}^{-} \le v_{i,j}$ or $c_{ij}^{+} \ge v_{i,j}$:

$$r_{ij}^{+} = 1, \forall i, j; \tag{3.17}$$

$$r_{ij}^{-} = Min\{(c_{ij}^{-} - v_{i,j-1}) / (v_{i,j} - v_{i,j-1}), (v_{i,j+1} - c_{ij}^{+}) / (v_{i,j+1} - v_{i,j})\}, \forall i, j; \tag{3.18}$$

where v_{ij} is the criterion for pollutant i at risk level j. Thus, we can obtain the following interval parameter fuzzy relation matrix:

$$\tilde{R} = \left\{ r_{ij} \middle| i = 0,1,..., m;\ j = 0,1,..., n \right\}. \tag{3.19}$$

Similarly, we have:

$$\tilde{A}^{\pm} = \{ a_{ij}^{\pm} | i = 1,2,..., m;\ j = 1,2,..., k \}. \tag{3.20}$$

Thus, the integrated risk level can be determined as follows:

$$\tilde{B}^{\pm} = \tilde{A}^{\pm} \circ \tilde{R}^{\pm}, \tag{3.21}$$

where $\circ$ cab be a max-min or max-* composition [16].

Let $b_k^{\pm} = \max(b_j^{\pm} | j = 1,2,..., n).$ (3.22)

Based on the principle of maximum membership degree [16], it can be determined that system has an integrated risk level k.

4 Case Study

4.1 Overview of the Study System

A hypothetical problem is developed to illustrate integrated fuzzy modeling and risk analysis approach. The study site contains a leaking underground gasoline storage tank. About 600 m away from the tank area, there is a deep bore well used for rural drinking water supply. The recent groundwater monitoring data indicate high concentrations of several chemical stemming from petroleum products. The main

contaminants in leaked petroleum products are benzene, toluene, ethyl-benzene and xylenes (BTEX). What makes BTEX a topic of concern is the their toxic effects on human's health. All these compounds are acutely toxic and have noticeable adverse health effects at high concentrations. The BTEX can enter the human body through ingestion of contaminated crops, inhalation of vapour from the soil, intake of contaminated drinking water, and skin exposure. Drinking and bathing in water containing these contaminants can put one at risk of exposure.

A multi-component transport problem, with a continuous point source of pollution in a porous media with uniform flow field has been studied. For this purpose, a finite element generated numerical solution has been used. Such solution generally requires extreme simplifications, but the results can be used for approximate solutions. They are also very useful to illustrate the sensitivity of different parameters in overall uncertainty.

A numerical model consisting of 40x30 nodal grid with a uniform grid spacing of 50 m in both direction was used to simulate the model. Zero concentration boundaries were placed at the left, upper and lower model boundaries with a constant source placed at 500 m from the surface and 750 m from the left boundary. Sample data contaminated water is collected from 600 m from the pollution point source on the longitudinal section.

Characteristics of the uncertain parameters and other data used in the simulation are shown in Table 1 and Table 2 respectively.

Table 1. Triangular fuzzy numbers for uncertain in*Parameters*

	Low 0 M-value	Medium 1 M-value	High 0 M-value
V(m/day)	0.3	0.6	1.0
α_L (m)	100	200	300
α_T (m)	20	40	60

Table 2. Other crisp input data used *simulation*

Parameters	Value
Thickness of flow, b	50 m
Source strength, M	120 kg/day
Effective porosity, p	0.17
Grid distance (Δx)	50 m
Grid distance (Δy)	50 m
Time increment	1 day

4.2 Solution Approach

Fuzzy mathematical modelling using TM were used to gain insight into site conditions and to simulate fate and transport of pollutants in the subsurface. The simulation study can provide technical bases and solid support for in-depth risk assessment for subsurface contamination problems and give a detailed uncertainty analysis of the problem. Prediction results as fuzzy numbers will be directly used as inputs for risk-analysis models where issues of integration between numerical simulation and risk-analysis are to be addressed.

4.2.1 Simulation of Flow and Transport System

The procedures for solving the coupled multi-phase flow and multi-component transport problem are as follows:

Step 1: Solve the fluid equations simultaneously for the current time-step using time- lagged phase densities and interphase mass transfer. The distributions of hydraulic head and Darcy's velocity in space will be obtained.
Step 2: Solve the phase-summed transport using current values of apparent partition coefficients, interphase mass transfer rates and phase densities.
Step 3: Back-calculate new interphase mass transfer rates, update densities and apparent partition coefficients and repeat step 2 until transport solution converges.
Step 4: Proceed to the next time-step.

More detailed formulation and solution processes for the multi-phase and multi-component transport model in porous media were provided by [5], [6]. And a detail discussion of simulation using Transport method has been discussed in [7].

4.2.2 Risk Assessment Using IFRA

The results of the fuzzy simulation along with other system components has been used for risk assessment using IFRA. General weight (W_i) for each pollutants has been decided according to the relative risk of the pollutants based on different health and ecological risk criteria. Weight has been assigned using AHP with SVD. Weight and source strength of different BTEX compounds has been shown in Table 3. Risk level criteria for all compounds under study has been shown in Table 4 which has been decided in consultation of experts, EPA's recommendation of Maximum Contaminant Level (MCL) for drinking water and documentation for Immediately Dangerous to Life or Health Concentrations (IDLHs). Weighting coefficient a_i of each pollutants has been calculated using equation 4.6. Fuzzy subset V has been build as membership grade of u_i (for pollutant i) versus the multifactorial space using

Table 3. General weight and source strength of each pollutants

Pollutant	Weight	Source Strength (kg/day)
Benzene	0.45	13.2
Toluene	0.2	31.2
Ethyl Benzene	0.25	13.2
Xylene (o,m,p)	0.1	62.4

Table 4. Risk level criteria for all compounds under study (amount in mg/L)

Risk level	Benzene	Toluene	Ethyl Benzene	Xylene(o,m,p)
Low	0-0.005	0-1	0-0.7	0-10
Moderate	0.005-0.05	1-5	0.7-3	10-20
Moderately High	0.05-1	5-50	3-30	20-100
High	1-50	50-250	30-150	100-400
Very High	50-500	250-500	150-800	400-900
Deadly	>500	>500	>800	>900

normalised concentration of pollutants which are fuzzy numbers. The membership grade of fuzzy relation between given $c_{ij}^{\pm}$ at membership level μ_j for fuzzy number $C^{\pm}$ and risk level j can be calculated according to conditions set in equations 3.10-3.18. And finally the integrated risk level has been determined using equation 3.21.

5 Results and Discussion

A common approach for estimating and assessing environmental risks is to reduce the complicated systems into mathematical models. Generally in a deterministic model, the model parameters have lot of associated uncertainty. The input data can not be determined precisely because the state of knowledge is not perfect or near perfect. Assessment of the parameters can be based on expert judgement and sometime expressed as linguistic terms. Crisp set is unable to express this sort of uncertain data which can be best expressed by fuzzy numbers.

In this study, integrated fuzzy environmental modelling and risk assessment has been used to show usability of fuzzy simulation technique. The fuzzy transformation method has been used for system modelling. A finite element generated numerical solution for multi-phase and multi-component transport problem, with a continuous point source of pollution in a porous media with uniform flow field has been used for predicting pollutants concentration in groundwater. The result of the simulations has been shown in Figure 2.

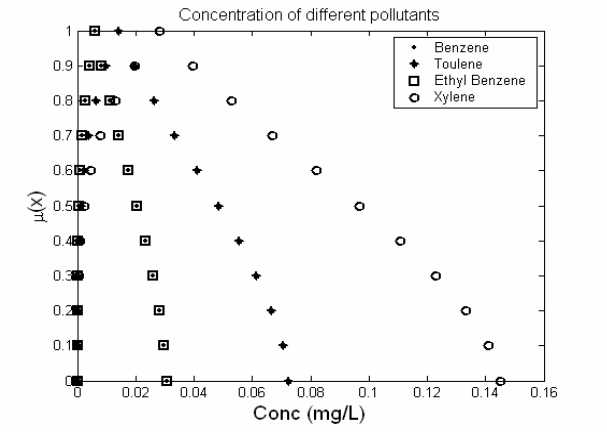


Fig.2. Concentration of different pollutants obtained from Fuzzy system simulation

Figure 3 is showing the concentration of solute at different time interval obtained from system simulation. The lower and upper bound of different membership level of fuzzy number has been mapped which show clear narrowing of width as it move to to higher lever of α-cut. Result has been compared with other fuzzy methods reported by [4]. The width of the concentration membership function obtained from Transformation method is narrower than other comparable fuzzy methods like vertex method in the same case study. The difference in the concentration output is mainly

due to interaction of the concentration variable in space and time dimensions. Neglecting this dependency of input variables result in overestimation of the imprecision of solute concentration. A detailed discussion of the effect of fuzzy number dependence can be found in [3].

For the purpose of quantifying uncertainty more effectively using fuzzy output, Integrated Fuzzy Relation Analysis has been used. The result of integrated risk analysis at different membership degree has been shown in Table 5. Degree of membership can be interpreted as confidence level. Uncertainty with the risk prediction is decreasing as level of confidence is increasing. Integrated risk at α-level 0 is 'Low to Moderately High' which become narrower at α-level 0.5 as 'Low to Moderate' which further narrowed to become 'Moderate' at α-level 0.8 and above. Average risk perception at this contaminated site can be quantified as 'Moderate'.

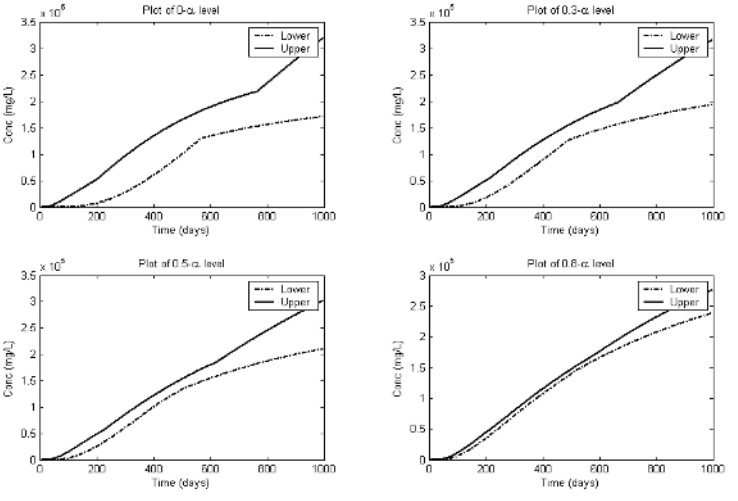


Fig. 3. Comparison of solute concentration outputs of solute transport at different α-levels obtained from Fuzzy Transformation method

Table 5. Integrated Risk at different membership levels

Membership level	Integrated Risk
0	Low to Moderately High
0.1	Low to Moderately High
0.2	Low to Moderately High
0.3	Low to Moderately High
0.4	Low to Moderate
0.5	Low to Moderate
0.6	Low to Moderate
0.7	Low to Moderate
0.8	Moderate
0.9	Moderate
1.0	Moderate

6 Conclusion

Proposed IFRA approach presents a new model to integrated risk assessment which contribute to the area of environmental risk assessment under uncertainty. Integration of system simulation and risk analysis using fuzzy approach allowed to incorporate system modelling uncertainty and subjective and inexact risk criteria. Case study from environmental domain has been considered in order to show its applicability in environmental engineering in general and environmental risk analysis in particular. Model can effectively incorporate different risk criteria and give an integrated assessment of risk at different confidence level. Assessment of risk at various confidence level present a comprehensive view of risk in uncertain environment and help decision maker with more choice.

Further exploration based on the proposed approach would be beneficial. For example it can be easily extended to incorporate effects of different pollutants and different remediation techniques and the cost of remediation within in a general framework. Trade-offs between environmental and economic objectives can be analyzed. In general, IFRA approach can be effectively incorporated in an environmental management decision-support system for remediation and managing contaminated sites or aquifers.

References

1. Abebe A.J., Guinot, V., Solomatine, D.P.: Fuzzy alpha-cut vs. Monte Carlo techniques in assessing uncertainty in model parameters, Proc. 4th Int. Conf. Hydroinformatics. Iowa (2000).
2. Dong W., Shah, H.C.: Vertex Method for Computing Functions of Fuzzy Variables, Fuzzy Sets and Systems, 24 (1987) 65-78.
3. Dou, C., Woldt, W., Bogardi, I., Dahab, M.: Steady state groundwater flow and simulation with imprecise parameters, Water Resour. Res., 31(11)(1995) 2709-2719.
4. Dou, C., Woldt, W., Bogardi, I., Dahab, M.: Numerical Solute Transport Simulation using Fuzzy Sets Approach, J. Contaminant Hydrology 27(1997) 107-126.
5. Kaluarachchi, J.J., Parker, J.C.: Modeling multicomponent organic chemical transport in three-fluid-phase porous media, J. Contam. Hydrol., 5(1990) 349 - 374.
6. Katyal, A.K., Parker, J.C.: An adaptive solution domain algorithm for solving multiphase flow equations, Computers & Geosciences, 18(1992) 1-9.
7. Kumar V., Schuhmacher, M.: Fuzzy uncertainty analysis in system modeling, Proceeding of ESCAPE –15, Barcelona, Spain(2005), 391-396.
8. Hanss, M., Willner, K.: On using fuzzy arithmetic to solve problems with uncertain model parameters. In Proc. of the Euromech 405 Colloquium, Valenciennes, France(1999) 85–92.
9. Hanss, M.: The transformation method for the simulation and analysis of systems with uncertain parameters, Fuzzy Sets and Systems 130 – 3(2002) 277–289.
10. Huang G. H., Chen Z., Tontiwachwuthikul P., Chakma A.: Environmental Risk Assessment for Underground Storage Tanks Through an Interval Parameter Fuzzy Relation Analysis", Energy Sources, in Vol. 21, No.1 (1999) 75-96
11. Pedrycz, W.: Fuzzy sets in pattern recognition, Pattern Recognition, 23(1990) 121-146

12. Sauty, J. P.: An analysis of hydrodispersive transfer in aquifers, Water Resour. Res. 16-1(1980) 145-158.
13. Zadeh, L.: Fuzzy algorithms, Information and Control 12 (1968) 94–102.
14. Zadeh, L. A.: Fuzzy Sets and Their Application to Pattern Classification And Clustering Analysis. In J. Van Ryzin (Ed.): Classification and Clustering. Academic Press (1977) 251-299.
15. Zadeh, L.: Fuzzy logic, IEEE Computer 1 , 14(1988) 83–93.
16. Zimmermann, H. J.: Fuzzy Set Theory and Its Applications, 2nd ed., Kluwer Academic Publishers, Boston, MA(1991).

Watermarking Non-numerical Databases

Agusti Solanas and Josep Domingo-Ferrer

Rovira i Virgili University of Tarragona,
Dept. of Computer Engineering and Maths,
Av. Països Catalans, 26, E-43007 Tarragona, Catalonia
{agusti.solanas, josep.domingo}@urv.net

Abstract. This paper presents a new watermarking method for protecting non-numerical databases. The proposed watermarking system allows the data owner to define a similarity function in order to reduce the distortion caused by watermark embedding while, at the same time, reducing the number of element modifications needed by the embedding process. A mathematical analysis is provided to justify the robustness of the mark against different types of malicious attacks. The usefulness of this extensible and robust method is illustrated by describing some application domains and examples.

Keywords: Private watermarking, Data hiding, Database security, Non-numerical databases.

1 Introduction

Watermarking systems have been widely studied for intellectual property protection (IPR) of multimedia data [1, 2, 3, 4]. It is common for watermarking systems to make use of well-known cryptographic techniques such as digital signatures [5] or signal processing techniques such as phase modulation [6] or spread spectrum [7, 8]. Most methods designed for multimedia data rely on the perceptual limitations of humans, , our inability to distinguish between very similar colors or sounds [9, 10]. However, over the last few years, researchers have realized that these limitations cannot be exploited when trying to protect other kinds of data, such as software [11] and databases [12].

Recent contributions on database IPR [12, 13] have clarified the main differences and singularities of typical database content (alphanumeric data) vs multimedia. Databases have very little redundancy as compared with multimedia data and this fact makes it very difficult to find enough bandwidth in which to embed the watermark. Moreover, databases can contain non-numerical or categorical data like city names, drug names, hair colors, etc. Such non-numerical data cannot be smoothly marked by increasing or reducing their value or modifying some of their bits. A non-numerical element must be completely altered in order to embed a mark and this limitation represents a great challenge that is addressed in this paper.

Some authors have tackled this problem before ([13]), but their proposals have some shortcomings regarding data distortion and watermark length. We present

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 239–250, 2006.

here a watermarking system s for non-numerical data that: i) minimizes the number of changes needed to embed the mark and; ii) reduces the distortion produced by the mark by allowing the user to customize the watermark embedding system through the definition of a similarity function related to the data.

The rest of the paper is organized as follows. In Section 2, we describe our model in detail and specify the notation and the assumptions used. Section 3 presents our watermarking system. Section 4 analyzes its properties of robustness against different types of malicious attacks. In Section 5, the usefulness of the similarity function is justified by presenting some examples and application domains. Finally, conclusions are listed in Section 6.

2 The Model

We first state a model to which all subsequent solutions refer. We consider the following elements to define our model: i) the data: what are we working with? ii) our target: what do we want to achieve? and, iii) the enemy: what kind of attacks are likely to be used by intruders to destroy our watermark? Moreover, at the end of this section, some brief comments about notation and assumptions are made.

2.1 The Data

The data we work with consist of a finite number T of non-numerical and discrete elements E stored in a database. It is assumed that all elements in E are known and can be ranked (, alphabetic ordering would do). We easily find lots of examples of this kind of data: for example, city names, carmaker names or drug names. The main characteristic of non-numerical data is that they cannot be smoothly modified, this is, any change is a complete change of value. If the data we have are so critical that they cannot be modified at all then no watermark system can be applied because –by its very nature– a watermarking system has to change some elements in order to embed the mark. In this paper, we assume that the data can be modified to a limited extent.

We consider a database organized in relations R, where each relation can be viewed as the union of a primary key $R.Pk$ and one or more attributes A. The proposed watermarking system can be applied to any relation in which modification of the primary key is not allowed: as argued in [12], we assume that modification of the primary key results in an unacceptable loss of information.

Without loss of generality, we consider a relation R formed by the union of a primary key $R.Pk$ and a single attribute A, that is

$$R \rightarrow \{R.Pk, A\}$$

2.2 Our Goal

We want to be able to hide a mark into the data without causing unacceptable data modifications while, at the same time, making the mark as robust as possible

against different types of malicious attacks. This challenging problem can be broken down into two goals:

- Minimize the number of changes caused by the watermark embedding system, while maintaining its robustness;
- Allow the user to extend the system by defining a specific similarity function for minimizing the impact of the changes.

2.3 The Intruder

The intruder wants to get hold of data D and, after a malicious modification, sell an unmarked version D' for which the owners of D are unable to show their intellectual property rights. Of course, data utility for D' and D should be similar for the attack to make sense (otherwise either D' is useful but still carries the mark or D' has become useless as a result of mark removal attacks). To destroy the mark embedded in D, the intruder can use different types of malicious attacks.

- **Horizontal sampling:** In this attack, the intruder randomly selects a set of tuples and discards the rest. Thus, if the mark depends on any kind of spatial relation, it will be lost. The mark has to be resilient to this attack, so that the intruder is forced to reduce the number of tuples selected from D to an extent such that the resulting D' is no longer very useful.
- **Vertical sampling:** Similar to the previous attack, this one is based on randomly selecting attributes in a tuple that will be erased. In order to resist this attack, the watermark should be recoverable from a single attribute.
- **Perturbation of randomly chosen elements:** Let a data element be the value taken by a specific attribute in a specific tuple. Perturbing a randomly selected subset of data elements is a very common attack for numerical data. The main difference when applying this attack to non-numerical data is that any modification is likely to be significant; it is not easy for the intruder to perturb non-numerical data without substantial utility loss. In order for the mark to be resilient against this attack, it should resist as many element perturbations as needed to render D' useless.
- **Horizontal and vertical re-ordering:** This attack consists of swapping pairs of tuples or attributes without modifying them. If the mark has to resist this attack, it cannot be based on any relative spatial position of the data elements.

2.4 Notation and Assumptions

In this section, we briefly enumerate some assumptions and notation used in the remainder of this article (see also Table 1).

- Hash function H. Secure one-way hash functions such as SHA [14] are used in our algorithm. Given a value z', we assume it is computationally unfeasible to find z such that $H(z) = z'$.

Table 1. Table of symbols

Symbols	Meaning
E	Set of all non-numerical elements
G	A pseudo-random generator
N	Number of markable elements
n	Number of actually marked elements
K_1	Secret key used for embedding the mark
K_2	Secret key used for computing the mark
P	One-dimensional table of products $x_i s_i$
R	A relation in the database
sf	A similarity function
T	Total number of elements in the data
t	A tuple in the relation
V	Binary vector of selected elements
γ	Fraction of selected elements for embedding

- Pseudo-random number generator G. This generator must be seeded with a combination of information taken from the data and a secret key only known to the owner of the data. We assume that the pseudo-random generator outputs numbers that are uniformly distributed between 0 and 1 Once the pseudo-random generator is initialized, a new number is obtained by using the $_{l}$ $_{\cdot\cdot}$ G function.
- Least significant bit $_{\cdot_{l}}$ $_{\cdot}(e)$. This function returns the value of the least significant bit of the elements $e \in E$ and is mainly used in the mark recovery process.

3 Our Watermarking System

The proposed watermarking system can be subdivided into two subsystems: embedding and recovery.

3.1 The Embedding Subsystem

Embedding faces three main problems. First, it is necessary to decide where the watermark has to be hidden, that is, the elements of the attribute A in the relation R which will be considered candidates for a modification. Second, a similarity function df can be defined by the user in order to minimize the impact of embedding a watermark W into the data D and the watermark embedding system must allow the user to do so. Third, the watermark W must be computed and embedded into D, while meeting the previous restrictions.

Selection of the embedding positions. Similarly to [12] and [13], we make the assumption that a primary key $r.P_k$ exists in the relation which cannot be modified without causing unacceptable damage to data. We want the selected elements to be picked independently of their relative position in the relation.

Algorithm 1. Selection of the elements to be modified

```
1)  function GetElementsToModify(K1, R) return V→
2)      for each tuple t ∈ R do
3)          seed G with r.Pk||K1
4)          if next(G) ≤ γ then
5)              Vt = 1
6)          else
7)              Vt = 0
8)          end if
9)      end for
10)      return V→
11)  end function GetElementsToModify
```

To that end, we use the primary key that uniquely addresses an element. In order to make this process secure, we use a secret key K_1 and we concatenate it with the primary key $r.P_k$ to obtain a value that is used to seed a pseudo-random generator G. The data elements for embedding are selected by using the pseudo-random generator as described in Algorithm 1.

After running Algorithm 1, a one-dimensional table $\vec{V}$ is obtained. The size of this table equals the number T of tuples in the relation. Each position of $\vec{V}$ contains a 0 or a 1 representing the selection result. All positions set to 1 represent the tuples in the relation that are selected for being modified in order to embed the mark.

In order to control the impact of watermark embedding, the relation between the selected elements and the total number of elements T is controlled through the γ parameter. The expected number of selected elements is $N = \gamma T$. Thus, if $\gamma = 0.25$ then 25% of tuples can be expected to be selected for being marked.

The similarity function. When a mark has to be hidden into numerical data, numerical data elements can be smoothly modified by slightly increasing or decreasing their values. On the contrary, hiding a mark into a non-numerical data element is often not smooth, as it implies substituting a categorical value for another. Previous approaches [13] assume that the replacement of a categorical value by another introduces the same distortion into the data independently of the new categorical value that replaces the original one. Even if we agree that changing the category of a non-numerical element is an important modification, we claim that the amount of distortion caused by this replacement depends on the similarity between the original category and the replacement category. In order to minimize the impact of watermark embedding on the data, we propose to resort to a ***user-defined*** similarity function, $sf(e_1, e_2) \rightarrow [0, 1]$

Given two elements e_1 and e_2, the similarity function returns a similarity value in $[0, 1]$. A 0 similarity is interpreted as "very different" and a 1 similarity as "very similar". Using such a similarity function, the distortion produced by swapping two data elements can be quantified and minimized. In Section 5 some example similarity functions applied to different domains are described.

Hiding the mark. The last step in the embedding process consists of: i) finding the elements that will replace the original ones in order to hide the watermark; ii) carrying out the replacement. This process can be denoted as:

$$Embed(R, K_2, sf, M) \rightarrow R'$$

To hide the watermark in the relation R we need: i) a secret key K_2 different from the one used to select the embedding position [1]; ii) a similarity function sf to minimize the impact of watermark embedding (optionally defined by the user); and iii) a security parameter M.

Once the elements that will be modified are selected using Algorithm 1, we specify the constraint below to be met by the elements that will replace the original ones:

$$\sum_{i=0}^{N} s_i x_i \geq M \tag{1}$$

where: $\vec{X} = \{x_i\}$ are pseudo-random numbers uniformly distributed in $[-\lambda, \lambda]$, where λ is a robustness parameter; $S = \{s_i\}$ are the least significant bits of the replacement data elements expressed as integers in $\{-1, 1\}$; M is a user-definable security parameter that determines the robustness and the impact of the mark. In the next paragraphs, details are given on the computation of x_i and s_i.

$\vec{X}$ A value x_i is computed for each selected tuple. that is, for indexes i such that $V_i = 1$ (in terms of Algorithm 1). This computation is performed by using a secret key K_2 and the primary relation key $R.Pk$ [2] to seed a pseudo-random number generator G. Then a set of N pseudo-random numbers are obtained using G and they are scaled in $[-\lambda, \lambda]$. In other words, we use a hash function that receives the concatenation of the primary key and a secret key K_2 as an input parameter and returns a number in $[-\lambda, \lambda]$.

$$H(R.Pk|K_2) \rightarrow [-\lambda, \lambda]$$

We require G to be such that $H(\cdot)$ is a secure one-way hash function: inferring the value of $R.Pk|K_2$ from $H(R.Pk|K_2)$ should be infeasible.

$\vec{S}$ Once the values $\vec{X} = \{x_i\}$ are fixed, we must determine values s_i satisfying Constraint (1). We want to minimize the impact of mark embedding on data, which translates to reducing the number and magnitude of changes to be made.

We initialize each s_i with the least significant bit $lsb(e_i)$ of the original element e_i to be marked. Specifically,

[1] It is possible to compute the watermark by using only one secret key, but we prefer to use two keys in order to avoid the risk of correlations between the generated pseudo-random numbers[13].

[2] Since we assume that the primary key cannot be modified, the values of X are only obtainable by the data owner and cannot be modified.

Algorithm 2. Computation of $\vec{S}$

```
1)  procedure SetSElements(S⃗, X⃗, M)
2)      P = ComputeProducts(S⃗, X⃗)
3)      SortInIncreasingOrder(P⃗)
4)      while M̂ < M do
5)          i=ObtainIndexOfMostNegativeProduct(P⃗)
6)          Swap(S⃗,i)
7)          RemoveIFromP(P⃗,i)
8)          M̂=ComputeMark(S⃗,X⃗)
9)      endwhile
10)  end procedure SetSElements
```

$$s_i = \begin{cases} -1 \text{ if } lsb(e_i) = 0 \\ \\ 1 \text{ if } lsb(e_i) = 1 \end{cases}$$

After the initialization of $\vec{S}$, we compute

$$\sum_{i=0}^{N} s_i x_i = \hat{M}$$

Next, if $\hat{M} > M$ then Constraint (1) is met; so, we take $\hat{M}$ as M and no changes are introduced to the data (minimum distortion). When $\hat{M} < M$ then it is necessary to change some values of $\vec{S}$ in order to satisfy the embedding constraint. The way in which the values of $\vec{S}$ are changed is described in Algorithm 2. The algorithms called within Algorithm 2 are described in the remainder of the Section (Algorithms 3 and 4).

Initially, the products $\vec{P}$ of each x_i and s_i are computed in order to find the impact of the i-th element in the computation of $\hat{M}$. Then the one-dimensional table $\vec{P}$ is sorted in order of increasing magnitude. To satisfy Constraint (1) with the minimum number of changes, the least significant bit s_i of the most negative product is inverted; in that way, with a single bit inversion, we obtain a maximum increase of $\hat{M}$. To perform the inversion of s_i, the element e_i in the relation R must be replaced by the most similar element of E with a different least significant bit (see Algorithm 4). A similarity function sf is used to determine the most similar element to e_i. This similarity function should be defined by the owner of the database. However, it is optional and when it is not given, a simple alphabetical comparison could be made to obtain a similarity value.

λ Note that the magnitude of the most negative product is related to the range $[-\lambda, \lambda]$ where the x_i are chosen. Thus, a larger λ will reduce the expected number of iterations of Algorithm 2 and therefore the expected

Algorithm 3. Computation of the watermark from a given $\vec{S}$ and $\vec{X}$

```
1)  function ComputeMark(S⃗, X⃗) return M̂
2)      M̂ = 0
3)      For i = 1 to N do
4)          M̂ = M̂ + s_i x_i
5)      end for
6)      return M̂
7)  end function ComputeMark
```

Algorithm 4. Replacement of an original element by its most similar substitute

```
1)  procedure Replace(S⃗, i)
2)      lsb = 0      //Initialize the least significant bit
3)      if s_i == 1 then      //change the s_i value
4)          s_i = −1
5)          lsb = 0
6)      else if s_i == −1 then
7)          s_i = 1
8)          lsb = 1
9)      endif
10)     newElement = getMostSimilarElement(e_i, lsb, sf)
11)     e_i = newElement
12)  end procedure Replace
```

number $n \leq N$ of actually marked elements. The drawback of taking λ too big is that, the larger λ, the less elements will carry the mark, so that we gain imperceptibility but lose robustness. Therefore, λ should be chosen so that the resulting n is not much smaller than the number of markable elements N.

It is easy to see that, following Algorithm 2, the [illegible] of changes made to satisfy Constraint (1) is minimal for a fixed value of λ. Using a similarity function sf capturing the semantics of data allows each individual change (replacement) to be minimal in [illegible]; this is done by the [illegible] user-definable function called in Algorithm 4. The result is minimal data alteration in watermark embedding.

3.2 The Recovery Subsystem

Watermark recovery must determine whether a watermark is embedded in a relation. To perform this task, the recovery subsystem receives as parameters: the relation $\hat{R}$ which presumably embeds the mark and may have been attacked; the security parameter M; and the secret keys K_1 and K_2 only known to the data owner. Thus, this subsystem can be denoted as:

$$Recovery(\hat{R}, K_1, K_2, M) \rightarrow (yes/no)$$

Similarly to the embedding process, it is first necessary to obtain the marked elements using Algorithm 1. Note that is not necessary to know the original R in order to apply Algorithm 1 because the primary key of R is supposed to remain unmodified in $\hat{R}$. Once the marked elements are located, the value of each x_i is computed in the same way as in the embedding process, using the secret key K_2. Finally, the value of each s'_i is obtained from the least significant bit of the elements by applying the $lsb()$ function. Note that, in general it can happen that $s'_i \neq s_i$, as a result of accidental/intentional distortion during the data lifecycle. Once all the above information is recovered, the recovery subsystem computes $\sum_{i=0}^{N} s'_i x_i = \hat{M}'$

The recovery subsystem decides that the data contain a watermark when $\hat{M}' \geq \frac{M}{2}$. Otherwise, no mark is recovered.

4 Robustness Analysis

The proposed watermarking system is robust against random alterations and vertical and horizontal sampling. The intruder can perform a broad range of different malicious attacks. We now describe how the watermark embedded by our watermarking system tolerates these attacks.

- **Vertical sampling:** Our system can be applied to any relation R with at least a primary key and an attribute A. The inserted mark does not depend on any relationship between attributes and can be embedded individually in as many attributes as desired. Thus, the attack based on selecting some attributes and erasing the rest has no effect on our watermark because, at least, one marked attribute remains.
- **Horizontal and vertical re-ordering:** The horizontal re-ordering attack consists of swapping the positions of pairs of tuples without modifying them. Our watermarking system is not vulnerable to this kind of attack because the relative position of the elements in the relation R is not used to determine whether they are marked.
  Similarly, vertical re-ordering consists in swapping the positions of pairs of attributes without modifying them. As argued in the previous sections, our method is applied to an attribute and it does not depend on its relative position in the relation.
- **Perturbation of randomly chosen elements:** The recovery system detects the existence of a mark when $\hat{M}' \geq \frac{M}{2}$. The intruder wants to destroy the mark by modifying the value of randomly chosen elements. If the intruder is able to destroy enough marked elements then the mark will not be recovered. Thus, a natural strategy that leads to arbitrary reduction of the probability of mark destruction is to increase the number n of marked elements, which can be done by decreasing λ and increasing the number N of markable elements.
- **Horizontal sampling:** This malicious attack is based on a random selection of a fraction of tuples of a relation R. This usually tricky attack is not

effective against our method because the amount of non-selected tuples has to be very big compared with the number of tuples modified by the watermark in order to destroy it. Considering that a non-selected tuple is like an altered tuple, the analysis is analogous to the one above for perturbation attack.

4.1 A Toy Example

To illustrate the robustness of our model against perturbation of randomly chosen elements, we take a toy database that consists of a relation R with 250 tuples or elements. We embed a mark that modifies $n \leq N = 25$ elements (10%) and a mark that modifies $n \leq N = 30$ elements (12%). We assume that, for data to stay useful, up to 30% of elements can be modified. This is up to 75 element modifications, about three times the number of modifications caused by watermark embedding. Also, we have chosen a value for M such that the mark is destroyed if more than half of the N markable elements are modified. If the intruder modifies P elements among the total T elements, the probability that she destroys the mark by randomly hitting more than $N/2$ markable elements is

$$P[Destruction] = \frac{\sum_{i=0}^{N/2} \binom{N}{\frac{N}{2}+i} \binom{T-N}{P-(\frac{N}{2}+i)}}{\binom{T}{P}}$$

Table 2 shows the probability of the intruder destroying the mark by modifying less than 30% of the elements, that is, up to $P = 75$ elements. It can be seen that, even if the intruder modifies three times as many elements as those modified by the mark embedding algorithm (75 vs 25) her probability of success is no more than 0.25.

Table 2. Destruction of the mark in the toy example

Modified elements	P(Destroy |25 marked elem.)	P(Destroy |30 marked elem.)
25	0.0000087	≈ 0
30	0.000077	0.0000009
35	0.000422	0.00001
40	0.0016	0.00007
45	0.0051	0.00034
50	0.0132	0.0012
55	0.029	0.0038
60	0.057	0.0099
65	0.102	0.022
70	0.16	0.045
75	0.25	0.081

5 Application Domains

The main application of the presented watermarking system is to protect non-numerical (i.e. categorical) databases from being copied and re-sold by an intruder. These databases are sold to companies which want to obtain information from the data, usually by applying data mining techniques.

We next illustrate how a similarity function could be defined in a couple of specific example databases.

5.1 Drugs Database

Imagine that we have a drugs database storing information about the drugs taken by a set of patients. We may have information about the composition of each drug and we can determine the similarity between them. In this case of study, the similarity function defined by the user may be based in the next considerations:

Similarity function: Coincidences in the number and proportion of components in a given drug. Following this similarity function we can replace the element "ASPIRIN 250g" by the generic element "acetylsalicylic acid 250g" without any distortion. Note that the element "acetylsalicylic acid 250g" must be in the database in order to be considered for replacing the element "ASPIRIN 250g". Similarly, if we try to replace "ASPIRIN 250g" by "CHLORHEXIDINE GLUCONATE 1g" the similarity function must return a value very close to 0.

5.2 Network Nodes Database

Let us consider a case where we have the database of an internet service provider. This database contains a set of network nodes determined by a discrete label (e.g. A2345-C, B3490-D). If we use alphabetical order and do not care about the similarity between nodes, the impact produced by watermark embedding could be important. However, if we consider a similarity function this impact could be clearly reduced.

Similarity function: In this case, the similarity function can be defined as the number of "hops" between nodes. This measurement gives a very concise idea about the location of the nodes. Thus, if two nodes are nearby, the similarity function will tend to 1. On the contrary, if the number of "hops" from one node to another is large, the similarity function returns a number close to 0.

6 Conclusions and Further Work

We have presented a new watermarking system for protecting non-numerical data. The system minimizes the number of modifications needed to embed the mark and allows the data owner to define a similarity function to guide each individual modification so that the utility loss it entails is minimal. The robustness analysis demonstrates the resiliency of our mark against different kind of malicious attacks. The similarity function is user-defined and depends on the particular database to be protected; this has been illustrated with two examples. Future work will involve a false positive rate analysis and extensive robustness tests in large databases with a broader range of attacks. Also, the definition of a similarity function that optimally (rather than reasonably) captures data utility loss in a specific database is a nontrivial issue for future research in artificial intelligence.

Acknowledgments

The authors are partly supported by the Catalan government under grant 2005 SGR 00446, and by the Spanish Ministry of Science and Education through project SEG2004-04352-C04-01 "PROPRIETAS".

References

1. Jordan, F., Vynne, T.: Motion vector watermarking. Patent (1997) Laboratoire de Traitement des Signaux École Polytechnique Fédérale de Lausanne.
2. Hartung, F., Girod, B.: Watermarking of uncompressed and compressed video. Signal Processing **66** (1998) 283–301
3. Katzenbeisser, S., Petitcolas, F.A.P.: Information Hiding: techniques for steganography and digital watermarking. Computer security series. Artech House (2000)
4. Domingo-Ferrer, J.: Anonymous fingerprinting of electronic information with automatic reidentification of redistributors. Electronics Letters **34** (1998) 1303–1304
5. Pitas, I., Kaskalis, T.H.: Applying signatures on digital images. In: IEEE Workshop on Nonlinear Signal and Image Processing, Thessaloniki, Greece (1995) 460–463
6. Ruanaidth, J.O., Downling, W., Boland, F.: Phase watermarking of digital images. In: Proceedings of the IEEE international Conference on Image Processing. Volume 3. (1996) 239–242
7. Cox, I.J., Kilian, J., Leighton, T., Shamoon, T.: Secure spread spectrum watermarking for multimedia. Technical report, NEC Research Institute, Technical Report 95 - 10 (1995)
8. Sebé, F., Domingo-Ferrer, J., Solanas, A.: Noise-robust watermarking for numerical datasets. Lecture Notes in Computer Science **3558** (2005) 134–143
9. Delaigle, J.F., Vleeschuwer, C.D., Macq, B.: Watermarking using a matching model based on the human visual system, Marly le Roi. (1997)
10. Ko, B.S., Nishimura, R., Suzuki, Y.: Time-spread echo method for digital audio watermarking. IEEE Transactions on Multimedia **7** (2005) 212–221
11. Collberg, C.S., Thomborson, C.: Watermarking, tamper-proofing and obfuscation: tools for software protection. IEEE Transactions on Software Engineering **28** (2002) 735–746 http://dx.doi.org/10.1109/TSE.2002.1027797.
12. Agrawal, R., Haas, P.J., Kiernan, J.: Watermarking relational data: framework, algorithms and analysis. VLDB journal **12** (2003) 157–169
13. Sion, R.: Proving ownership over categorical data. In: Proceedings of the 20th International Conference on Data Engineering (ICDE04), Boston (2004) 584–595
14. NIST: Proposed federal information processing standard for secure hash standard. Federal Register **57** (1992) 41727

New Approach to the Re-identification Problem Using Neural Networks

Jordi Nin and Vicenç Torra

IIIA-CSIC
Campus UAB s/n
08193 Bellaterra (Catalonia, Spain)
jnin@iiia.csic.es
vtorra@iiia.csic.es

Abstract. Schema and record matching are tools to integrate files or databases. Record linkage is one of the tools used to link those records that while belonging to different files correspond to the same individual.

Standard record linkage methods are applied when the records of both files are described using the same variables. One of the non-standard record linkage methods corresponds to the case when files are not described using the same variables.

In this paper we study record linkage for non common variables. In particular, we use a supervised approach based on neural networks. We use a neural network to find the relationships between variables. Then, we use these relationships to translate the information in the domain of one file into the domain of the other file.

Keywords: Database integration, record linkage, re-identification algorithms, neural networks, data mining, information fusion, information privacy.

1 Introduction

Information systems are currently pervasive and information about individuals is scattered in databases. Such information is normally distributed and stored in an heterogeneous way.

In this scenario, database integration techniques (, data cleaning and information fusion techniques) are crucial. Such techniques intend to make possible the retrieval of relevant information of a single individual even when the required data for answering a query was distributed and represented in a non homogeneous way.

Actually there are many techniques for doing these tasks. Re-identification algorithms [17] are one of the most important methods for information fusion and data cleaning. They are used to identify those objects that appear in more than one file. For example, there are algorithms (record linkage / record matching algorithms) used to identify those records that while found in different files

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 251–261, 2006.

correspond to the same individual [8]. Also, there are algorithms (schema matching) that are used to identify variables that under different names correspond to the same semantic concept [13].

In this work we will focus on record linkage algorithms. As said, the goal of record linkage algorithms is to link the records that while belonging to different files or databases correspond to the same individual.

Such algorithms are used for several different purposes. One of their uses is information fusion (information integration or database integration). In this case, record linkage permits to join two, or more, files putting in a common framework all the individuals represented in the files. In this case, record linkage algorithms are combined with other techniques for data consolidation. This is, techniques to remove inconsistencies and improve data accuracy. So, the resulting database has a better quality than the original ones.

Data integration and data consolidation can either be considered as a final process (as in databases), or as one data preprocessing technique (as in multidatabase data mining).

Another use of record linkage tools is risk assessment in privacy preserving data mining (PPDM) [1] and statistical disclosure control (SDC) [15]. In this framework, the algorithms permit to evaluate whether a protection mechanism provides an adequate protection to the providers of sensitive information (˛ to the respondents of an economic survey). This is, record linkage is a tool that permits to evaluate whether disclosure is avoided when sensitive data is perturbed before its public release.

Classical record linkage methods need that files to be joined have some variables in common, ˛ the age of an individual appears in both files. The difficulties of this approach are caused by the fact that the files to be linked usually contain errors. These errors may be accidental (˛ , due to an incorrect manipulation of the data) or intentional (˛ to protect sensitive information as in PPDM).

The goal of advanced record linkage methods is the re-identification of records when files are not described using common variables. In this case, standard record linkage algorithms cannot be applied. So, alternative approaches are required. One of the approaches is based on the so-called structural information [12].

In [8], [9] we show that record re-identification is possible using OWA operators [18] and twofold integrals [5], [6]. In such works, the structural information corresponds to computing some numerical representatives of the data. Fuzzy integrals are used for extracting several representatives for each record, and then standard re-identification algorithms can be applied.

Another approach for dealing with such problem is to study the relationship or dependence among the variables in the different files. In this case, in the case that only two files A and B are considered, the goal is to build a model between the variables of A and the variables of B. In this way, it is latter possible to translate the values on the domain of A into valuess on the domain of B. Then, after such translation, re-identification is possible using standard record linkage. This is done using the new translated file, say A', and the original file B.

In this work we consider that the construction of such model is done in a supervised way. This is, we consider that there are a set of records of both files A and B for which we know the correct re-identifications. Such records are used to build the model between the two files.

In this paper we propose the use of neural networks (with a backpropagation algorithm) for learning such model. We show that re-identification based on this approach leads to good results for several problems. These results are significatively better than previous approaches. So, the use of a supervised tool permits to increase the accuracy and performance of record linkage, even in the case that files do not share variables.

A related work for re-identification that also use neural networks is SEMINT (SEMantic INTegrator). SEMINT, described in [4], is a tool based on neural networks, to identify relationships between attributes or classes in different database schemas. [4] shows that it is possible, using schema information and data content, to produce rules for matching corresponding attributes automatically. Although this work is related in the sense of using neural networks for re-identification, their goal is different from the one described in this paper. Here, we focus on record matching (record linkage) instead of schema matching.

The structure of the paper is as follows. In Section 2 we describe some elements that are needed latter on. Then, in Section 3 we introduce our approach to record linkage. Section 4 describes some of the experiments performed. Then, the paper finishes with some conclusions and description of future work.

2 Preliminaries

This section describes the backpropagation algorithm using gradient descent as well as classical re-identification methods. These methods are used by our approach when we transform the problem of re-identification without non-common variables into the standard problem of re-identification with common variables.

We start with the description of the backpropagation algorithm, and then, in Section 2.2 we revise the re-identification methods.

2.1 The Backpropagation Algorithm

Neural networks are one of the most common pattern recognition techniques. In some problems, when there is no clear method to establish a model, neural networks are an alternative approach.

Neural networks can learn the relationships among attributes directly from its instances/examples and empirically infer solutions from the data without prior knowledge.

Unlike traditional computing, neural network are trained, not programmed. One of the learning algorithm for training neural networks is the backpropagation algorithm.

There are two ways for applying backpropagation algorithm, one of them is using analytical derivations [3], and the other is using the method of gradient descent [10]. We have used the second approach.

The backpropagation algorithm using the method of the gradient descent looks for the minimum of the error function in a weight space. So, the goal is to find a combination of weights which minimizes the error function. Then, the method computes for each example the gradient of the error function, and with this gradient the weights are updated. This process is repeated several times for all examples. A complete description of this method is given in [10].

To apply this algorithm we need to use the so-called activation function which guarantees the continuity and differentiability. In our implementation we have used the Sigmoid function (see Section 4.2 for details).

2.2 Re-identification Methods

The goal of classical re-identification methods is to link records that while belonging to different files correspond to the same individual. In the classical approach records are described using the same variables. , we have one file with the census of a city or country and another file with the income tax. Both files contain the fields and others. Besides both files contain other data specific to either census information or income tax information. Record linkage algorithms permit to link specific information about census and income taxes when they refer to the same individual using the common variables. Difficulties in the re-identification process are due to the errors in names, address, and so on. These errors can be due to misspellings (, John vs. Jonh) or changes of location (, Clark Kent previously living in Smallville has moved to Metropolis).

Two main approaches have been used for re-identification in the standard case. See [11, 16, 17] for more details:

Distance-Based Record Linkage: Records of two files A and B are compared, and each record in A is linked to the nearest record in B.

Probabistic Record Linkage: Conditional probabilities of coincidence (and non-coincidence) of values among records given correct matching are obtained. From these conditional probabilities an index is computed for each pair of records (a,b) with a in A and b in B. This index is used to classify pairs as linked (a and b correspond to the same individual).

2.3 Record Linkage Evaluation

The comparison of non-standard record linkage methods is not easy. The most straightforward way is to compare the results of new approaches with existing ones. We will use this approach in Section 4 to evaluate the results of our new approach. In particular, we will compare with the results in [8], [9].

An alternative approach used in [8], [9] was to compare the performance with a . This approach shows that the probability of correctly linking more than one or two records in a file of more than 100 records is negligible. We do not include such comparison here as the number of re-identified records is clearly significant.

3 Record Linkage Using Neural Networks

In this paper, as we have explained in the introduction, we consider the re-identification problem when files do not share variables. We will consider the case of only two files.

To deal with this problem, we reduce the non-standard situation into one in which we have two files with common variables. So, we are able to apply the methods described in Section 2.2. To do so, we use here a supervised approach, following machine learning jargon. This is, we have a set of examples from which we are able to learn some models.

Formally speaking, we will consider in our approach that we have two files A and B, and that these files have different variables but that their records contain information on the same individuals. Then, in order to apply traditional record linkage methods we transform the original file A into a new file A' that have the same variables of file B. For achieving this, under the supervised learning assumption, we construct a neural network for each variable in B in terms of the variables in A. Then, using such networks we can calculate for each record a in A a new record a' that contains the same variables than the records in B.

Let A^{sl} and B^{sl} be two files with the data for the supervised learning process, and let A and B be the two files for re-identification. Let V_A be the variables in files A and A^{sl} and let V_B be the variables in files B and B^{sl}. As files A^{sl} and B^{sl} are used in the learning process, both files should contain the same number of records and it is known the correct matches between records in A^{sl} and the records in B^{sl}.

Considering this notation, our approach can be defined as follows:

1. For each variable v in V_B do:
   (a) Define a neural network N with input variables V_A and output variable v. Let $NN(a)$ denote the output of this neural network for inputs a.
   (b) Apply the backpropagation algorithm to learn the parameters of the model NN using the file A^{sl} and a projection of the file B^{sl} on the variable v.
   (c) For each record a in A, find $a' = NN(a)$, and use these values to define the projection of A' on variable v.
2. Apply standard record linkage (either distance-based or probabilistic record linkage) to files A' and B. Note that the file A' contains as much variables as B.

The next section describes the experiments that we have performed to validate our approach.

4 Experiments

To analyze our approach for record linkage, we have tested it using eight different problems. These problems, that were previously studied in [8], [9], are defined using data obtained from the UCI repository [14] and the Data Extraction System (DES) from the U. S. Census Bureau [2].

To test record linkage in the case of non-common variables, we have split the downloaded files into two different files in such a way that both files contained exactly the same records but only some of the variables. The partition has been done considering only the numerical attributes, and among them those with a high correlation coefficient (greater than 0.7). All other attributes have been discarded.

Below, we list the problems studied. For each problem, we give two lists of attributes. They correspond to the attributes considered in each of the files. All attributes are numerical, and detailed description of the meaning of the attributes can be found in the original files.

Problems studied:

- Iris Plants Database: {sepal length, petal length}, {sepal width, petal width}
- Abalone Database: {Height, Whole weight, Viscera weight}, {Length, Diameter, Shucked weight, Shell weight}
- Ionosphere Database: {V5, V7, V9, V11, V13, V20}, {V15, V17, V19, V21, V23, V30}
- Dermatology Database: {polygonal papules, follicular papules, oral mucosal involvement, knee and elbow involvement, scalp involvement, melanin incontinence, exocytosis, focal hypergranulosis, follicular horn plug}, {clubbing of the rete ridges, elongation of the rete ridges, thinning of the suprapapillary epidermis, vacuolisation and damage of basal layer, saw-tooth appearance of retes, perifollicular parakeratosis, band-like infiltrate}
- Boston Housing Data: {INDUS, RM, AGE, RAD}, {NOX, TAX, MEDV}
- Faults in a urban waste water treatment plant (Water-treatment): {PHE, DBO-E, SS-E, SSV-E, SED-E, COND-E, DBO-D, SSV-D, DBO-S, RD-DBO-S, RD-DQO-S}, {PH-P, DBO-P, SS-P, SSV-P, SED-P, CONDP, PH-D, DQO-D, COND-D, SS-S, SED-S, COND-S, RD-DBO-G, RDDQO-G}
- Wisconsin Diagnostic Breast Cancer (WDBC): {V2-V4, V6-V8, V10, V12, V13, V18, V20, V26, V29, V32}, {V5, V9, V15-V16, V19, V22-V25, V27-V28, V30}
- 1995 - Current Population Survey (Census): {AFNLWGT, EMCONTRB, PTOTVAL, TAXINC, POTHVAL, PEARNVAL, WSALVAL}, {AGI, FEDTAX, STATETAX, INTVAL, FICA, ERNVAL}

The number of records in each file is not homogeneous. For example, the Iris Plants Database contains only 150 records while the Abalone database contains 4177 records.

4.1 Defining the Re-identification Problem

For applying the supervised approach, we have splitted the files into two parts. One for learning and the other for testing. Testing has been done using for all cases, 100 records. Such records have been taken at random from the original set. Then, the remaining records have been used for learning. For example, in the case of the Iris Plants Database we have used 50 records for learning an 100 records for experiments, and in the case of the Census data (that contains 1080 records), we use 100 for test and 980 for learning.

In order that the range of the variables do not affect the performance of record linkage, we have included a pre-processing step consisting on the normalization of the raw data. We have considered both ranging (translation of the values into the $[0, 1]$ range) and standardization (values are normalized using the mean and the standard deviation).

4.2 The Neural Network

As it is known, a neural network has several parameters: number of hidden layers, number of hidden neurons, learning coefficient, activation function, initial values and number of iterations over the training set.

In our experiments we have fixed all of them as constants except for the number of hidden neurons. For this latter parameter, we have considered different values between 5 and 50.

The fix parameters are the next:

Initial values: The initial weights of the network are calculated using Nguyen-Widrow method [7].

Number of iterations over the training set: 100 iterations.

Learning coefficient: It is constant in the whole learning process and equal to 0.3.

Number of hidden layers: All neural networks have one single hidden layer.

Activation function: It is the sigmoid function ($f(x) = 1/(1+e^{-x})$) in all the experiments.

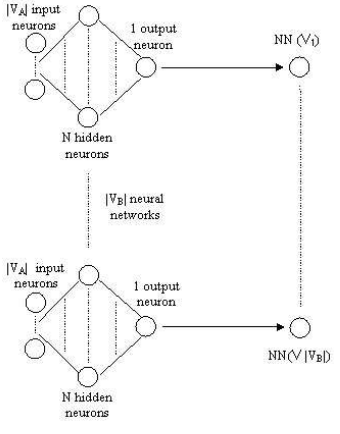


Fig. 1. Graphical scheme of the neural networks array

As described in Section 3, we have one neural network for each output variable. Each one has $|V_A|$ input variables (the number of variables in the files A and A^{sl}) and one output variable. Due to this, we have not one single neural network in our experiments, but an array of them. The length of this array equals the number of variables in files B or B^{sl}.

4.3 Results

The experiments performed show that the supervised approach to record-linkage for non-common variables lead to good results. Figures 2 and 3 represent the rate of success (the percentage of hits) obtained for each problem and for each configuration of hidden neurons. As the outcome of each experiment depends on the random selection of some records, and to the randomness present in the initialization of the neural network, we define the rate of success as the average of 5 different executions with the same problem and the same parameters.

Results are given also comparing probabilistic (figures on the right) and distance-based record linkage (left). Each chart in Figures 2 and 3 includes several lines, each one representing the success obtained for a problem for different sizes of the hidden layer. Two of the problems described in this section are not included as only a few records were re-identified.

These figures show that the best configurations are obtained when the normalization corresponds to ranging, and the record-linkage method selected is distance-based record linkage. This configuration leads to the best results in half of the tests: abalone with 14% hits using 45 hidden neurons, census with 47% hits using 35 hidden neurons, water-treatment with 40% hits using 10 hidden neurons and the wdbc problem with 80% hits using 15 hidden neurons.

The results obtained with the rest of configurations are not so good. Nevertheless, we obtain average results for the ionosphere problem with 12% hits using standardization and distance-based record linkage with 25 hidden neurons.

We can also observe that the larger is the number of hidden neurons, it is not true that the larger the success of the method. For example, the best result for

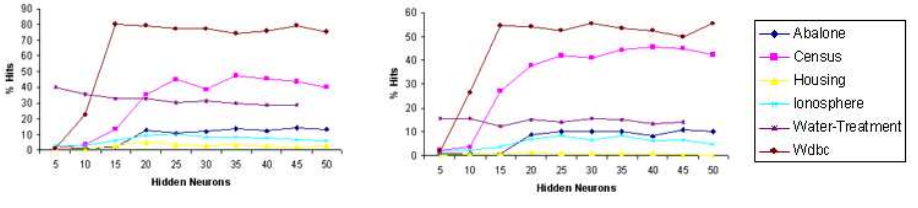


Fig. 2. Graphical results using ranging pre-processing, left results are using distance based record linkage and right results are using Probability record linkage

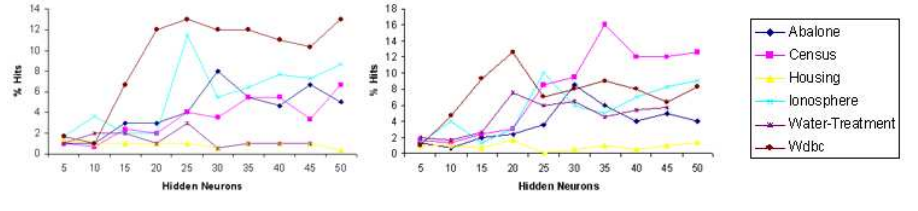


Fig. 3. Graphical results using standardization pre-processing, left results are using distance based record linkage and right results are using Probability record linkage

the wdbc problem corresponds to 80% of hits while this solution was achieved using 15 hidden neurons. This is probably due to an overfitting problem, when the number of hidden neurons is too large. According to the figures, it seems that an appropriate number of neurons is around 15.

Our method does not lead to good results for two different problems, [illegible] and [illegible] problem. From our point of view, there are different causes to these bad results. One of them corresponds to the number of records as in both cases the number of records are not much (150 records for the [illegible] problem and 358 for the [illegible] problem. Another cause seems to be a low number of attributes, the [illegible] problem has only two attributes. Finally, it can be observed that the [illegible] problem has a lot of missing values, that can distort the results.

Up to now, the study of record linkage for files with non-common variables was focused on unsupervised approaches. Supervised methods have been used in the case of common variables. In this setting, neural networks were used for schema matching. This approach is not applicable here, as we assume that the variables used to describe records in the two files are different. Thus, it is not possible to find a one-to-one correspondence between them.

Then, when comparing the performance of our approach with respect to unsupervised methods, we naturally obtain better performance. This is the case, for example, when we compare the present approach with the ones based on OWA operators [8] or fuzzy integrals [9]. In such situation, we observe that the rate of success (number of re-identifications) of the new approach is larger than in previous research.

For illustration, we can underline that in the wdbc problem we obtain now 80% of success over 26% with OWA operator, or 11% hits with the Twofold integral; in the census problem we get 47% over 11% with OWA operator and 8% with the Twofold integral; and in the water-treatment problem we get 40% over 17% with OWA operator and 9% with the Twofold integral.

4.4 Alternative Experiments

In the previous sections we have described the results when the splitting between the files for learning and test was done considering a constant number of records for the testing phase. Alternatively, we have also considered the same problem but splitting the files having 80% of the records for learning and the remaining 20% for testing.

For some of the files the results do not change significantly. This is the case of [illegible] and [illegible]. The last one changes from 80% to 78.5% while the number of hits goes from 80 (over 100 records in the testing set) to 88.7 (over 113 records).

Some of the results are rather different. Significant differences are obtained for [illegible] and [illegible]. Nevertheless they are caused by the fact that the number of records in the training set has increased while the size of the testing set has decreased. So, an increase of percentage is not of any interest here.

It is worth to mention the differences on the results obtained for the [illegible] and [illegible]. While in relation to the proportion of hits, the [illegible] has 14%

of hits in the first approach, now, in the second one the proportion is only 1.9%. Nevertheless, while the number of records is 100 in the first approach, this number was much increased in the second approach to 836. So, in absolute value, the number of records re-identified is slightly larger in the second approach (16 over 14 hits). Similarly, in the case of the , the proportion of hits in the first approach is 45% and in the second approach is only 28.3%. Nevertheless, in the first approach the number of hits is 45 and in the second one is 61. This increase was obtained when the number of records considered moved from 100 to 216. Note that this increment corresponds to a decrement of the number of records considered in the learning process. Therefore, the results are still rather significant for both files.

Now we summarize the best results obtained in the second set of experiments, giving details on the record-linkage method used as well as the normalization method (ranging and standardization) for the best results. Here PRL stands for probabilistic record linkage, DBRL stands for distance based record linkage .

- : 15.6 hits or 1.9% (using ranging and DBRL, with 40 hidden neurons)
- : 61.2 hits or 28.3% (using ranging and DBRL, with 25 hidden neurons)
- : 6.8 hits or 9.7% (using ranging and DBRL, with 30 hidden neurons)
- : 5 hits or 5% (using ranging and DBRL, with 15 hidden neurons)
- : 7 hits or 10% (using standardization and PRL, with 20 hidden neurons)
- : 2.6 hits or 8.7% (using standardization and PRL, with 35 hidden neurons)
- : 37 hits or 48.7% (using ranging and DBRL, with 25 hidden neurons)
- : 88.7 hits or 78.5% (using ranging and DBRL, with 25 hidden neurons)

5 Conclusions and Future Work

In this paper, we have studied an alternative method for record linkage based on neural networks when information is available for using supervised machine learning techniques. The method is particularly suited for numerical information.

We have proved that the use of neural networks to model the relationships between the variables is a suitable tool for such re-identification as they have lead to good results in eight different problems.

In this paper, we have compared this new approach with others, based on structural information. We show that when supervised approaches can be applied, this new approach is better than the previous existing ones. Nevertheless, the application of this method requires a training record set.

As future work, we consider the realization of more experiments, and the study of alternative types of neural networks. Finally, we consider the possibility of extending this approach in the case that no learning data is available.

Acknowledgements

This work was partly funded by the Spanish Ministry of Education and Science under project SEG2004-04352-C04-02 "PROPRIETAS".

References

1. Agrawal, R., Srikant, R., (2000), Privacy Preserving Data Mining, Proc. of the ACM SIGMOD Conference on Management of Data, 439-450.
2. Data Extraction System, U.S. Census Bureau, http://www.census.gov/DES/www/welcome.html
3. Freeman, J.A., Skapura, D.M., (1991), Neural Networks. Algorithms, Applications and Programming Techniques, Addison-Wesley.
4. Li, W., Clifton C., (2000), SEMINT: A tool for identifying correspondences in heterogeneus databases using neural networks, Data & Knowledge Engineering, 33 49-84.
5. Narukawa, Y., Torra, V., (2004), Twofold integral and Multi-step Choquet integral, Kybernetika, 40:1 39-50.
6. Narukawa, Y., Torra, V.,(2005), Graphical interpretation of the twofold integral and its generalization, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 13, No. 4 415-424.
7. Nguyen, D., Widrow, B., (1990), Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights, Proc. of the Int'l Joint Conference on Neural Networks, 3:21.26.
8. Nin, J., Torra, V., (2005), Towards the use of OWA operators for record linkage, Proc. of the European Soc. on Fuzzy Logic and Technologies, in press.
9. Nin, J., Torra, V., (2005), Empirical analysis of database privacy using twofold integrals, Lecture Notes in Artificial Intelligence, 3801 1–8.
10. Rojas, R., (1996), Neural Networks - A Systematic Introduction, Springer
11. Torra, V., Domingo-Ferrer, J., (2003), Record linkage methods for multidatabase data mining, in V. Torra (Ed.), Information Fusion in Data Mining, Springer, 101-132.
12. Torra, V., (2000), Towards the re-identification of individuals in data files with non-common variables, Proc. of the 14th European Conference on Artificial Intelligence (ECAI2000) (IOS Press, ISBN 1 58603 013 2), 326-330, Berlin, Germany.
13. Torra V., OWA operators in data modeling and re-identification, IEEE Trans. on Fuzzy Systems, 12:5 (2004) 652-660.
14. Murphy, P., M., Aha, D. W., (1994), UCI Repository machine learning databases, http://www.ics.uci.edu/ mlearn/MLRepository.html, Irvine, CA: University of California, Department of Information and Computer Science.
15. Willenborg, L., de Waal, T., (2001), *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, Springer-Verlag.
16. Winkler, W. E., (2003), Data Cleaning Methods, Proc. SIGKDD 2003, Washington.
17. Winkler, W. E., (2004), Re-identification methods for masked microdata, Privacy in Statistical Databases 2004, Lecture Notes in Computer Science 3050 216-230.
18. Yager, R. R., (1988), On ordered weighted averaging aggregation operators in multi-criteria decision making, IEEE Trans. Syst., Man, Cybern., 18 183-190.

Bayesian Correction for SNP Ascertainment Bias

María M. Abad-Grau[1] and Paola Sebastiani[2]

[1] Software Engineering Department, University of Granada, Granada 18071, Spain
mabad@ugr.es
[2] Department of Biostatistics, Boston University, Boston MA 02118, USA
sebas@bu.edu

Abstract. Genomewide analysis of linkage disequilibrium (LD) is commonly based in the maximum likelihood estimator. This estimator of LD suffers of a well known bias toward disequilibrium that becomes particularly serious in small samples with SNPs that are not very common in the population. Algorithms able to identify LD patterns, such as haplotype blocks or LD decay maps do a non-random selection of SNPs to be included in the analysis in order to remove this bias. However, they introduce ascertainment bias that can mask the real decay of disequilibrium in the population, with several consequences on the validity and reproducibility of genetic studies. In this work, we use a new Bayesian estimator of LD that greatly reduces the effect of ascertainment bias in the inference of LD decay. We also provide a software that use the Bayesian estimator to compute pairwise LD from SNP samples.

1 Introduction

Variations in the DNA structure can help to determine genetic susceptibility to disease, as well as to reconstruct the history of human populations, and natural selection [1]. Single nucleotide polymorphism (SNPs) — variations in human DNA that occur in at least 1% of the population — have become nowadays an invaluable resource to map the human genome. Each of the two possible variants of an SNP on a chromosome pair, called here alleles, acts as a flag on the chromosomal region. Evidence that consecutive alleles on the same chromosome are inherited together in the same [illegible] can help to discover DNA regions associated with disease, and eventually to identify functional loci that, by themselves or in groups, modulate disease risk. A key to this process is [illegible] (LD), defined as association between SNP variants that results in alleles at two distinct loci being inherited together.

Several pairwise measures of LD to capture the strength of association between alleles of SNPs have been proposed [2], and are widely used because of their simplicity. Among them, the measure D and its normalized version D' are the most popular [3], and are usually estimated in a sample by maximum likelihood. The main problem of the maximum likelihood estimates (MLE) of D and D' is that they exhibit a strong bias toward disequilibrium that becomes particularly

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 262–273, 2006.

serious in small samples with SNPs that are not very common in the population [4]. Several solutions have been proposed, including non-parametric approaches [4, 5] but none of these solutions appears to be able to remove the bias without imposing some "ad hoc" threshold on the minimum allele frequency (MAF) of those SNPs that can be used to infer the pattern of LD in the population. For example, the default setting of HaploView to identify haplotype blocks by using confidence intervals imposes a lower bound of 0.05 on the MAF [6, 7].

However, imposing this threshold leads to a non-random selection of SNPs and introduces a different type of bias called [illegible]. The ascertainment bias can be introduced when SNPs identified in small samples —called [illegible]— are typed in larger samples of chromosomes. In this case, SNPs with small MAF are not typed in the larger samples, thus potentially biasing the pattern of LD. Ascertainment bias is also introduced when SNPs are identified and typed by direct sequencing but the information about invariant sites has been lost [8, 9], or when not all the variant loci are chosen but only a non-random selection of them for economic reasons. As an example, SNPs typed in the International Hapmap Project[10] have an average density of one marker per 5 kb, and the SNP selection was previously done by using smaller samples. In this situation, an ascertainment bias has been shown in [11, 12].

We have recently introduced a Bayesian measure of LD that finds a better approximation of the real magnitude of LD in the population, and maintains the same linear computational complexity of the maximum likelihood estimator of D' [13]. The main feature of our Bayesian estimator is to use a prior distribution on pairwise associations between different SNPs that is a decreasing function of their physical distance. The effect of the prior distribution is to drastically reduce the bias toward disequilibrium even in small samples, without imposing any threshold on the MAF. In this paper we explore the efficacy of our Bayesian solution to determine robust maps of LD decay, and show the lack of ascertainment bias compared to LD decay maps based on the traditional estimators of LD.

The remaining of this paper is as follow. We describe briefly the common pairwise measures of LD and our Bayesian solution in Section 2. In Section 3, we examine the effect of SNPs selection on real data taken from the Hapmap Project[10]. Conclusions and future work are in Section 4.

2 Measures of Linkage Disequilibrium

Given two loci L_1 and L_2, with allele variants A/a and B/b, we define the probability of transmission of the haplotype pair $A = i, B = j$ by $p_{ij} = p(L_1 = i, L_2 = j)$, $i = A, a$, $j = B, b$. These probabilities can be displayed into a 2×2 table summarizing the transmission rates of the four possible haplotype pairs:

Locus 1	Locus 2 B	Locus 2 b	
A	p_{AB}	p_{Ab}	p_A
a	p_{aB}	p_{ab}	p_a
	p_B	p_b	1

The margins of the table are the allele frequencies, and we assume A and B denote the major alleles. Without loss of generality, we can also assume that $p_A \geq p_B$.

The two loci are in [illegible] when all the allele variants are transmitted independently so that $p_{ij} = p_i p_j$ for all $i = A, a$, $j = B, b$. On the other hand, LD implies some form of stochastic dependency between the alleles and hence some form of departure from independence of the probabilities p_{ij}. A widely used measure of LD is the parameter D defined by $D = p_{AB} - p_A p_B$.

Because the magnitude of D is a function of the allele frequencies thus complicating its interpretation, several authors have proposed different forms of normalization [2]. The most common one is the measure D' in $[0, 1]$ that was suggested by Lewontin [14] and is defined as $|D|/\max D$:

$$D' = \begin{cases} -\dfrac{p_{AB} - p_A p_B}{p_a p_b} & \text{if } D \leq 0 \\ \dfrac{p_{AB} - p_A p_B}{p_a p_B} & \text{if } D > 0 \end{cases} \tag{1}$$

with $D' = 0$ in linkage equilibrium, $0 < D' < 1$ measuring increasing magnitude of disequilibrium, and $D' = 1$ in perfect disequilibrium, when there is perfect dependency between alleles.

Suppose now we have a data set of N individuals, that correspond to $n = 2N$ known haplotype pairs for the two loci (we assume here known phase for all haplotype pairs):

Locus 1	Locus 2: B	Locus 2: b	
A	n_{AB}	n_{Ab}	n_A
a	n_{aB}	n_{ab}	n_a
	n_B	n_b	n

Then it is well known that the maximum likelihood estimate (MLE) of D is

$$\hat{D} = \hat{p}(AB) - \hat{p}(A)\hat{p}(B)$$

where $\hat{p}_{AB} = n_{AB}/n$, $\hat{p}_A = n_A/n$, and $\hat{p}_B = n_B/n$ are the MLE of the parameters p_{AB}, p_A and p_B. The MLE of D' is then obtained by replacing the parameter values by their MLEs in (1). A critical feature of $\hat{D}'$ is that it achieves the maximum value 1 whenever $n_{ab} = 0$ or $n_{aB} = 0$, even if the frequencies of haplotype pairs are consistent with a situation of perfect equilibrium. This bias has been shown to be particularly severe in small samples [4].

To reduce the bias without imposing thresholds on the MAF, in [13] we introduced a Bayesian estimator of D' that is defined by:

$$\tilde{D}'|\tilde{p}_{AB}, \tilde{p}_A, \tilde{p}_B = \frac{\tilde{D}}{\tilde{p}_a \tilde{p}_B} I(\tilde{D} \geq 0) - \frac{\tilde{D}}{\tilde{p}_a \tilde{p}_b} I(\tilde{D} < 0) \tag{2}$$

where $I(x \in X)$ is the Dirac function defined as $I(x \in X) = 1$ if $x \in X$ and 0 otherwise and

$$\begin{aligned}\tilde{D}|\tilde{p}_{AB}, \tilde{p}_A, \tilde{p}_B &= \tilde{p}(AB) - \tilde{p}(A)\tilde{p}(B) \\ \tilde{p}_{AB} &= (n_{AB} + \alpha_{AB})/(n + \alpha_T) \\ \tilde{p}_A &= (n_A + \alpha_{AB} + \alpha_{Ab})/(n + \alpha_T) \\ \tilde{p}_B &= (n_B + \alpha_{AB} + \alpha_{aB})/(n + \alpha_T)\end{aligned}$$

for appropriate choices of the hyper-parameters α_{ij}, $i = A, a; j = B, b$ and $\alpha_T = \sum_{ij} \alpha_{ij}$. In the original manuscript [13], we discuss several choices for the hyper-parameters that encode prior knowledge about LD as a function of the physical distance d between the two SNPs. In this paper, we focus on uniform priors defined as:

$$\alpha_{ij} = \alpha(1 - \exp(-d))$$

with $\alpha > 0$ and d representing the physical distance between the two loci in 10^5 bases, or 100kb. This prior distribution imposes a correction on the transmission rates of haplotype pairs that approach the uniform distribution for loci at very large distance: essentially for loci that are far apart by more than 500kb the prior distribution is uniform with hyper-parameters $\alpha_{ij} = \alpha$ and prior transmission rates $p_{ij} = p_i p_j$. This correction is consistent for example with default settings of HaploView [6], in which pairs of loci that are at a distance larger than 500kb are not include in the LD analysis. It can be shown that the estimator in Equation (2) is the maximum a posteriori (MAP) estimator of D', conditional on MAP estimates of p_{AB}, p_A and p_B.

An implementation of this method, able to compute $\tilde{D}$ and $\tilde{D}'$ from SNPs genotype data sets, can be downloaded from http:// bios.ugr.es / DB3. If family trio genotypes instead of unrelated individuals are provided, the software infers ambiguous haploypes from families using a modification of the method in GeneHunter[15] or HaploView [16], which is based in the transmission disequilibrium test (TDT) [17]. Following the TDT and assuming no recombination in the transmission of haplotypes from parents to offspring, haplotypes with more than 2 loci cannot be completely reconstructed if the other 2 members of the trio are heterozygous[18]. However, when only bilocus haplotype counters are required, there is a special condition without any ambiguity for the problem of counting biloci haplotypes, which holds in a trio when both parents are heterozygous at both positions and the offspring is heterozygous at one position, AB/Ab for example.

The software utilizes an iterative proportional algorithm in order to estimate bi-locus haplotypes when the phase, even using pedigree data, is unknown. Although there is a maximum number of iterations for this algorithm to end, it usually reach a stable maximum in a very little number of iterations. Once haplotype frequencies are estimated by using the Bayesian approach, $\tilde{D}$ and $\tilde{D}'$ can be obtained. Computational complexity for the algorithm to obtain $\tilde{D}$ and $\tilde{D}'$ for a pair of loci is $O(n)$, with n being the number of individuals in the sample. It is straightforward to show that it has the same computational complexity of the MLE.

3 Evaluation

In the introduction, we identified three sources of ascertainment bias: the inclusion of invariant sites, imposing an "ad hoc" threshold on the MAF, and non-random selection of SNPs to be typed. In this section, we describe data sets that we generated to artificially reproduce these situations, and the effect on the MLE and MAP estimators.

3.1 Evaluation Measures

In order to observe the effect of the ascertainment bias on the ML and MAP estimators of D', we create LD decay maps in different situations that are described below. The feature of LD decay maps is that they can easily highlight variations of the ascertainment bias as a function of the physical distance between SNPs. Each point is the average estimation of D' for all SNPs within a physical distance $d \pm 10$ kb. Overlapping distances were used. Therefore, an average value is plotted for each 10 kb. The maximum distance used was 500 kb, due to the fact that no additional LD decay can be observed for larger distances [19, 20].

The original data are genotypes of 19120 SNPs uniformly distributed over chromosome 22 from the 90 families of the CEPH population in the first phase of the HapMap Project [10]. The data are publicly available from http://hapmap.org. We have chosen this chromosome because it is the shortest one and then it has been widely studied [21, 22, 23], and this population because the sample contains genotypes from [illegible] (both parents and one child) instead of unrelated individuals, so that most of the transmitted haplotype pairs can be reconstructed using families. Once haplotype are reconstructed using familial information, only the data from the 60 unrelated parents have been used.

We then created artificial samples from this data in order to assess the effect of the ascertainment bias in the ML and MAP estimator under three different scenarios.

[illegible] In order to check the effect of this bias on the estimators, we computed LD decay maps for the 60 unrelated individuals using (i) all the SNPs in these samples (hence all the SNPs for which we observed some allelic variation); (ii) the same SNPs augmented by those loci without allelic variation. Therefore while in the former data we include only those SNPs that have variability in the sample, in the second we are also using invariant sites.

[illegible] A common cause of ascertainment bias is due to the use of a small sample, panel, to identify the SNPs[11] that are then typed in a larger sample. Sometimes, several panels are used, as in the HapMap Project [10]. The effect of this pre-selection is that rare alleles are disregarded more frequently than less rare ones. We have randomly selected subsamples of 5, 20 and 25 family trios and then used SNPs with allelic variations in those small samples to create the LD decay maps. To emphasize that the bias is not due to lower SNPs density but to the non-random selection, we also

created three sub-samples from the original data with 1/2, 1/3 and 1/4 of the total number of SNPs. Although SNP density influences gene-scale patterns of linkage disequilibrium and definition of blocks [24] so that the eventual utility of the maps with sparse markers has been questioned [25], this is not a case of ascertainment bias and no bias should be observed.

The bias effect of using a panel sample should be equivalent to the non random selection of SNPs based on the use of a threshold on the MAF. To support this conjecture, we generated three sub-samples from the original data, with SNP selection determined by these thresholds for the MAF: 0.05, 0.1 and 0.2. There is a well-known upward bias for the MLE of D' (see for example [4]).

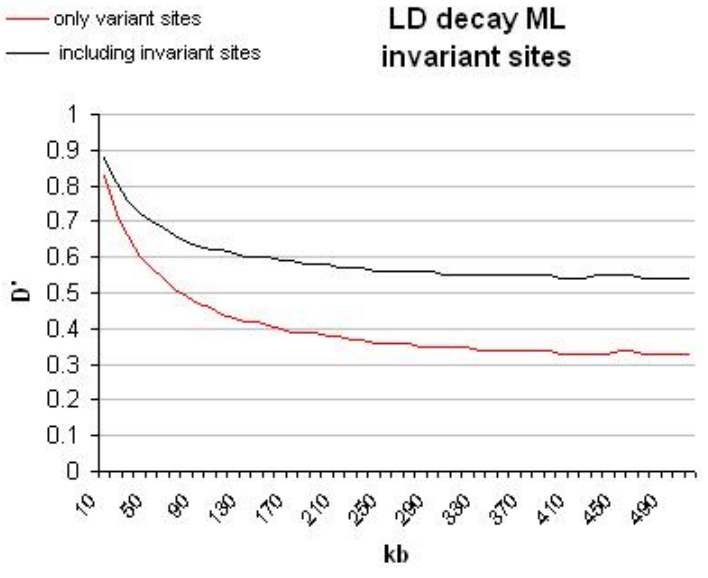
3.2 Results

Figures 1 to 4 show the LD decay maps describing the effect of different patterns of SNP selection. In each figure, the x-axis reports the distance between pairs of SNPs. For each x value, the y-axis reports the estimate of the average D' for all SNPs within a distance $x \pm 10$kb. We consider bias an average difference between LD decay maps generated in different conditions of 0.05 or more. Results for each experiment are described below.

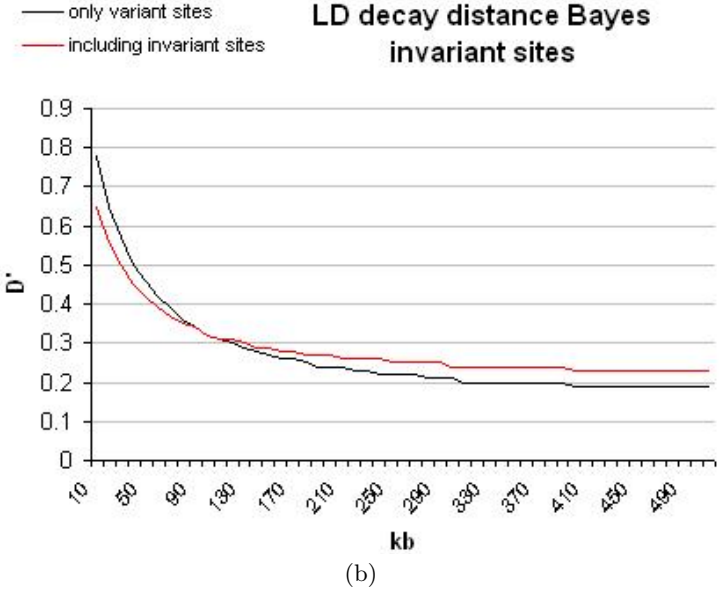
The inclusion of invariant sites has a high impact in the reconstructed pattern of LD decay (see Figure 1). If there is not variation in a locus for a given sample, the MLE of D' is 1, while the MAP estimator is smaller than 1. Both estimators are biased although the bias of the MAP estimator is much smaller. As expected, results are very different depending on the estimator. While the upward bias of the MLE increases with the distance (see Figure1(a)), the Bayesian estimator shows a smaller bias but in different ways depending on distance (see Figure 1(b)). We conjecture that invariant sites should not be used because there is a complete uncertainty in this information.

Consider first the effect of SNPs identification in a small panel. In agreement with results published in [11], there is a downward bias for the MLE of D', see Figure 2 (a), whereas the MAP estimator does not shown any significant bias, as can be observed in plot (b) of the same Figure. On the other hand, the results displayed in Figure 3 show no bias in either the MLE or the MAP estimator, and therefore suggest that the bias is due to the non-random selection of SNPs induced by the identification of SNPs in small samples.

Our results confirm the already reported upward bias of the MLE of D' (Figure 4(a)). On the other hand, the MAP estimator is able to correct the bias for not very large distances –see Figure 4(b)–. Even for larger distances, the bias is significantly smaller than the one of the MLE of D'.

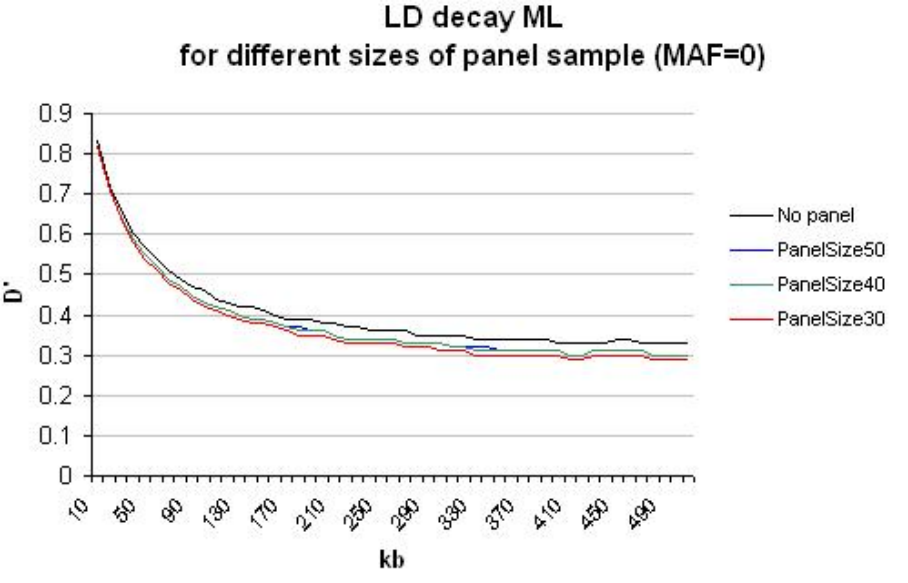


(a)

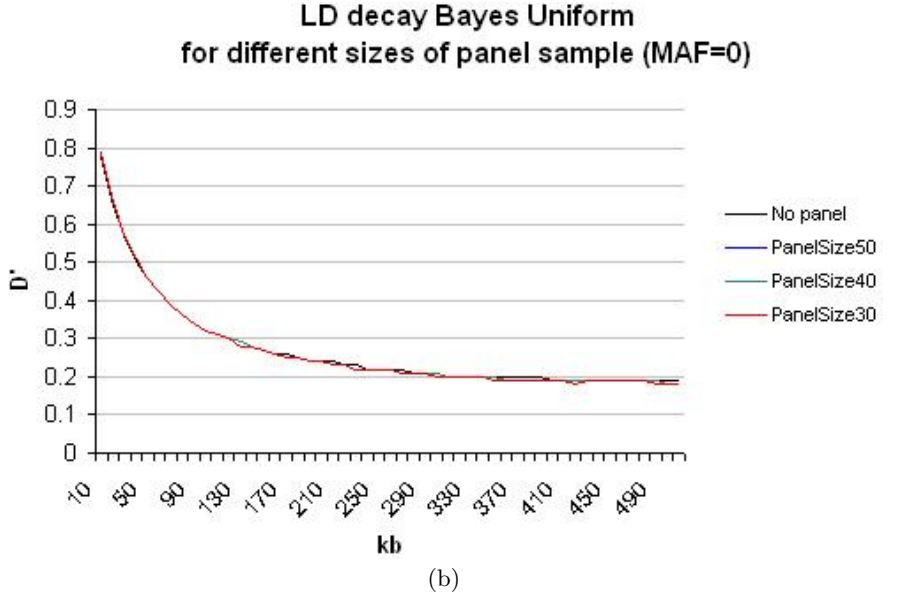


(b)

Fig. 1. *Ascertainment bias due to invariant sites.* LD decay maps for MLE(a) and Bayesian estimator(b) of D'. Each plot shows the LD decay in two conditions: (1) using only those those SNPs with allelic variations are retained in the analysis; (2) all SNPs in the sample are retained in the analysis.

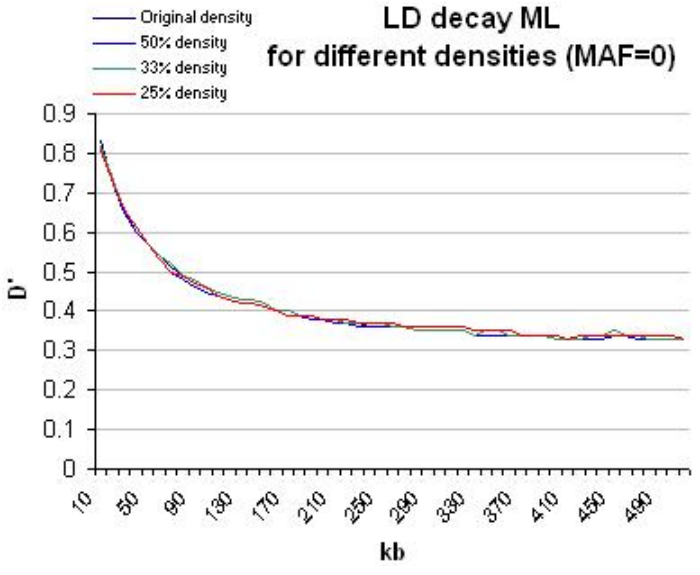


(a)

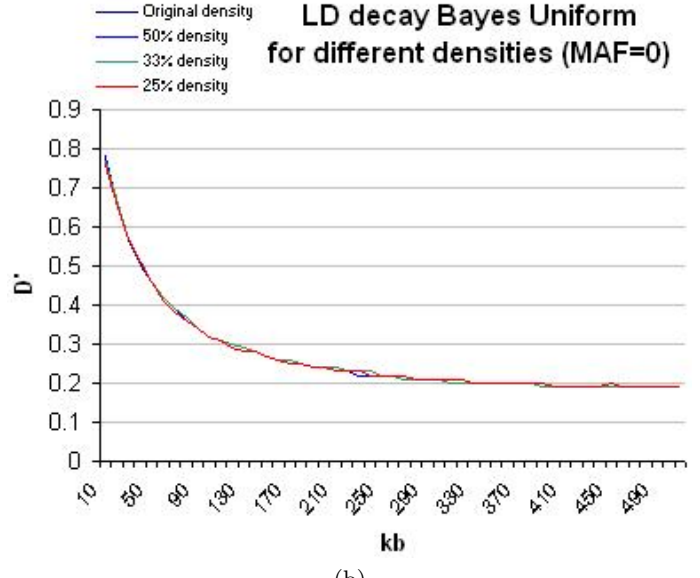


(b)

Fig. 2. *Ascertainment bias due to SNPs density: SNPs identification in small panels.* LD decay maps for MLE (a) and MAP estimator (b) of D'. Each plot shows the LD decay for the sample with all the SNPs for the 60 unrelated individuals and 3 more maps in which panel samples of 50, 40 and 30 individuals are used to select the SNPs.

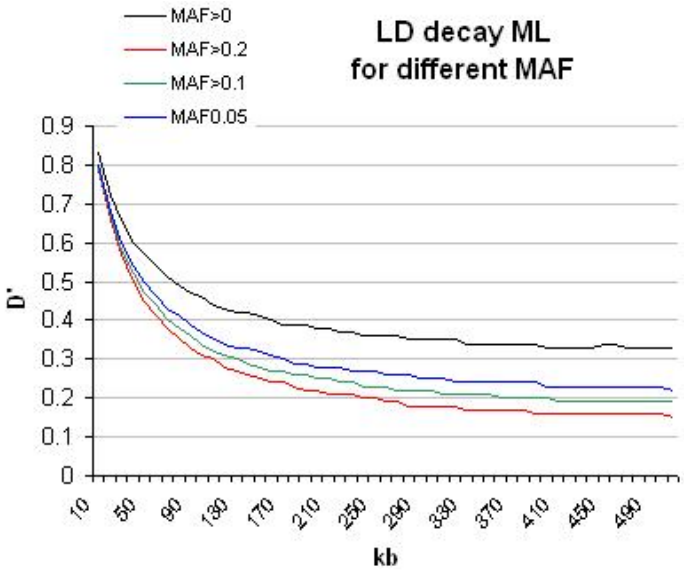


(a)

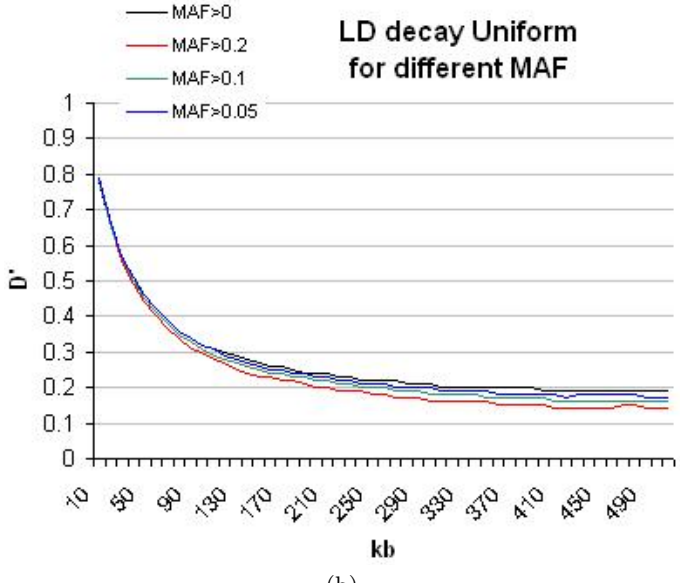


(b)

Fig. 3. *Ascertainment bias due to SNPs density: Random selection of SNPs.* LD decay maps for MLE (a) and MAP estimator (b) of D'. Each plot shows the LD decay map for the sample with all the SNPs for the 60 unrelated individuals and 3 more maps in which a random selection of SNP has been generated, with 50%, 33% and 25% of the original SNPs.



(a)



(b)

Fig. 4. *Ascertainment bias due to MAF thresholding.* LD decay maps for MLE (a) and MAP estimator (b) of D'. Each plot shows the LD decay map for the sample with all the SNPs for the 60 unrelated individuals and 3 more maps in which only SNPs with MAF greater than 5%, 10% and 20% are used.

4 Conclusions

Although preliminary and limited to only one chromosome, our results suggest an overall superiority of the use of our MAP estimator to infer pattern of LD decay. The advantages of the Bayesian estimator is to avoid the use of thresholds on the MAF so that all the SNPs with observed allelic variability can be used to infer patterns of LD decay in the human population. However, there are other sources of bias that are not controllable after the data collections (such as non-random selection of SNPs to be genotyped). A reduction in the number of SNPs can drastically influence gene-scale patterns of linkage disequilibrium and definition of blocks based on MLE of D'[24]. However, our results show that the Bayesian estimator of D' produces patterns of LD decays that are essentially invariant to SNP density or thresholds imposed on the MAF. We plan to extend this preliminary study to all chromosomes from the CEPH and Yoruba populations in the Hapmap project.

Acknowledgements

Research supported by the US federal grant R21 HL080463 (NIH/NHLBI) and by the Spanish Research Program under project TIN2004-06204-C03-02.

References

1. Abecasis, G.R., Ghosh, D., Nichols, T.E.: Linkage Disequilibrium: Ancient history drives the new genetics. Hum Hered **59** (2005) 118–124
2. Hedrick, P.W.: Gametic disequilibrium measures: proceed with caution. Genetics **117** (1987) 331–341
3. Lewontin, R.C.: On measures of gametic disequilibrium. Genetics **120** (1988) 849–852
4. Teare, M.D., Dunning, A.M., Durocher, F., Rennart, G., Easton, D.F.: Sampling distribution of summary linkage disequilibrium measures. Ann Hum Genet **66** (2002) 223–233
5. Devlin, B., Roeder, K., Tiobech, C.S., Byerley, W.: Genome-wide distribution of linkage disequilibrium in the population of Palau and its implications for gene flow in remote Oceania. Nat Genet **29** (2001) 229–232
6. Barrett, J.C., Maller, J., Daly, M.J.: Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics **21** (2005) 263–265
7. Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M., Altshuler, D.: The structure of haplotype blocks in the human genome. Science **296** (2002)
8. Nielsen, R.: Estimation of population parameters and recombination rates using single nucleotide polymorphisms. Genetics **154** (2000) 931–942
9. Kuhner, M.K., Beerli, P., Yamato, J., Felsenstein, J.: Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics **156** (2000) 1393–1401

10. HapMap-Consortium, T.I.: The international hapmap project. Nature **426** (2003) 789–796
11. Nielsen, R., Signorovitch, J.: Correcting for ascertainment biases when analyzing snp data: applications to the estimation of linkage disequilibrium. Theor Popul Biol **63** (2003) 245–255
12. Ronald, J., Akey, J.M.: Genome-wide scans for loci under selection in humans. Hum Genom **2** (2005) 113–125
13. Abad, M.M., Sebastiani, P.: A Bayesian measure of linkage disequilibrium, submitted. Technical report, Boston University School of Public Health, 715 Albany Street, Boston MA 02118 (2005)
14. Lewontin, R.C.: The interaction of selection and linkage. 1: general considerations: heterotic models. Genetics **49** (1964) 49–67
15. Daly, M.J., Kruglyak, L., Pratt, S., Houstin, N., Reeve, M.P., Kirby, K., Lander, E.S.: Genehunter 2.0–a complete linkage aalysis system. American Journal of Human Genetics **63** (1998) A286
16. Barret, J.C., Fry, B., Maller, J., Daly, M.J.: Haploview: analysis and visualization of ld and haplotype maps. Bioinformatics **21** (2005) 263–265
17. Clayton, D.: A generalization of the Transmision/Disequilibrium Test for uncertain-haplotype transmission. American Journal of Human Genetics **65** (1999) 1170–1177
18. Sebastiani, P., Grau, M.M.A., Alpargu, G., Ramoni, M.F.: Robust Transmission Disequilibrium Test for incomplete family genotypes. Genetics **168** (2004) 2329–2337
19. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., Lander, E.S.: High-resolution haplotype structure in the human genome. Nat Genet **29** (2001) 229–232
20. Dunning, A., Durocher, F., Healey, C., Teare, M.D., McBride, S.E., Carlomagno, F., Xu, C.F., Dawson, E., Rhodes, S., Ueda, S., Lai, E., Luben, R.N., Rensburg, E.J.V., Mannermaa, A., Kataja, V., Rennart, G., Dunham, I., Purvis, I., Easton, D., Ponder, B.A.: The extent of linkage disequilibrium in four populations with distinct demographic histories. Am J Hum Genet **67** (2000) 1544–54
21. ad N. Shimizu, I.D., Roe, B.A., et. al., S.C.: The DNA sequence of human chromosome 22. Nature **402** (1999) 489–495
22. Mullikin, J.C., Hunt, .E., Cole, C.G., Mortimore, B.J., Rice, C.M., Burton, J., Matthews, L.H., Pavitt, R., plumb et. al, R.W.: An snp map of human chromosome 22. Nature **407** (2000) 516–520
23. Dawson, E., Abecasis, G., Bumpstead, S., Chen, Y., Hunt, S., Beare, D.M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganskek, R., Lohmussaar, E., Zernant, J., Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D.R., Cardon, L.R., Dunham, I.: A first-generation linkage disequilibrium map of human chromosome 22. Nature **418** (2002) 544–548
24. Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J.: The impact of snp density on fine-scale patterns of linkage disequilibrium. Human Molecular Genetics **13** (2004) 577–578
25. Weiss, K.M., Terwilliger, J.D.: How many diseases doe it take to map a gene with snps. NatureGenetics **26** (2000) 151–157

An Application of Support Vector Machine to Companies' Financial Distress Prediction*

Xiao-Feng Hui and Jie Sun

School of Management, Harbin Institute of Technology, Harbin 15 00 01,
HeiLongJiang Province, China
xfhui@hit.edu.cn
sjhit@sina.com

Abstract. Because of the importance of companies' financial distress prediction, this paper applies support vector machine (SVM) to the early-warning of financial distress. Taking listed companies' three-year data before special treatment (ST) as sample data, adopting cross-validation and grid-search technique to find SVM model's good parameters, an empirical study is carried out. By comparing the experiment result of SVM with Fisher discriminant analysis, Logistic regression and back propagation neural networks (BP-NNs), it is concluded that financial distress early-warning model based on SVM obtains a better balance among fitting ability, generalization ability and model stability than the other models.

1 Introduction

Bankruptcy of enterprise not only makes stockholders, creditors, managers, employees, and other interest parts suffer individual economic loss, but also if many enterprises run into bankruptcy the economic development of the whole country will be greatly shocked. In general, most enterprises that ran into bankruptcy had experienced a condition of financial distress, but they could not detect financial distress at an early stage and timely take effective measures to prevent bankruptcy. Therefore, in the perspective of management, it is important to explore a more effective financial distress prediction model to signal early-warning for enterprises which will possibly get into financial distress, so that managers can take strategic actions to avoid deterioration of financial state and bankruptcy. Besides, from the view of financial institution, an effective financial distress prediction model can help them detect customers with high default risk at an early stage so as to improve their efficiency of commercial credit assignment [1].

Financially distressed enterprises usually have some characteristic symptoms, which generally can be indicated by the data in financial statements and the financial ratios derived from financial statements. With the development of all kinds of classification and prediction techniques, from univariate analysis to multi discriminate analysis (MDA), from statistical methods to machine learning methods, research literatures on enterprises' financial distress prediction become more and more abundant.

* Sponsored by National Natural Science Foundation of China (No. 70573030).

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 274 – 282, 2006.

Beaver (1966), one of the first researchers to study bankruptcy prediction, investigated the predictability of the 14 financial ratios using 158 samples consisting of failed and non-failed firms [2]. Beaver's study was followed by Altman's model (1968) based on the MDA to identify the companies into known categories. According to Altman, bankruptcy could be explained quite completely by using a combination of five (selected from an original list of 22) financial ratios [3]. Logit model is widely used to deal with two classes classification problems, and Ohlson was the first to apply it to predicting financial distress in 1980 [4].

The most widely used machine learning method in the field of financial distress prediction is neural networks (NNs), which has strong capability of identifying and representing nonlinear relationships in the data set. Odom and Sharda (1990) made an early attempt to use NNs for financial distress prediction. He used the same financial ratios as Altman's study and took MDA model as the benchmark [5]. From then on, many scholars (Fletcher and Goss, 1993; Carlos Serrano-Cinca, 1996; Parag C. P., 2005; etc.) were dedicated to compare NNs with MDA and logit, which brought a lot of positive support for the conclusion that NNs can predict financial distress more accurate than those benchmarks [6][7][8].

Generally, statistical methods have the advantages of simple model structure and easiness to understand and use, but they have restrictive assumptions such as linearity, normality and independence of input variables, which limits the effectiveness and validity of prediction. In contrary, NNs is not constrained by those assumptions and have strong ability of fitting nonlinear relationships between descriptive variables and conclusive variables. But NNs also has the disadvantages such as unfixed structure, over-fitting, needing a lot of samples, and black-box effect.

Support vector machine (SVM) is a relatively new machine learning technique, originally developed to resolve local minima and over-fitting problems which are the main sources of trouble to NNs [9], [10], [11]. Shin K.-S. (2005) and Min J. H. (2005) respectively made an attempt to use SVM to predict corporate bankruptcy with Korean data and got satisfying results [12], [13]. Other applications of SVM by Kim K. J. (2003) and Tay F. E. H. etc. (2001) also showed that it is a promising classification and prediction method [14], [15].

This paper attempts to apply SVM to predicting the financial distress of Chinese listed companies and compare the result of SVM with the results got by the methods of Fisher discriminant analysis, Logistic regression and NNs. The rest of the paper is divided into five sections. Section 2 is the brief description of SVM theory. Section 3 is about data collection and preprocessing. Section 4 gives the modeling process and experiment results. Section 5 discusses and analyzes the experiment results. Section 6 makes conclusion.

2 Theory of SVM

SVM, put forward by Vapnik in 1990's, is a relatively new machine learning technique, which is developed on the basis of statistical learning theory [9]. Former researches have shown that SVM has the following merits according to learning ability and generalization ability.

1) SVM is based on the principle of structural risk minimization, not on the principle of empirical risk minimization, so SVM can better avoid the problem of over-fitting.
2) SVM algorithm is uneasy to get into local optimization, because it is a convex optimization problem and its local optimal solution is just the global optimal solution.
3) In practice, when the number of samples is relatively small, SVM can often get better result than other classification and prediction techniques.

A simple description of the SVM algorithm is provided as follows [9], [10], [11], [12], [13]. Suppose $D=\{x_i, y_i\}_{i=1}^{N}$ is a training data set with input vectors $x_i=(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^{\mathrm{T}} \in \mathrm{R}^{\mathrm{n}}$ and target labels $y_i \in \{-1,+1\}$, in which N is the number of training samples. In the condition that the training samples are linearly separable, SVM algorithm is to find an optimal separating plane $w \bullet x+b=0$, which can not only separate training samples without error but also make the margin width between the two parallel bounding planes at the opposite side of the separating plane get a biggest value.

In the nonlinearly separable case, SVM firstly uses a nonlinear function $\Phi(x)$ to map input space to a high-dimensional feature space. Then a nonlinear optimal separating hyperplane $w \bullet \Phi(x)+b=0$ with the biggest margin width can be found by the same technique as linear model. Those data instances which are nearest to the separating hyperplane are called support vectors, and other data instances are irrelevant to the bounding hyperplanes. Because most problems are nonlinear separable and linearly separable case is the special situation of nonlinearly separable case, i.e. $\Phi(x)=x$, here only the SVM theory under nonlinearly separable case is stated. So according to Vapnik's original formulation, the SVM classifier should satisfy the following conditions

$$\begin{cases} w^{\mathrm{T}}\Phi(x_i)+b \geq +1 & if \quad y_i=+1 \\ w^{\mathrm{T}}\Phi(x_i)+b \leq -1 & if \quad y_i=-1 \end{cases} \tag{1}$$

which is equivalent to

$$y_i[w^{\mathrm{T}}\Phi(x_i)+b] \geq 1 \qquad (i=1,\cdots,N) \tag{2}$$

where w^{T} represents the weight vector and b is the bias.

Because the margin width between both bounding hyperplanes equals to $2/(\|w\|^2)$, the construction of optimal separating hyperplane with biggest margin width can be defined as the following optimization problem

$$\begin{aligned} &\min \quad \frac{1}{2}w^{\mathrm{T}}w + C\sum_{i=1}^{N}\xi_i \\ &\text{s.t.} \\ &\begin{cases} y_i[w^{\mathrm{T}}\Phi(x_i)+b] \geq 1-\xi_i & (i=1,2,\cdots,N) \\ \xi_i \geq 0 & (i=1,2,\cdots,N) \end{cases} \end{aligned} \tag{3}$$

in which ξ_i are slack variables. Feature space generally can not be linearly separated, if the separating hyperplane is constructed perfectly without even one training error, it is easy to appear over-fitting phenomenon, so slack variables are needed to allow a small part of misclassification. $C \in R^+$ is a tuning parameter, weighting the importance of classification errors with the margin width.

This problem is transformed into its dual problem because it is easier to interpret the results of the dual problem than those of the primal one.

$$\begin{aligned} &\max \quad \frac{1}{2}\partial^{\mathrm{T}} Q\partial - e^{\mathrm{T}}\partial \\ &\text{s.t.} \\ &\begin{cases} 0 \le \partial_i \le C \qquad (i = 1,2,\cdots,N) \\ y^{\mathrm{T}}\partial = 0 \end{cases} \end{aligned} \tag{4}$$

In the optimization problem above, e^{T} is the N-dimension vector of all ones. Q is a $N \times N$ positive semi-definite matrix, and $Q_{ij} = y_i y_j K(x_i, x_j)$, in which $K(x_i, x_j) = \Phi(x_i)^{\mathrm{T}} \Phi(x_j)$ is called kernel function. The three common types of kernel function are polynomial kernel function, radial basis kernel function and sigmoid kernel function. ∂_i are Lagrange multipliers. A multiplier exits for each training data instance and data instances corresponding to non-zero ∂_i are support vectors.

Do this optimization problem and the ultimate SVM classifier is constructed as following

$$\operatorname{sgn}\left(\sum_{i=1}^{N} \partial_i y_i K(x, x_i) + b\right) \tag{5}$$

3 Data Collection and Preprocessing

3.1 Data Collection

The data used in this research was obtained from China Stock Market & Accounting Research Database. Companies that are specially treated (ST)[1] by China Securities Supervision and Management Committee (CSSMC) are considered as companies in financial distress and those never specially treated are regarded as healthy ones. According to the data between 2000 and 2005, 135 pairs of companies listed in Shenzhen Stock Exchange and Shanghai Stock Exchange are selected as initial dataset. Suppose the year when a company is specially treated as the benchmark year

[1] The most common reason that China listed companies are specially treated by CSSMC is that they have had negative net profit in continuous two years. Of course they will also be specially treated if they purposely publish financial statements with serious false and misstatement, but the ST samples chosen in this study are all companies that have been specially treated because of negative net profit in continuous two years.

(t-0), then (t-1), (t-2) and (t-3) respectively represent one year before ST, two years before ST and three years before ST. After eliminating companies with missing and outlier data, the final numbers of sample companies are 75 pairs at year (t-1), 108 pairs at year (t-2) and 91 pairs at year (t-3).

3.2 Data Scaling

To avoid features in greater numeric ranges dominating those in smaller numeric ranges and also to avoid numerical difficulties during the calculation [16], all data are scaled to the range [-1,1] according to the formula (6).

$$x' = a + \frac{x - \min_x}{\max_x - \min_x} \times (b - a) \tag{6}$$

where $\min_x$ and $\max_x$ are respectively the minimum and maximum value of feature x, and a and b are respectively the minimum and maximum value after scaling, here a= -1, b=1.

3.3 Choice of Financial Ratios

Companies at the different stages before financial distress usually have different symptoms which are indicated by different financial ratios. Different from other researches which use the same financial ratios set to construct predictive models for different years before financial distress, aiming at improving the predictive ability, this study use different financial ratios set respectively for the three years before ST. Each year's financial ratios set are selected from 35 original financial ratios by the statistical method of stepwise discriminant analysis. According to the sample data, the chosen financial ratios sets for year (t-1), (t-2) and (t-3) are listed in Table1.

From Table 1, it is known that financially distressed companies at the stage (t-3) mainly showed abnormality in activity ratios and debt ratios. At the stage (t-2) the profitability of financially distressed companies began to evidently differ from that of healthy ones. At the stage of (t-1) activity ratios, debt ratios, profitability and growth ability all start to further deteriorate.

Table 1. Financial ratios sets of the three years before ST

Year	Financial ratios set	
(t-1)	Total asset turnover	Asset-liability ratio
	Earning per share	Total asset growth rate
(t-2)	Account payable turnover	Current asset turnover
	Fixed asset turnover	Asset-liability ratio
	Return on total asset	Return on current asset
	Return on equity	
(t-3)	Current asset turnover	Fixed asset turnover
	The ratio of cash to current liability	Asset-liability ratio
	The proportion of current liability	

4 Model Construction and Experiment Results

4.1 Construction of SVM Model

Construction of SVM model is to choose the kernel function and search for the values of model parameters. This study uses radial basis kernel function because it was the most widely used kernel function and usually got better result than other kernel functions (see reference [13] for detailed reason). The radial basis kernel function is as formula (7).

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2) \qquad \gamma > 0 \tag{7}$$

Then construction of SVM needs two parameters to be identified, the tuning parameter C and the kernel parameter γ. Improperly setting these parameters will lead to training problems such as over-fitting or under-fitting. So aiming at searching for proper values for these parameters, this study follows the cross-validation and grid-search techniques recommended by Chih-Wei Hsu etc. Trying exponentially growing sequences of C and γ ($C = 2^{-5}, 2^{-4}, \ldots, 2^{15}$; $\gamma = 2^{-15}, 2^{-14}, \ldots, 2^{5}$), it is designed to find a good pair of parameter values which make the 5-fold cross-validation accuracy highest. Though grid-search is an exhaustive search method, its computational time is not much more than other advanced search methods, for example, approximating the cross-validation rate [16].Grid-search on C and γ for SVM model at year (t-1) is as Fig. 1.

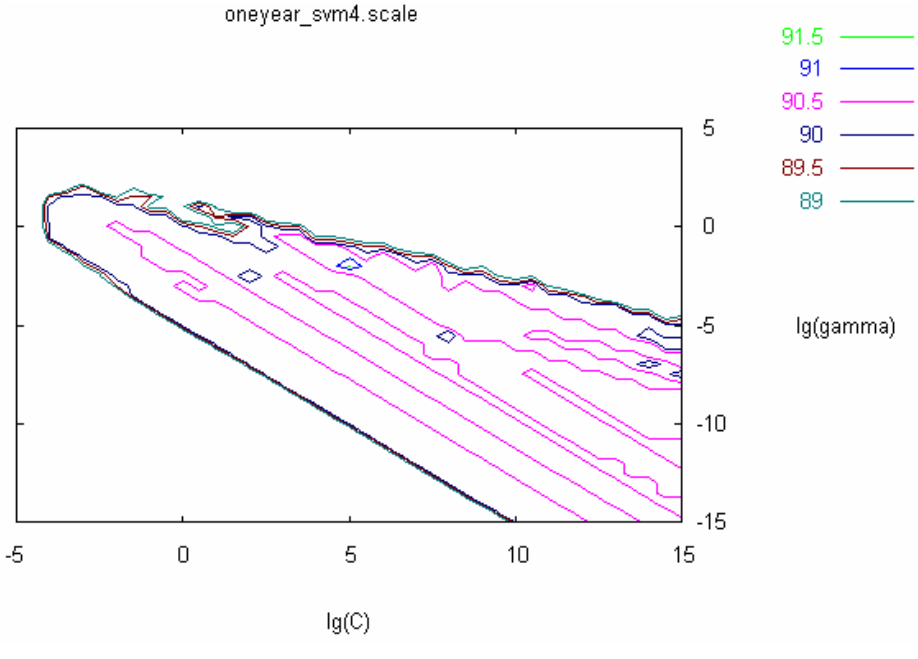


Fig. 1. Grid-search on C and γ for SVM model at year (t-1)

So the good pair of model parameters for year (t-1) is (2^5, 2^{-2}), i.e. C =32, γ =0.25. By the same method, the model parameters for year ((t-2) and (t-3) are respectively defined as (0.5, 1) and (128, 0.0625).

4.2 Experiment Results

In order to make comparative study, Fisher discriminant analysis, Logistic regression and BP-NNs were also carried out in the experiment. Leave-one-out method was used to test the validity of different models, because leave-one-out accuracy can objectively reflect the models' ability to predict financial distress for companies outside the training samples. If the total number of samples is *N*, to calculate the leave-one-out accuracy, *N* times of training and testing are needed for each predictive method. SPSS 11.5 was utilized for Fisher discriminant analysis and Logistic regression. MATLAB 6.5 was used for BP-NNs, and its structure for year (t-1), (t-2) and (t-3) are respectively 4-8-1, 7-11-1 and 5-8-1, and the learning rate was set as 0.1. LIBSVM software developed by Prof. Lin Chih-Jen in Taiwan University was used for SVM modeling and testing. The experiment results are list in Table 2.

Table 2. The training and testing accuracy of different models at different years

Year	Models	MDA (%)	Logit (%)	NNs (%)	SVM (%)
(t-1)	Training accuracy	89.5	87.6	92.2	91.5
	Leave-one-out	88.2	86.9	88.2	88.9
	Drop percentage	1.45	0.80	4.34	2.84
(t-2)	Training accuracy	86.1	86.1	89.6	87.5
	Leave-one-out	85.2	84.7	84.3	85.6
	Drop percentage	1.05	1.63	5.92	2.17
(t-3)	Training accuracy	77.5	79.1	80.9	80.2
	Leave-one-out	75.3	77.5	74.2	78.6
	Drop percentage	2.84	2.02	8.28	2.00

5 Discussion and Analysis

From table 2, it is clear that the predictive ability of each model declines from year (t-1) to year (t-3), which indicates that the nearer to the time when financial distress breaks out, the more information content the financial ratios contain, so that the more strong predictive ability each model has.

Besides, whichever year it is, SVM has the highest leave-one-out accuracy. Whether from the perspective of training accuracy or leave-one-out accuracy, SVM performs better than Fisher discriminant analysis and Logistic regression. The training accuracy of SVM is a little lower than BP-NNs at each year, but its leave-one-out accuracy is higher than BP-NNs, which shows that SVM has better generalization ability than BP-NNs and can better avoid over-fitting phenomenon.

Furthermore, from the point of view of model stability, Fisher discriminant analysis, Logistic regression and SVM have better stability but BP-NNs has relatively

worse stability according to the drop percentage, which equals the difference between training accuracy and leave-one-out accuracy divides training accuracy. Compared with Fisher discriminant analysis and Logistic regression, SVM has worse model stability than them at year (t-1) and (t-2), but at year (t-3) SVM performs a little better than them. Compared with BP-NNs, SVM has much better model stability.

So, SVM has a better balance among fitting ability, generalization ability and model stability. By finding out support vectors for financial distress prediction from training samples, SVM is suitable to predict financial distress for companies outside training sample, and it can keep the predictive accuracy relatively stable when the training samples change within a certain range.

6 Conclusion

SVM is a relatively new machine learning and classification technique with strict theoretical basis. This paper applies SVM to companies' financial distress prediction on the basis of stating its basic theory. In the empirical experiment, three years' data of 135 pairs of Chinese listed companies were selected as initial sample data, and stepwise discriminant analysis method was used to select financial ratios set, and cross-validation and grid-search techniques are utilized to define good parameters for SVM model. By comparing the experiment results of SVM financial distress prediction model with Fisher discriminant analysis, Logistic regression and BP-NNs, it is concluded that SVM gets a better balance among fitting ability, generalization ability and model stability than the other three models. So SVM, which is not only a theoretically good classifier but also has satisfying empirical application result, is a promising method in the practice of financial distress prediction and should be more widely used in the domain of financial decision.

Acknowledgements

The authors gratefully thank anonymous reviewers for their comments and editors for their work.

References

1. Gestel T. V., Baesens B., Suykens J. A.: Bayesian Kernel Based Classification for Financial Distress Detection, Vol. 1. European Journal of Operational Research (2005) 1-2
2. Beaver W.: Financial Ratios as Predictors of Failure. Journal of Accounting Research (1966) 71-111
3. Altman E. I.: Financial Ratios Discriminant Analysis and the Prediction of Corporate Bankruptcy, Vol. 23. Journal of Finance (1968) 589–609
4. Ohlson J. A.: Financial Ratios and Probabilistic Prediction of Bankruptcy, Vol. 18. Journal of Accounting Research (1980) 109-131
5. Odom M., Sharda R.: A Neural Networks Model for Bankruptcy Prediction. Proceedings of the IEEE International Conference on Neural Network (1990) 163–168

6. Fletcher D., Goss E.: Forecasting with Neural Networks: an Application Using Bankruptcy Data, Vol. 24. Information and Management (1993) 159–167
7. Carlos Serrano-Cinca: Self Organizing Neural Networks for Financial Diagnosis, Vol. 17. Decision Support Systems (1996) 227-238
8. Parag C. P.: A Threshold Varying Artificial Neural Network Approach for Classification and Its Application to Bankruptcy Prediction Problem, Vol. 32. Computers & Operations Research (2005) 2561-2582
9. Cristianini N., Shawe-Taylor J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, England (2000)
10. 10.Vapnik V.: The Nature of Statistical Learning Theory. Springer, Berlin (1995)
11. 11.Vapnik V.: Statistical Learning Theory. Wiley, New York (1998)
12. 12.Shin K.-S., Lee T. S., Kim H.-J.: An Application of Support Vector Machines in Bankruptcy Prediction Model, Vol. 28. Expert Systems with Applications (2005) 127–135
13. 13.Min J. H., Lee Y.-C.: Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters, Vol. 28. Expert Systems with Applications (2005) 128-134
14. 14.Kim K. J.: Financial Time Series Forecasting Using Support Vector Machines, Vol. 55. Neurocomputing (2003) 307–319
15. 15.Tay F. E. H., Cao L.: Application of Support Vector Machines in Financial Time Series Forecasting, Vol. 29. Omega (2001) 309–317
16. 16.Hsu C.-W., Chang C.-C., Lin C.-J.: A Practical Guide to Support Vector Classification. Technical Report. http://www.csie.ntu.edu.tw/~cjlin/papers /guide/guide.pdf

Probabilistic Verification of Uncertain Systems Using Bounded-Parameter Markov Decision Processes

Di Wu and Xenofon Koutsoukos

Institute for Software Integrated Systems,
Department of Electrical Engineering & Computer Science,
Vanderbilt University, Nashville, TN 37235, USA
{Di.Wu, Xenofon.Koutsoukos}@Vanderbilt.Edu

Abstract. Verification of probabilistic systems is usually based on variants of Markov processes. For systems with continuous dynamics, Markov processes are generated using discrete approximation methods. These methods assume an exact model of the dynamic behavior. However, realistic systems operate in the presence of uncertainty and variability and they are described by uncertain models. In this paper, we address the problem of probabilistic verification of uncertain systems using Bounded-parameter Markov Decision Processes (BMDPs). Proposed by Givan, Leach and Dean [1], BMDPs are a generalization of MDPs that allow modeling uncertainty. In this paper, we first show how discrete approximation methods can be extended for modeling uncertain systems using BMDPs. Then, we focus on the problem of maximizing the probability of reaching a set of desirable states, we develop a iterative algorithm for probabilistic verification, and we present a detailed mathematical analysis of the convergence results. Finally, we use a robot path-finding application to demonstrate the approach.

1 Introduction

Verification of probabilistic systems is usually based on variants of Markov processes. For systems with continuous dynamics, Markov processes are generated using discrete approximation methods. Probabilistic verification aims at establishing bounds on the probabilities of certain events. Typical problems include the maximum and the minimum probability reachability problems, where the objective is to compute the control policy that maximizes the probability of reaching a set of desirable states, or minimize the probability of reaching an unsafe set. Algorithms for verification of MDPs have been presented in [2, 3]. These methods assume an exact model of the dynamic behavior for defining the transition probabilities. However, realistic systems operate in the presence of uncertainty and variability and they are described by uncertain models. Existing verification methods are insufficient for dealing with such uncertainty.

In this paper, we address the problem of probabilistic verification of uncertain systems using Bounded-parameter Markov Decision Processes (BMDPs). Proposed by Givan, Leach and Dean [1], BMDPs are a generalization of MDPs that

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 283–294, 2006.

allows modeling uncertainty. A BMDP can be viewed as a set of exact MDPs (sharing the same state and action space) specified by intervals of transition probabilities and rewards. Policies are compared on the basis of interval value functions. Optimistic and pessimistic criteria for optimality are used to define partial order relations between pairs of interval value functions.

Our approach is motivated by a robot path-finding application. Under uncertainty, the motion of the robot can be described by a set of stochastic differential equations with uncertain parameters. We show how discrete approximation methods can be extended for modeling such uncertain systems using BMDPs. Although we focus on a robotic example, the approach can be extended for probabilistic verification of stochastic hybrid (discrete-continuous) systems [4].

The paper focuses on the problem of maximizing the probability of reaching a set of desirable states. We develop a iterative algorithm for probabilistic verification, and we present a detailed mathematical analysis of the convergence results. The results presented in [1] are for dynamic programming methods assuming a discounted reward criterion. A discount factor of less than 1 ensures the convergence of the iterative methods for the interval value functions. Probabilistic verification for BMDPs can be formulated based on the Expected Total Reward Criterion (ETRC) for MDPs [5]. Under ETRC, the discount factor is set to 1, and the convergence of the iterative algorithms for BMDPs is more involved because the iteration operators are no longer contraction mappings. Furthermore, the interval value function may be not well defined unless proper restrictions on the intervals of transition probabilities and rewards are applied. Based on the ETRC, we solve the maximum probability reachability problems for BMDPs. Finally, we demonstrate our approach using a robot path-finding application.

Variants of uncertain MDPs have been also studied in [6, 7, 8, 9]. These approaches consider a discounted reward. In addition, the work [10] considers an average performance criterion. Probabilistic verification of uncertain systems is a significant problem which requires an undiscounted criterion and cannot be treated with these algorithms.

In the next section, we review the basic notions of BMDPs. In Section 3, we explain how we can use BMDPs to model uncertain systems. In Section 4, we formulate and solve the maximum probability reachability problem for BMDPs. In Section 5, we present a robot path-finding application to demonstrate our approach. We close with conclusions and some future directions in Section 6.

2 Bounded-Parameter Markov Decision Processes

We first review some basic notions of BMDPs, establish the notation that we use in this paper, and briefly summarize the main results in [1].

A BMDP is a four tuple $\mathcal{M} = \langle \mathcal{Q}, \mathcal{A}, \hat{F}, \hat{R} \rangle$ [1] where $\mathcal{Q}$ is a set of states, $\mathcal{A}$ is a set of actions, $\hat{R}$ is an interval reward function that maps each $q \in \mathcal{Q}$ to a

[1] In this paper, we use $\hat{X}$ to denote an interval, i.e. $\hat{X} = [\underline{X}, \overline{X}] \subseteq \mathbb{R}$.

closed interval of real values $[\underline{R}(q), \overline{R}(q)]$, and $\hat{F}$ is an interval state-transition distribution so that for $p, q \in \mathcal{Q}$ and $\alpha \in \mathcal{A}$,

$$\underline{F}_{p,q}(\alpha) \leqslant Pr(X_{t+1} = q | X_t = p, U_t = \alpha) \leqslant \overline{F}_{p,q}(\alpha).$$

For any action α and state p, the sum of the lower bounds of $\hat{F}_{p,q}(\alpha)$ over all states q is required to be less than or equal to 1, while the sum of the upper bounds is required to be greater than or equal to 1.

A BMDP $\mathcal{M}$ defines a set of exact MDPs. Let $M = \langle \mathcal{Q}', \mathcal{A}', F', R' \rangle$ be an MDP. If $\mathcal{Q} = \mathcal{Q}'$, $\mathcal{A} = \mathcal{A}'$, $R'(p) \in \hat{R}(p)$ and $F'_{p,q}(\alpha) \in \hat{F}_{p,q}(\alpha)$ for any $\alpha \in \mathcal{A}$ and $p, q \in \mathcal{Q}$, then we say $M \in \mathcal{M}$. To simplify discussions, in the following paragraphs the rewards are assumed to be tight, i.e. a single value. The approaches in this paper can be easily generalized to the case of interval rewards.

A policy is a mapping from states to actions, $\pi : \mathcal{Q} \to \mathcal{A}$. We use Π to denote the set of stationary Markov policies. The policy in this paper is restricted to be in Π. For any policy π and state p, the interval value function of π at p is a closed interval defined as

$$\hat{V}_\pi(p) = [\min_{M \in \mathcal{M}} V_{M,\pi}(p), \max_{M \in \mathcal{M}} V_{M,\pi}(p)]$$

where

$$V_{M,\pi}(p) = R(p) + \gamma \sum_{q \in \mathcal{Q}} F^M_{p,q}(\pi(p)) V_{M,\pi}(q)$$

where $0 < \gamma < 1$ is called the discount factor.

An MDP $M \in \mathcal{M}$ is π-maximizing if for any $M' \in \mathcal{M}$, $V_{M,\pi} \geq_{dom} V_{M',\pi}$[2]. Likewise, M is π-minimizing if for any $M' \in \mathcal{M}$, $V_{M,\pi} \leq_{dom} V_{M',\pi}$. For any policy $\pi \in \Pi$, there exist a π-maximizing MDP $\overline{M}(\pi)$ and a π-minimizing MDP $\underline{M}(\pi)$ in $\mathcal{M}$.

The interval policy evaluation operator $\widehat{IVI}_\pi$ for each state p is defined as

$$\widehat{IVI}_\pi(\hat{V})(p) = [\underline{IVI}_\pi(\underline{V})(p), \overline{IVI}_\pi(\overline{V})(p)]$$

where

$$\underline{IVI}_\pi(\underline{V}) = \min_{M \in \mathcal{M}} VI_{M,\pi}(\underline{V}) = VI_{\underline{M}(\pi),\pi}(\underline{V}),$$

$$\overline{IVI}_\pi(\overline{V}) = \max_{M \in \mathcal{M}} VI_{M,\pi}(\overline{V}) = VI_{\overline{M}(\pi),\pi}(\overline{V})$$

and $VI_{M,\pi} : \mathcal{V} \to \mathcal{V}$ is the policy evaluation operator for the exact MDP M and policy π

$$VI_{M,\pi}(V)(p) = R(p) + \gamma \sum_{q \in \mathcal{Q}} F^M_{p,q}(\pi(p)) V(q).$$

To define the π-minimizing MDP $\underline{M}(\pi)$, we only need to compute its transition function $F_{\underline{M}(\pi)}$. Let q_1, q_2, ..., q_k $(k = |\mathcal{Q}|)$ be an ordering of $\mathcal{Q}$ so that $\underline{V}(q_i) \leqslant \underline{V}(q_j)$ for any $1 \leqslant i < j \leqslant k$. Let r be the index $1 \leqslant r \leqslant k$ which

[2] $V \geq_{dom} U$ if and only if for all $q \in \mathcal{Q}$, $V(q) \geqslant U(q)$.

maximizes $\sum_{i=1}^{r-1} \overline{F}_{p,q_i}(\alpha) + \sum_{i=r}^{k} \underline{F}_{p,q_i}(\alpha)$ without letting it exceed 1. Then the transition function of the π-minimizing MDP $\underline{M}(\pi)$ is given by

$$F_{p,q_j}^{\underline{M}(\pi)}(\alpha) = \begin{cases} \overline{F}_{p,q_j}(\alpha) & \text{if } j < r \\ \underline{F}_{p,q_j}(\alpha) & \text{if } j > r \end{cases} \text{ and } F_{p,q_r}^{\underline{M}(\pi)}(\alpha) = 1 - \sum_{i=1, i\neq r}^{i=k} F_{pq_i}(\alpha).$$

The definition of the π-maximizing MDP is similar.

In order to define the optimal value function for a BMDP, two different orderings on closed real intervals are introduced: $[l_1, u_1] \leq_{opt} [l_2, u_2] \iff (u_1 < u_2 \vee (u_1 = u_2 \wedge l_1 \leqslant l_2))$ and $[l_1, u_1] \leq_{pes} [l_2, u_2] \iff (l_1 < l_2 \vee (l_1 = l_2 \wedge u_1 \leqslant u_2))$. In addition, $\hat{U} \leq_{opt} \hat{V}$ ($\hat{U} \leq_{pes} \hat{V}$) if and only if $\hat{U}(q) \leq_{opt} \hat{V}(q)$ ($\hat{U}(q) \leq_{pes} \hat{V}(q)$) for each $q \in \mathcal{Q}$. Then the optimistic optimal value function $\hat{V}_{opt}$ and the pessimistic optimal value function $\hat{V}_{pes}$ are given by

$$\hat{V}_{opt} = \sup_{\pi \in \Pi, \leq_{opt}} \hat{V}_\pi \text{ and } \hat{V}_{pes} = \sup_{\pi \in \Pi, \leq_{pes}} \hat{V}_\pi,$$

respectively. The value interation for $\hat{V}_{opt}$ is used when the agent aims at maximizing the upper bound $\overline{V}$ while $\hat{V}_{pes}$ is used when the agent aims at maximizing the lower bound $\underline{V}$. In the subsequent sections, we focus on the optimistic case for the optimal interval value functions. Unless noted, results for the pessimistic case can be inferred analogously.

The interval value iteration operator $\widehat{IVI}_{opt}$ for each state p is defined as

$$\widehat{IVI}_{opt}(\hat{V})(p) = \max_{\alpha \in \mathcal{A}, \leq_{opt}} [\min_{M \in \mathcal{M}} VI_{M,\alpha}(\underline{V})(p), \max_{M \in \mathcal{M}} VI_{M,\alpha}(\overline{V})(p)]. \tag{1}$$

Due to the nature of $\leq_{opt}$, $\widehat{IVI}_{opt}$ evaluates actions primarily based on the interval upper bounds, breaking ties on the lower bounds. For each state, the action that maximizes the lower bound is chosen from the subset of actions that equally maximize the upper bound. Hence (1) can be rewritten as

$$\widehat{IVI}_{opt}(\hat{V}) = [\underline{IVI}_{opt}(\hat{V}), \overline{IVI}_{opt}(\overline{V})] \tag{2}$$

where

$$\underline{IVI}_{opt}(\hat{V}) = \underline{IVI}_{opt,\overline{V}}(\underline{V})$$

and for any $q \in \mathcal{Q}$,

$$\overline{IVI}_{opt}(\overline{V})(q) = \max_{\alpha \in \mathcal{A}} \max_{M \in \mathcal{M}} VI_{M,\alpha}(\overline{V})(q),$$

$$\underline{IVI}_{opt,\overline{V}}(\underline{V})(q) = \max_{\alpha \in \rho_{\overline{V}}(q)} \min_{M \in \mathcal{M}} VI_{M,\alpha}(\underline{V})(q)$$

where

$$\rho_W(p) = \arg\max_{\alpha \in \mathcal{A}} \max_{M \in \mathcal{M}} VI_{M,\alpha}(W)(p). \tag{3}$$

Methods similar to those used in proving the convergence of total discounted reward optimality for exact MDPs can be used to prove that iterating $\widehat{IVI}_{opt}$ converges to $\hat{V}_{opt}$. Detailed proofs of convergence results can be found in [1].

3 Modeling Uncertain Systems by BMDPs

In this section, we describe how BMDPs can be generated for uncertain systems and we illustrate the approach using a robot-path finding application. Consider a continuous system with dynamics described by a stochastic differential equation (SDE) $dx = f(x,u)dt + \sigma(x)dw$ where $x \in X$ is the state of the system, $u \in U$ is the control action, $\sigma(x)$ is a diffusion term of appropriate dimensions, and $w(t)$ is a Wiener process. The SDE is approximated by a controlled Markov process that evolves in a state space that is a discretization of the state space X. The criterion which must be satisfied by the approximating MDP is [illegible] [11]. Local consistency means that the conditional mean and covariance of the MDP are proportional to the local mean and covariance of the original process. An approximation parameter h analogous to a "finite element size" parameterizes the approximating Markov process. As h goes to zero, the local properties of the MDP resemble the local properties of the original stochastic process.

The transition probabilities of the MDP can be computed systematically from the parameters of the SDE (details can be found in [11]). If the diffusion matrix $a(x) = \sigma(x)\sigma^T(x)$ is diagonal and we consider a uniform grid with e_i denoting the unit vector in the i^{th} direction, the transition probabilities are

$$F_{x,x\pm he_i}(u) = \frac{a_{ii}(x)/2 + hf_i^{\pm}(x,u)}{Q(x,u)}, \tag{4}$$

where $\Delta t(x,u) = h^2/Q(x,u)$, $Q(x,u) = \sum_i [a_{ii}(x) + h|f_i(x,u)|]$ and $a^+ = \max\{a,0\}$ and $a^- = \max\{-a,0\}$.

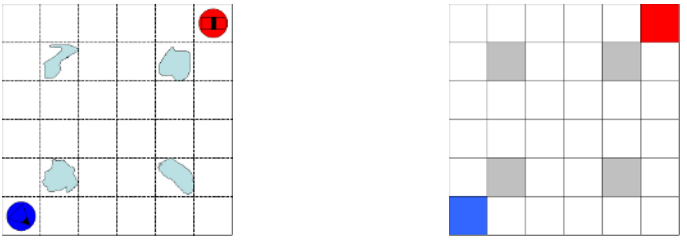
The approximation described above assumes that the system model is known exactly. For many practical systems, however, model parameters are not known exactly. Uncertain continuous systems are usually modeled assuming that some parameters take values in a pre-defined (usually convex) set. In this case, the approximation outlined above will result in BMDPs where the transition probabilities are replaced by interval transition probabilities.

In the following, we illustrate the approximation approach with a robot-path finding example. For simplicity, we assume that mobile robots operate in planar environments and we do not model the orientation or any nonholonomic constraints. The behavior of the robot is described by

$$dx = u_1 dt + \sigma_1 dw$$
$$dy = u_2 dt + \sigma_2 dw$$

where $(x,y)^{\mathrm{T}}$ is the coordinate of the robot, $(u_1,u_2)^{\mathrm{T}}$ is the control input representing the command velocity, and $w(t)$ is a Wiener process modeling noise.

Figure 1(a) shows the original model of the operating environment of the robot. The robot is initially at the lower left corner and the destination is at the upper right corner. We discretize the robot's operating environment using a uniform grid and we assume that there are only 4 control actions, {Up, Down, Left, Right}. As shown in Figure 1(b), we also approximate the position of the robot, the destination, and the obstacles as MDP states. Consider a fixed control



(a) (b)

Fig. 1. Robot path-finding problem. (a) Original model of the path-finding problem. (b) The approximated MDP model.

action denoted by u. Because of uncertainty in the system such as motor friction or an unknown workload, it is reasonable to assume that the control action corresponds to the command velocity $u = (u_1, u_2)^{\mathrm{T}}$ where u_i $(i = 1, 2)$ is not exact but takes values in the interval $[\underline{u}_i, \overline{u}_i]$. Define the set $\tilde{\mathcal{U}} = \{(u_1, u_2)^{\mathrm{T}} : u_i \in [\underline{u}_i, \overline{u}_i]\}$, then the interval transition probabilities can be computed by

$$\hat{F}_{x,x\pm he_i}(u) = \left[\min_{u\in\tilde{\mathcal{U}}} F_{x,x\pm he_i}(u), \max_{u\in\tilde{\mathcal{U}}} F_{x,x\pm he_i}(u)\right]. \tag{5}$$

Assuming that $\mathcal{U}$ is compact, the function F has well-defined extrema. The intervals can be computed either analytically if the functions F are monotone with respect to the uncertain parameters or using numerical optimization methods. Thus, for each state-action pair we obtain an interval for the transition probabilities and by repeating for all state-action pairs we obtain a BMDP model.

Note that in our approach, (4) and (5) are applied to one dimension of the uncertain system at a time, so the approach is actually applicable to more general systems in higher dimensions than the above robotic example.

4 Maximum Probability Reachability Problem

In this section, we formulate the maximum probability reachability problem, we present a value iteration algorithm, and we analyze its convergence.

4.1 Interval Expected Total Reward for BMDPs

In this paper, we are primarily interested in the problem of maximizing the probability that the agent will reach a desirable set of states. By solving this problem, we can establish bounds on the probabilities of reaching desirable configurations used in probabilistic verification of discrete systems. This problem can be formulated using the Expected Total Reward Criterion (ETRC) for BMDPs (see Section 4.3). Under the ETRC, we compare policies on the basis of the interval expected total reward $\hat{V} = [\underline{V}_\pi, \overline{V}_\pi]$ where for any $q \in \mathcal{Q}$.

$$\overline{V}_\pi(q) = E_{\overline{M}(\pi),\pi}\left\{\sum_{t=1}^{\infty} R(X_t(q))\right\} \text{ and } \underline{V}_\pi(q) = E_{\underline{M}(\pi),\pi}\left\{\sum_{t=1}^{\infty} R(X_t(q))\right\}.$$

We may regard these as the expected total discounted reward with a discount factor $\gamma = 1$. However, for $\gamma = 1$ the convergence results in [1] no longer hold, because the iteration operators $\widehat{IVI}_\pi$, $\widehat{IVI}_{opt}$ and $\widehat{IVI}_{pes}$ are not contraction mappings. Furthermore, the interval value function may not be well defined unless proper restrictions on the intervals of the transition probabilities and rewards are applied.

For simplicity, we use vector notation. For example, R and V are column vectors, whose i-th element is respectively the scalar reward and value function of the i-th state p_i; F_M is the transition probability function of MDP M and $F_{M,\pi}$ is the transition probability matrix of the Markov Chain reduced from M when given a policy π, whose (i,j)-th element is the probability of transitioning from state p_i to state p_j when executing action $\pi(p_i)$.

Let $R^+(q) = \max\{R(q), 0\}$ and $R^-(q) = \max\{-R(q), 0\}$ and define

$$\overline{V}^{\pm}_\pi(q) \equiv \lim_{N\to\infty} E_{\overline{M}(\pi),\pi}\left\{\sum_{t=1}^{N-1} R^{\pm}(X_t(q))\right\}.$$

Since the summands are non-negative, both of the above limits exist[3]. The limit defining $\overline{V}_\pi(q)$ exists whenever at least one of $\overline{V}^+_\pi(q)$ and $\overline{V}^-_\pi(q)$ is finite, in which case $\overline{V}_\pi = \overline{V}^+_\pi(q) - \overline{V}^-_\pi(q)$. $\underline{V}^+_\pi(q)$, $\underline{V}^-_\pi(q)$ and $\underline{V}_\pi(q)$ can be similarly defined. Noting this, we impose the following finiteness assumption which assures that $\hat{V}_\pi$ is well defined.

Assumption 1. [illegible] $\pi \in \Pi$ [illegible] $q \in \mathcal{Q}$ [illegible] $\overline{V}^+_\pi(q)$, $\overline{V}^-_\pi(q)$, [illegible] $\underline{V}^+_\pi(q)$, $\underline{V}^-_\pi(q)$, [illegible]

Consider the optimal interval value functions $\hat{V}_{opt}$ defined in Section 2. The following theorem establishes the optimality equation for the ETRC and shows that the optimal interval value function is a solution of the optimality equation.

Theorem 1. [illegible] $\overline{V}_{opt}$ [illegible]

$$V = \sup_{\pi\in\Pi}\max_{M\in\mathcal{M}} VI_{M,\pi}(V) = \sup_{\pi\in\Pi}\{R + F_{\overline{M}(\pi),\pi}V\} \equiv \overline{IVI}_{opt}(V),$$

[illegible] $\underline{V}_{opt,W}$ [illegible]

$$V = \sup_{\pi\in\rho_W}\min_{M\in\mathcal{M}} VI_{M,\pi}(V) = \sup_{\pi\in\rho_W}\{R + F_{\underline{M}(\pi),\pi}V\} \equiv \underline{IVI}_{opt,W}(V)$$

[illegible] W [illegible][4]

[3] This includes the case when the limit is $\pm\infty$.

[4] Proofs are omitted due to length limitation, and can be found in [12].

Based on Theorem 1, the value iteration operator $\widehat{IVI}_{opt}$ can be defined as in Equation (1). The following lemma establishes the monotonicity of the iteration operators.

Lemma 2. [illegible] $U \leq V$ [illegible] $\mathcal{V}$ [illegible] $U \leq_{dom} V$ [illegible] $\overline{IVI}_{opt}(U) \leq_{dom} \overline{IVI}_{opt}(V)$ [illegible] $\underline{IVI}_{opt,W}(U) \leq_{dom} \underline{IVI}_{opt,W}(V)$ [illegible] W [illegible]

Lemma 2 also suggests that the iteration operator $\widehat{IVI}_{opt}$ has the following property: for any $\hat{U} \leq_{opt} \hat{V}$ in $\hat{\mathcal{V}}$, $\widehat{IVI}_{opt}(\hat{U}) \leq_{opt} \widehat{IVI}_{opt}(\hat{V})$. These properties are essential in the proof of the convergence results of the interval value iteration algorithm.

Clearly, Assumption 1 is necessary for any computational approach. In the general case for the expected total reward criterion (ETRC), we cannot validate that the assumption holds. However, in the maximum probability reachability problem, the (interval) value function is interpreted as (interval) probability and therefore Assumption 1 can be easily validated as shown in Section 4.3.

4.2 Interval Value Iteration for Non-negative BMDP Models

In order to prove convergence of the value iteration, we consider the following assumptions in addition to Assumption 1:

Assumption 2. [illegible] $q \in Q$ $R(q) \geqslant 0$

Assumption 3. [illegible] $q \in Q$ [illegible] $\pi \in \Pi$ $\overline{V}_{\pi}^{+}(q) < \infty$ [illegible] $\underline{V}_{\pi}^{+}(q) < \infty$

If a BMDP is consistent with both Assumption 2 and 3, it is a non-negative BMDP model, and its value function under the ETRC is called non-negative interval expected total reward. Note that Assumption 3 implies Assumption 1, so Theorem 1 and Lemma 2 hold for non-negative BMDP models. Lemma 3 suggests that $\hat{V}_{opt}$ is the minimal solution of the optimality equation, and Theorem 4 establishes the convergence result of interval value iteration for non-negative BMDPs.

Lemma 3. [illegible] $\overline{V}_{opt}$ [illegible] $V = \overline{IVI}_{opt}(V)$ [illegible] $\mathcal{V}^{+}$ [illegible] $\mathcal{V}^{+} = \mathcal{V} \cap [0, \infty]$ [illegible] $\underline{V}_{opt,W}$ [illegible] $V = \underline{IVI}_{opt,W}(V)$ [illegible] $\mathcal{V}^{+}$ [illegible] W [illegible]

Theorem 4. [illegible] $\hat{V}^0 = [0,0]$ [illegible] $\{\hat{V}^n\}$ [illegible] $\hat{V}^n = \widehat{IVI}_{opt}^{n}(\hat{V}^0)$ [illegible] $\hat{V}_{opt}$

It can be shown that the initial value of the interval value function is not restricted to be $[0,0]$. By choosing a $\hat{V}^0$ with $0 \leqslant \underline{V}^0 \leqslant \underline{V}_{opt}$ and $0 \leqslant \overline{V}^0 \leqslant \overline{V}_{opt}$, interval value iteration converges to $\hat{V}_{opt}$ for non-negative BMDPs. For BMDP models consistent with Assumption 2 and Assumption 3, convergence of the iterative algorithm is guaranteed by Theorem 4 for $\hat{V}^0 = [0,0]$.

4.3 Verification Based on Non-negative BMDP Models

An instance of the maximum probability reachability problem for BMDPs consists of a BMDP $\mathcal{M} = \langle \mathcal{Q}, \mathcal{A}, \hat{F}, R \rangle$ together with a destination set $\mathcal{T} \subseteq \mathcal{Q}$. The objective of maximum probability reachability problem is to determine, for all $p \in \mathcal{Q}$, the maximum interval probability of starting from p and finally reaching any state in $\mathcal{T}$, i.e.

$$\hat{U}^{max}_{\mathcal{M},opt}(p) = \sup_{\pi \in \Pi, \leq_{opt}} [\underline{U}_{\mathcal{M},\pi}(p), \overline{U}_{\mathcal{M},\pi}(p)]$$

where

$$\underline{U}_{\mathcal{M},\pi}(p) = \min_{M \in \mathcal{M}} Pr_{M,\pi}(\exists t.X_t(p) \in \mathcal{T}), \tag{6}$$

$$\overline{U}_{\mathcal{M},\pi}(p) = \max_{M \in \mathcal{M}} Pr_{M,\pi}(\exists t.X_t(p) \in \mathcal{T}). \tag{7}$$

$\underline{U}_{\mathcal{M},\pi}$ and $\overline{U}_{\mathcal{M},\pi}$ are probabilities and therefore by definition take values in $[0, 1]$. Thus, the interval value function satisfies Assumption 1. Note that $\underline{U}_{\mathcal{M},\pi}(p)$ can be computed recursively by

$$\underline{U}_{\mathcal{M},\pi}(p) = \begin{cases} \min_{M \in \mathcal{M}} \sum_{q \in \mathcal{Q}} F^M_{p,q}(\pi(p)) \underline{U}_{\mathcal{M},\pi}(q) & \text{if } p \in \mathcal{Q} - \mathcal{T} \\ 1 & \text{if } p \in \mathcal{T} \end{cases} \tag{8}$$

In order to transform the Maximum Probability Reachability Problem to a problem solvable by interval value iteration, we add a terminal state r with transition probability 1 to itself on any action, let all the destination states in $\mathcal{T}$ be absorbed into the terminal state, i.e., transition to r with probability 1 on any action, and set the reward of each destination state to be 1 and of every other state to be 0. Thus we form a new BMDP model $\widetilde{\mathcal{M}} = \langle \tilde{\mathcal{Q}}, \tilde{\mathcal{A}}, \tilde{F}, \tilde{R} \rangle$, where $\tilde{\mathcal{Q}} = \mathcal{Q} \cup \{r\}$, $\tilde{\mathcal{A}} = \mathcal{A}$ and for any $p, q \in \tilde{\mathcal{Q}}$, and $\alpha \in \mathcal{A}$

$$\tilde{R}(p) = \begin{cases} 1 \text{ if } p \in \mathcal{T} \\ 0 \text{ if } p \notin \mathcal{T} \end{cases},$$

$$\tilde{F}_{p,q}(\alpha) = \begin{cases} \hat{F}_{p,q}(\alpha) & \text{if } p \notin \mathcal{T} \cup \{r\} \\ [0,0] & \text{if } p \in \mathcal{T} \cup \{r\} \text{ and } q \neq r \\ [1,1] & \text{if } p \in \mathcal{T} \cup \{r\} \text{ and } q = r \end{cases}. \tag{9}$$

Since $\tilde{R}(r) = 0$, by the structure of $\tilde{F}_{p,q}$, it is clear that $\underline{V}_{\widetilde{\mathcal{M}},\pi}(r)$ will not be affected by the values of any states. For any $p \in \mathcal{Q}$

$$\underline{V}_{\widetilde{\mathcal{M}},\pi}(p) = \min_{M \in \widetilde{M}} \left\{ \tilde{R}(p) + \sum_{q \in \tilde{Q}} F^M_{p,q}(\pi(p)) \underline{V}_{M,\pi}(q) \right\}. \tag{10}$$

Specifically, for $p \in \mathcal{T}$

$$\underline{V}_{\widetilde{\mathcal{M}},\pi}(p) = \min_{M \in \widetilde{M}} \left\{ \tilde{R}(p) + \sum_{q \in \tilde{Q}} F^M_{p,q}(\pi(p)) \underline{V}_{M,\pi}(q) \right\} = \tilde{R}(p) + \underline{V}_{\widetilde{M},\pi}(r) = 1. \tag{11}$$

From (9), (10) and (11), it follows that $\underline{U}_{\mathcal{M},\pi}$ is equivalent to $\underline{V}_{\widetilde{\mathcal{M}},\pi}$. Similarly, $\overline{U}_{\mathcal{M},\pi}$ is equivalent to $\overline{V}_{\widetilde{\mathcal{M}},\pi}$. Therefore

$$\hat{V}_{\widetilde{M},opt} = \sup_{\pi \in \Pi, \leq_{opt}} [\underline{V}_{\widetilde{M},\pi}, \overline{V}_{\widetilde{M},\pi}] = \sup_{\pi \in \Pi, \leq_{opt}} [\underline{U}_{\mathcal{M},\pi}, \overline{U}_{\mathcal{M},\pi}] = \hat{U}_{\mathcal{M},opt}. \tag{12}$$

The BMDP $\widetilde{M}$ constructed as described above is consistent with Assumption 3, so the interval value function for each state exists, which suggests that the MPRP for $\mathcal{M}$ can be solved using the algorithm presented in Section 4.1. Further, $\widetilde{\mathcal{M}}$ satisfies Assumption 2, and therefore the convergence is characterized by Theorem 4.

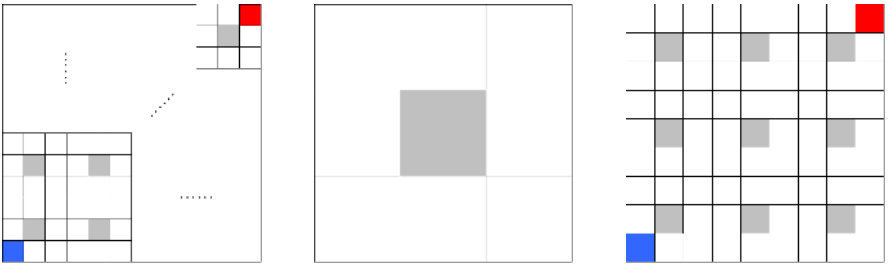
Note that we don't assume the existence of a proper policy. Convergence is guaranteed without this assumption. In the case of the maximum probability reachability problem, if there is not proper policy (for a particular state) then the algorithm will simply compute the corresponding interval value function (probability) as $[0, 0]$. The approach can be used to validate the existence of a proper policy and actually this is one of the ways that probabilistic verification algorithms can be used in practice.

5 Experimental Results

This section illustrates the approach using a robot path-finding application. In our model, an action succeeds with interval probability $[0.75, 0.9]$ and moves in any other direction with interval probability $[0.05, 0.1]$. For instance, if the robot choose the action "Up", the probability of reaching the adjacent grid to its north is within $[0.75, 0.9]$, the probability of reaching each of the other adjacent grids is within $[0.05, 0.1]$. We also assume the robot will stay where it is with a probability in the same interval probability as if it is not out of bound. Obstacle grids are treated as absorbing states, i.e. transition to itself with interval probability $[1, 1]$ on any action. The goal is to find a policy that maximizes the interval probability that the robot will reach the destination from the initial position.

The layout of the gridworld used in our simulation is shown in Figure 2(a). The (blue) cell in the lower left corner is the initial position of the robot. The (red) grid in the upper right corner is the destination. The (grey) cells represent obstacles. In order to evaluate the computational complexity and scalability of our algorithm, the environment is made up of the same 3×3 tiles as shown in Figure 2(b). For instance, the 9×9 gridworld shown in Figure 2(c) is made up of 9 such tiles, while the 6×6 gridworld in Figure 1(b) in Section 2 is made up of 4 such tiles.

Table 1 shows the interval maximum probabilities for the robot to reach the destination from the initial position, number of iterations and time needed for the iterative algorithm to converge. For example, the optimistic maximum reachability probability for the 9×9 gridworld is $[0.2685, 0.6947]$, the pessimistic maximum reachability probability for the 18×18 gridworld is $[0.1067, 0.4806]$. We can see that the larger the size of the gridworld, the lower the reachability probability. This is because larger gridworld suggests a longer path for the robot to reach



(a) (b) (c)

Fig. 2. Robot path-finding problem. (a) The operating environment of the robot. b) 3×3 tile – the basic component of the environment model. c) The 9×9 environment model that is made up of the 3×3 tiles.

Table 1. Interval maximum reachability probabilities ($\hat{U}_{opt}^{max}$, $\hat{U}_{pes}^{max}$) for the robot path-finding problem, Number of iterations (I_{opt}, I_{pes}) and time (t_{opt}, t_{pes}, in seconds) needed for the iterative algorithms to converge

Size	**States**	$\hat{U}_{opt}^{max}$	I_{opt}	t_{opt}	$\hat{U}_{pes}^{max}$	I_{pes}	t_{pes}
9×9	81	$[0.2685, 0.6947]$	43	3.98	$[0.4156, 0.6947]$	43	3.98
12×12	144	$[0.1707, 0.6145]$	54	15.04	$[0.2645, 0.6145]$	54	14.94
15×15	225	$[0.1083, 0.5435]$	63	42.10	$[0.1681, 0.5435]$	63	41.84
18×18	324	$[0.0686, 0.4807]$	71	98.34	$[0.1067, 0.4806]$	71	98.54
21×21	441	$[0.0434, 0.4251]$	79	201.49	$[0.0434, 0.4251]$	79	201.55
24×24	576	$[0.0275, 0.3760]$	87	374.92	$[0.0275, 0.3760]$	87	375.95

the destination, and greater chance to collide with obstacles. All the simulations are carried out on a Windows XP laptop, 1.60GHz, with 768 MB of RAM, using MATLAB 7.0. Our experimental results suggest that the time complexity of the interval value iteration is polynomial. The exact complexity characterization is a subject of current work.

6 Conclusions

The results described in this paper show that BMDPs can be used for probabilistic verification of uncertain systems. With proper restrictions on the reward and transition functions, the interval value function is well defined and bounded. We also analyze the convergence of iterative methods for computing the interval value function. These results allow us to solve a variety of new problems for BMDPs. The paper focuses on the maximum reachability probability problem. Additional verification problems are subject of current and future work.

Acknowledgments. This work was partially supported by the NSF CAREER grant CNS-0448554.

References

1. Givan, R., Leach, S., Dean, T.: Bounded-parameter Markov decision process. Artificial Intelligence **122** (2000) 71–109
2. Courcoubetis, C., Yannakakis, M.: Markov decision processes and regular events. IEEE Transaction on Automatic Control **43** (1998) 1399–1418
3. de Alfaro, L.: Computing minimum and maximum reachability times in probabilistic systems. In: CONCUR '99: Proceedings of the 10th International Conference on Concurrency Theory, London, UK, Springer-Verlag (1999) 66–81
4. Koutsoukos, X.D.: Optimal control of stochastic hybrid systems based on locally consistent Markov decision processes. International Journal of Hybrid Systems **4** (2004) 301–318
5. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, Inc., New York, NY, USA (1994)
6. Satia, J.K., Lave, R.E.: Markovian decision processes with uncertain transition probabilities. Operations Research **39** (1953) 1095–1100
7. White, C.C., Eldeib, H.K.: Parameter imprecision in finite state, finite action dynamic programs. Operations Research **34** (1986) 120–129
8. White, C.C., Eldeib, H.K.: Markov decision processes with imprecise transition probabilities. Operations Research **43** (1994) 739–749
9. Tang, H., Liang, X., Gao, J., Liu, C.: Robust control policy for semi-Markov decision processes with dependent uncertain parameters. In: WCICA 04': Proceedings of the Fifth World Congress on Intelligent Control and Automation. (2004)
10. Kalyanasundaram, S., Chong, E.K.P., Shroff, N.B.: Markovian decision processes with uncertain transition rates: Sensitivity and robust control. In: CDC '02: Proceedings of the 41th IEEE Conference on Decision and Control. (2002)
11. Kushner, H.J., Dupuis, P.: Numerical Methods for Stochastic Control Problems in Continuous Time, 2nd ed. Springer-Verlag, New York, NY, USA (2001)
12. Wu, D., Koutsoukos, X.D.: Probabilistic verification of bounded-parameter Markov decision processes. Technical Report ISIS-05-607, Institute for Software Integrated Systems, Vanderbilt University, Nashville, TN, USA (2005)
13. Dean, T., Givan, R., Leach, S.: Model reduction techniques for computing approximately optimal solutions for Markov decision processes. In: Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI-97), San Francisco, CA, Morgan Kaufmann Publishers (1997) 124–131
14. Rudin, W.: Real and Complex Analysis, 3rd Edition. McGraw-Hill, New York, NY, USA (1994)

Modality Argumentation Programming

Juan Carlos Nieves and Ulises Cortés

Universitat Politècnica de Catalunya,
Knowledge Engineering and Machine Learning Group,
Departament de Lleguatges i Sistemes Informàtics,
c/Jordi Girona 1-3, E08034, Barcelona, Spain
{jcnieves, ia}@lsi.upc.edu

Abstract. This work is focus in the following critical questions:

1. *How to incorporate modalities in the process of argumentation reasoning?*
2. *Is it possible to build arguments faced with incomplete information?*

Our proposal is based in a specification language which has the following properties: a) it permits to give specifications of modalities in a natural way; b) it defines a process of argumentation reasoning considering modalities; and c) it permits to build arguments from incomplete information.

Keywords: Decision-Making, Argumentation, Logic Programming, Rational Agents.

1 Introduction

Argumentation has proved to be a useful tool for representing and dealing with domains in which rational agents are not able to decide by themselves about something, and may encounter other agents with different preference values. The ability to reason effectively about what is the or appropriate course of action to take in a given situation is an essential activity for a rational agent. A simple rational agent may also use argumentation techniques to perform its individual reasoning as it needs to make rational decisions under complex preferences policies, or to reason about its commitments, its goals, .

A critical question about to carry out argumentation theory to implementation systems still exists. For instance, one of the main objectives of the EU funded project ASPIC[1] is to provide a strong foundation for the design and implementation of a set of generic argument software components which can be used by 3^{rd} party applications.

1.1 Motivation

Since Aristotle, modalities have been an object of study for logicians especially in relation with the construction of arguments. Modalities are terms which indicate the level of with which a claim can be made. One possible definition of is [3]:

[1] Consortium for argumentation technology. http://www.argumentation.org/

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 295–306, 2006.

> "The classification of logical propositions according to their asserting or denying the possibility, impossibility, contingency, or necessity of their content".

Research on rational agents has raised further questions about modalities in the context of argumentation, and the roles that arguments play in the pursuit of an agent's goals and plans.

In our own work on medical decision-making we have very different sources of examples of argumentation [1, 13]. The main objective is to discover the acceptable set of arguments that support a given claim in a given context. This is a purposeful and purposed process where the validity of arguments and the evidence of premises are both approached. One import point of our particular medical domain (organ transplant) is that there is small amount of information available , the viability criterions which are applied whether a particular organ is viable to be transplanted. However, we have a hight-level of detail and quality information , each medical case. Usually, the medical information (in our particular medical domain) is supported by a set of clinal tests.

Lattices have been used to model a wide range of problems. For instance, lattice domains are useful to perform aggregate operations which are a great tool for modeling decision-making in Artificial Intelligence. The use of lattice domains in declarative programming have shown high level of expressiveness. For example, the use of partial-order clauses and lattice domains in partial-order programming is particularly useful for expressing concise solutions to problems for graph theory, program analysis, and database querying [8, 10].

Osorio [7] showed to perform aggregate operations using negation as failure, also Nieves [6] showed how to perform relaxation in optimization problems using aggregate operations and negations as failure.

In this paper, we introduce a declarative language to handle arguments with modalities like , , and . Modality is a category of linguistic meaning having to do with the expression of possibility and necessity. In [16] a study of the kinds of modal meaning can be found.

In §2 we put forward the syntax to be used and give a brief introduction to lattices and order. In §3 we introduce our framework and present some examples. In §4, we present the declarative semantics of our framework. Finally, in §5 we offer our conclusions.

2 Background

2.1 Syntax

The language of a propositional logic has an alphabet consisting of

(i) proposition symbols: $p_0, p_1, ...$
(ii) connectives : $\vee, \wedge, \leftarrow, \neg, \bot, \top$
(iii) auxiliary symbols : (,).

Where $\vee, \wedge, \leftarrow$ are 2-place connectives, $\neg$ is 1-place connective and $\bot, \top$ are 0-place connectives. The proposition symbols and $\bot$ stand for the indecomposable propositions, which we call [illegible], or [illegible]. A literal is an atom, a, or the negation of an atom $\neg a$. The [illegible] of a literal is defined as $(a)^c = \neg a$ and $(\neg a)^c = a$.

A general clause, [illegible], is denoted:

$$l_1 \vee \ldots \vee l_m \leftarrow l_1, \ldots, l_j, not\ l_{j+1}, \ldots, not\ l_n$$ [2]

where $m \geq 0$, $n \geq 0$, each l_i is a literal. When $n = 0$ and $m > 0$ the clause is an abbreviation of $l_1 \vee \ldots \vee l_m \leftarrow \top$[3], where $\top$ is $\neg\bot$. When $m = 0$ the clause is an abbreviation of $\bot \leftarrow l_1, \ldots, l_n$[4]. Clauses of theses forms are called constraints (the rest, non-constraint clauses). A logic program is finite set of general clauses.

A signature $\mathcal{L}$ is a finite set of elements that we call atoms. By $\mathcal{L}_P$ we understand it to mean the signature of P, i.e. the set of atoms that occurs in P. Given a signature $\mathcal{L}$, we write $Prog_{\mathcal{L}}$ to denote the set of all programs defined over $\mathcal{L}$. A general semantics [illegible] is a function on $Prog_{\mathcal{L}}$ which associates with every program a partial interpretation.

We point out that we understand the negation $\neg$ as the classical negation and the negation *not* as the negation as failure [5].

2.2 Lattices and Order

In this section, we present some fundamental definitions of lattice theory in order to make this paper self contained (see [2] for more details).

Definition 1. [illegible] $\leq$ [illegible] $x, y, z \in P$

(i) $x \leq x$
(ii) $x \leq y$ [illegible] $y \leq x$ [illegible] $x = y$
(iii) $x \leq y$ [illegible] $y \leq$ [illegible] $x \leq z$

[illegible]

A set P equipped with an order relation $\leq$ is said to be an order set (or partial ordered set).

Definition 2. [illegible] $S \subseteq P$ [illegible] $x \in P$ [illegible] $s \leq x$ [illegible] $s \in S$ [illegible] S^u [illegible] S^l [illegible]

If S^u has a least element x, then x is called the least upper bound (LUB) of S. Equivalently, x is the least upper bound of S if

[2] $l_1, \ldots, l_n$represents the formula $l_1 \wedge \ldots \wedge l_n$.
[3] Or simply $l_1 \vee \ldots \vee l_m$.
[4] In fact $\bot$ is used to define $\neg A$ as $A \rightarrow \bot$.

(i) x is an upper bound of S, and
(ii) $x \leq y$ for all upper bound y of S.

The least upper bound of S exists iff there exists $x \in P$ such that

$$(\forall y \in P)[((\forall s \in S)s \leq y) \Longleftrightarrow x \leq y],$$

and this characterizes the LUB of S. Dually, if S^l has a greatest element, x, then x is called the greatest lower bound (GLB) of S. Since least element and greatest elements are unique, LUB and GLB are unique when they exist.

The least upper bound of S is called the supremum of S and is denoted by $sup\ S$; the greatest lower bound of S is also called the infimum of S is denoted by $inf\ S$.

Definition 3. [illegible] P [illegible]

(i) [illegible] $sup\{x, y\}$ [illegible] $inf\{x, y\}$ [illegible] $x, y \in P$ [illegible]
(ii) [illegible] $sup\ S$ [illegible] $inf\ S$ [illegible] $S \subseteq P$ [illegible]

[illegible] Let us consider the set of labels $S := \{$[illegible]$\}$ and let $\preceq$ be a partial order such that the following set of relations holds :

$\{Certain \preceq Confirmed, Confirmed \preceq Probable,$
$Confirmed \preceq Plausible, Probable \preceq Support,$
$Plausible \preceq Supported, Supported \preceq Open\}$

A graphic representation of S according to $\preceq$ is showed in Figure 1. It is not difficult to see that S is a lattice and even more it is complete lattice.

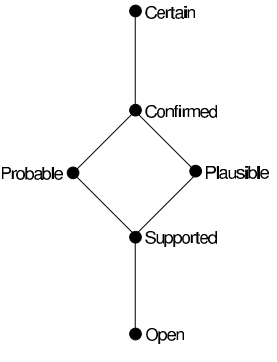


Fig. 1. A lattice

The labels given in Example 1 could be qualifiers of a knowledge base.

3 Modality Argumentation Programming

In this section, we are going to present our framework. We start by defining the syntax.

Definition 4 (Modality clause).

$$Modality : C.$$

$Modality \in Q$ C

Notice that by using a complete lattice , a modality clause categorizes the sentence expressed in the general clause . This means that a modality clause locates a sentence in .

We understand a modality as a category of certain meaning having to do with the expression of possibility. Therefore, a set of possibilities could be categorized by a complete lattice. For instance, let $S := \{$ $\}$ be a set of labels where each label is a possible category of believes, so this set could be categorized as it is shown in Figure 1.

Definition 5 (Modality logic program). $\langle Q, Modality\_Clauses \rangle$ $Modality\_Clauses$ $Modality : C \in Modality\_Clauses$ $Modality \in Q$

Let Q be the lattice presented in Example 1 and let us consider the following propositions atoms which represent medical knowledge for organ transplantation.

- = 'donor has streptococcus viridans endocarditis'.
- = 'recipient infected with streptococcus viridans'.
- = 'non viable'.

One possible modality logic program with its intuitive meaning could be described as follows:

: . (It is confirmed that the donor has been infected by streptococus viridas)

: $\leftarrow$. (If the donor has been infected by streptococus viridas, then it is plausible that the recipient could be infected too.)

: $\leftarrow$. (If it is plausible that the recipient could be infected by streptococus viridas, then it is probable that his organs are not viable for transplanting)

Following the definition of argument presented in [11], we are going to define our concept of argument.

Definition 6 (Argument). $P := \langle Q, Modality\_Clauses \rangle$

$$Arg = \langle Claim, Support, Modality\_Qualifier \rangle$$

$Support$

$Support$
$Support \models_I Claim$
$Support$ $Support$

$Modality\_Qualifier \in Q$

The symbol $\models_I$ denotes logic consequence in Intuitionistic Logic (see [14] for details). Intuitionistic Logic has studied in the context of logic programming, specially in Answer Set Programming, with two kinds of negations [12, 9]. Notice that is minimal set inclusion and is not unique.

By definition, an argument has a . The modality qualifier has the objective of quantify the level of of an argument. There are two kinds of quantifiers: , and . So, we can define two kinds of arguments.

Definition 7 (Pessimistic argument).
$\langle Claim, Support, Modality\_Qualifier \rangle$

$$Modality\_Qualifier := GLB\{Modality\_Qualifier |$$

$$(Modality\_Qualifier : formula) \in Support\}$$

Definition 8 (Optimistic argument).
$\langle Claim, Support, Modality\_Qualifier \rangle$

$$Modality\_Qualifier := LUB\{Modality\_Qualifier |$$

$$(Modality\_Qualifier : formula) \in Support\}$$

Let us consider again the lattice of Example 1 and the modality logic program presented in Example 2. One possible argument is:

$\langle nv, \{(Confirmed : dsve), (Plausible : risv \leftarrow dsve.), (Probable : nv \leftarrow risv.)\}$, **?**$\rangle$

So, a pessimistic argument is:

$\langle nv, \{(Confirmed : dsve), (Plausible : risv \leftarrow dsve.), (Probable : nv \leftarrow risv.)\}$, **Supported**$\rangle$

an optimistic argument:

$\langle nv, \{(Confirmed : dsve), (Plausible : risv \leftarrow dsve.), (Probable : nv \leftarrow risv.)\}$, **Confirmed**$\rangle$

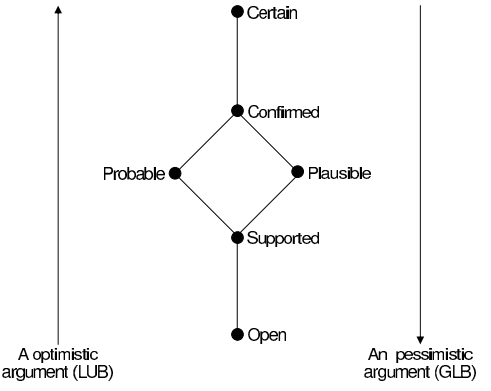


Fig. 2. A lattice of modalities

In the Figure 2 is presented the lattice with the levels of , that an argument could be defined.

Two arguments constructed from two different knowledge bases (two different rational agents) could be in conflict. We are going to define two well known kinds of conflicts: and .

Definition 9. Let $Arg_1 = \langle Claim_1, Support_1, Modality\_Qualifier_1 \rangle$ and $Arg_2 = \langle Claim_2, Support_2, Modality\_Qualifier_2 \rangle$ Arg_1 Arg_2 if $Claim_1 = l$ and $Claim_2 = (l)^c$

Definition 10. Let $Arg_1 = \langle Claim_1, Support_1, Modality\_Qualifier_1 \rangle$ and $Arg_2 = \langle Claim_2, Support_2, Modality\_Qualifier_2 \rangle$ Arg_1 Arg_2 if $Claim_1 = l_i$ and

$$(Modality : l_1 \vee \ldots \vee l_m \leftarrow l_1, \ldots, l_j, not\ l_{j+1}, \ldots, not\ l_i, \ldots, l_n) \in Support_2$$

By considering the concepts of and , we define our concept of .

Definition 11. Let $Arg_1 = \langle Claim_1, Support_1, Modality\_Qualifier_1 \rangle$ and $Arg_2 = \langle Claim_2, Support_2, Modality\_Qualifier_2 \rangle$ Arg_1 Arg_2 if Arg_1 and Arg_2 and $LUB\{Modality\_Qualifier_1, Modality\_Qualifier_2\} = Modality\_Qualifier_1$

Notice that, if Arg_1 defeats Arg_2, then Arg_1's claim has a support with more evidence/certainty that Arg_2.

In order to illustrate those definitions, let us consider the following example.

Let us consider the lattice presented Example 1, and the following proposition atoms:

- = 'donor is HIV positive'.
- = 'organ is viable for transplanting'.
- = 'organ has correct functions and correct structure '.
- = 'positive clinical test'.

Let P_1 be the following modality logic program:

: a. (It is probable that donor is HIV positive)
: $\neg b \leftarrow a, not\ c$. (If donor is HIV positive and there is not evidence that the organ has correct functions and correct structure, then the organ is not viable for transplanting)

One possible argument Arg_1 from P_1 is :

$$\langle \neg b, \{(Probable : a.), (Probable : \neg b \leftarrow a, not\ c.)\}, Probable \rangle$$

This argument suggests that the organ is not viable for transplanting ($\neg b$). Now, let P_2 be the following modality logic program:

: q. (It is confirmed that the organ has positive clinical tests)
: $c \leftarrow q$. (If the organ has positive clinical tests, then it is plausible that the organ has correct functions and correct structure.)

One possible argument Arg_2 from P_2 is :

$$\langle c, \{(confirmed : q.), (Plausible : c \leftarrow q)\}, Confirmed \rangle$$

One can see that Arg_2 undercuts Arg_1 and even more Arg_2 defeats Arg_1 because $LUB\{Probable, Confirmed\} = Confirmed$. So, one can not say explicitly that an organ is not viable for transplanting ($\neg b$).

This example is controversial in the medical domain, because usually an organ from a donor who is HIV positive is not viable for transplanting. However, there are cases where the recipient is also HIV positive then he could be receptor of an acceptable organ from a donor HIV positive.

4 Declarative Semantics

In this section, we are going to present the declarative semantics for our framework. This semantics is characterized in two parts. The first part determines the models of the modality logic program without considering the modality qualifier and the second one determines the modality qualifiers of the arguments using aggregate operations which are implemented by negation as failure.

Definition 12. $\Delta(P)$

$$\Delta(P) := \{C | (Modality : C) \in P\}.$$

Definition 13. $\Gamma(S)$

$$\Gamma(S) := \{Qualifier(Modality) | (Modality : C) \in S\}$$

The following two definitions are similar to Definition 4.11 of [7].

Definition 14 (GLB_basic_ext). $\preceq$

$$\begin{aligned}
&f_{\preceq}(X) \leftarrow Qualifier(X)\\
&f_{\preceq}(X) \leftarrow f_{\preceq}(X1), X1 \prec X\\
&f_{\prec}(X) \leftarrow f_{\preceq}(X1), X1 \prec X\\
&f_{=}(X) \leftarrow f_{\preceq}(X), \neg f_{\prec}(X)\\
&f_{\preceq}(Z) \leftarrow f_{\preceq}(X), f_{\preceq}(Y), GLB_{\preceq}(X, Y, Z)
\end{aligned}$$

Definition 15 (LUB_basic_ext). $\preceq$

$$\begin{aligned}
&f_{\preceq}(X) \leftarrow Qualifier(X)\\
&f_{\preceq}(X) \leftarrow f_{\preceq}(X1), X1 \succ X\\
&f_{\prec}(X) \leftarrow f_{\preceq}(X1), X1 \succ X\\
&f_{=}(X) \leftarrow f_{\preceq}(X), \neg f_{\prec}(X)\\
&f_{\preceq}(Z) \leftarrow f_{\preceq}(X), f_{\preceq}(Y), LUB_{\preceq}X, Y, Z)
\end{aligned}$$

Now, we present how to build arguments from a modality logic program considering its models.

Definition 16. $M(P) := SEM(\Delta(P))$ $Arg := \langle Claim, Support, Modality\_Qualifier\rangle$ $Claim \in M(P)$ $Modality\_Qualifier := modality$

$$f_{=}(modality) \in WFS(\Gamma(Support) \cup GLB\_basic\_ext)$$

$$f_{=}(modality) \in WFS(\Gamma(Support) \cup LUB\_basic\_ext)$$

[5]

In order to illustrate the above definitions, we present the following example.

Let us consider again the lattice of Example 1 and the following proposition atoms which represent, like in Example 2, medical knowledge.

[5] See [4] for a formal definition of the well-founded semantics.

- [illegible] = 'donor is brain-dead'
- [illegible] = 'discard metastatic abscess'
- [illegible] = 'determine bacteria causing endocarditis'
- [illegible] = 'bacteria is streptoccocus viridians'

Let Π be the following modality logic program:

[illegible]: $dbd.$ (The donor is brain-dead)
[illegible]: $dma \vee dbce \leftarrow dbd.$ (It is probable that if a donor is brain-dead, then it is discarded a metastatic abscess or there is a bacteria causing endocarditis)
[illegible]: $dbce \leftarrow bsv.$ (It is confirmed that if a donor has been infected by streptoccocus viridians, then it is diagnosed endocarditis)

Then

$$\Delta(\Pi) := \{(dbd.), (dma \vee dbce \leftarrow dbd.), (dbce \leftarrow bsv)\}$$

In this example, we consider SEM(P) as the stable models semantics [5]. Let us consider the stable models of $\Delta(\Pi)$ which are $\{dbd, dma\}$,$\{dbd, dbce\}$. This means that we can construct three different arguments:

1. $\langle dbd, \{(Confirmed : dbd.)\}, Qualifier_1\rangle$
2. $\langle dbce, \{(Confirmed : dbd.), (Probable : dma \vee dbce \leftarrow dbd)\}, Qualifier_2\rangle$
3. $\langle dma, \{Confirmed : dbd.\}, (Probable : dma \vee dbce \leftarrow dbd)\}, Qualifier_3\rangle$

These arguments have not defined their modality quantifiers yet. Let us consider the support of Argument 2. $S := \{(Confirmed : dbd.), (Probable : dma \vee dbce \leftarrow dbd)\}$, so

$$\Gamma(S) := \{(Qualifier(Confirmed).), (Qualifier(Probable).)\}$$

By considering $WFS(\Gamma(S) \cup GLB\_basic\_ext)$, we can infer the pessimist modality qualifier of Argument 2, it is not difficult to see that $f_{=}(Probable) \in WFS(\Gamma(S) \cup GLB\_basic\_ext)$, this means that $Qualifier_2 := Probable$.

$$\langle dbce, \{(Confirmed : dbd.), (Probable : dma \vee dbce \leftarrow dbd)\}, Probable\rangle$$

This means, that in this context, we have an argument that suggests that if an donor is brain-dead, then it is [illegible] that he could be infected by a bacteria which is causing endocarditis.

Notice that, the use of disjunctive clauses allows to build arguments under incomplete information and also the quantification of the knowledged base permits to quantify the arguments. By using this kind of arguments, it is possible to support decisions taken under incomplete information.

5 Conclusions and Future Work

In this work we introduced an argumentation framework which allows to incorporate modalities during the process of argumentation reasoning. We understand a modality as a category of certain meaning having to do with the expression of possibility. Therefore, a set of possibilities could be categorized by a complete lattice.

Our argumentation framework is based in a specification language which permits to provide specifications with levels of , , in a natural way. Also, the specification language allows to use disjunctive clauses, so it allows specifications in situations where the available information is incomplete, as in the medical domain showed in the examples. The declarative semantics of our language permits to build arguments such that any argument is supported by a set of modality clauses and the argument's claim is quantified , its support. We present a couple of examples from our real application to manage the assignation process of human organs for transplantation [1, 13], although the examples are simple they permit to see the potential of our framework.

Among the future work, we have planned to deploy this framework in the context of multi-agent systems, in particular to our real multi-agent system called CARREL [15].

Acknowledgements

Thanks to Mauricio Osorio for his fruitful commends and discussions. We thank Antonio López-Navidad for his assessment to formulate the clinical information related with examples. Also we would like to thank the anonymous referees for their comments.

This research was partially supported by the grant FP6-IST-002307 (ASPIC). The views expressed in this paper are not necessarily those of ASPIC consortium.

References

1. U. Cortés, P. Tolchinsky, J. C. Nieves, A. López-Navidad, and F. Caballero. Arguing the discard of organs for tranplantation in CARREL. In *CATAI 2005*, pages 93–105, 2005.
2. B. A. Davey and H. A. Priestly. *Introduction to Lattices and Order*. Cambridge University Press, second edition, 2002.
3. M.-W. O. Diccionary. http://www.m-w.com/.
4. A. V. Gelder, K. A. Ross, and J. S. Schlipf. The well-founded semantics for general logic programs. *Journal of the ACM*, 38(3):620–650, 1991.
5. M. Gelfond and V. Lifschitz. Classical Negation in Logic Programs and Disjunctive Databases. *New Generation Computing*, 9:365–385, 1991.
6. J. C. Nieves and B. Jayaraman. Relaxation in partial order programming. In *Workshop on Logic and Computation: MICAI'2002*, 2002.
7. M. Osorio and B. Jayaraman. Relating aggregation and negation-as-failure. *New Generation Comput.*, 17(3):255–284, 1999.

8. M. Osorio, B. Jayaraman, and D. A. Plaisted. Theory of partial-order programming. *Sci. Comput. Program.*, 34(3):207–238, 1999.
9. M. Osorio, J. A. Navarro, and J. Arrazola. Applications of Intuitionistic Logic in Answer Set Programming. *Theory and Practice of Logic Programming (TPLP)*, 4:325–354, May 2004.
10. M. Osorio, J. C. Nieves, and B. Jayaraman. Aggregation in functional query languages. *Journal of Functional and Logic Programming*, 2004(2), August 2004.
11. S. Parsons and P. McBurney. Argumentation-based dialogues for agent coordination. *Group Decision and Negotiation*, 12(5):415–439, 2003.
12. D. Pearce. Stable Inference as Intuitionistic Validity. *Logic Programming*, 38:79–91, 1999.
13. P. Tolchinsky, U. Cortés, J. C. Nieves, A. López-Navidad, and F. Caballero. Using arguing agents to increase the human organ pool for transplantation. In *Proc. of the Third Workshop on Agents Applied in Health Care (IJCAI 2005)*, 2005.
14. D. van Dalen. *Logic and structure*. Springer-Verlag, Berlin, 3rd., aumented edition edition, 1994.
15. J. Vázquez-Salceda, U. Cortés, J. Padget, A. López-Navidad, and F. Caballero. Extending the CARREL System to mediate in the organ and tissue allocation processes: A first approach. *Artificial Intelligence in Medicine*, 3:233–258, 2003.
16. K. von Fintel. Modality and language. In D. M. Borchert, editor, *Encyclopedia of Philosophy – Second Edition*. MacMillan, 2005.

Feature Selection in SVM Based on the Hybrid of Enhanced Genetic Algorithm and Mutual Information

Chunkai Zhang and Hong Hu

Department of Mechanical Engineering and Automation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, China
ckzhang@hotmail.com

Abstract. Feature selection is a well-researched problem, which can improve the network performance and speed up the training of the network. In this paper, we proposed an effective feature selection scheme for SVM using the hybrid of enhanced genetic algorithm and mutual information, in which mutual information between each input and each output of the data set is employed in mutation in evolutionary process to purposefully guide search direction based on some criterions. In order to avoid the noise fitness evaluation, in evaluating the fitness of an input subset, a SVM should adaptively adjust its parameters to obtain the best performance of network, so an enhanced GA is used to simultaneously evolve the input features and the parameters of SVM. By examining two real financial time series, the simulation of three different methods of feature selection shows that the feature selection using the hybrid of GA and MI can reduce the dimensionality of inputs, speed up the training of the network and get better performance.

1 Introduction

Unlike most of the traditional methods, support vector machine (SVM) implements the Structural Risk Minimization Principal which seeks to minimize an upper bound of the generalization error rather than minimize the training error [1]. This eventually results in remarkable characteristics such as good generalization performance, the absence of local minima and the sparse representation of solution.

In the modeling, all possible inputs to a SVM, include relevant features and irrelevant features, can be quite large. According to the definitions of John, Kohavi and Pfleger [2], irrelevant features do not participate in defining the unknown concepts, weakly relevant features possess redundant information and can be eliminated if other features subsuming this information are included, and strongly relevant features are indispensable. So selecting significant features from all candidate features as an optimum input subset (feature selection) is necessary to decrease the size of network, speed up the training and improve the generalization performance of a SVM. In the framework of SVM, several approaches for feature selection are available [3,4,5,6]. In [3, 4], feature selection using mutual information between input and output to select features for SVM, in which 1) it assumes that the inputs are essentially independent and that no output is in fact a complex function of two or more of the input variables;

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 307 – 316, 2006.

2) the objective function disregards SVM with which the selected features are to be used. So there is no guarantee that the optimal subset of features will be processed in the optimal way by the learning algorithm and by the operating SVM [7]. In [5, 6], genetic algorithms are proved to be useful for selecting important features in SVM, in which the objective function for selection is a measure of SVM performance. Although it is considered superior because they incorporate the inductive bias of the SVM, they typically involve expensive search routines.

In this paper, we address both these shortcomings simultaneously. The proposed method employs the mutual information (MI) between input and output in mutation in GA to purposefully guide the evolutionary search direction based on some criterions for SVM, which can suggest substantial gains in efficiency. In order to avoid the noise fitness evaluation [14], in evaluating the fitness of an input subset, a SVM should adaptively adjust its parameters to obtain the best performance of network, which represents the fitness of this input subset, so an enhanced GA is used to simultaneously evolve the input features and the parameters of SVM. By examining two real financial time series, the simulation of three different methods of feature selection shows that the feature selection using the hybrid of the enhanced GA and MI can improve the network performance, reduce the network complexity, and speed up the training of the network.

The rest of this paper is organized as follows. Section 2 describes SVM, mutual information (MI) and GA. Section 3 the hybrid of enhanced GA and MI is used to evolve an optimum input subset for a SVM. Section 4 presents experimental results in two financial time series problem. The paper is concluded in Section 5.

2 Background

2.1 The Description of SVM

Support vector machine (SVM) can trace their roots back to statistical learning theory, as introduced by Vapnikin the late 1960s [8]. On the basis of the VC dimension concept, constructive distribution-independent bounds on the rate of convergence of learning processes can be obtained and the structural risk minimization principle has been found. SVM has high generalization ability and is capable of learning in high-dimensional spaces with a small number of training examples.

SVM nonlinearly maps inner product of feature space to the original space via a kernel. Training SVM is equivalent to solving a linearly constrained quadratic programming, the solution of SVM is unique global, and it is only dependent on a small subset of training data points which are referred to as support vectors.

To control generalization capability of SVM, there are a few free parameters like limiting term C and the kernel parameters like RBF width σ. Before training the network, these parameters should be decided. The numerical implementation of SVM is mainly based on QP with options of decomposing a large-scale QP problems into a series of smaller-size QP problems [9]. In the present work, the SVM is trained using an adapted version of decomposition methods and working set selection strategies similar to that of Joachims [10].

2.2 Mutual Information

In the information theory founded by Shannon [11], the uncertainty of a random variable C is measured by entropy $H(C)$. For two variables X and C, the conditional entropy $H(C \mid X)$ measures the uncertainty about C when X is known, and MI, $I(X;C)$, measures the certainty about C that is resolved by X. Apparently, the relation of $H(C)$, $H(C \mid X)$ and $I(X;C)$ is:

$$H(C) = H(C \mid X) + I(X;C) \tag{1}$$

or, equivalently,

$$I(X;C) = H(C) - H(C \mid X),$$

As we know, the goal of training classification model is to reduce the uncertainty about predictions on class labels C for the known observations X as much as possible. In terms of the mutual information, the purpose is just to increase MI $I(X;C)$ as much as possible, and the goal of feature selection is naturally to achieve the higher $I(X;C)$ with the fewer features.

With the entropy defined by Shannon, the prior entropy of class variable C is expressed as

$$H_s(C) = -\sum_{c \in C} P(c) \log P(c) \tag{2}$$

where $P(c)$ represents the probability of C, while the conditional entropy is $H(C \mid X)$ is

$$H_s(C \mid X) = -\int_x p(x)(\sum_{c \in C} p(c \mid x) \log p(c \mid x))dx \tag{3}$$

The MI between X and C is

$$I_s(X;C) = -\sum_{c \in C} \int_x p(c,x) \log \frac{p(c,x)}{P(c)p(x)} dx \tag{4}$$

Mutual information can, in principle, be calculated exactly if the probability density function of the data is known. Exact calculations have been made for the Gaussian probability density function. However, in most cases the data is not distributed in a fixed pattern and the mutual information has to be estimated. In this study, the mutual information between each input and each output of the data set is estimated using Fraser & Swinney's method [12].

The mutual information of independent variables is zero, but is large between two strongly dependent variables with the maximum possible value depending on the size

of the data set. And this assumes that all the inputs are independent and that no output is in fact a complex function of two or more of the input variables.

Although the MI-based feature selection techniques are widely used [3,4,13], they suffer from many limitations. Firstly, their objective functions disregard the classifier with which the selected features are to be used. So there is no guarantee that the optimal subset of features will be processed in the optimal way by the learning algorithm and by the operating classifier. Secondly, the selected feature subset cannot be guaranteed optimal. For example, selecting a fixed number of inputs from a ranked list consisting of combinations along with single entries is somewhat problematical, and once a feature is added at an early step, it cannot be removed although it may not constitute the best subset of features in conjunction with the later selected features. Finally, there are a number of parameters that need to be set a priori. For example, the number of features added or removed, the significance level for selecting features and the final feature size.

2.3 Genetic Algorithm

GA is an efficient search method due to its inherent parallelism and powerful capability of searching complex space based on the mechanics of natural selection and population genetics. Because the problem of feature selection can be formulated as a search problem to find a nearoptimal input subset, so the artificial intelligence techniques, such as genetic algorithm (GA), is used to selects the optimal subset of features.

The method of using GA to select input features in the neural network is straightforward. In GA, every candidate feature is mapped into individual (binary chromosomes) where a bit "1" (gene) denotes the corresponding feature is selected and a bit of "0" (gene) denotes the feature is eliminated. Successive populations are generated using a breeding process that favors fitter individuals. The fitness of an individual is considered a measure of the success of the input vector. Individuals with higher fitness will have a higher probability of contributing to the offspring in the next generation ('Survival of the Fittest').

There are three main operators that can interact to produce the next generation. In replication individual strings are copied directly into the next generation. The higher the fitness value of an individual, the higher the probability that that individual will be copied. New individuals are produced by mating existing individuals. The probability that a string will be chosen as a parent is fitness dependent. A number of crossover points are randomly chosen along the string. A child is produced by copying from one parent until a crossover point is reached, copying then switching to the other parent and repeating this process as often as required. An N bit string can have anything from 1 to N-1 crossover points. Strings produced by either reproduction or crossover may then be mutated. This involves randomly flipping the state of one or more bits. Mutation is needed so new generations are more than just a reorganization of existing genetic material. After a new generation is produced, each individual is evaluated and the process repeated until a satisfactory solution is reached. The procedure of GA for feature selection is expressed as follows:

Procedure of genetic Algorithm for feature selection

Initialization

$N \rightarrow$ *Population size*

$P \rightarrow$ *Initial population with* N *subsets of* Y

$P_c \rightarrow$ *Crossover probability*

$P_m \rightarrow$ *Mutation probability*

$T \rightarrow$ *Maximum number of generations*

$k \rightarrow 0$

Evolution

Evaluation of fitness of P

while ($k < T$ *and* P *does not converge) do*

Breeder Selection

Crossover with P_c

Mutation with P_m

Evaluation of fitness of P *Replication*

Dispersal

$k+1 \rightarrow k$

In contrast with MI, GA is more attractive, because it incorporate the inductive bias of the SVM and the search process involves no user selectable parameters, such as the final feature size and the signification level etc.. In addition, it has the potential to simultaneously evolve the input features and the parameters of SVM. But GA typically involves expensive search routines.

3 The Proposed Method

3.1 The Feature Selection Problem for a SVM

We consider the general feature selection problem for a SVM:

$$f_{G^*}^* = \min_{G:|G|=K} \min_{f \in H_G} [\frac{1}{M} \sum_{i=1}^{M} (y_i - y_{ri})^2]^{1/2} \tag{5}$$

where y_r is the desired target value, and y is the output of network. G is a subset of features, i.e. $G \in \{x_1, x_2, ..., x_N\}$. f is a function $f: X \rightarrow Y$. H_G denotes the restriction of H on G, i.e. the class of function in H that map G to Y, here $Y = \{y_1, y_2, ..., y_N\}$. The SVM learning framework aims to 1) discover a function $f: X \rightarrow Y$, from a hypothesis H class of functions; 2) select K features out of N. In this paper, we choose the *RBF* kernel function in SVM.

3.2 The Hybrid of GA and Mutual Information

In the proposed method, we use mutual information between each candidate input and each output to guide the mutation in GA. First, using Fraser & Swinney's method, the mutual information x_i between each candidate input and each output is estimated. In order to reduce time of calculating the mutual information between single input and output in the whole data set, we randomly select some data from data set with probability 0.5 to construct a data set named *MI* set, and then calculate mutual information in this data set. So data set $D = \{x_i, i = 1,..., N\}$ is constructed, where x_i represents the mutual information of i th candidate input, and N means there are N candidate inputs.

Then calculate the mathematical statistics of x_i: the mean $\bar{x}$ and standard deviation s_N

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{6}$$

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{7}$$

And define the three sets which satisfy $D = D_1 \bigcup D_2 \bigcup D_3$:

$$D_1 = \{x_i \mid x_i - \bar{x} > \frac{s_N}{2}\},$$

$$D_2 = \{x_i \mid -\frac{s_N}{2} \le x_i - \bar{x} \le \frac{s_N}{2}\}$$

$$D_3 = \{x_i \mid x_i - \bar{x} < -\frac{s_N}{2}\}$$

In GA, we use mutual information between each candidate input and each output to guide the mutation based on some criterions, as follows:

$$g_i = \begin{cases} 1 & x_i \in D_1 \\ 0 & x_i \in D_3 \\ rand & x_i \in D_2 \end{cases} \tag{8}$$

where g_i represents i th gene in a binary chromosome, it means i th candidate input. If the mutual information x_i of i th candidate input belongs to D_1, it means it is a strongly relevant input for each output, so include it into input feature subset; if the mutual information x_i of i th candidate input belongs to D_2, it means it is a weakly

relevant input for each output, so randomly include it into input feature subset; If the mutual information x_i of i th candidate input belongs to D_3, it means it is almost irrelevant input for each output, so exclude it from input feature subset.

3.3 An Enhanced GA

In order to avoid the noise fitness evaluation [14], in evaluating the fitness of an input subset, a SVM should adaptively adjust its parameters to obtain the best performance of network, which represents the fitness of this input subset. In order to accurately evaluate the fitness of the candidate input subset, the candidate input subset and the parameters of SVM should not be synchronously adjusted. Here, we propose an enhance GA to select the most suitable features and the parameters related to SVM as follows:

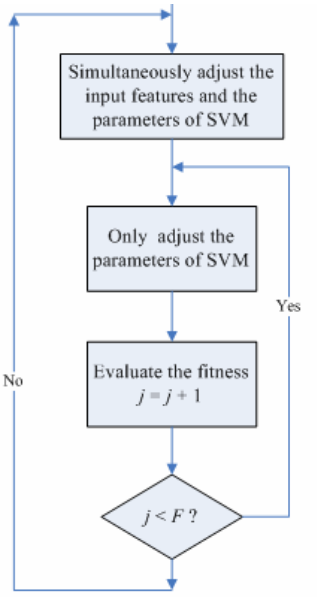


Fig. 1. The enhance GA

In an enhanced GA, the parameter F reperesents the iterative number of only evolving the parameters of SVM with no change of input features.

In a training run, needing *K* different inputs to be selected from a set of *N* possible inputs, the genome string would consist of (*K*+1) bits. The first *K* bits ($d_i, i = 1,2,...,K$) in the genome represent the candidate inputs for SVM which are constrained to be in the range {0, 1}; the d_{K+1} is a real number which represents a parameter of SVM, the RBF kernel width (σ) which has to be within the range $l_{\min} \le d_{K+1} \le l_{\max}$. The parameters $l_{\max}$ and $l_{\min}$ represent, respectively, the lower and the upper bounds on the SVM parameter.

4 Experimental Studies

Two financial futures contracts collated from the Chicago Mercantile Market are examined, which are the Standard&Poor 500 stock index futures (CME-SP) and United Sates 30-year government bond (CBOT-US) [5]. The daily closing prices are used as the data set. The original closing price is transformed into a five-day relative difference in percentage of price (RDP). The input variables are constructed from 14 lagged RDP values based on 5-day periods ($x_4, x_5, ..., x_{17}$) and 3 lagged transformed closing prices which is obtained by subtracting a 15-day exponential moving average from the closing price (x_1, x_2, x_3). All the data points are scaled into the range of [–0.9, 0.9] as the data points include both positive values and negative values. There are a total of 907 data patterns in the training set, and 400 data patterns in the test set in all the data sets.

The RBF kernel is used for SVM, and the kernel parameters σ should be evolved due to different features of the input subsets.

Three different methods of feature selection: SA, GA and the proposed method, are respectively applied in CBOT-SP and CBOT-US. In GA and the proposed method, N =50, T =60, P_c =0.6, P_m =0.02 and F =5. And a SVM with *K* inputs is trained using an adapted version of decomposition methods and working set selection strategies similar to that of Joachims [10]. And the fitness function is defined to be $1/RMSE$, and the root mean square error (*RMSE*) is calculated by

$$RMSE = (\sum (Y - Y_r)^2)^{1/2} \tag{9}$$

where Y_r is the desired target value, and Y is the output of network. And the fitness value is defined to be $1/RMSE$.

After selecting an optimal input subset using one of the above techniques, these inputs were assessed by means of an evaluation SVM whose architecture was chosen based on initial experiments. And a SVM with *K* inputs is trained using an adapted version of decomposition methods and working set selection strategies similar to that of Joachims [10]. The overall performance of this testing network was assumed to reflect the appropriateness of this particular selection of inputs.

The performance of SVM using the selected features is present in table 1. And the input subsets of features selected by SA, GA and the proposed method are indicated in table 2.

Table 1. Result of feature selection in CME-SP and CME-US

Method	CBOT-SP		CBOT-US	
	RMSE	TIME(s)	RMSE	TIME(s)
Full features	0.9778	*	1.1620	*
GA	0.8931	933	1.0685	274
SA	0.8410	425	1.0422	341
The proposed method	0.8297	392	1.0316	179

Table 2. Selected features in CME-SP and CME-US

Method	Features (CBOT-SP)	Features (CBOT-US)
GA	$x_1, x_3, x_5, x_6, x_7, x_9, x_{12}, x_{31}, x_{14}, x_{16}, x_{17}$	$x_1, x_2, x_3, x_8, x_{10}, x_{11}, x_{16}$
SA	$x_1, x_2, x_3, x_7, x_{12}$	$x_1, x_2, x_3, x_{10}, x_{12}, x_{16}$
The proposed method	x_1, x_2, x_3, x_{12}	$x_1, x_2, x_3, x_{10}, x_{12}$

Table 1 indicates that the proposed method exhibited better performance than the other techniques at RMSE and training time. In addition, it was found that there was very little increase in performance after 51 generations for the proposed method in CME-SP and 39 generations for the proposed method in CME-US, but 79 generations for GA in CME-SP and 48 generations for GA in CME-US.

Table 2 indicates that although there is considerable similarity between GA and the proposed method there are differences between the inputs selected, and GA and the proposed method selected the different number of inputs, which explains the effect of mutual information. In contrast to SA, the proposed method get small number input features, which explains the better performance at RMSE and training time. In addition, it was found that there are common inputs x_1, x_3, x_{12} in CME-SP and x_1, x_2, x_3, x_{10} in CME-US for three techniques, it means that they are strongly relevant to the output.

5 Conclusion

We proposed an effective feature selection scheme for SVM using the hybrid of the enhanced genetic algorithm and mutual information, in which mutual information between each input and each output of the data set is employed in mutation in evolutionary process to purposefully guide search direction based on some criterions. In order to avoid the noise fitness evaluation, in evaluating the fitness of an input subset, a SVM should adaptively adjust its parameters to obtain the best performance of network, so an enhanced GA is proposed to simultaneously evolve the input features and the parameters of SVM. By examining a real financial time series forecasting, the simulation of three different methods of feature selection shows that the feature selection using the hybrid of GA and MI can reduce the dimensionality of inputs, speed up the training of the network and get better performance.

References

[1] 1. V. N. Vapnik: The Nature of Statistical Learning Theory. Springerp-Verlag, New York, (1995).

[2] G. H. John, R. Kohavi, and K. Pfleger, “Irrelevant features and the subset selection problem,” in Proc. 11th Int. Conf. Machine Learning, San Mateo, CA, pp. 121–129, 1994.

[3] Taira, Hirotoshi: Feature selection in SVM text categorization. Proceedings of the National Conference on Artificial Intelligence, p 480-486, 1999.

[4] D. Huang,Tommy W.S. Chow: Effective feature selection scheme using mutual information. Neurocomputing vol. 63, pp. 325 – 343, (2005).
[5] F.E.H. Tay, L.J. Cao, A comparative study of saliency analysis and genetic algorithm for feature selection in support vector machines, Intell. Data Anal. 5 (3) 191–209, (2001).
[6] B. Samanta, K.R. Al-Balushi, S.A. Al-Araimi: Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection. Engineering Applications of Artificial Intelligence, Vol.16, 657-665, (2003).
[7] R. Battiti, "Using mutual information for selecting features in supervised neural net training," IEEE Trans. Neural Networks, vol. 5, pp. 537–550, (1994).
[8] V. N. Vapnik: On the uniform convergence of relative frequencies of events to their probabilities. Soviet Mathematics: Doklady 9, 915–918, (1968).
[9] Campbell, C.: Algorithmic approaches to training support vector machines: a survey. Proceedings of ESANN2000 Bruges, Belgium, Paper #ESANN2000-355, (2000).
[10] Joachims, T.: Making large-scale SVM learning practical. In: Scholkopf, B., Burges, C.J., Simola, A. (Eds.), Advances in Kernel Methods-Support Vector Learning. MIT Press, Cambridge, MA, USA, pp. 169-184, 1999.
[11] T.M. Cover, J.A. Thomas: Elements of Information Theory, Wiley, New York, 1991.
[12] A.M. Fraser & H.L. Swinney: Independent Coordinates for Strange Attractors from Mutual Information. Physical Review A, Vol. 33/2, pp 1134 – 1140. 1986.
[13] Belinda Choi, Tim Hendtlass, and Kevin Bluff: A Comparison of Neural Network Input Vector Selection Techniques, LNAI 3029, pp. 1-10, Springer-Verlag Berlin Heidelberg (2004).
[14] X. Yao. A review of evolutionary artificial neural networks. Int. J. Intell. Syst., vol. 8, no. 4, pp. 539-567, 1993.

Improving Fuzzy Rule-Based Decision Models by Means of a Genetic 2-Tuples Based Tuning and the Rule Selection*

R. Alcalá[1], J. Alcalá-Fdez[1], F. J. Berlanga[2], M. J. Gacto[1], and F. Herrera[1]

[1] University of Granada, Dept. Computer Science and A.I., E-18071 Granada, Spain
{alcala, jalcala, herrera}@decsai.ugr.es, mjgacto@ugr.es
[2] University of Jaén, Dept. Computer Science, E-23071 Jaén, Spain
berlanga@ujaen.es

Abstract. The use of knowledge-based systems can represent an efficient approach for system management, providing automatic control strategies with Artificial Intelligence capabilities. By means of Artificial Intelligence, the system is capable of assessing, diagnosing and suggesting the best operation mode. One important Artificial Intelligence tool for automatic control is the use of fuzzy logic controllers, which are fuzzy rule-based systems comprising the expert knowledge in form of linguistic rules. These rules are usually constructed by an expert in the field of interest who can link the facts with conclusions. However, this way to work sometimes fails to obtain an optimal behavior. To solve this problem, within the framework of Machine Learning, some artificial intelligence techniques could be applied to enhance the controller behavior.

In this work, a post-processing method is used to obtain more compact and accurate fuzzy logic controllers. This method combines a new technique to perform an evolutionary lateral tuning of the linguistic variables with a simple technique for rule selection (that removes unnecessary rules). To do so, the tuning technique considers a new rule representation scheme by using the linguistic 2-tuples representation model which allows the lateral variation of the involved linguistic labels.

1 Introduction

The use of knowledge-based systems can represent an efficient approach for system management, providing automatic control strategies with Artificial Intelligence capabilities. By means of Artificial Intelligence, the system is capable of assessing, diagnosing and suggesting the best operation mode. One important Artificial Intelligence tool for automatic control is the use of Fuzzy Logic Controllers (FLCs). FLCs are Fuzzy Rule-Based Systems (FRBSs) comprising the expert knowledge in form of linguistic rules. These rules are usually constructed by an expert in the field of interest who can link the facts or evidence with conclusions. When a real-world situation is presented to the computer, it can

* Supported by the Spanish Ministry of Science and Technology under Project TIC-2002-04036-C05-01.

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 317–328, 2006.

use these rules to draw conclusions based on different situations in the way an expert would. However, this way to work sometimes fails to obtain an optimal behavior. To solve this problem, within the framework of Machine Learning, some Artificial Intelligence techniques could be successfully applied to enhance the controller behavior.

Recently, to improve the behavior of FRBSs, a new linguistic rule representation model was proposed to perform a genetic lateral tuning of Membership Functions (MFs) [2]. This new approach was based on the linguistic 2-tuples representation [10], that allows the symbolic translation of a label by considering an unique parameter per label. It involves a reduction of the search space that eases the derivation of optimal models respect to the classical tuning.

On the other hand, rule selection methods directly select a subset of rules from a given fuzzy rule set in order to minimize the number of rules, maintaining the system performance [9, 12, 13, 14]. The combination of the lateral tuning with rule selection methods can present a positive synergy, reducing the tuning search space, easing the system readability and even improving the system accuracy.

In this work, we present a study of how this new tuning approach can be combined with a rule selection method to improve fuzzy rule-based decision models obtained from the experts's experience. To do that, we propose an evolutionary method combining these two approaches to obtain more compact and accurate FLCs. Additionally, we analyze the positive synergy between both techniques, showing its behavior by solving a real-world problem for the control of a Heating, Ventilating and Air Conditioning (HVAC) system.

This paper is arranged as follows. The next section presents the lateral tuning and rule selection techniques. Section 3 describes the evolutionary algorithm for the global lateral tuning and rule selection. Section 4 presents the HVAC system control problem. Section 5 shows an experimental study of the method behavior applied to the HVAC problem. Finally, Section 6 points out some conclusions.

2 Lateral Tuning and Rule Selection

This section introduces the global lateral tuning of MFs and presents the basics and characteristics of the rule selection technique.

2.1 Lateral Tuning of Membership Functions

In [2], a new model of tuning of FRBSs was proposed considering the linguistic 2-tuples representation scheme introduced in [10], which allows the lateral displacement of the support of a label and maintains the interpretability associated to the obtained linguistic FRBSs. This proposal also introduces a new model for rule representation based on the concept of symbolic translation [10].

The symbolic translation of a linguistic term is a number within the interval [-0.5, 0.5), expressing this interval the domain of a label when it is moving between its two adjacent lateral labels (see Figure 1.a). Let us consider a set of labels S

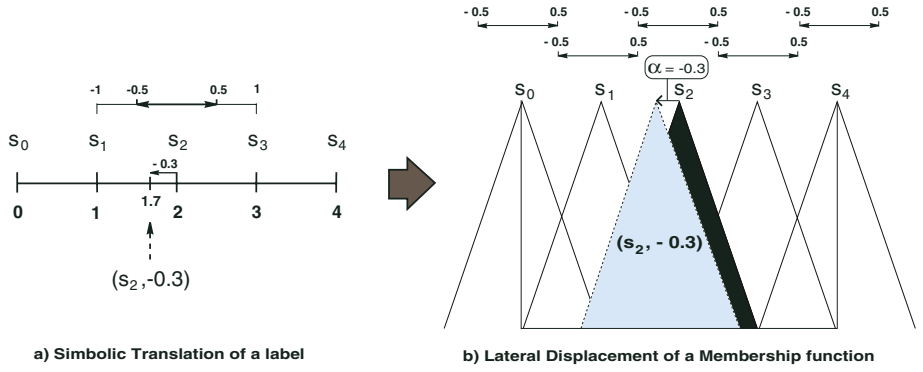


Fig. 1. Symbolic Translation of a Linguistic Label and Lateral Displacement of the involved MF

representing a fuzzy partition. Formally, to represent the symbolic translation of a label in S we have the 2-tuple,

$$(s_i, \alpha_i), \;\; s_i \in S, \; \alpha_i \in [-0.5, 0.5).$$

Actually, the symbolic translation of a label involves the lateral displacement of its associated MF. As an example, Figure 1 shows the symbolic translation of a label represented by the pair $(s_2, -0.3)$ together with the lateral displacement of the corresponding MF.

In the context of the FRBSs, we are going to see its use in the linguistic rule representation. Let us consider a control problem with two input variables, one output variable and a Data Base (DB) defined from experts determining the MFs for the following labels:

$$Error, \; \bigtriangledown Error \rightarrow \{N, Z, P\}, \; Power \rightarrow \{L, M, H\} \; .$$

Based on this DB definition, an example of classical rule and linguistic 2-tuples represented rule is:

If **error** is Zero and $\bigtriangledown$**Error** is Positive then **Power** is High.

If **error** is (Zero, 0.3) and $\bigtriangledown$**Error** is (Positive, -0.2) then **Power** is (High, -0.1).

In [2], two different rule representation approaches were proposed, a global approach and a local approach. In our particular case, the learning is applied to the level of linguistic partitions (global approach). In this way, the pair (X_i, label) takes the same α value in all the rules where it is considered, i.e., a global collection of 2-tuples is considered by all the fuzzy rules. For example, X_i is (High, 0.3) will present the same value for those rules in which the pair "X_i is High" was initially considered. Since the 3 parameters usually considered

per label are reduced to only 1 symbolic translation parameter, this proposal decreases the learning problem complexity easing indeed the derivation of optimal models. Other important issue is that, from the parameters α applied to each label, we could obtain the equivalent triangular MFs, by which a FRBS based on linguistic 2-tuples could be represented as a classical Mamdani FRBS.

2.2 The Rule Selection Technique

Rule set reduction techniques try to minimize the number of rules of a given FRBS while maintain (or even improve) the system performance. To do that, erroneous and conflicting rules that degrade the performance are eliminated, obtaining a more cooperative fuzzy rule set and therefore involving a potential improvement of the system accuracy. Furthermore, in many cases the accuracy is not the only requirement of the model but also the interpretability becomes an important aspect. Reducing the model complexity is a way to improve the system readability, i.e., a compact system with few rules requires a minor effort to be interpreted.

Fuzzy rule set reduction is generally applied as a post-processing stage, once an initial fuzzy rule set has been derived. One of the most known fuzzy rule set reduction techniques is the rule selection. This approach involves obtaining an optimal subset of fuzzy rules from a previous fuzzy rule set by selecting some of them. We may find several methods for rule selection, with different search algorithms that look for the most successful combination of fuzzy rules [9, 12, 13]. In [14], an interesting heuristic rule selection procedure is proposed where, by means of statistical measures, a relevance factor is computed for each fuzzy rule composing the FRBSs to subsequently select the most relevant ones.

These kinds of techniques for rule selection could be easily combined with other post-processing techniques to obtain more compact and accurate FRBSs. In this way, some works have considered the selection of rules together with the tuning of MFs by coding all of them (rules and parameters) in the same chromosome [5, 7]. In this work, we propose the combination of the rule selection with the lateral tuning of MFs.

3 Algorithm for the Lateral Tuning and Rule Selection

To perform the lateral tuning together with the rule selection we consider a Genetic Algorithm (GA) based on the well-known steady-state approach. The steady-state approach [15] consists of selecting two of the best individuals in the population and combining them to obtain two offspring. These two new individuals are included in the population replacing the two worst individuals if the former are better adapted than the latter. An advantage of this technique is that good solutions are used as soon as they are available. Therefore, the convergence is accelerated while the number of evaluations needed is decreased.

In the following, the components needed to design this process are explained. They are: coding scheme and initial gene pool, chromosome evaluation, the genetic operators and a restarting approach to avoid premature convergence.

3.1 Coding Scheme and Initial Gene Pool

To combine the rule selection with the global lateral tuning, a double coding scheme for both [illegible] (C_S) and [illegible] (C_T) is used:

- For the C_S part, the coding scheme generates binary-coded strings of length m (with m being the number of fuzzy rules in the existing FRBS, obtained from expert knowledge). Depending on whether a rule is selected or not, the alleles '1' or '0' will be respectively assigned to the corresponding gene. Thus, the corresponding part C_S^p for the p-th chromosome will be a binary vector representing the subset of rules finally obtained,

$$C_S^p = (c_{S1}^p, \ldots, c_{Sm}^p) \mid c_{Si}^p \in \{0, 1\} \ .$$

- For the C_T part, a real coding is considered, i.e., the real parameters are the GA representation units (genes). This part is the joint of the α parameters of each fuzzy partition. Let us consider the following number of labels per variable: $(m^1, m^2, \ldots, m^n)$, with n being the number of system variables. Then, a chromosome has the following form (where each gene is associated to the tuning value of the corresponding label),

$$C_T = (c_{11}, \ldots, c_{1m^1}, c_{21}, \ldots, c_{2m^2}, \ldots, c_{n1}, \ldots, c_{nm^n}).$$

Finally, a chromosome C^p is coded in the following way:

$$C^p = C_S^p C_T^p \ .$$

To make use of the available information, the FRBS previously obtained from expert knowledge is included in the population as an initial solution. To do so, the initial pool is obtained with first individual having all genes with value '1' in the C_S part and having all genes with value '0.0' in the C_T part. The remaining individuals are generated at random.

3.2 Evaluating the Chromosome

The fitness function depends on the problem being solved (see Section 4.1).

3.3 Genetic Operators

The crossover operator will depend on the chromosome part where it is applied:

- For the C_T part, the BLX-α crossover [6] and a hybrid between a BLX-α and an arithmetical crossover [8] are considered. In this way, if two parents, $C_T^v = (c_{T1}^v, \ldots, c_{Tk}^v, \ldots, c_{Tg}^v)$ and $C_T^w = (c_{T1}^w, \ldots, c_{Tk}^w, \ldots, c_{Tg}^w)$, are going to be crossed, two different crossovers are considered,
  1. Using the BLX-α crossover [6] (with α being a constant parameter chosen by the GA designer), one descendent $C_T^h = (c_{T1}^h, \ldots, c_{Tk}^h, \ldots, c_{Tg}^h)$ is obtained, with c_{Tk}^h being randomly generated within the interval $[I_{L_k}, I_{R_k}] = [c_{min} - I \cdot \alpha, c_{max} + I \cdot \alpha]$, $c_{min} = min(c_{Tk}^v, c_{Tk}^w)$, $c_{max} = max(c_{Tk}^v, c_{Tk}^w)$ and $I = c_{max} - c_{min}$.

2. The application of the arithmetical crossover [8] in the wider interval considered by the BLX-α, $[I_{L_k}, I_{R_k}]$, results in the next descendent:
$$C_T^h \text{ with } c_{Tk}^h = aI_{L_k} + (1-a)I_{R_k},$$
with $a \in [0,1]$ being a random parameter generated each time this crossover operator is applied. In this way, this operator performs the same gradual adaptation in each gene, which is a good characteristic.

– In the C_S part, the standard two-point crossover is used.

Finally, four offspring are generated by combining the two ones from the C_S part with the two ones from the C_T part. The mutation operator flips the gene value in the C_S part and no mutation is considered in C_T part, in order to improve the algorithm convergence. In this way, once the mutation operator is applied over the four offspring obtained from the crossover, the resulting descendents are the two best of these four individuals.

3.4 Restarting Approach

Finally, to get away from local optima, this algorithm uses a restart approach. Thus, when the population of solutions converges to very similar results (practically the same fitness value in all the population), the entire population but the best individual is randomly generated within the corresponding variation intervals. It allows the algorithm to perform a better exploration of the search space and to avoid getting stuck at local optima.

4 The HVAC System Control Problem

In EU countries, primary energy consumption in buildings represents about 40% of total energy consumption and more than a half of this energy is used for indoor climate conditions. On a technological point of view, it is estimated that the consideration of specific technologies like Building Energy Management Systems (BEMSs) can save up to 20% of the energy consumption of the building sector, i.e., 8% of the overall Community consumption. With this aim, BEMSs are generally applied only to the control of active systems, i.e., HVAC systems.

An HVAC system is comprised by all the components of the appliance used to condition the interior air of a building. The HVAC system is needed to provide the occupants with a comfortable and productive working environment which satisfies their physiological needs. In Figure 2, a typical office building HVAC system is presented. This system consists of a set of components to be able to raise and lower the temperature and relative humidity of the supply air.

The energy consumption as well as indoor comfort aspects of ventilated and air conditioned buildings are highly dependent on the design, performance and control of their HVAC systems and equipments. Therefore, the use of appropriate automatic control strategies, as FLCs, for HVAC systems control could result in important energy savings when compared to manual control [1, 11].

Some artificial intelligence techniques could be successfully applied to enhance the HVAC system capabilities [4, 11]. However, most works apply FLCs

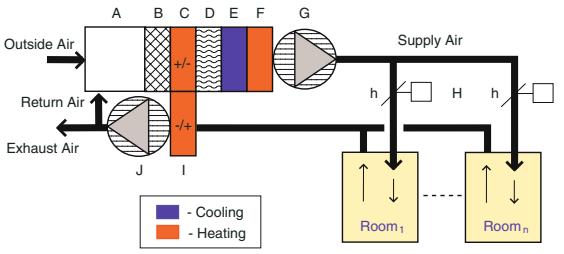


Fig. 2. Generic structure of an office building HVAC system

to individually solve simple problems such as thermal regulation (maintaining a temperature setpoint), energy savings or comfort improvements. On the other hand, the initial rule set is usually constructed based on the operator's control experience using rules of thumb, which sometimes fail to obtain satisfactory results [11]. Therefore, the different involved criteria should be optimized for a good performance of the HVAC System. Usually,

In our case, five criteria should be optimized improving an initial FLC obtained from human experience (involving 17 variables) by the application of the proposed technique for the lateral tuning of the MFs and rule selection. To do so, we consider the calibrated and validated models of a real test building. Both, the initial FLC and the simulation model were developed within the framework of the JOULE-THERMIE programme under the GENESYS [1] project. From now on, this test building will be called the GENESYS test site.

In the following subsections the five different objectives and the final fitness function to be optimized will be presented together with the initial FLC architecture and variables (see [1] for more information on this problem).

4.1 Objectives and Fitness Function

. In this way, the global objective is to the following five criteria:

O_1 Upper thermal comfort limit: $if\ PMV > 0.5,\ O_1 = O_1 + (PMV - 0.5)$, where PMV is the more global Predicted Mean Vote thermal comfort index 7730 selected by the international standard organization ISO, incorporating relative humidity and mean radiant temperature[2].

O_2 Lower thermal comfort limit: $if\ PMV < -0.5,\ O_2 = O_2 + (-PMV - 0.5)$.

O_3 Indoor air quality requirement: $if\ CO_2\ conc. > 800ppm,\ O_3 = O_3 + (CO_2 - 800)$.

[1] GENESYS Project: Fuzzy controllers and smart tuning techniques for energy efficiency and overall performance of HVAC systems in buildings, European Commission, Directorate-General XII for Energy (contract JOE-CT98-0090).

[2] http://www.iso.org/iso/en/ISOOnline.frontpage

O_4 Energy consumption: $O_4 = O_4+$ Power at time t.

O_5 System stability: $O_5 = O_5+$ System change from time t to $(t-1)$, where system change states for a change in the system operation, e.g., a change in the fan speed or valve position.

These criteria are combined into one overall objective function by means of a vector of weights. When trustworthy weights are available, this approach reduces the size of the search space providing the adequate direction into the solution space and its use is highly recommended. In our case, trusted weights were obtained by the experts for the objective weighting fitness function: $w_1^O = 0.0083022$, $w_2^O = 0.0083022$, $w_3^O = 0.00000456662$, $w_4^O = 0.0000017832$ and $w_5^O = 0.000761667$. Finally, the fitness function to be minimized was computed as:

$$F = \sum_{i=1}^{5} w_i^O \cdot O_i \ .$$

4.2 FLC Variables and Architecture

A hierarchical FLC architecture considering the PMV, CO_2 concentration, previous HVAC system status and outdoor temperature was proposed by the BEMS designer for this site. This architecture, variables and initial Rule Base (RB) are presented in Figure 3.

The DB is composed of symmetrical fuzzy partitions with triangular-shaped MFs labeled from $L1$ to Ll_i (with l_i being the number of labels of the i-th

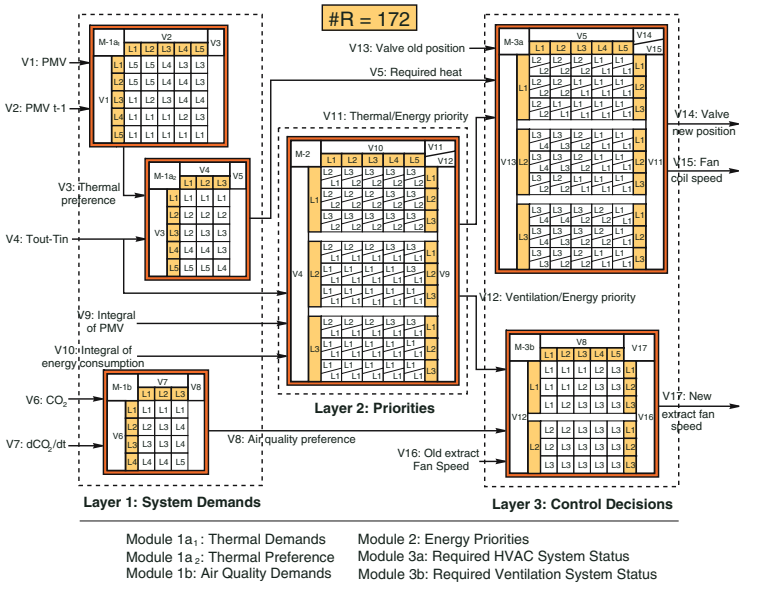


Fig. 3. Initial RB and generic structure of the GENESYS FLC

variable). The initial DB is depicted in Figure 4 together with the tuned DB. Figure 3 represents the decision tables of each module of the hierarchical FLC in terms of these labels. Each cell of the table represents a fuzzy subspace and contains its associated output consequent(s), i.e., the corresponding label(s). The output variables are denoted in the top left square for each module. Both, the initial RB and the DB, were provided by the BEMS designer.

5 Experiments

To evaluate the goodness of the approach proposed (global lateral tuning with rule selection), the HVAC problem is considered to be solved. The FLCs obtained from the proposed approach will be compared to the performance of a classic On-Off controller and to the performance of the initial FLC (provided by experts). [illegible][3][illegible]. The intention from experts was to try to have 10% energy saving (O_4) together with a global improvement of the system behavior compared to On-Off control. Comfort parameters could be slightly increased if necessary (no more than 1.0 for criteria O_1 and O_2). The methods considered in this study are shown in Table 1.

Table 1. Methods Considered for Comparison

Method, Ref.	Year	Description
S, [3]	2005	Rule Selection (C_S part of GL-S)
CL, [1]	2003	Classical Tuning
GL, [2]*	2004	Global Lateral-Tuning (C_T part of GL-S)
CL-S, –	–	Classical Tuning (CL) + Rule Selection (S)
GL-S, –	–	The proposed method

* The global lateral tuning proposed in [2] adapted to this problem

The values of the parameters used in all of these experiments are presented as follows: 31 individuals, 0.2 as mutation probability per chromosome (except for the lateral tuning which has no mutation), 0.3 for the factor α in the hybrid crossover operator and 0.35 as factor a in the max-min-arithmetical crossover in the case of CL. The termination condition will be the development of 2000 evaluations, in order to perform a fair comparative study. In order to evaluate the GA good convergence, three different runs have been performed considering three different seeds for the random number generator.

The results presented in Table 2, where % stands for the improvement rate with respect to the On-Off controller for each criterion and #R for the number of fuzzy rules, correspond to averaged results obtained from the three different runs. The results obtained with the On-Off and the initial FLC controller are also included in this table. No improvement percentages have been considered in the table for $O_1 \ldots O_3$, since these objectives always met the experts requirements and the On-Off controller presents zero values for these objectives.

Table 2. Comparison among the different methods

MODEL	#R	PMV O_1	PMV O_2	CO_2 O_3	Energy O_4	Energy %	Stability O_5	Stability %
ON-OFF	–	0.0	0	0	3206400	–	1136	–
FLC	172	0.0	0	0	2901686	9.50	1505	-32.48
$\overline{S}$	160	0.1	0	0	2886422	9.98	1312	-15.52
$\overline{C}$	172	0.0	0	0	2586717	19.33	1081	4.84
$\overline{GL}$	172	0.9	0	0	2325093	27.49	1072	5.66
$\overline{C-S}$	109	0.1	0	0	2536849	20.88	1057	6.98
$\overline{GL-S}$	113	0.7	0	0	2287993	**28.64**	800	**29.58**

A good trade-off between energy and stability was achieved for all the models obtained from GL-S, maintaining the remaining criteria within the optimal values. GL-S presents improvement rates of about a 28.6% in energy and about a 29.6% in stability, with the remaining criteria for comfort and air quality within the requested levels. Moreover, the proposed algorithm presents a good convergence and seems to be independent of random factors.

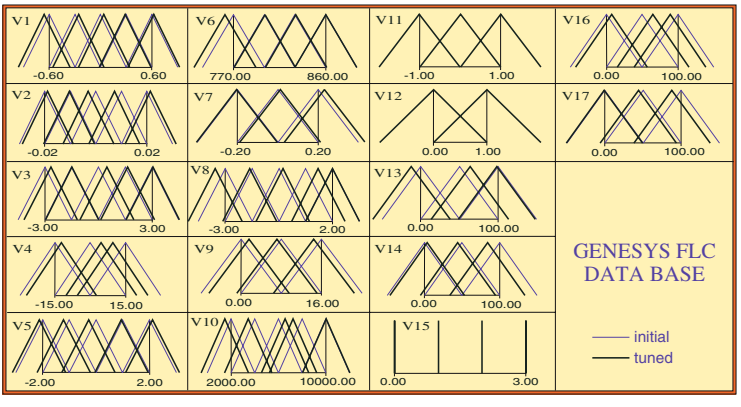


Fig. 4. Initial and Tuned DB of a Model Obtained with GL-S (Seed 1)

Figure 4 represents the initial and the final DB of the FLC obtained by GL-S with seed 1. It shows that small variations in the MFs cause large improvements in the FLC performance. Figure 5 represents the decision tables of the FLC obtained from GL-S1 (see Section 4.2). In this case, a large number of rules have been removed from the initial FLC, obtaining much simpler models (more or less 59 rules were eliminated). This fact improves the system readability, and allows us to obtain simple and accurate FLCs.

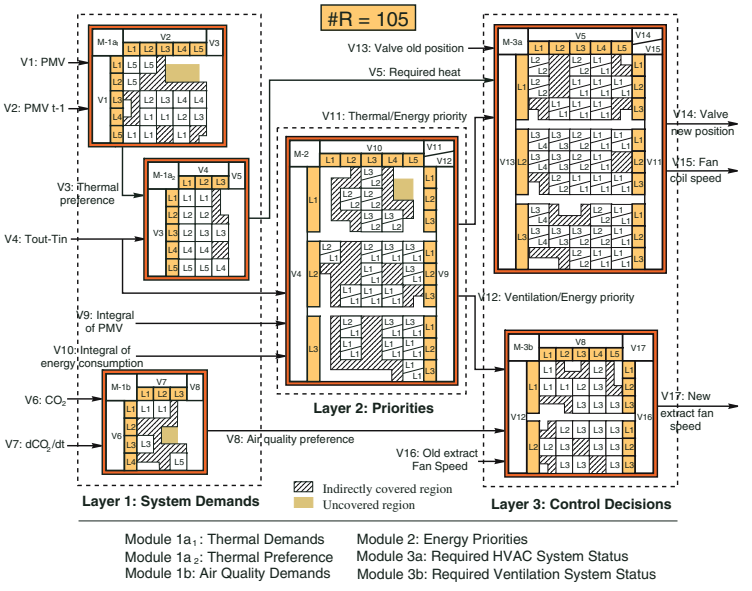


Fig. 5. RB and final structure of a FLC Obtained with GL-S (seed 1)

6 Concluding Remarks

In this work, we propose the use of a global lateral tuning together with the rule selection to obtain FRBSs to aid the BEMS expert in the control of HVAC Systems. The techniques based on lateral tuning, specially that including rule selection, have yielded much better results than the remaining approaches showing their good behavior on these kinds of complex problems. It is due to the following reasons:

- The search space reduction that the lateral tuning involves in complex problems. It allows to these techniques to obtain more optimal FLCs.
- The complementary characteristics that the use of the tuning approaches and the rule selection approach present. The ability of the rule selection to reduce the number of rules by only selecting the rules presenting a good cooperation is combined with the tuning accuracy improvement, obtaining accurate and compact FLCs.

As further work, we propose the use of multiobjective GAs in order to obtain even simpler FLCs and maintaining a similar accuracy.

References

1. Alcalá R., Benítez J.M., Casillas J., Cordón O., Pérez R.: Fuzzy control of HVAC systems optimized by genetic algorithms. Applied Intelligence **18** (2003) 155–177
2. Alcalá R., Herrera F.: Genetic tuning on fuzzy systems based on the linguistic 2-tuples representation. Proc. of the IEEE Int. Conf. on Fuzzy Syst. **1** (2004) 233–238

3. Alcalá R., Casillas J., Cordón O., González A., Herrera F.: A genetic rule weighting and selection process for fuzzy control of HVAC systems. Engineering Applications of Artificial Intelligence **18:3** (2005) 279–296
4. Calvino F., Gennusa M.L., Rizzo G., Scaccianoce G.: The control of indoor thermal comfort conditions: introducing a fuzzy adaptive controller. Energy and Buildings **36** (2004) 97–102
5. Casillas J., Cordón O., del Jesus M.J., Herrera F.: Genetic tuning of fuzzy rule deep structures preserving interpretability and its interaction with fuzzy rule set reduction. IEEE T. Fuzzy Syst. **13:1** (2005) 13–29
6. Eshelman L.J., Schaffer J.D.: Real-coded genetic algorithms and interval-schemata. Foundations of Genetic Algorithms **2** (1993) 187–202.
7. Gómez-Skarmeta A.F., Jiménez F.: Fuzzy modeling with hybrid systems. Fuzzy Sets Syst. **104** (1999) 199–208
8. Herrera F., Lozano M., Verdegay J.L.: Fuzzy connectives based crossover operators to model genetic algorithms population diversity. Fuzzy Sets Syst. **92:1** (1997) 21–30
9. Herrera F., Lozano M., Verdegay J.L.: A learning process for fuzzy control rules using genetic algorithms. Fuzzy Sets Syst. **100** (1998) 143–158
10. Herrera F., Martńez L.: A 2-tuple fuzzy linguistic representation model for computing with words. IEEE T. Fuzzy Syst. **8:6** (2000) 746–752
11. Huang S., Nelson R.M.: Rule development and adjustment strategies of a fuzzy logic controller for an HVAC system - Parts I and II (analysis and experiment). ASHRAE Trans. **100:1** (1994) 841–850, 851–856.
12. Ishibuchi H., Murata T., Türksen I.B.: Single-objective and two objective genetic algorithms for selecting linguistic rules for pattern classification problems. Fuzzy Sets Syst. **89:2** (1997) 135–150
13. Krone A., Krause H., Slawinski T.: A new rule reduction method for finding interpretable and small rule bases in high dimensional search spaces. Proc. of the IEEE Int. Conf. on Fuzzy Syst. **2** (2000) 693–699
14. Krone A., Taeger H.: Data-based fuzzy rule test for fuzzy modelling. Fuzzy Sets Syst. **123:3** (2001) 343–358
15. Whitley D., Kauth J.: GENITOR: A different genetic algorithm. Proc. of the Rocky Mountain Conf. on Artificial Intelligence (1988) 118–130

Path Bitmap Indexing for Retrieval of XML Documents

Jae-Min Lee and Byung-Yeon Hwang

Department of Computer Engineering, Catholic University, Korea
{likedawn, byhwang}@catholic.ac.kr

Abstract. The path-based indexing methods such as the three-dimensional bitmap indexing have been used for collecting and retrieving the similar XML documents. To do this, the paths become the fundamental unit for constructing index. In case the document structure changes, the path extracted before the change and the one after the change are regarded as totally different ones regardless of the degree of the change. Due to this, the performance of the path-based indexing methods is usually bad in retrieving and clustering the documents which are similar. A novel method which can detect the similar paths is needed for the effective collecting and retrieval of XML documents. In this paper, a new path construction similarity which calculates the similarity between the paths is defined and a path bitmap indexing method is proposed to effectively load and extract the similar paths. The proposed method extracts the representative path from the paths which are similar. The paths are clustered using this, and the XML documents are also clustered using the clustered paths. This solves the problem of existing three-dimensional bitmap indexing. Through the performance evaluation, the proposed method showed better clustering accuracy over existing methods.

1 Introduction

XML is positioned as a core element in modern research and technologies. Therefore, storing, retrieving, and using the XML documents effectively have important meaning as fundamental technology. The XML researches are classified into two generally. The first focuses on predefining the data structure so that it prevents the XML documents from having different structures. The second stores and retrieves documents having different structures [1, 2, 3]. The existing RDBMS vendors already proposed the methods to retrieve the XML documents inside their systems [4, 5, 6] and many XML retrieval systems have been proposed [7, 8, 9]. The existing bitmap indexing methods [10, 11] maps the XML into three-dimensional bitmap index which consists of documents, paths, and words. It showed good performance by extracting the information very fast using bit-wise operation.

The existing methods, however, have problems that cannot detect the similar paths by using the entire path as fundamental index unit to collect and retrieve the similar documents. Consequently, the retrieval of similar documents using bitmap indexing only covers the cases that the paths are exactly matched. Therefore a new indexing method to detect the similar paths is needed. In the paper, a path construction similarity to measure the similarity of paths is defined. A new path bitmap indexing method is proposed also using the path construction similarity. The proposed method extracts

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 329 – 339, 2006.

the representative path from the similar paths loaded. Using the representative path extracted, the clustering is performed over the similar paths and then clustering of XML documents is done. This solves the problem of existing three-dimensional bitmap indexing.

This paper is organized as follows. Section 2 introduces the three-dimensional bitmap indexing which is a representative path-based indexing method. Its problems are also discussed. A new path similarity measure is defined in Section 3 and a new path bitmap indexing method is proposed using the measure. Section 4 discusses the extraction of representative path from similar paths. Section 5 describes the XML document clustering using the proposed method. Section 6 discusses the performance of XML document clustering using the proposed method. Finally Section 7 presents the conclusion and further research.

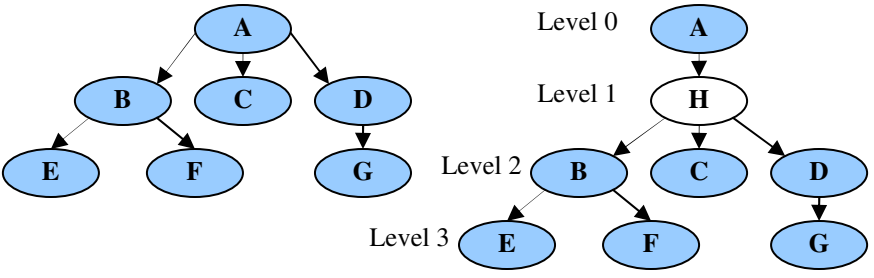
2 Related Work

BitCube, a three-dimensional indexing method, is a technique which effectively extracts information from XML documents. BitCube constructs three-dimensional index by using documents, paths, and words as three axis. The index has one-bit field where the bit-wise operations are possible. BitCube clusters the similarly structured documents based on the paths of the documents. The criterion which determines the similarities of documents is that how evenly the documents contain the same path. BitCube demonstrated outstanding performance by showing fast retrieval speed [10, 11] through the systems such as XQEngine [8] and XYZFind [9].

The three-dimensional index of BitCube has the problem of memory usage and operation speed when large number of documents are loaded. xPlaneb solved this problem by reconstructing index using the linked lists instead of BitCube's three-dimensional array index. xPlaneb, however, has performance degradation in memory usage when small amount of documents are loaded in single cluster [12]. X-Square, which selectively employs the two methods mentioned above, has better performance in memory usage. The operation speed, however, is worse than xPlaneb [13].

The three-dimensional bitmap indexing is a path-based method to collect and retrieve similar documents. In other words, the documents sharing many same paths are saved in the same cluster because they have the similar structures. Since the three-dimensional bitmap indexing extracts the paths as it is, it can detect the documents with same paths but cannot detect the documents with similar paths. Therefore, if a document undergoes structural change, the path which is extracted before the change and the path which is extracted after the change are regarded as totally different ones. Due to this, the three-dimensional bitmap indexing suffers significant performance problem in clustering and retrieving similar documents.

Figure 1 shows two XML documents and the list of their paths which have similar structures. The two documents have four paths respectively. Path 1 and 5, path 2 and path 6, path 3 and path 7, and path 4 and path 8 are similar paths, respectively. The existing three-dimensional bitmap indexing treats these as different paths because they are not exactly the same. Consequently, the similarity of two documents becomes 0 and they are loaded in different clusters. This causes a significant problem in



PID	Path
1	A_B_E
2	A_B_F
3	A_C
4	A_D_G

PID	Path
5	A_H_B_E
6	A_H_B_F
7	A_H_C
8	A_H_D_G

Fig. 1. Similarly structured documents and the extracted paths

clustering similarly structured documents. The same is applied to the retrieval of similar documents.

3 Path Bitmap Indexing

3.1 Path Bitmap Index

The path bitmap index is the index to cluster the similar paths. The existing three-dimensional bitmap indexing uses three-dimensional index consisting of documents, paths, and words. The path bitmap indexing, however, uses two-dimensional index consisting of paths and nodes which constructs the paths. The index collects the similar paths and then loads into the same index. One system has multiple indexes where similar paths are collected into one index. These indexes are called clusters. The new index differs from the existing index in that the new index contains the order information of nodes that consists of path. This is because the new index clusters based on the nodes while the existing index clusters based on the paths. The order between the paths has no meaning. For the nodes, however, the order of nodes has significant meaning.

Figure 2 shows the structure of the path bitmap index. The path bitmap index has the path name list and the node name list since it consists of the paths and their nodes. The paths loaded in the index become true-and-false one-dimensional Boolean array according to the existence of the nodes. They form the base bitmap index. Each path records its order information of the nodes using the next node pointer. The center of the clusters becomes true-and-false one-dimensional Boolean array. The decision of true and false depends on how many paths include the node. The nodes that have true values are called the representative nodes and they form the representative paths. The representative paths, as in the path, use the next node pointer to record the order information of the nodes.

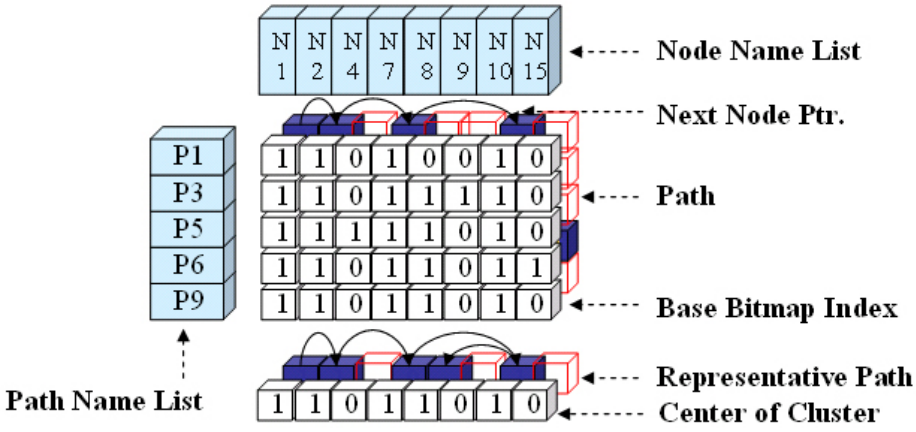


Fig. 2. The path bitmap index

Each index where the similar paths are collected should maintain the information that represents the index. The existing bitmap index achieves this by keeping the center of the clusters. The path bitmap index also keeps the information of the representative nodes through the center of the cluster. However, as mentioned above, since the order of nodes is important in the case of paths, the order information between the nodes should be included in addition to the simple node information. Consequently, the information that represents one cluster can be extracted through the representative nodes and the order information. This information is called the representative path of the cluster. The election of the representative node and representative path is discussed in the next section.

3.2 Path Construction Similarity

In order to cluster the similar paths in the path bitmap index which is proposed in the previous section, some new criteria to measure the degree of similarity of paths is needed. This section discusses the path construction similarity. A path is actually the set of nodes which construct the path. The nodes have some particular order. The paths are similar if the sets of nodes are similar. But more importantly, the orders of nodes should be similar for the paths to be similar. The path construction similarity measures the degree of similarity of node sequences that construct the paths. More specifically, the path construction similarity is measured as follows: among the two paths, the path that has more nodes is selected; the average of the node values in the path is calculated.

Definition 1. The node value *NodeValue(P, N)* of the node *N* which constructs the path *P* is defined as follows. *d* is the number of nodes that are not matched sequentially).

$$NodeValue(P,N) = \frac{1}{2^d}$$

Definition 2. The path construction similarity $P.C.Sim(P_1, P_2)$ of the two different paths P_1 and P_2 is defined as follows. $NodeNum(P)$ is the number of all nodes that the path P contains. $NodeNum(P_1)$ is always less than or equal to $NodeNum(P_2)$.

$$P.C.Sim(P_1, P_2) = \frac{\sum_{i=1}^{NodeNum(P_2)} NodeValue(P_2, N_i)}{NodeNum(P_2)}$$

Figure 3 shows how the path construction similarity is measured between the path 1 in Figure 1 and the two paths A_H_I_B_E and A_H_I_B_J_E. The node values of A, B, E in the path A_H_I_B_E become 1 because the nodes A, B, and E match the path 1 sequentially. Since the nodes H and I do not exist in the path 1 and appear sequentially, the node values are 0.25. The path construction similarity between the path A_H_I_B_E and the path 1 become ((1+0.25+0.25+1+1)/5=0.7). In the same way the path construction similarity between the path A_H_I_B_J_E and the path 1 become ((1+0.25+0.25+1+0.5+1)/6=0.67).

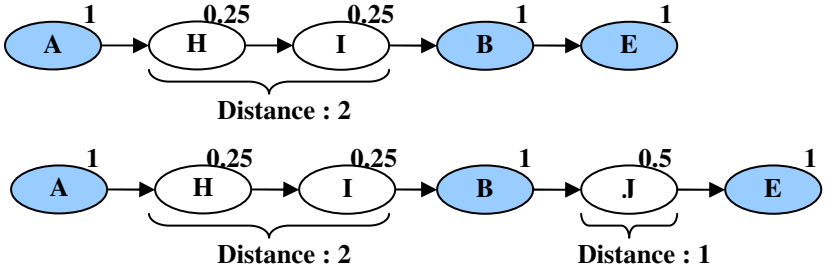


Fig. 3. Calculation of the path construction similarity

4 Extraction of the Representative Path

4.1 The Election of Representative Nodes

In order to extract the representative path, the nodes that have common paths are elected. These nodes are called the representative nodes. The representative nodes are determined through the representativeness and it is defined in Definition 3.

Definition 3. The node representativeness Rep(C, N) of the node N in the cluster C is defined as follows. PathNum(C, N) is the number of paths in the cluster C that contains the node N and PathNum(C) is the number of all paths in the cluster C.

$$Rep(C, N) = \frac{PathNum(C, N)}{PathNum(C)}$$

For instance, compared with the path 1 in Figure 1, since the path construction similarity of the paths 2, 5, and 6 are 0.83, 0.87, and 0.87 respectively, they are

determined as similar paths. Therefore, if the paths 1, 2, 5, and 6 exist in the cluster C1, the node representativeness of the nodes A, B, E, F, and H that construct the paths are (4/4=1), (4/4=1), (2/4=0.5), (2/4=0.5), (2/4=0.5) respectively. If the threshold 0.4 is assumed, all nodes are elected as the representative nodes. If the threshold is 0.8, only the node A and B are elected as the representative nodes.

To elect the representative nodes, the nodes that construct similar paths should be checked. If the checking and counting is progressed from the lower level nodes to the upper level nodes, an ordered set of nodes is obtained. This is called the temporary representative path.

Table 1. Accumulated counting for electing the representative nodes

Node Level / Node Name	0	1	2	3
A	**4**	-	-	-
B	0	**2**	**4**	-
E	0	0	1	**2**
F	0	0	1	**2**
H	0	**2**	-	-

In the example above, the threshold is assumed 0.4. In order to elect the representative node from the cluster C_1, all the nodes that construct the paths 1, 2, 5, and 6 should be checked. Here the level 0 nodes from the paths 1, 2, 5, and 6 are checked first and then the level 1 nodes are checked. In the same way level 2 or 3 nodes are checked sequentially while the accumulated counts of nodes are recorded. During the count, when a node exceeds the threshold, the node is declared as a representative node. If the nodes are arranged as the order of election, an ordered set of nodes is obtained. Table 1 shows the count values and their change to elect the representative nodes. The table shows that the paths 1, 2, 5, and 6 elect all the nodes as representative nodes. Since the nodes are elected in the order of A, B, H, E, and F, the temporary representative path is A_B_H_E_F. If the threshold is set to 0.8, since the representative nodes are A and B, and the order is also A and then B, the temporary representative path is A_B.

4.2 The Determination of Order for Representative Nodes

In order to elect the representative path, the order of elected representative nodes should be determined. The order is determined through the relationship between the nodes. The relationship is defined in Definition 4.

Definition 4. The relativeness $Rel(C, N_1, N_2)$ between the nodes N_1 and N_2 in the cluster C is defined as follows. In the formula, $PostNodesNum(C, N_1, N_2)$ is the number of node N_2 appearing after the node N_1 in the cluster C. $PathNum(C, N_1, N_2)$ is the number of paths in all the paths in the cluster C that contain the node N_1 or N_2.

$$Rel(C, N_1, N_2) = \frac{|\, PostNodesNum(C, N_1, N_2) - PostNodesNum(C, N_2, N_1) \,|}{PathNum(C, N_1, N_2)}$$

If the relativeness between the two nodes is high, the order of nodes has significant meaning. If the relativeness is low, however, ordering the nodes may be harmful to the representativesness of the representative path. Generally, the relativensss between the nodes become low in two cases. The first case is when a node *A* seldom appears in the paths in which another node *B* appears. In this case, the node *A* cannot be a representative node for the paths that contain *B*.

The second case is that the nodes *A* and *B* appear in the same paths but the number of cases of the node *A* after the node B and that of node *B* after the node *A* are nearly the same. In this case, the representative path in which the node *A* appears before the node *B* cannot represent the path in which the node *B* appears before the node *A*. The reverse is true also. The representative path in which the node *B* appears before the node *A* cannot represent the path in which the node *A* appears before the node *B*. Therefore, if $Rel(C, N_1, N_2)$ is lower than some threshold, the election of the representative node is cancelled even though the representativeness of the nodes N_1 and N_2 is higher than the threshold. If $Rel(C, N_1, N_2)$ is higher than some threshold, the order of two nodes is determined. If $PostNodesNum(C, N_1, N_2) - PostNodesNum(C, N_2, N_1)$ is positive, it is determined that N_1 precedes N_2. In the contrary, if $PostNodesNum(C, N_1, N_2) - PostNodesNum(C, N_2, N_1)$ is negative, it is determined that N_2 precedes N_1. In this way the orders of all the representative nodes are determined.

The post node table is used to measure the relativeness of the representative nodes and to determine the order of the representative nodes fast. The table records how many times a node appears after another node. In other words, the table keeps $PostNodesNum(C, N_1, N_2)$ values for all the representative nodes. Table 2 shows the post node table for the representative nodes *A, B, E, F,* and *H* in the cluster C_1 when the threshold is 0.4. Using the table, the relativeness of the representative nodes is measured that construct the temporary representative path that was extracted in Section 4.1. Then the order of nodes is adjusted using $PostNodesNum(C, N_1, N_2) - PostNodesNum(C, N_2, N_1)$. The adjustment is similar to the insertion sort.

Table 2. The post node table for the representative nodes of the paths 1, 2, 5, and 6

Post Node Name / Node Name	A	B	E	F	H
A	-	4	2	2	2
B	0	-	2	2	0
E	0	0	-	0	0
F	0	0	0	-	0
H	0	2	1	1	-

For instance, when the threshold 0.4 for the node relativeness is assumed, the order of temporary representative path A_B_H_E_F that was extracted from the cluster C_1 is adjusted as follows. The temporary representative path is scanned from the second node to the last node while searching the position for the node to be inserted. First the representative node *B* measures the relativeness with the node A which exists in front of B. Since $Rel(C_1, A, B)$ is (|4 - 0| / 4=1), it is determined to be meaningful. If $PostNodesNum(C_1, A, B) - PostNodesNum(C_1, B, A)$ is negative, the representative node B

is inserted before the node *A*. Since *PostNodesNum(C_1, A, B) – PostNodesNum(C_1, B, A)* is positive, however, the order remain the same.

Next, the representative node *H* measures the relativeness with the nodes *A* and *B*, respectively. Since *Rel(C_1, B, H)* is (|0 - 2| / 4=0.5), it is determined to be meaningful. *PostNodesNum(C_1, B, H) – PostNodesNum(C_1, H, B)* is negative. Since *Rel(C_1, A, H)* is (|2 - 0| / 4=0.5), it is determined to be meaningful. Since *PostNodesNum(C_1, A, H) – PostNodesNum(C_1, H, A)* is positive, the representative node *H* is inserted between the nodes *A* and *B*. The same applies up to the last node *F* and a complete representative path is obtained. One thing to be noticed is the relativeness on the representative nodes *E* and *F*. Since *Rel(C_1, E, F)* is (0 – 0 / 4=0), the election of representative nodes for the two nodes is cancelled during the order adjustment. Consequently, the representative path A_H_B is obtained after the order adjustment for the temporary representative path A_B_H_E_F.

5 Clustering Similar XML Documents Using Path Bitmap Index

Sections 3 and 4 discussed the path bitmap indexing which clusters the similar paths. The existing three-dimensional bitmap indexing uses the documents, paths, and word IDs to cluster the XML documents. The new method, however, uses the index ID of the corresponding path instead of the path ID. This solves the problem of the existing three-dimensional bitmap indexing that it cannot detect the similar paths.

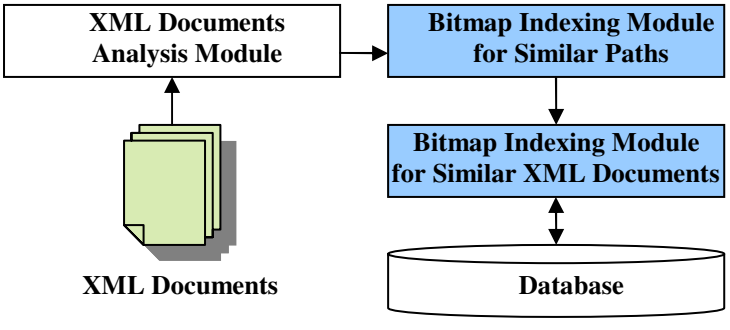


Fig. 4. The architecture of xPlaneb+

Figure 4 shows the architecture of xPlaneb+ that clusters XML documents using the path bitmap indexing. A new XML document is converted into the set of paths through the XML document analysis module and then transferred to the bitmap indexing module for similar paths. The bitmap indexing module for similar paths searches the index ID of the most similar representative path and then transfers the result to the bitmap indexing module for similar XML documents. The bitmap indexing module for similar XML documents performs the clustering using the transferred index ID.

6 Performance Evaluation

This section shows the clustering of similar XML documents using the representative paths which are extracted through the proposed method, and also the performance evaluation is done. The proposed method was implemented under the MS Windows 2000 Server and MS Visual C#.NET. MS SQL Server was employed for the database management system.

The total number of clusters is measured after the clustering is done on a group of similar documents using an existing method and the proposed method. The transformation rate measures how a document or a group of documents differ from some target document structurally. This is defined in Definition 5.

Definition 5. The transformation rate *TransRate(D_1, D_2)* between the documents D_1 and D_2 is defined as follows. The Nodenum(D1,D2) is the number of the nodes which are in the union of the documents D1 and D2, and the Transnodenum(D1,D2) is that of the nodes which are exclusively in one of D1 and D2.

$$TransRate(D_1, D_2) = \frac{TransNodeNum(D_1, D_2)}{NodeNum(D_1, D_2)}$$

The number of clusters is measured on a group of the similar documents with the transformation rates 0.1 and 0.4. Figure 5 shows the result of the performance evaluation. The existing three-dimensional bitmap indexing generates large number of clusters as the transformation rate increases. When the transformation rate reaches 40%, the number of clusters is almost the same as the number of documents. The proposed method, however, generates a lot less number of clusters as the transformation rate increase which can be said to be more accurate clustering. This is because the proposed method detects the similar paths so that the similar documents go to less number of clusters.

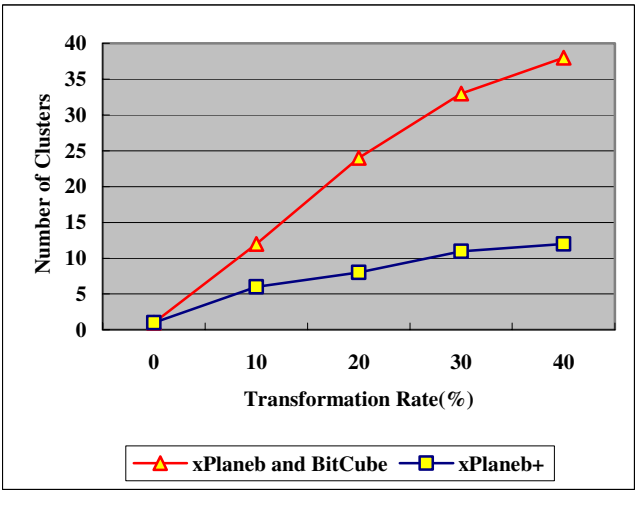


Fig. 5. The number of clusters for the transformation rate

7 Conclusion

The paths are the basic units for the index construction in the path-based indexing such as the existing three-dimensional bitmap indexing for the similar XML document retrieval. When the structure of a document changes, the extracted paths before the change and after the change are regarded totally different one. Consequently, the path-based indexing suffers from the significant performance degrade in the retrieval and clustering of similar documents. Therefore, a new method to detect the similar paths is needed for the effective storage and retrieval of XML documents.

The paper defines the path construction similarity which becomes the basis for measuring the similarity of the paths. Using this, the path bitmap indexing is proposed which enables effective load and retrieval of the similar paths. The method extracts the representative nodes and the temporary representative paths from the similar paths loaded. Using the post node table, the order of representative nodes is adjusted which construct the representative paths, and then the representative paths are extracted. Through the extracted representative paths, the clustering is performed for the similar paths and XML documents. This solves the problem of the existing three-dimensional bitmap indexing.

The performance evaluation shows that the existing three-dimensional bitmap indexing generates large number of clusters as the transformation rate increases. When the transformation rate reaches 40%, the number of clusters is almost the same as the number of documents. The proposed method, however, generates a lot less number of clusters as the transformation rate increase which can be said to be more accurate clustering.

The proposed method has some problem in the condition that the similar words or synonyms exist. The further research includes the solution for the problem and more accurate clustering.

References

1. Fuhr, N., Grossjohann, K.:XIRQL: An Extension of XQL for Information Retrieval. ACM SIGIR Workshop, Athens, Greece (2000)
2. Cooper, B., Sample, N., Franklin, M., Shadmon, M.:A Fast Index for Semistructured Data. Proc. of the 27th VLDB Conference, Roma, Italy (2001)
3. Garafalalos, M., Rastogi, R., Seshadri, S., Shim, K.:XTRACT: A System for Extracting Document Type Descriptors from XML Documents. Proc. of the ACM SIGMOD (2000)
4. Banerjee, S.: Oracle XML DB. An Oracle Technical White Paper (2003)
5. Ennser, L., Delporte, C., Oba, M., Sunil, K.: Integrating XML with DB2 XML Extender and DB2 Text Extender. IBM Redbook (2000)
6. Howlett, S., Jennings, D.: SQL Server 2000 and XML: Developing XML-Enabled Data Solutions for the Web. MSDN Magazine, Vol.17, No.1 (2002)
7. Egnor, D., Lord, R.: XYZFind: Structured Searching in Context with XML. ACM SIGIR Workshop, Athens, Greece (2000)
8. http://www.fatdog.com
9. Delobel, C., Reynaud, C., Rousset, M.:Semantic Integration in Xyleme: A Uniform Tree-based Approach. Data & Knowledge Engineering, Vol.44 (2003) 267-298

10. Yoon, J., Raghavan, V., Chakilam, V.: BitCube: Clustering and Statistical Analysis for XML Documents. 13th International Conference on Scientific and Statistical Database Management, Virginia (2001)
11. Yoon, J., Raghavan, V., Chakilam, V., Kerschberg, L.: BitCube: A Three-Dimensional Bitmap Indexing for XML Documents. J. of Intelligent Information System, Vol.17 (2001) 241-254
12. Lee, J., Hwang, B.: xPlaneb: 3-Dimensioal Bitmap Index for XML Document Retrieval. J. of Korea Information Science Society, Vol.31, No.3 (2004) 331-339
13. Lee, J., Hwang, B., Lee, B.: X-Square: Hybrid 3-Dimensioal Bitmap Indexing for XML Document Retrieval. Lecture Notes in Computer Science, Vol. 3307 (2004) 221-232

A Modified Fuzzy C-Means Algorithm for Differentiation in MRI of Ophthalmology

Wen-Liang Hung[1] and Yen-Chang Chang[2]

[1] Department of Applied Mathematics, National Hsinchu University of Education, Hsin-Chu, Taiwan, ROC
wlhung@mail.nhcue.edu.tw
[2] Department of Business Administration, Vanung University, Chung-Li, Taiwan, ROC
yenchang@vnu.edu.tw

Abstract. In this paper we propose an algorithm, called the modified suppressed fuzzy c-means (MS-FCM), that simultaneously performs clustering and parameter selection for the suppressed FCM (S-FCM) proposed by Fan et al. [2]. Numerical examples illustrate the effectiveness of the proposed MS-FCM algorithm. Finally, the S-FCM and MS-FCM algorithms are applied in the segmentation of the magnetic resonance image (MRI) of an ophthalmic patient. In our comparisons of S-FCM, MS-FCM and alternative FCM (AFCM) proposed by Wu and Yang [14] for these MRI segmentation results, we find that the MS-FCM provides better detection of abnormal tissue than S-FCM and AFCM when based on a window selection. Overall, the MS-FCM clustering algorithm is more efficient and is strongly recommended as an MRI segmentation technique.

1 Introduction

In real data analysis, the large data sets are unavoidable. In this situation, the computation time of the FCM algorithms will grow rapidly and the issue of slow convergence has been dealt with by various authors ([1], [4], [9], and [15]). Recently, Wei and Xie [13] proposed the rival checked FCM (RCFCM) clustering algorithm to speed up FCM algorithm. But the RCFCM algorithm only pays attention to the largest and next largest degree of membership. To overcome this shortcoming, Fan et al. [2] proposed the S-FCM clustering algorithm which established a relationship between the hard c-means (HCM) and FCM clustering algorithms. However, the problem of selecting a suitable parameter α in S-FCM constitutes an important part of implementing the S-FCM algorithm for real applications. It is therefore important to select a suitable α such that the S-FCM algorithm can take on the advantages of the fast convergence speed of the HCM as well as the superior partition performance of the FCM. Fan et al. [2] failed to provide a selection method for the parameter α in S-FCM.

Based on the concept of machine learning with the capability of learning to improve the performance of a task on the basis of the previous experience, we propose a modified S-FCM (MS-FCM) algorithm that performs clustering and selects the parameter α in S-FCM simultaneously. Numerical experiments illustrate the effectiveness of the proposed MS-FCM with the ability of performing clustering and also

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 340 – 350, 2006.

selecting the parameter α with a prototype-driven learning. Finally, we use the MS-FCM algorithm to an MRI segmentation in ophthalmology.

2 A Modified S-FCM Algorithm Based on Prototype-Driven Learning

Let $X = \{x_1, \cdots, x_n\}$ be a data set in an s-dimensional Euclidean space R^s with its ordinary Euclidean norm $\|\bullet\|$ and let c be a positive integer greater than one. The FCM algorithm is an iterative algorithm using the necessary conditions for a minimizer of the objective function J_{FCM} with

$$J_{FCM}(\mu, \upsilon) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} \left\| x_j - \upsilon_i \right\|^2, \ m > 1,$$

where $\mu = \{\mu_1, \cdots \mu_c\}$ is a fuzzy c-partition with $\mu_{ij} = \mu_i(x_j)$, $\upsilon = \{\upsilon_1, \upsilon_2, \cdots, \upsilon_c\}$ is the c cluster centers and the weighting exponent m has an influence on the clustering performance of FCM (cf. [8]). The necessary conditions for a minimizer (μ, υ) of J_{FCM} are the following update equations:

$$\upsilon_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} x_j}{\sum_{j=1}^{n} \mu_{ij}^{m}}, \ i = 1, \cdots, c \tag{1}$$

and

$$\mu_{ij} = \left(\sum_{k=1}^{c} \frac{\left\| x_j - \upsilon_i \right\|^{2/(m-1)}}{\left\| x_j - \upsilon_k \right\|^{2/(m-1)}} \right)^{-1}, \ i = 1, \cdots, c; \ j = 1, \cdots, n. \tag{2}$$

Recently, these FCM clustering algorithms have found frequent use in segmenting MRI ([5], [10], [11], [17]). The convergence speed of FCM becomes important as MRI data sets are generally large. To speed up FCM algorithm, Fan et al. [2] proposed the S-FCM clustering algorithm based on the idea of RCFCM, which magnifies the largest membership degree and suppresses the second largest membership degree. The main change from FCM is to modify memberships μ_{ij} as follows:

If the data point x_j has membership in the p th cluster and the p th is the largest of all c clusters, the value is noted as μ_{pj}. (Annotate: If there are two or more largest memberships, randomly choose one.) The memberships are then modified as:

$$\mu_{pj} = 1 - \alpha \sum_{i \neq p} \mu_{ij} = 1 - \alpha + \alpha \mu_{pj} \text{ and } \mu_{ij} = \alpha \mu_{ij},\ i \neq p,\ 0 \leq \alpha \leq 1. \quad (3)$$

But the selection of α was not addressed in their paper. Next, we will provide a simple method that can perform clustering and also select the parameter α in S-FCM with a prototype-driven learning approach.

The idea of selecting α is based on the separation strength given by $\min_{i \neq k} \|v_i - v_k\|^2$. We make note that the term $\min_{i \neq k} \|v_i - v_k\|^2$ was used in the validity index proposed by Xie and Beni [16], defined as

$$V_{XB} = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^2 \|x_j - v_i\|^2}{n(\min_{i \neq k} \|v_i - v_k\|^2)}. \quad (4)$$

We know that $\min_{i \neq k} \|v_i - v_k\|^2$ indicates the separation strength between clusters. If the value of $\min_{i \neq k} \|v_i - v_k\|^2$ is large, the result is a lower degree of overlapping and a greater separation between the clusters is produced. In this case, HCM is superior to FCM. That is, a small value of α will be a better selection for S-FCM. On the other hand, FCM should outperform HCM when the borders between clusters are not sharp. That is, clusters are not clearly separated. In this case, the value of $\min_{i \neq k} \|v_i - v_k\|^2$ is small and a large value of α will be a better selection for S-FCM. According to the above, we know that α is better assigned as a monotone decreasing function of $\min_{i \neq k} \|v_i - v_k\|^2$. The problem here is how to select a useful and suitable function. Adopting the Cauchy distribution concept, we propose an Cauchy-type function to select α with

$$\alpha = \frac{1}{1 + \min_{i \neq k} \frac{\|v_i - v_k\|^2}{\beta}}, \quad (5)$$

where β is a normalized term so that we choose β as a sample variance. That is, we define β as

$$\beta = \frac{\sum_{j=1}^{n} \|x_j - \bar{x}\|^2}{n}, \text{ where } \bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}.$$

Therefore, we propose the modified S-FCM (MS-FCM) algorithm as follows:

MS-FCM algorithm

S1: Fix $m > 1$ and $2 \leq c \leq n-1$ and give c initial cluster centers v_i.

REPEAT

S2: Compute α with v_i by equation (5).

S3: Compute μ_{ij} with v_i by equation (2).

S4: Modify μ_{ij} by equation (3).

S5: Update v_i with the modified μ_{ij} by equation (1).

UNTIL(cluster centers stabilized)

Note that the differences between S-FCM and MS-FCM are : (i) In S-FCM, the parameter α is prior given but the value of α in MS-FCM is updated at each iteration using Eq. (5); (ii) The update of α in MS-FCM is a prototype- driven learning approach to match the cluster behavior of the given data set.

Next, we compare the performance of MS-FCM to S-FCM with $\alpha = 0.5$ according to the normal mixtures of two classes under the accuracy and computational efficiency criteria. The accuracy of an algorithm is measured by the mean squared error (MSE) that is the average sum of squared error between the true parameter and the estimated in N repeated trials. The computational efficiency of an algorithm is measured by the average numbers of iterations (NI) in N repeated trials. We also design various normal mixture distributions shown in Table 1. Algorithms are processed with the same specifying initial values, $\varepsilon = 0.0001$ and $m = 2$ (cf. [7], [12]). The MSE is calculated using

$$\frac{\sum_{k=1}^{N}\sum_{i=1}^{2}(\hat{u}_i^{(k)} - u_i)^2}{Nc} \tag{6}$$

where $\hat{u}_i^{(k)}$ is the estimated subpopulation mean of the k th trial, u_i is the true subpopulation mean and $N = 50$.

Table 1. Normal mixtures for the numerical experiments

$\gamma N(0,1) + (1-\gamma)N(2,1)$	Test A	A1	A2	A3
γ	0.1	0.3	0.5	
$0.5N(0,1) + 0.5N(u,1)$	Test B	B1	B2	B3
u	1	2	3	
$0.5N(0,1) + 0.5N(2,\sigma^2)$	Test C	C1	C2	C3
σ^2	0.5	1	2	
$\gamma N(0,1) + (1-\gamma)N(u,\sigma^2)$	Test D	D1	D2	D3
γ	0.1	0.3	0.5	
u	2	1	3	
σ^2	2	1	0.5	

We analyze Tests A, B and C with an acceptable sample size $n=100$ and large sample size $n=500$. The results for $n=100$ and 500 are shown in Table 2, respectively. From Table 2, SFCM has good accuracy in all tests. Although the average number of iterations of MS-FCM is larger than other algorithms in some cases, the difference is small.

As supported by the above experiments, the MS-FCM algorithm is computationally simple and produces satisfactory results. In the next section, we will apply the S-FCM and MS-FCM clustering algorithms to an ophthalmological MRI segmentation.

Table 2. Accuracy and computational efficiency for the normal mixtures

Test	$n=100$			$n=500$		
	FCM	S-FCM	MS-FCM	FCM	S-FCM	MS-FCM
	MSE NI	MSE NI	MSE NI	MSE NI	MSE NI	MSE NI
A1	0.7386 7.20	0.7084 6.28	0.4624 6.78	0.6825 7.24	0.6684 6.12	0.3946 5.04
A2	0.1714 5.82	0.1604 5.70	0.1298 5.16	0.1626 6.56	0.1539 5.88	0.1180 6.36
A3	0.0606 3.42	0.0590 3.12	0.0569 3.14	0.0421 3.04	0.0435 2.64	0.0346 2.68
B1	0.1909 4.06	0.1896 3.54	0.1696 3.22	0.1701 4.02	0.1803 3.06	0.1605 2.80
B2	0.0650 3.54	0.0607 3.26	0.0559 3.20	0.0417 3.04	0.0417 2.70	0.0324 2.54
B3	0.0285 3.06	0.0293 2.80	0.0275 3.00	0.0155 2.48	0.0131 2.20	0.0068 2.20
C1	0.0484 3.44	0.7325 5.96	0.0342 3.06	0.0362 2.92	0.0304 2.44	0.0208 2.70
C2	0.0564 3.42	0.0527 3.26	0.0504 3.34	0.0451 3.00	0.0432 2.84	0.0332 2.76
C3	0.1405 3.94	0.1289 3.80	0.1254 3.04	0.0451 3.00	0.0432 2.84	0.0332 2.76
D1	0.8076 6.36	0.7325 5.96	0.3738 5.80	0.7850 6.00	0.6136 5.98	0.3248 5.60
D2	0.2211 4.76	0.2227 4.12	0.2144 4.02	0.2602 4.74	0.1993 4.34	0.1672 3.48
D3	0.0240 2.60	0.0200 2.28	0.0156 3.10	0.0159 2.50	0.0113 2.00	0.0063 3.02

3 Application to Ophthalmological MRI Segmentation

Segmentation of the medical images obtained from MRI is a primary step in most applications of computer vision to medical image analysis. Yang et al. [17] applied AFCM, proposed by [14], in a real case study of MRI segmentation to differentiate between normal and abnormal tissue in ophthalmology. Here we also use S-FCM and MS-FCM algorithms to analyze these MRI data sets (cf. [17]). The first MRI data set is illustrated in Figs. 1 and 2. The second MRI data set is shown in Fig. 3. We first attempt to cluster the full size images (Figs. 1 and 2) into the same five clusters used by [17]. We also apply S-FCM and MS-FCM to a window segmentation illustrated in Fig. 3.

From the red circle on the full size two dimensional MRI in Fig. 1, we can clearly detect white tumor tissue at the chiasma. Since we have no information on the structure of the first data set, we set $\alpha=0.5$ in S-FCM. The segmentation results by AFCM with $m=2$ (Fig. 1.1), S-FCM with $m=2$ and 5 (Figs. 1.2 and 1.5) and MS-FCM with $m=2$ and 5 (Figs. 1.3 and 1.6) are able to distinguish the tumor from the healthy tissue. However, AFCM with $m=5$ (Fig. 1.4) yields incorrect segmentation. It means that the performance of AFCM is sensitive to m that may cause a bad segmentation result when m is not properly chosen. We mention that this phenomenon of AFCM with $m=5$ was not observed and studied in [17].

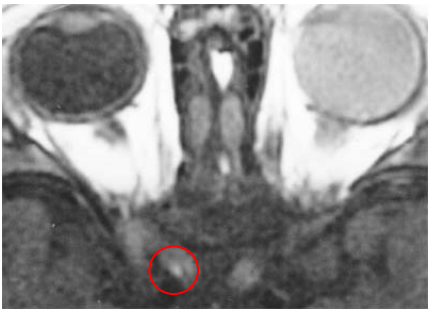


Fig. 1. Original MR image

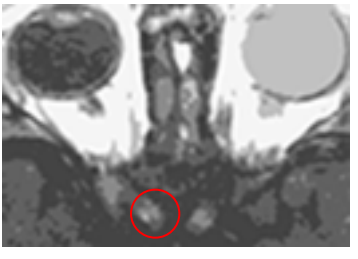


Fig. 1.1. Segmentation result of AFCM (m=2, S=1)

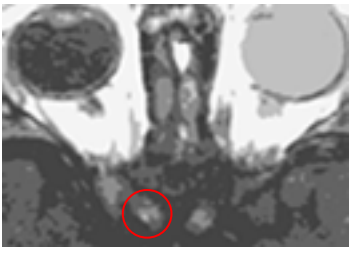


Fig. 1.2. Segmentation result of S-FCM (m=2, S=0.9073, FN=0.0927, FP=0)

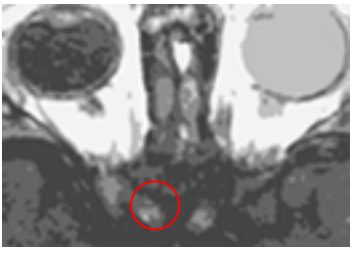


Fig. 1.3. Segmentation result of MS-FCM (m=2, S=0.9073, FN=0.0927, FP=0)

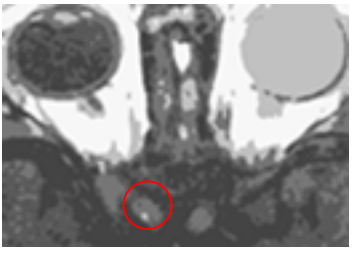


Fig. 1.4. Segmentation result of AFCM (m=5, S=0.5633, FN=0, FP=0.0024)

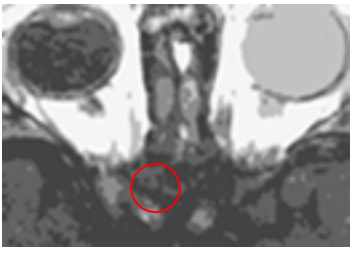


Fig. 1.5. Segmentation result of S-FCM (m=5, S=0.9073, FN=0.0927, FP=0)

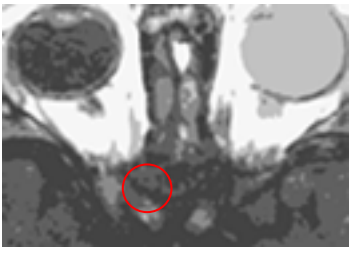


Fig. 1.6. Segmentation result of MS- FCM (m=5, S=1, FN=0, FP=0)

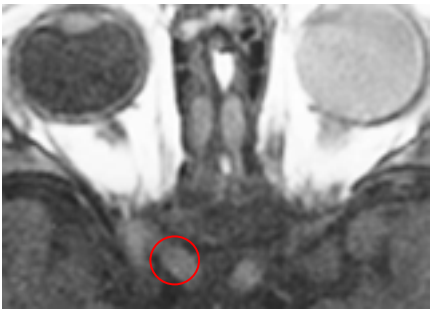


Fig. 2. Distorted MR image

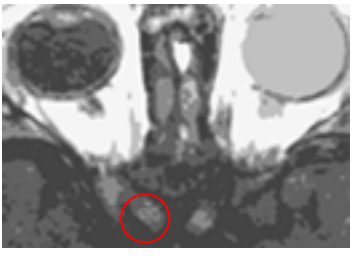


Fig. 2.1. Segmentation result of AFCM (m=2, S=1)

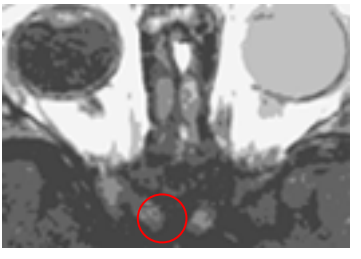


Fig. 2.2. Segmentation result of S-FCM (m=2, S=0.7374, FN=0.2626, FP=0)

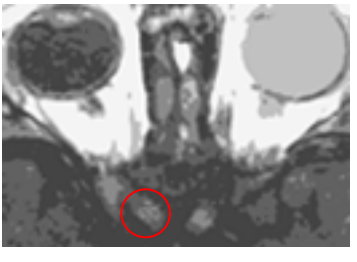


Fig. 2.3. Segmentation result of MS-FCM (m=2, S=0.7374, FN=0.2626, FP=0)

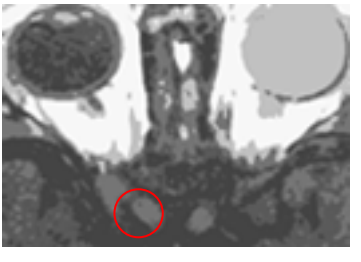


Fig. 2.4. Segmentation result of AFCM (m=5, S=0, FN=1, FP=0)

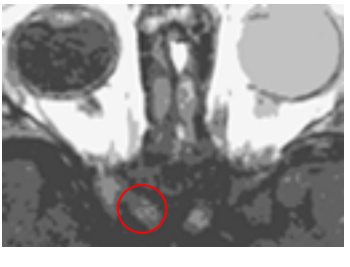


Fig. 2.5. Segmentation result of S-FCM (m=5, S=0.7374, FN=0.2626, FP=0)

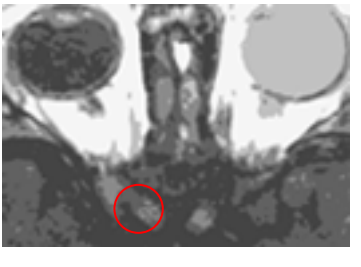


Fig. 2.6. Segmentation result of MS- FCM (m=5, S=1, FN=0, FP=0)

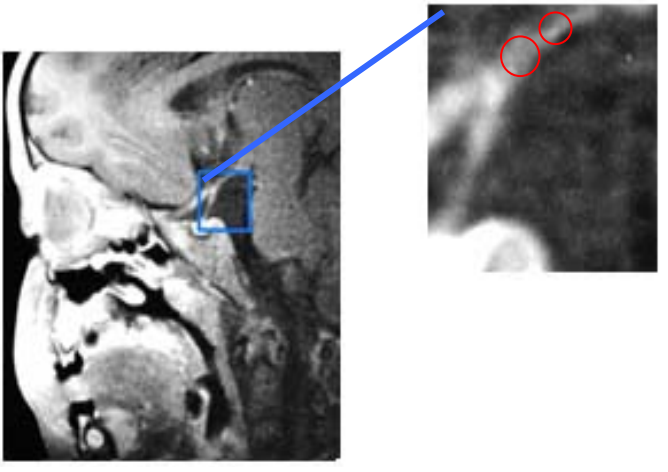


Fig. 3. Original MR image and its window selection

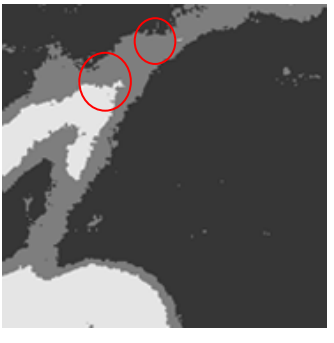


Fig. 3.1. Segmentation result of AFCM (m=2, S=1)

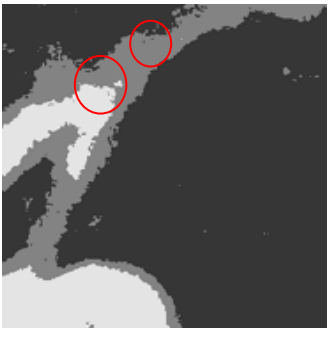


Fig. 3.2. Segmentation result of S-FCM (m=2, S=0.7685, FN=0.2315, FP=0)

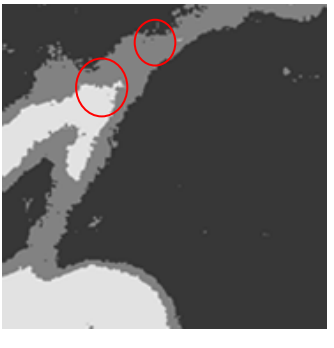


Fig. 3.3. Segmentation result of MS-FCM (m=2, S=0.8981, FN=0.1019, FP=0)

Fig. 3.4. Segmentation result of AFCM (m=5, S=0, FN=1, FP=0)

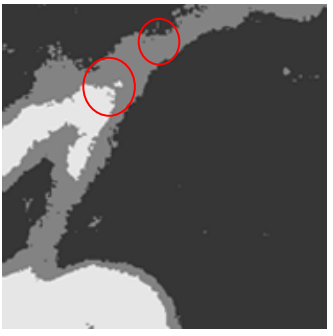


Fig. 3.5. Segmentation result of S-FCM (m=5, S=0.6759, FN=0.3241, FP=0)

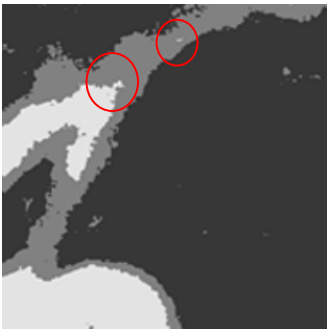


Fig. 3.6. Segmentation result of MS-FCM (m=5, S=0.8981, FN=0.1019, FP=0)

Table 3. Computational performance of MS-FCM and S-FCM for Fig. 1-3

		$m=2$			$m=5$		
criterion	dataset	AFCM	S-FCM	MS-FCM	AFCM	S-FCM	MS-FCM
Iteration number	Fig. 1	130	23	21	339	25	29
	Fig. 2	140	19	22	360	26	26
	Fig. 3	141	23	22	199	29	24
CPU time(s)	Fig. 1	3535.6	255.9	249.9	9638.4	292.9	349.9
	Fig. 2	3795.1	264.7	264.7	10258.0	307.1	354.7
	Fig. 3	1656.5	107.9	107.9	2452.9	149.6	127.1

Table 4. Final output values of α from MS-FCM and S-FCM for Fig. 1-3

m	Fig. 1	Fig. 2	Fig. 3
2	0.817	0.817	0.402
5	0.830	0.830	0.399

A distorted MR image, shown in Fig. 2, is used here to illustrate how an algorithm is able to detect tumorous tissue, despite uncertainty. With the use of the AFCM with $m=2$ (Fig. 2.1) and S-FCM with $m=2$ and 5 (Figs. 2.2 and 2.5) and MS-FCM with $m=2$ and $m=5$ (Figs. 2.3 and 2.6), we are still able to detect the tumorous tissue. But, AFCM with $m=5$ (Fig. 2.4) yields incorrect segmentation. Table 3 shows the computational performance of these three clustering algorithms. One can see that the number of iterations and the CPU time for MS-FCM in Figs. 1 and 2 are less than for AFCM but slightly larger than for S-FCM ($\alpha=0.5$) in some cases. However, the MS-FCM provides more accurate segmentation results than S-FCM for Figs. 1 and 2. From Table 4, it is seen that these output values of α are stable and robust to m.

From Fig. 3, one lesion was clearly seen in the MR image. Since we have not any information on the structure of the second data set, we set $\alpha=0.5$ in S-FCM. First, we applied AFCM, S-FCM and MS-FCM with $m=2$ and 5 to the window selection as illustrated in Figs. 3.1-3.6. We can see occult lesions (red circles) clearly enhanced with AFCM ($m=2$) in Fig. 3.1 and MS-FCM in Figs. 3.3 and 3.6. However, AFCM

(m=5) in Fig. 3.4 fails to indicate these occult lesions. And S-FCM in Figs. 3.2 and 3.5 does not clearly show these occult lesions. This indicates a poor performance on the parts of AFCM (m=5) and S-FCM ($\alpha = 0.5$).

To evaluate detection of abnormal tissue, it is necessary to make a quantitative comparison of the image, segmented by each algorithm, with a reference image. Because the segmentation results with AFCM (m=2) in Figs. 1-3 can successfully differentiate the tumor from the normal tissues (cf. [17]), Figs. 1.1, 2.1 and 3.1 are considered as reference images. The comparison score S (cf. [6], [18]) for each algorithm is defined as

$$S = \frac{\left| A \cap A_{ref} \right|}{\left| A \cup A_{ref} \right|}$$

where A represents the set of pixels belonging to the tumor tissue found by the i th algorithm and A_{ref} represents the set of pixels belonging to the tumor tissue in the reference segmented image. Moreover, adopting the similar idea of false negative and false positive from [3], we may also define the following two error types based on A and A_{ref} :

$$\text{False Negative (FN)} = \frac{\left| A_{ref} \cap A^c \right|}{\left| A_{ref} \right|}, \text{False Positive (FP)} = \frac{\left| A_{ref}^c \cap A \right|}{\left| A_{ref}^c \right|},$$

where $A_{ref} \cap A^c$ represents the set of pixels in A_{ref} has not been detected to be tumor tissue by the i th algorithm and $A_{ref}^c \cap A$ represents the set of pixels in A_{ref}^c has been detected to be tumor tissue by the i th algorithm and A^c and A_{ref}^c represent the complements of A and A_{ref} , respectively. From the values of S , FN and FP corresponding to Figs. 1.2-1.6, 2.2-2.6 and 3.2-3.6, we can see that the overall performance of MS-FCM is better than the others. Furthermore, the performance of MS-FCM is also robust to the weighting exponent m .

4 Conclusion

In this paper we use a learning technique to search for the parameter α in S-FCM and then created the MS-FCM clustering algorithm. The advantage of MS-FCM algorithm is to perform clustering and select the parameter α simultaneously. Though each learning iteration in MS-FCM would take slightly more computation time than S-FCM, the MS-FCM could provide more accurate clustering results than S-FCM. Finally, the MS-FCM algorithm is applied in the segmentation of the MRI of an ophthalmic patient. The results show that the MS-FCM provides better detection of

abnormal tissue than S-FCM and AFCM. The MS-FCM is actually a good algorithm for real applications. Overall, we recommend those concerned with applications in cluster analysis try the proposed MS-FCM algorithm.

References

1. Cannon, R.L., Dave, J.V., Bezdek, J.C.: Efficient Implementation of the Fuzzy *c*-means Clustering Algorithms. IEEE Trans. Pattern Anal. Machine Intelligence 8 (1986) 248-255.
2. Fan, J.L., Zhen, W.Z., Xie, W.X.: Suppressed Fuzzy *c*-means Clustering Algorithm. Pattern Recognition Letters 24 (2003) 1607-1612.
3. Fernández-García, N.L., Medina-Carnicer, R., Carmona-Poyato, A., Madrid-Cuevas, F.J., and Prieto-Villegas, M.: Characterization of Empirical Discrepancy Evaluation Measures. Pattern Recognition Letters 25 (2004) 35-47.
4. Kamel, M.S., Selim, S.Z.: A Relaxation Approach to the Fuzzy Clustering Problem. Fuzzy Sets Systems 61 (1994) 177-188.
5. Lin, J.S., Cheng, K.S., Mao, C.W.: Segmentation of Multispectral Magnetic Resonance Image Using Penalized Fuzzy Competitive Learning Network. Computers and Biomedical Research 29 (1996) 314-326.
6. Masulli, F., Schenone, A.: A Fuzzy Clustering Based Segmentation System as Support to Diagnosis in Medical Imaging. Artificial Intelligence in Medicine 16 (1999) 129-147.
7. Pal, N.R., Bezdek, J.C., Tsao, E.C.K.: Generalized Clustering Networks and Kohonen Seif-organizing Scheme. IEEE Trans. Neural Networks 4 (1993) 549-557.
8. Pal, N.R., Bezdek, J.C.: On Cluster Validity for Fuzzy *c*-means Model. IEEE Trans. Fuzzy Syst. 1 (1995) 370-379.
9. Pei, J.H., Fan, J.L., Xie, W.X.: A New Efficient Fuzzy Clustering Method: Cutset of Fuzzy *c*-means Algorithm. Acta Electronica Sinica 26 (1998) 83-86 (in Chinese).
10. Phillips, W.E., Velthuizen R.P., Phuphanich, S., Hall, L.O., Clarke, L.P., Silbiger, M.L.: Application of Fuzzy *c*-means Segmentation Technique for Differentiation in MR Images of a Hemorrhagic Glioblastoma Multiforme. Magnetic Resonance Imaging 13 (1995) 277-290.
11. Suckling, J., Sigmundsson, T., Greenwood, K., Bullmore, E.T.: A Modified Fuzzy Clustering Algorithm for Operator Independent Brain Tissue Classification of Dual Echo MR Images. Magnetic Resonance Imaging 17 (1999) 1065-1076.
12. Tsao, E.C.K., Bezdek, J.C., Pal, N.R.: Fuzzy Kohonen Clustering Networks. Pattern Recognition 27 (1994) 757-764.
13. Wei, L.M., Xie, W.X.: Rival Checked Fuzzy *c*-means Algorithm. Acta Electronica Sinica 28 (2000) 63-66 (in Chinese).
14. Wu, K.L, Yang, M.S.: Alternative *c*-means Clustering Algorithms. Pattern Recognition 35 (2002) 2267-2278.
15. Xie, W.X., Liu, J.Z.: A Combine Hard Clustering Algorithm and Fuzzy Clustering Algorithm-Fuzzy *c*-means Clustering Algorithm with Two Layers. Fuzzy Systems Math. 2 (1991) 77-85 (in Chinese).
16. Xie, X.L., Beni, G.: A Validity Measure for Fuzzy Clustering. IEEE Trans. Pattern Anal. Machine Intelligence 13 (1991) 841-847.
17. Yang, M.S., Hu, Y.J., Lin, K.C.R., Lin, C.C.L.: Segmentation Techniques for Tissue Differentiation in MRI of Ophthalmology Using Fuzzy Clustering Algorithms. Magnetic Resonance Imaging 20 (2002) 173-179.
18. Zhang, D.Q., Chen, S.C.: A Novel Kernelized Fuzzy c-means Algorithm with Application in Medical Image Segmentation. Artificial Intelligence in Medicine 32 (2004) 37-50.

On Fuzzy *c*-Means for Data with Tolerance

Ryuichi Murata[1], Yasunori Endo[2], Hideyuki Haruyama[1], and Sadaaki Miyamoto[2]

[1] Graduate School of Systems and Information Engineering, University of Tsukuba Ibaraki 305-8573, Japan
muraryu@soft.risk.tsukuba.ac.jp
[2] Department of Risk Engineering Faculty of Systems and Information Engineering, University of Tsukuba Ibaraki 305-8573, Japan

Abstract. This paper presents two new clustering algorithms which are based on the entropy regularized fuzzy *c*-means and can treat data with some errors. First, the tolerance which means the permissible range of the error is introduced into optimization problems which relate with clustering, and the tolerance is formulated. Next, the problems are solved using Kuhn-Tucker conditions. Last, the algorithms are constructed based on the results of solving the problems.

1 Introduction

Fuzzy *c*-means(FCM) [1] is one of the known fuzzy clusterings and many FCM variants has been proposed after FCM. In this variants, FCM algorithm with based on the concept of the regularization by entropy has been proposed by one of the authors [2]. This algorithm is called regularized entropy FCM(eFCM) and discussed not only its usefulness but also the mathematical relations of other techniques.

In general, any data that is represented by numeric have some errors. We quote three examples as follows.

1. The error which is introduced when the object is mapped the real space to the pattern space.
2. The uncertainty which is introduced in real space.
3. The original range with data.

In this paper, we assume that such errors have some bound and we call the errors tolerance. In addition, we assume that the bound is given preliminarily. In the past, many clustering algorithms which classify the data with the tolerance are examined and one of the authors also proposed one of such algorithms [3, 4]. In these algorithms, the tolerance of the data is represented by a interval value and the nearest distance or Hausdorff are introduced to calculate the dissimilarity between the data with the tolerance. In the case to use these distances as dissimilarity, only the boundary of the interval value which represents the tolerance is needed for calculation of the value of the dissimilarity. On the other hand, when the usual distances, e.g. Euclidean or L_1, are used, the methods to calculate the

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 351–361, 2006.

representative values of the data with the tolerance are determined in advance, for example, average or center. But we consider that not only the boundary but also all the region of the tolerance should be considered and the tolerance should be formulated in the frame of the optimization.

In this paper, we consider new optimization problems and we construct new clustering algorithms based on standard FCM and eFCM for the data with the tolerance through the derivation of the optimal solutions for these problems. Alternatively, we consider two constraint conditions and proposed two algorithms in relation to these conditions.

1.1 Optimization Problem

We define some symbols at the beginning. $X = \{x_1, \ldots, x_n\}$ is a subset of p dimensional vector space $\mathcal{R}^p$ and we write $x_k = (x_k^1, \ldots, x_k^p)^T \in \mathcal{R}^p$. We classify the data set X into c clusters $C_i (i = 1, \ldots, c)$. Let $v_i = (v_i^1, \ldots, v_i^p)^T \in \mathcal{R}^p$ be the center for cluster C_i and $V = \{v_1, \ldots, v_c\}$ be the set of centers for clusters. In fuzzy clustering, the result for clustering is decided on the membership grade. We denote the partition matrix $U = [u_{ik}]$. In addition, $E = \{\varepsilon_1, \ldots, \varepsilon_n\}$ is the set of the tolerance. We assume that $\varepsilon_1, \ldots, \varepsilon_n$ have the upper bound of the tolerance $\kappa_1, \ldots, \kappa_n$, respectively.

We consider the objective functions :

The version of standard FCM:

$$J_s(U, E, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m \|x_k + \varepsilon_k - v_i\|^2. \tag{1}$$

The version of entropy regularized FCM:

$$J_e(U, E, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \|x_k + \varepsilon_k - v_i\|^2 + \lambda^{-1} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik}. \tag{2}$$

In (1), m is the parameter of power which satisfies the condition $m > 1$. In (2), secondly term in the right side is called by the term for regularization by entropy [2]. In this paper, $\|\cdot\|$ stands for the Euclidean norm. Now, we introduce two constraint conditions :

Constraint conditions:

$$\sum_{i=1}^{c} u_{ik} = 1. \tag{3}$$

$$\|\varepsilon_k\|^2 \leq \kappa_k^2 \quad (\kappa_k > 0). \tag{4}$$

In the following subsections, we will derive the solutions u_{ik}, ε_k and v_i which optimize the objective functions J_s and J_e under the constraint conditions (3) and (4).

1.2 The Version of Standard FCM

In this subsection, we derive the solutions which optimize the objective function J_s under the constraint conditions (3) and (4). We introduce the Lagrange function to solve the optimization problem :

$$\begin{aligned} L_1(U,E,W) &= J_s(U,E,W) + \sum_{k=1}^{n}\gamma_k(\sum_{i=1}^{c}u_{ik}-1) + \sum_{k=1}^{n}\delta_k(\|\varepsilon_k\|^2-\kappa_k^2) \\ &= \sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^m\|x_k+\varepsilon_k-v_i\|^2 + \sum_{k=1}^{n}\gamma_k(\sum_{i=1}^{c}u_{ik}-1) \\ &\quad + \sum_{k=1}^{n}\delta_k(\|\varepsilon_k\|^2-\kappa_k^2). \end{aligned} \tag{5}$$

From the Kuhn-Tucker condition, the necessary conditions are as follows.

$$\begin{cases} \dfrac{\partial L_1}{\partial v_i}=0, \quad \dfrac{\partial L_1}{\partial u_{ik}}=0, \quad \dfrac{\partial L_1}{\partial \varepsilon_k}=0, \\ \dfrac{\partial L_1}{\partial \gamma_k}=0, \quad \dfrac{\partial L_1}{\partial \delta_k}\le 0, \\ \delta_k\dfrac{\partial L_1}{\partial \delta_k}=0, \\ \delta_k\ge 0. \end{cases} \tag{6}$$

First, the optimal solution v_i has no constraint condition. So, from

$$\frac{\partial L_1}{\partial v_i} = -\sum_{k=1}^{n}2u_{ik}^m(x_k+\varepsilon_k-v_i)=0, \tag{7}$$

we have

$$v_i = \frac{\sum_{k=1}^{n}u_{ik}^m(x_k+\varepsilon_k)}{\sum_{k=1}^{n}u_{ik}^m}. \tag{8}$$

In the next place, we consider the optimal solution u_{ik}. From

$$\frac{\partial L_1}{\partial u_{ik}} = mu_{ik}^{m-1}\|x_k+\varepsilon_k-v_i\|^2+\gamma_k=0, \tag{9}$$

we have

$$u_{ik} = \left(-\frac{\gamma_k}{m\|x_k+\varepsilon_k-v_i\|^2}\right)^{\frac{1}{m-1}}. \tag{10}$$

In addition, from the constraint condition (3), we have

$$\sum_{j=1}^{c}\left(-\frac{\gamma_k}{m\|x_k+\varepsilon_k-v_j\|^2}\right)^{\frac{1}{m-1}}=1, \tag{11}$$

where we replace i by j to keep from the duplication of symbols. From (9), we have

$$-\gamma_k = m u_{ik}^{m-1} \|x_k + \varepsilon_k - v_i\|^2. \tag{12}$$

When we assign (12) to (11), we obtain

$$\sum_{j=1}^{c} \left(\frac{m u_{ik}^{m-1} \|x_k + \varepsilon_k - v_i\|^2}{m \|x_k + \varepsilon_k - v_j\|^2} \right)^{\frac{1}{m-1}} = 1. \tag{13}$$

The transformation of this equation is

$$u_{ik} = \left(\sum_{j=1}^{c} \left(\frac{\|x_k + \varepsilon_k - v_i\|^2}{\|x_k + \varepsilon_k - v_j\|^2} \right)^{\frac{1}{m-1}} \right)^{-1}. \tag{14}$$

In general, the u_{ik} is not the optimal solution because it is derived from the necessary solution. However, in this case, as the objective function is convex with respect to U, the u_{ik} becomes the optimal solution.

Finally, we derive the optimal solution ε_i. We have

$$\frac{\partial L_1}{\partial \varepsilon_k} = \sum_{i=1}^{c} 2 u_{ik}^m (x_k + \varepsilon_k - v_i) + 2\delta_k \varepsilon_k = 0. \tag{15}$$

The transformation of (15) is

$$\varepsilon_k = \frac{-\sum_{i=1}^{c} u_{ik}^m (x_k - v_i)}{\sum_{i=1}^{c} u_{ik}^m + \delta_k}. \tag{16}$$

On the contrary, from (6), we have

$$\delta_k (\|\varepsilon_k\|^2 - \kappa_k^2) = 0. \tag{17}$$

So, we have to consider two cases, $\delta_k = 0$ and $\|\varepsilon_k\|^2 = \kappa_k^2$ from this equation.

First, let's consider the case of $\delta_k = 0$. In this case, the optimization problem does not have the constraint condition (4). In fact, as $\frac{\partial L_1}{\partial \varepsilon_k} = \frac{\partial J_s}{\partial \varepsilon_k}$, we have

$$\frac{\partial J_s}{\partial \varepsilon_k} = \sum_{i=1}^{c} 2 u_{ik}^m (x_k + \varepsilon_k - v_i) = 0. \tag{18}$$

Notice $\sum_{i=1}^{c} u_{ik} = 1$, then we obtain

$$\varepsilon_k = \frac{-\sum_{i=1}^{c} u_{ik}^m (x_k - v_i)}{\sum_{i=1}^{c} u_{ik}^m}. \tag{19}$$

Next, let's consider the case of $\|\varepsilon_k\|^2 = \kappa_k^2$. When we assign (16) to $\|\varepsilon_k\|^2 = \kappa_k^2$, we have

$$\|\varepsilon_k\|^2 = \left\| \frac{-\sum_{i=1}^{c} u_{ik}^m (x_k - v_i)}{\sum_{i=1}^{c} u_{ik}^m + \delta_k} \right\|^2 = \kappa_i^2. \tag{20}$$

We deform this equation :

$$\sum_{i=1}^{c} u_{ik}^m + \delta_k = \pm \frac{\| \sum_{i=1}^{c} u_{ik}^m (x_k - v_i) \|}{\kappa_k}. \tag{21}$$

When we assign this equation to (16), we have

$$\varepsilon_k = \pm \frac{\kappa_k \sum_{i=1}^{c} u_{ik}^m (x_k - v_i)}{\| \sum_{i=1}^{c} u_{ik}^m (x_i - v_i) \|}. \tag{22}$$

(22) correspond to (19) in the case of $\|\varepsilon_k\|^2 = \kappa_k^2$. Because κ_k and numerator of (22) are positive, the sign of ε_k is negative. So we obtain

$$\varepsilon_k = \frac{-\kappa_k \sum_{i=1}^{c} u_{ik}^m (x_k - v_i)}{\| \sum_{i=1}^{c} u_{ik}^m (x_k - v_i) \|}. \tag{23}$$

From the mentioned, we obtain the following optimal solutions :

$$\begin{cases} v_i = \dfrac{\sum_{k=1}^{n} u_{ik}^m (x_k + \varepsilon_k)}{\sum_{k=1}^{n} u_{ik}^m}, & (24) \\ u_{ik} = \left(\displaystyle\sum_{j=1}^{c} \left(\frac{\|x_k + \varepsilon_k - v_i\|^2}{\|x_k + \varepsilon_k - v_j\|^2} \right)^{\frac{1}{m-1}} \right)^{-1}, & (25) \\ \varepsilon_k = -\alpha_k \displaystyle\sum_{i=1}^{c} u_{ik}^m (x_k - v_i), & (26) \end{cases}$$

where

$$\alpha_k = \min \left\{ \frac{\kappa_i}{\| \sum_{i=1}^{c} u_{ik}^m (x_k - v_i) \|}, \ \frac{1}{\sum_{i=1}^{c} u_{ik}^m} \right\}. \tag{27}$$

1.3 The Version of Entropy Regularized FCM

In this subsection, we derive the solutions which optimizes the objective function J_e under the constraint conditions (3) and (4) as same as the above subsection. To solve the optimization problem, we introduce the Lagrange function :

$$\begin{aligned} L_2(U, E, V) &= J(U, E, V) + \sum_{k=1}^{n} \gamma_k \Big(\sum_{i=1}^{c} u_{ik} - 1 \Big) + \sum_{k=1}^{n} \delta_k (\|\varepsilon_k\|^2 - \kappa_k^2) \\ &= \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \|x_k + \varepsilon_k - v_i\|^2 + \lambda^{-1} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik} \\ &\quad + \sum_{k=1}^{n} \gamma_k \Big(\sum_{i=1}^{c} u_{ik} - 1 \Big) + \sum_{k=1}^{n} \delta_k (\|\varepsilon_k\|^2 - \kappa_k^2). \end{aligned} \tag{28}$$

From the Kuhn-Tucker condition, the necessary conditions are as follows.

$$\begin{cases} \dfrac{\partial L_2}{\partial v_i} = 0, \quad \dfrac{\partial L_2}{\partial u_{ik}} = 0, \quad \dfrac{\partial L_2}{\partial \varepsilon_k} = 0, \\ \dfrac{\partial L_2}{\partial \gamma_k} = 0, \quad \dfrac{\partial L_2}{\partial \delta_k} \leq 0, \\ \delta_k \dfrac{\partial L_2}{\partial \delta_k} = 0, \\ \delta_k \geq 0. \end{cases} \tag{29}$$

First, the solution v_i has no constraint condition. So, from

$$\frac{\partial L_2}{\partial v_i} = -\sum_{k=1}^{n} 2u_{ik}(x_k + \varepsilon_k - v_i) = 0, \tag{30}$$

we obtain the optimal solution:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}(x_k + \varepsilon_k)}{\sum_{k=1}^{n} u_{ik}}. \tag{31}$$

In the next place, we consider the optimal solution u_{ik}. From

$$\frac{\partial L_2}{\partial u_{ik}} = \|x_k + \varepsilon_k - v_i\|^2 + \lambda^{-1}(\log u_{ik} + 1) + \gamma_k = 0, \tag{32}$$

we obtain

$$u_{ik} = e^{\lambda(-\gamma_k - \|x_k + \varepsilon_k - v_i\|^2) - 1}. \tag{33}$$

In addition, from the constraint condition (3), we have

$$\sum_{i=1}^{c} u_{ik} = \sum_{i=1}^{c} e^{\lambda(-\gamma_k - \|x_k + \varepsilon_k - v_i\|^2) - 1} = 1. \tag{34}$$

The transformation of this equation is

$$e^{-\lambda\gamma_k} = \frac{1}{\sum_{j=1}^{c} e^{-\lambda\|x_k + \varepsilon_k - v_j\|^2 - 1}}, \tag{35}$$

where we replace i by j to keep from the duplication of symbols. When we assign (35) to (33), we obtain

$$u_{ik} = \frac{e^{-\lambda\|x_k + \varepsilon_k - v_i\|^2}}{\sum_{j=1}^{c} e^{-\lambda\|x_k + \varepsilon_k - v_j\|^2}}. \tag{36}$$

Last, we derive the optimal solution ε_k. We have

$$\frac{\partial L_2}{\partial \varepsilon_k} = \sum_{i=1}^{c} 2u_{ik}(x_k + \varepsilon_k - v_i) + 2\delta_k\varepsilon_k = 0. \tag{37}$$

Notice $\sum_{i=1}^{c} u_{ik} = 1$, then we obtain

$$\varepsilon_k = -\frac{\sum_{i=1}^{c} u_{ik}(x_k - v_k)}{\delta_k + 1}. \tag{38}$$

From the same discussion in the above subsection, we have

$$\varepsilon_k = \frac{\kappa_k(x_k - \sum_{i=1}^{c} u_{ik} v_i)}{\|x_k - \sum_{i=1}^{c} u_{ik} v_i\|}. \tag{39}$$

From the mentioned, we obtain the optimal solutions:

$$\begin{cases} v_i = \dfrac{\sum_{k=1}^{n} u_{ik}(x_k + \varepsilon_k)}{\sum_{k=1}^{n} u_{ik}}, & (40) \\ u_{ik} = \dfrac{e^{-\lambda\|x_k + \varepsilon_k - v_i\|^2}}{\sum_{j=1}^{c} e^{-\lambda\|x_k + \varepsilon_k - v_j\|^2}}, & (41) \\ \varepsilon_k = -\alpha_k(x_k - \sum_{i=1}^{c} u_{ik} v_i), & (42) \end{cases}$$

where

$$\alpha_k = \min\left\{\frac{\kappa_k}{\|x_k - \sum_{i=1}^{c} u_{ik} v_i\|},\ 1\right\}. \tag{43}$$

2 Algorithms

In this section, we construct new clustering algorithms called sFCM-T and eFCM-T. Two algorithms correspond with the discussions of the subsection 1.2 and 1.3 respectively. These algorithms work out clustering by the alternative optimization procedure.

Algorithm 1 (sFCM-T).

Step 1 [illegible] m [illegible] $\kappa_k\ (k = 1, \ldots, n)$ [illegible] E [illegible] V

Step 2 [illegible] U [illegible]

$$u_{ik} = \left(\sum_{j=1}^{c}\left(\frac{\|x_k + \varepsilon_k - v_i\|^2}{\|x_k + \varepsilon_k - v_j\|^2}\right)^{\frac{1}{m-1}}\right)^{-1}.$$

Step 3 [illegible] E [illegible]

$$\varepsilon_k = -\alpha_k \sum_{i=1}^{c} u_{ik}^m (x_k - v_i),$$

[illegible]

$$\alpha_k = \min\left\{\frac{\kappa_i}{\|\sum_{i=1}^{c} u_{ik}^m (x_k - v_i)\|},\ \frac{1}{\sum_{i=1}^{c} u_{ik}^m}\right\}. \tag{44}$$

Step 4 [illegible] V [illegible]

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m (x_k + \varepsilon_k)}{\sum_{k=1}^{n} u_{ik}^m}.$$

Step 5 [illegible] (U, E, V) [illegible] Step 2

Algorithm 2 (eFCM-T).

Step 1 [illegible] λ [illegible] κ_k $(k = 1, \ldots, n)$ [illegible] E [illegible] V

Step 2 [illegible] U [illegible]

$$u_{ik} = \frac{e^{-\lambda \|x_k + \varepsilon_k - v_i\|^2}}{\sum_{j=1}^{c} e^{-\lambda \|x_k + \varepsilon_k - v_j\|^2}}.$$

Step 3 [illegible] E [illegible]

$$\varepsilon_k = -\alpha_k (x_k - \sum_{i=1}^{c} u_{ik} v_i),$$

[illegible]

$$\alpha_k = \min \left\{ \frac{\kappa_k}{\|x_k - \sum_{i=1}^{n} u_{ik} v_i\|}, \ 1 \right\}.$$

Step 4 [illegible] V [illegible]

$$v_i = \frac{\sum_{k=1}^{n} u_{ik} (x_k + \varepsilon_k)}{\sum_{k=1}^{n} u_{ik}}.$$

Step 5 [illegible] (U, E, V) [illegible] Step 2

3 Numerical Examples

In this section, we show some examples of classification by using sFCM-T and eFCM-T. A classified data set is the star chart of the Polaris and its neighboring stars[1]. This data set is constructed by 51 elements in the two dimensional Euclidean space(Fig.1), that is, the data set is

$$X = \{x_k \mid k = 1, \ldots, 51\} \subset \mathcal{R}^2.$$

We assume that $\varepsilon_1, \ldots, \varepsilon_{51}$ have the upper bound of tolerances $\kappa_1, \ldots, \kappa_{51}$, respectively :

$$E = \{\varepsilon_k \mid k = 1, \ldots, 51\} \subset \mathcal{R}^2.$$

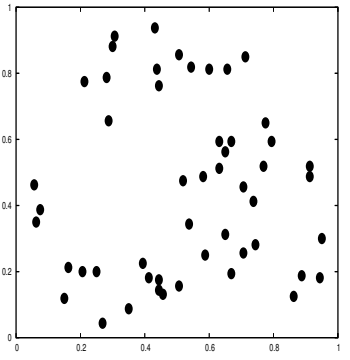


Fig. 1. The star chart of the Polaris and its neighboring stars

We classify the data set into three groups C_i $(i = 1, 2, 3)$:

$$V = \{v_i \mid i = 1, 2, 3\} \subset \mathcal{R}^2,$$
$$U = [u_{ik}], \quad i = 1, 2, 3, \quad k = 1, \ldots, 51.$$

In the following subsections, we show the results by using sFCM-T and eFCM-T. We classify the data set in the following two cases :

Case 1. The upper bound of tolerances $\kappa_1 = \cdots = \kappa_{51} = 0.01$,
Case 2. The upper bound of tolerances $\kappa_1 = \cdots = \kappa_{51} = 0.05$.

In each algorithm, we generate random initial clusters and classify the data set into three clusters. We run this trail 1000 times and show the most frequently pattern of result.

3.1 sFCM-T

In this subsection, we show two examples of classification using sFCM-T. We set $m = 2$ in all algorithms. First we show the result of classification by using sFCM

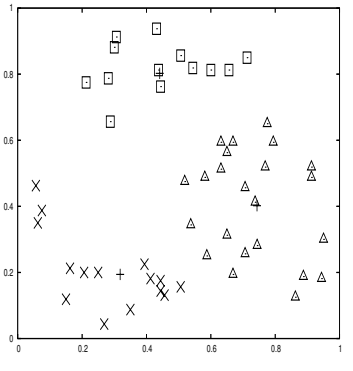


Fig. 2. This pattern is the result by using sFCM

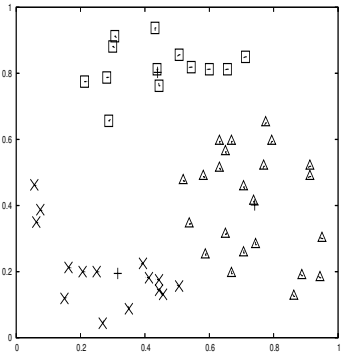


Fig. 3. This pattern is the result by using sFCM-T in the case of the data set have the tolerance $\kappa = 0.01$

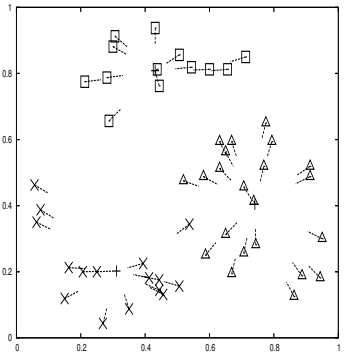


Fig. 4. This pattern is the result by using sFCM-T in the case of the data set have the tolerance $\kappa = 0.05$

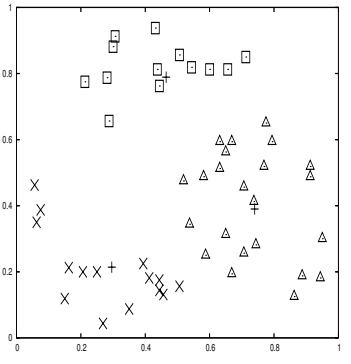


Fig. 5. This pattern is the result by using eFCM

to compare sFCM with sFCM-T(Fig.2). This pattern of result is obtained 1000 out of 1000 trials.

Fig.3 and Fig.4 shows results of the case of $\kappa_k = 0.01$ and $\kappa_k = 0.05$, respectively. The pattern of Fig.3 is obtained 1000 out of 1000 trials. This pattern is the same as Fig.2. On the other hand, the pattern of Fig.4 is obtained 474 out of 1000 trials.

3.2 eFCM-T

In this subsection, we show two examples of classification using eFCM-T. We set $\lambda = 25$ in all algorithms. First we show the result of classification by using sFCM(Fig.5). This pattern of result is obtained 1000 out of 1000 trials.

Fig.6 and Fig.7 shows results of the case of $\kappa_k = 0.01$ and $\kappa_k = 0.05$, respectively. The pattern of Fig.6 is obtained 1000 out of 1000 trials. This pattern is

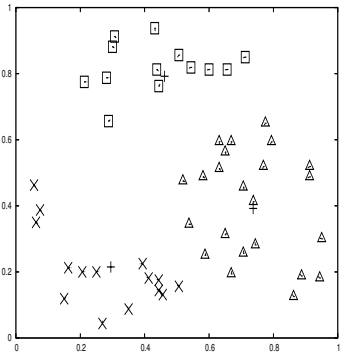


Fig. 6. This pattern is the result by using eFCM-T in the case of the data set have the tolerance $\kappa = 0.01$

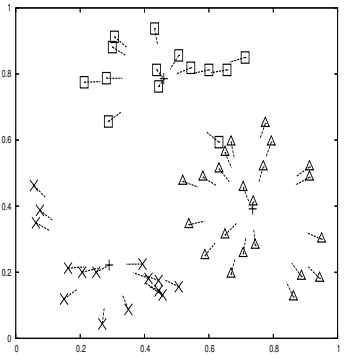


Fig. 7. This pattern is the result by using eFCM-T in the case of the data set have the tolerance $\kappa = 0.05$

the same as Fig.5. On the other hand, the pattern of Fig.7 is obtained 582 out of 1000 trials.

4 Conclusion

In this paper, we formulated the optimization problems for the data with tolerance and derivate the optimal solutions for two object functions. From these results, we have constructed two clustering algorithms.

The proposed techniques are essentially different from the past techniques which treat the interval value as the tolerance. Specifically, we introduced the inequality constraint for the tolerance and formulated it in the frame of the optimization.

We have a lot of problems to discussion : changing the tolerance to the probability or fuzzy by addition of the grade, expanding the functions of the algorithms by some kernel functions and so on.

References

[1] Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
[2] Liu, Z.Q., Miyamoto, S., eds. In: Soft computing and human-centered machines. Springer–Verlag Tokyo, Japan (2000) 85–129
[3] Endo, Y., Horiuchi, K.: On clustering algorithm for fuzzy data. In: Proc. 1997 International Symposium on Nonlinear Theory and Its Applications. (1997) 381–384
[4] Endo, Y.: Clustering algorithm using covariance for fuzzy data. In: Proc. 1998 International Symposium on Nonlinear Theory and Its Applications. (1998) 511–514

On the Use of Variable-Size Fuzzy Clustering for Classification

Vicenç Torra[1] and Sadaaki Miyamoto[2]

[1] Institut d'Investigació en Intel·ligència Artificial,
Campus de Bellaterra, 08193 Bellaterra, Catalonia, Spain
vtorra@iiia.csic.es
[2] Department of Risk Engineering,
School of Systems and Information Engineering,
University of Tsukuba, 305-8573 Ibaraki, Japan
miyamoto@esys.tsukuba.ac.jp

Abstract. Hard c-means can be used for building classifiers in supervised machine learning. For example, in a n-class problem, c clusters are built for each of the classes. This results into $n \cdot c$ centroids. Then, new examples can be classified according to the nearest centroid.

In this work we consider the problem of building classifiers using fuzzy clustering techniques. In particular, we consider the use of fuzzy c-means, as well as some variations. Namely, fuzzy c-means with variable size and entropy based fuzzy c-means.

Keywords: Clustering, Classification, Fuzzy c-means, Variable-size fuzzy c-means, entropy-based fuzzy c-means.

1 Introduction

Clustering [6] and classification [3] are common tools in machine learning [8]. In both cases, sets of examples are considered. In supervised machine learning there is a highlighted attribute that classifies the examples into categories. This attribute determines the class of the examples. This is not the case of unsupervised machine learning. In such framework, all attributes are considered as equal, when knowledge is extracted.

Then, in supervised machine learning, tools have been developed for finding models for the relevant attribute. That is, models are built that permit to assign a class (, assign a value to the relevant attribute) to each new example for which such class is not known. Several different types of models exist based on different assumptions. Examples include, neural networks, (fuzzy) rule-based systems, statistical models, etc.

In unsupervised machine learning, methods have been developed to extract knowledge from the data. Clustering is one of the tools used for extracting such knowledge, as it permits to build structures in which similar objects are put together in clusters.

Besides of this use of clustering as an unsupervised machine learning tool. Clustering can also be used for supervised learning. For example, clustering has been

V. Torra et al. (Eds.): MDAI 2006, LNAI 3885, pp. 362–371, 2006.

used for building models based on the k-means [5] and for building fuzzy rules. In this latter case, clustering is used to build (fuzzy) partitions of the examples.

In this paper we will consider the use of some fuzzy clustering techniques [6, 11] and their use in a classification tool. We propose a method to be used with fuzzy c-means [2] and then we develop variations for the case of entropy-based c-means and their variations with variable size.

The approach presented here includes the advantages that, from the conceptual view, present the alternative clustering methods against fuzzy c-means, and this one with respect to k-means [11].

The structure of the paper is as follows. In Section 2 we review some clustering techniques and an approach to build classification models. Then, in Section 3 we introduce our approach for using fuzzy clustering techniques. Then, in Section 4 we present an application and the experiments performed.

2 Clustering and Classification

In this section we give a review of some aspects of clustering and classification that will be used later on in this work. First we consider a few clustering algorithms, and then we show how to build classifiers using clustering methods.

Here, we consider n examples in a given p dimensional space. We will denote these examples by $x_k \in \mathbb{R}^p$ for $k = 1, \ldots, n$. When the class of each example is known, we will denote this information as follows: $\kappa = \{\kappa_1, \ldots, \kappa_{|\kappa|}\}$ corresponds to the classes; $|\kappa|$ is the number of classes and $\kappa(x_k)$, or simply κ_k, denotes the class for example x_k.

2.1 Fuzzy Clustering

As explained in the introduction, clustering methods are to obtain a set of clusters from a set of examples. In this case, the only information considered is a set of examples x_k in a p dimensional space. Therefore, no information on the class of x_k is required here (although, as we will show latter, this information might be available).

Some of the most well-known algorithms for clustering are the [illegible] (also known as k-means) and the [illegible].

Both methods assume that we know a priori the number of clusters to be built. Such number will be denoted here by the parameter c. Then, the algorithms find a partition of the set of examples (the set $\{x_k\}$) into c different clusters. In hard c-means the partition is a classical one. That is, examples are assigned to only one cluster. Instead, in fuzzy c-means the partition is fuzzy. That is, examples can belong at the same time to different clusters. In this case, the membership to the clusters is not complete but only partial. This situation is modelled using the so-called fuzzy sets and fuzzy memberships.

Here we will use u_{ik} to denote the membership of element x_k to the i-th cluster. In the case of hard c-means as elements are either in the cluster or not in the cluster, we have that u_{ik} is either 0 or 1 (boolean membership). Moreover, as the elements can only belong to one cluster, we have that $\sum_{i=1}^{n} u_{ik} = 1$.

Instead, in fuzzy clustering we have that as membership is partial we have that u_{ik} is in the interval $[0, 1]$. In this latter case, $u_{ik} = 0$ corresponds to non-membership and $u_{ik} = 1$ corresponds to full membership to cluster i. Values in-between correspond to partial membership (the largest the value, the greatest the membership). Nevertheless, in this latter case, the constraint $\sum_{i=1}^{n} u_{ik} = 1$ is maintaned. If this equality holds for all examples k, and we have that $u_{ik} \in [0, 1]$ for all i and k, we say that u defines a fuzzy partition.

In this section we will review first the fuzzy c-means (FCM) algorithm. Then, we will consider one of its variations: fuzzy c-means with variable size (VFCM). And finally, we will also describe an alternative method for fuzzy clustering known as entropy-based fuzzy c-means (EFCM).

Most fuzzy clustering algorithms are defined in terms of a minimization problem with some constraints. In the case of fuzzy c-means [2, 9], the minimization problem is the following one:

$$J_{FCM}(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m ||x_k - v_i||^2 \tag{1}$$

with constraints:

- $u_{ik} \in [0, 1]$
- $\sum_{i=1}^{c} u_{ik} = 1$ for all k

For conciseness, we will denote the values u that satisfy these two constraints by M.

With respect to the notation used above, we have that v_i is recalled as the centroid of the i-th cluster (cluster center/cluster representative), and that m is a parameter ($m \geq 1$) that expresses the desired level of fuzziness. This is, m determines the degree of fuzziness in the membership functions. With values of m near to 1, solutions tend to be crisp (with the particular case that $m = 1$ corresponds to the crisp c-means). Instead, larger values of m yield to clusters with increasing fuzziness in their boundaries.

Local optimal solutions of the fuzzy c-means problem are obtained using an iterative process, that interleaves two steps. The first one that estimates the optimal membership functions of elements to clusters (considering the centroids as fixed) and another that estimates the centroids for each cluster (having the memberships as constant). This process is defined as follows:

Step 1: Generate an initial U and V
Step 2: Solve $min_{U \in M} J(U, V)$ computing:

$$u_{ik} = \Big(\sum_{j=1}^{c} \Big(\frac{||x_k - v_i||^2}{||x_k - v_j||^2} \Big)^{\frac{1}{m-1}} \Big)^{-1}$$

Step 3: Solve $min_V J(U, V)$ computing:

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}$$

Step 4: If the solution does not converge, go to step 2; otherwise, stop

As the method leads to a local optimal, different initial values can lead to different solutions.

The so-called entropy-based fuzzy c-means (EFCM) is an alternative fuzzy clustering method (proposed in [10], see also [11]). The main difference between fuzzy c-means and entropy-based fuzzy c-means is the way in which fuzziness is introduced. In this case, a parameter λ ($\lambda \geq 0$) is used to force a fuzzy solution. Formally speaking, the method is defined in terms of the optimization of the following objective function:

$$J_{EFCM}(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}\{u_{ik}||x_k - v_i||^2 + \lambda^{-1}u_{ik}logu_{ik}\} \tag{2}$$

Again, the objective function is subject to the constraints $u_{ik} \in [0,1]$ and $\sum_{i=1}^{c} u_{ik} = 1$ for all k.

The parameter λ plays a role similar to m in fuzzy c-means. Here, the smaller the λ, the fuzzier the solutions. Instead, when λ tends to infinity, the second term becomes negligible and the algorithm yields to a crisp solution.

The way to solve EFCM is an iterative process, as for the FCM, but with different expressions for computing the memberships u_{ik} and the centroids v_i. More concretely, the following expressions are considered:

$$u_{ik} = \frac{e^{-\lambda||x_k - v_i||^2}}{\sum_{j=1}^{c} e^{-\lambda||x_k - v_j||^2}} \tag{3}$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}x_k}{\sum_{k=1}^{n} u_{ik}} \tag{4}$$

FCM and EFCM lead to different solutions. A relevant difference is that the centroids have a membership equal to one in the FCM while in the EFCM it might have a lower membership. It can be easily observed that given a unique set of centers, the memberships and the shape of the clusters would be different in both cases due to the way memberships u_{ik} are computed.

A variation of these clustering methods was introduced [12] so that the size of each cluster is variable. The variation consists on a variable for each cluster roughly corresponding to its size. The rationale of such introduction was to reduce misclassification when there are clusters of different size. In standard FCM, two adjacent clusters have equal membership function (equal to 0.5) in the mid-point between the two centroids.

Formally speaking, the size of the i-th cluster is represented with the parameter α_i (the largest is α_i, the largest is the proportion of elements that belong to the i-th cluster). A similar approach was given by Ichihashi, Honda and Tani in [7].

When such parameters for variable size are considered, the expressions to minimize for FCM and EFCM are as follows:

$$J_{FCM}(\alpha,U,V) = \sum_{i=1}^{c}\alpha_i\sum_{k=1}^{n}(\alpha_i^{-1}u_{ik})^m||x_k - v_i||^2$$

$$J_{EFCM}(\alpha, U, V) = \sum_{k=1}^{n}\sum_{i=1}^{c}\{u_{ik}||x_k - v_i||^2 + \lambda^{-1}u_{ik}log(\alpha_i^{-1}u_{ik})\}$$

Both objective functions are minimized considering the constraints given above for the membership values u_{ik}, and adding additional constraints for α_i. The new constraints are the following ones:

- $\sum_{i=1}^{c}\alpha_i = 1$
- $\alpha_i \geq 0$ for all $i = 1, \ldots, c$

These fuzzy clustering algorithms are also solved by an iterative process, but now including an additional step for estimating the parameters α_i. In the case of the FCM, the values of α are estimated by (this corresponds to [,] in the algorithm for FCM):

$$\alpha_i = \Big[\sum_{j=1}^{c}\Big(\frac{\sum_{k=1}^{n}(u_{jk})^m||x_k - v_j||^2}{\sum_{k=1}^{n}(u_{ik})^m||x_k - v_i||^2}\Big)^m\Big]^{-1}$$

In such algorithm, the expression for u_{ik} in Step 2 should be replaced by (the expression for v_i in Step 3 is valid):

$$u_{ik} = \Big(\sum_{j=1}^{c}\Big(\frac{\alpha_j}{\alpha_i}\Big)\Big(\frac{||x_k - v_i||^2}{||x_k - v_j||^2}\Big)^{\frac{1}{m-1}}\Big)^{-1} \tag{5}$$

In the case of variable-size EFCM, the expression of v_i is still valid but the following expressions are required for u_{ik} and α_i:

$$u_{ik} = \frac{\alpha_i e^{-\lambda||x_k - v_i||^2}}{\sum_{j=1}^{c}\alpha_j e^{-\lambda||x_k - v_j||^2}} \tag{6}$$

$$\alpha_i = \frac{\sum_{k=1}^{n}u_{ik}}{n}$$

2.2 Classification

In this section we review the use of hard c-means for building classifiers. For this purpose, we consider a set of examples x_k in a p dimensional space, and for each example its class $\kappa(x_k)$.

Then, given a set of examples x_k, the classification model is built considering the following two steps:

1. For each $\kappa_i \in \kappa$, define X_κ as those x_k such that its class is κ (, $\kappa(x_k) = \kappa$):

$$X_\kappa := \{x|\kappa(x_k) = \kappa\}$$

2. Apply hard c-means to each X_κ, and construct for each X_κ c different clusters. Therefore, we obtain $c \cdot |\kappa|$ centroids. We will use v_κ^r for $r = 1, \ldots, c$ to denote the c centroids obtained for class κ.

Then, using the centroids (v_κ^r for $r = 1, \ldots, c$) obtained in the previous step, the classification of new examples ex into the classes κ in κ is done applying the following algorithm:

$minDist = \infty$
For all r do
    For all κ do
        $minDist = \min(minDist, d(ex, v_\kappa^r))$ (distance between ex and the centroid v_κ^r)
    end loop
end loop
Assign ex to the class κ if $minDist = v_\kappa^r$ for some r.

This method corresponds to building a voronoi map in two steps. First, the examples of each class are partitioned and, second, their centroids are put together to define the map.

3 Classification Model Based on Fuzzy Clustering

We have extended the model described in Section 2.2 to incorporate fuzziness. Now, as in the case of crisp partitions, we start grouping the objects according to its class. This is, computing $X_\kappa := \{x | \kappa(x_k) = \kappa\}$. Then, the fuzzy clustering algorithm is applied to each class X_κ. This leads to a set of centroids for each class. We will use v_κ^r for $r = 1, \ldots, c$ to denote the c centroids obtained for class κ. As in the case of the c-means, we obtain $c \cdot |\kappa|$ centroids.

The computation of the class of new objects in the p dimensional space differs from the case of the c-means. In that case, the nearest centroid was considered as the most rellevant issue. Now, we will consider the membership value of the object into each class. Then, for testing we will select the class with the largest membership.

Nevertheless, the consideration of membership functions is not straightforward. As fuzzy c-means (and its variations) results into a fuzzy partition for each class κ, we have that the membership of a new object into the classes can be computed in, at least, two different ways. The two alternatives are described below. We use ex, as in the previous section, to denote the new example to be classified.

1. Consider the c fuzzy partitions as separated partitions. Then, compute for each class κ, and for each cluster r ($r = 1, \ldots, c$), the membership of ex to v_κ^r. Compare memberships and assign to ex the cluster and class with the largest membership.
2. Consider the c fuzzy partitions as a single partition (combined partition). This is, put all centroids together and compute a new fuzzy partition that encompass all the existing clusters. Then, determine the membership of ex to all the clusters and assign to ex the cluster and the class with the largest membership.

The computation of the new fuzzy partition in the second alternative is simple in the case of FCM. In this case, we can compute the fuzzy partition inferred from any set of centroids using the expression u_{ik} in Step 2 (Section 2). Thus, defining the set $\mathcal{V} := \cup_{r=1,\ldots,c} \cup_{\kappa} v_{\kappa}^{r}$ and then using:

$$u(ex, v_i) = \Big(\sum_{v \in \mathcal{V}} \Big(\frac{||ex - v_i||^2}{|ex - v||^2} \Big)^{\frac{1}{m-1}} \Big)^{-1}$$

for all $v_i \in \mathcal{V}$ we can determine the membership of ex to the new partition.

Similarly, when the clustering algorithm used is the EFCM, a similar process can be applied. In this case, Expression 3 should be used for computing the new memberships. So, the membership of ex to clusters v_i in $\mathcal{V}$ is defined as:

$$u(ex, v_i) = u_{ik} = \frac{e^{-\lambda||ex - v_i||^2}}{\sum_{v \in \mathcal{V}}^{c} e^{-\lambda||ex - v||^2}}$$

Instead, when variable size fuzzy c-means is considered, it is not enough to define the union of the centroids and apply the corresponding expressions for computing the membership values. As can be observed in Expressions 5 and 6, these expressions depend on the values of α (the size of the clusters), and when merging the two sets of centroids, the values of α are no longer valid. To solve this drawback we have defined a new vector of α' in terms of the previous values of α. These new values are computed as follows:

$$\alpha'_{\kappa,i} := \alpha_{\kappa,i} * |learningSetClass(\kappa)|/|learningSet|$$

where i is for all $i \in 1, \ldots, c$.

This definition of α' satisfies the constraints that $\alpha_i \geq 0$ and $\sum \alpha_i = 1$.

3.1 Analysis

Methods based on k-means assume that clusters are crisp and that in the resulting model the region under study is splitted (in a crisp way) following a Voronoi tessellation. This tessellation is based on the centroids of the clusters obtained by the k-means.

When fuzzy clustering algorithms are used, these assumptions are changed. First of all, we soften the crisp constraint on the boundaries of each region. Thus, objects can belong at the same time to different clusters. Nevertheless, when for a given point only the largest membership value is considered, the resulting tessellation is still the Voronoi one.

Variable-size fuzzy clustering permits clusters to have different size. Roughly speaking, the larger the number of objects associated to a centroid, the larger the region of the corresponding cluster. In this case, the tessellation changes its shape and not only the centroids come into consideration but also the dimension of the cluster (the parameter α using the notation given above).

Thus, using other fuzzy clustering methods than the k-means for building a classifier, the differences on the clustering model are exported to the classifier. This is for example the case of using a fuzzy clustering method that considers variable size.

4 Experiments

We have applied our approach to the classification of gene expressions in the budding yeast [illegible]. In particular, we used the data downloaded from [13] and described in [4]. Each gene is described in terms of numerical values, and most of them include a label with its name and function. The file contains information on 6221 genes. This data has been used in several studies as in [1].

In our case, we have used these data to compare the four approaches described in Section 3. This is, the classification based on the FCM and the EFCM, with and without variable size.

Some preprocessing was applied to the data as there are missing values. First, data was normalized to avoid scaling problems among variables. Normalization was achieved substracting the mean of each variable and dividing by the corresponding deviation. After normalization, missing values have been replaced by zero (this corresponds to replace the original data by its mean).

In this paper we report the results obtained for the case of the gene labels equals to "mitosis" and "protein degradation".

For testing, we have splitted the data into two sets: one for learning the model and the other for testing. Given a label, we have considered two classes: (i) those genes belonging to the class (positive examples) and (ii) those that do not belong to the class. Then, we selected at random 20% of the records of each class for learning and the rest was used for testing.

For each training/test pair, we have tested the four algorithms FCM and EFCM with and without variable size. Also, in each case we have compared the two alternatives of constructing the membership function (considering the partitions as separated entities or putting them together). For each of the algorithms, several values of m, λ and c were considered. In particular, we have considered the following values for m and λ: $m = \{1.05, 1.2, 1.4\}$, $\lambda = \{40, 20, 10\}$. With respect to c, we have considered two cases, one with $c = 3$ for both positive and negative examples and another with $c = 3$ for positive examples and $c = 8$ for negative examples. The consideration of a larger number of clusters for negative examples was due to the fact that the number of negative examples is much larger than those for positive examples.

Table 1. Rate of success considering separated partitions and combined partitions. In the upper part of the table results corresponds to the "mitosis" problem and the lower part corresponds to the "protein degradation" problem.

Clustering	parameter	c	Separated	Combined
FCM	$m = 1.05$	3	0.5163	0.8426
VFCM	$m = 1.05$	3	0.6355	0.9430
ENT	$\lambda = 40.0$	3	0.0176	0.8117
FCM	$m = 1.05$	3	0.5439	0.8099
VFCM	$m = 1.05$	3	0.5417	0.9639
ENT	$\lambda = 40.0$	3	0.0262	0.7966

Table 2. The rate of success for FCM and variable size fuzzy c-means (FCM) for the "mitosis" and the "protein degradation" problem, for several executions

c	FCM-mitosis	VFCM-mitosis	FCM-P.D.	CFCM-P.D.
3	0.823	0.99051	0.806	0.98392
8	0.867	0.99051	0.845	0.98392
3	0.823	0.99051	0.884	0.98408
8	0.882	0.99051	0.926	0.98408
3	0.851	0.99035	0.808	0.98408
8	0.828	0.99051	0.836	0.98408
3	0.841	0.99035	0.854	0.98424
8	0.914	0.99051	0.829	0.98424
3	0.900	0.99019	0.784	0.98392
8	0.920	0.99019	0.823	0.98408
3	0.832	0.99051	0.864	0.98408
8	0.856	0.99051	0.910	0.98408
3	0.878	0.99067	0.837	0.98392
8	0.889	0.99051	0.877	0.98408
3	0.842	0.99051	0.875	0.98408
8	0.848	0.99051	0.900	0.98408

Each combination of algorithm/parameters was executed 8 times, selecting each time the 20% of the records at random using a different seed.

The results show that considering a single partition lead to better results than considering two separated partitions. Table 1 shows the number of records classified correctly for the algorithms considered for some of the tests. It can be seen that the difference between separated and combined partitions is significant, being the combined partitions better than the separated ones.

Additionally, we can see that when the partitions are combined, the FCM with variable size is the method that obtains better results. These results are valid for both the experiments on "mitosis" and "protein degradation". The results of the 8 executions (with $m = 1.2$ and c either 3 or 8) are given in Table 2.

5 Conclusions and Future Work

In this paper we have studied the use of fuzzy c-means and some of its variations for building classifiers. We have proposed a way to deal with fuzziness and fuzzy partitions when several classes are present. We have analysed the characteristics of our method with respect to the one based on k-means. We have underlined the differences between them. The approach has been applied to data from bioinformatics. In particular, we have applied our method to classify gene expressions from the yeast Saccharomyces cerevisiae.

Although the results obtained by our approach are significant for the two problems studied, the use of fuzzy clustering is not necessarily better than crisp clustering for all sets of examples. The appropriateness of the method depends on

the data and its structure. Nevertheless, we consider that the better performance of variable-size fuzzy c-means with respect to standard fuzzy c-means is rellevant.

As future work we consider the implementation of new experiments, and the study of new methods to combine the fuzzy sets resulting from several fuzzy clustering algorithms (or several executions of the same clustering method with different objects).

Acknowledgements

Partial support by Generalitat de Catalunya (AGAUR, 2004XT 00004) and by the Spanish MEC (project "PROPRIETAS", SEG2004-04352-C04-02 and grant "Salvador de Madariaga" PR2005-0337) is acknowledged.

References

1. Barreto Bezerra, G., Nunes de Castro, L., (2003), Bioinformatics Data Analysis Using an Artificial Immune Network, ICARIS 2003, Lecture notes in Computer Science 2787 22-33.
2. Bezdek, J. C., (1981), Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
3. Duda, R. O., Hart, P. E., Stork, D. G., (2000), Pattern Classification, Wiley.
4. Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D., (1998), Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA, 95 14863-14868.
5. Hastie, T., Tibshirani, R., Friedman, J., (2001), The Elements of Statistical Learning, Berlin: Springer.
6. Höppner, F., Klawonn, F., Kruse, R., Runkler, T., (1999), Fuzzy cluster analysis, Wiley.
7. Ichihashi, H., Honda, K., Tani, N., (2000), Gaussian mixture PDF approximation and fuzzy c-means clustering with entropy regularization, Proc. of the 4th Asian Fuzzy System Symposium, May 31-June 3, Tsukuba, Japan, 217–221.
8. Mitchell, T., (1997), Machine Learning, McGraw Hill.
9. Miyamoto, S., Umayahara, K., (2000), Methods in Hard and Fuzzy Clustering, pp 85–129 in Z.-Q. Liu, S. Miyamoto (Eds.), Soft Computing and Human-Centered Machines, Springer-Tokyo.
10. Miyamoto, S., Mukaidono, M., (1997), Fuzzy c - means as a regularization and maximum entropy approach, Proc. of the 7th International Fuzzy Systems Association World Congress (IFSA'97), June 25-30, Prague, Chech, Vol.II 86–92.
11. Miyamoto, S., (1999), Introduction to fuzzy clustering, Ed. Morikita, Japan. (in Japanese)
12. Miyamoto, S., Umayahara, K., (2000), Fuzzy c-means with variables for cluster sizes, 16th Fuzzy System Symposium, Akita, Sept.6-8, 537–538 (in Japanese).
13. http://rana.lbl.gov/EisenData.htm

Author Index