Studies in Computational Intelligence 537

Rafael Alejandro Espín Andrade
Rafael Bello Pérez · Angel Cobo Ortega
Jorge Marx Gómez · Ariel Racet Valdés
Editors

Soft Computing for Business Intelligence

Springer

Studies in Computational Intelligence

Volume 537

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/7092

About this Series

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

Rafael Alejandro Espín Andrade · Rafael Bello Pérez
Angel Cobo Ortega· Jorge Marx Gómez
Ariel Racet Valdés
Editors

Soft Computing for Business Intelligence

Springer

Editors
Rafael Alejandro Espín Andrade
Instituto Superior Politécnico
José Antonio Echeverría
Centro de Estudios de Técnicas
de Dirección
La Habana
Cuba

Rafael Bello Pérez
Centro de Estudios de Informática Carretera
a Camajuani Villa Clara
Universidad Central Marta Abreu de
Las Villas
Santa Clara
Cuba

Angel Cobo Ortega
Department of Applied Mathematics and
Computer Science
Universidad de Cantabria
Santander
Spain

Jorge Marx Gómez
Department für Informatik
Carl Von Ossietzky Universität Oldenburg
Oldenburg
Germany

Ariel Racet Valdés
Instituto Superior Politécnico
José Antonio Echeverría
La Habana
Cuba

ISSN 1860-949X ISSN 1860-9503 (electronic)
ISBN 978-3-642-53736-3 ISBN 978-3-642-53737-0 (eBook)
DOI 10.1007/978-3-642-53737-0
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013956524

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Prologue

Business Intelligence (BI) is generically based on the combination of data collection from internal and external sources, data integration and applying data analysis methodologies. Originally, the approach focused on the problems of firms. BI generates domain insight, information and special purpose knowledge. This is not an absolute new approach as the BI mainstream may make us believe. For instance, long ago Buddha has made clear: "But after observation and analysis, when you find that anything agrees with reason and is conducive to the good and benefit of one and all, then accept it and live up to it."

The book *Softcomputing for Business Intelligence* is the (successful) joint venture of five editors. It tears down the traditional focus on business, and extends Business Intelligence techniques in an impressive way to a broad range of fields like medicine, environment, wind farming, social collaboration and interaction, car sharing and sustainability. In so far the book is the remarkable output of a program based on the idea of joint trans-disciplinary research as supported by the Eureka IberoAmerica Network and the University of Oldenburg. It is the second publication on BI produced by the network, however, with an interestingly broader view.

The book contains twenty-seven papers allocated to three sections*: Softcomputing*, *Business Intelligence and Knowledge Discovery*, and *Knowledge Management and Decision Making*. Although the contents touch different domains they are similar in so far as they follow the BI principle "Observation and Analysis" while keeping an eye on sound methodologies, mainly Fuzzy Logic, Compensatory Fuzzy Logic (CFL) and extensions.

According to C. Read's slogan "Better to be vaguely right than to be exactly wrong" section I *Softcomputing* include five (slightly overlapping) stimulating papers on compensatory fuzzy logic, fuzzy inference, prospect theory, rough set theory and the related mathematical morphological theory.

"I would rather discover one causal law than be king of Persia" is a statement issued by Democritus, and may be considered as the essential of section II on *Business Intelligence and Knowledge Discovery*. The span of subjects delivered in eleven papers is quite impressive: A fuzzy logic application to assess the vulnerability of protected areas using scorecards, a Balanced Scorecard based framework for IT

management, a classification of time-series of medical data, a fuzzy logic approach to linear regression problems, a review of taxonomies for BI in the tourism industry for improving decision making, assessment of the fuzzy relationships between hypertension and diabetes by compensatory fuzzy logic, CFL based categorization of customer data, a survey of relating metaheuristics and data mining techniques, dashboard based wind farm control specifying key performance measures, portfolio selection of stocks based on neural networks, and, finally, educational data mining by applying soft clustering.

Section III is devoted to *Knowledge Management and Decision Making*, and includes eleven papers. It may be framed by the slogan "What there is anything I do not need ", which is dedicated to Aristotle. This means that raw data is not enough. They should be interpreted by embedding them into a context, analyzing them by sound methods and visualizing them by tables, graphics and – really useful for human beings – footnotes. Such generated information is of an explicit and pragmatic type, and is called knowledge.

The first paper makes a contribution to sustainability indicators using Compensatory Fuzzy logic, the second one generalize FL operators followed by contributions devoted to scheduling problems in case of imprecise data and users' preferences, evaluation of two multi-objective ranking methods and multi-criteria water restoration management, fuzzy data mining recommender systems, and decision making under fuzziness of decision makers' preferences. Next, R&D efficiency is analyzed by various similarity measures and visualized using Kohonen's self organizing maps, QoS of car sharing is studied by a multi-agent framework, multi-criteria decision making is picked up in a further paper, which presents details of ELECTRE III. The final paper makes use of CFL for customer segmentation.

The book is a very good example that BI techniques like Compensatory Fuzzy Logic or, more general, soft computing, are not only useful for enterprises but can be successfully applied to a broad range of domains like environmental studies, assets sharing, water resourcing, social interaction etc.The truth of this is simply coined by Immanuel Kant "There exists nothing more practical than a good theory".

Hopefully, the book will attract more young researchers around the world and not only in Ibero-America to soft computing concerning areas of human beings' vital interest. I wish the book the attention it really deserves.

Hans-J. Lenz
Freie Universität Berlin, Germany

Editors Preface

It is a renowned fact that Knowledge is the principal productive force in the framework of the Knowledge Society's ideas.

In the above context, new scientific paradigm emerges. It deals with transdisciplinary science with non-linear dynamic systems in evolution, being able at the same time to take into account and make advantage of not equilibrium states, uncertainty and vagueness.

Awareness of accelerated interactions among people, environment and economy is recognized throughout the society. In this sense, Sustainable Development paradigm gains on importance. The knowledge base for Sustainable Development grows rapidly with a main characteristic – interconnectivity among heterogenic variables. All this underlines the need of new ways for solving social, ecologic and economic problems.

The new scientific and development pattern claims for holistic analysis according transdisciplinary science. This new kind of analysis should be used for decision making in organizations. Consequently, Business Informatics should turn into position to answer this organizational need through holistic and transdisciplinary analysis.

Business Intelligence as part of Business Informatics has evolved as disciplinarian science utilizing disciplinarian toolboxes and disciplinarian modules. Important elements of Business Intelligence like Management Control, Strategic Analysis, and Data Mining used to be separated modules, without a systemic connection, where Data Mining systems are composed for a lot of very different disciplinarian solutions. Such characteristics determine Business Intelligence systems as partially used, because of the disciplinarian background of the users. This lack of connections between the systems interrupts the automatic-analysis-flows, required for good decision making support.

"Softcomputing for Business Intelligence" is the new initiative of Eureka Iberoamerica Network `www.eurekaiberoamerica.com` and its extra-regional international partner, the University of Oldenburg, Germany. It is a further step after the first joint book "Towards a transdisciplinary technology for Business Intelligence: Gathering Knowledge Discovery, Knowledge Management and Decision Making".

Research program based on transdisciplinary scientific strategy, has been developed by Eureka Iberoamerica Network and the University of Oldenburg. Its objective is to create transdisciplinary technology for Business Intelligence. The program is fostered on a wide understanding of Knowledge Discovery and Business Intelligence with the following objectives:

- Developing of a new transdisciplinary technology for Business Intelligence, joining Knowledge Discovery, Knowledge Management and Decision Making.
- Developing of transdisciplinary theories and practical solutions and strategies towards modeling of different intellectual activities studied from rational thinking such as: Evaluation, Estimation, Reasoning, Learning, Knowledge Discovery and Decision Making, by mixing rationally different disciplines into Operations Research and Computational Intelligence.
- Creating a multidisciplinary repository of knowledge, valuable for organizations' management, in view of all functions, processes and categories involved and using all kinds of sources.

This new book is a contribution to the above stated strategy, gathering interesting mix of papers on the crucial topics: Softcomputing, Business Intelligence, Knowledge Discovery and Decision Making.

It offers selected chapters, joined in three parts: I. Softcomputing, II. Business Intelligence and Knowledge Discovery, and III. Decision Making. Some of them are result of the work according the program; the rest is included in accordance to the strategy pursued.

The open vocation of the network offers information and proposals to scientists and institutions for cooperation, joining research backgrounds and competences. The first part includes some fundamental efforts in the creation of transdisciplinary mathematical science, according fields like Fuzzy Logic, Rough Sets, Probability Theory and Morphology. They contribute to advance towards the transdisciplinary science required, for technology's transformation from boxes of disciplinarian tools and modules, towards holistic systemic analysis, based on users' background knowledge.

In the second part new general methods of Knowledge Discovery based on Fuzzy Logic predicates and Metaheuristics are presented, as well as some classical methods of Data Mining. Furthermore, chapters describing interesting approaches useful for common Business Intelligence tasks are incorporated.

The final third part joins approaches to Decision Making, setting an emphasis on their knowledge use through diverse methods.

Contents

Part I
Soft Computing

Compensatory Fuzzy Logic: A Frame for Reasoning and Modeling Preference Knowledge in Intelligent Systems

Rafael Alejandro Espín Andrade, Eduardo Fernández, and Erick González

Abstract. This paper presents a new approach to designing multivalued logic systems, called Compensatory Fuzzy Logic that besides constituting a formal system with logic properties of remarkable interest represents a bridge between Logic and Decision-Making. The main aim of this proposal is to use the language as key element of communication in the construction of semantic models that make easier the evaluation, decision-making and knowledge discovery. The axioms that constitute the base of this proposal gather actual characteristics of the decision-making processes, and the way of reasoning of people who intervene in them. Some of these axioms are inspired by approaches that adopt a descriptive position in supporting decision-making. Most axioms contain elements of a rational thought. Hence, this logical approach for decision-making may be considered as a third position that combines normative and descriptive components. This approach enters to make part of the arsenal of methods for multicriteria evaluation, adapting itself especially to those situations in which a decision-maker can verbally describe, often in an ambiguous way, the heuristic it uses when executing actions of multicriteria evaluation/classification. Principal kind of operators of Fuzzy Logic are studied according the introduced axioms. Quasi-Arithmetic Based Compensatory Logic is introduced and its particular case, the Geometric Mean Based Compensatory Logic too. The Universal and Existential quantifiers are defined according the definition of this last logic, for discrete and continues sets. An illustration example using Geometric Mean Based Compensatory Logic is used to explain the Compensatory Fuzzy Logic properties.

Rafael A. Espín Andrade · Erick Gonzalez
"Jose Antonio Echeverria" Higher Technical Institute, Cuba
e-mail: rafaelespin@yahoo.com, erickgc@cemat.cujae.edu.cu

Eduardo Fernández
Autonomous University of Sinaloa, Mexico
e-mail: eddyf171051@gmail.com

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_1, © Springer-Verlag Berlin Heidelberg 2014

1 Introduction

Knowledge, reasoning and decision-making have a deep relationship. The process of decision-making for human agents carries a way of reasoning using a body of knowledge about decision alternatives, simultaneously incorporating its subjective reflection in the decision-maker's mind. The modeling of decision-making implies the simulation of this specific way of reasoning, and as such it connects with Artificial Intelligence. Nevertheless, Artificial Intelligence has concentrated its principal effort on diagnosis and knowledge representation; most of its technologies do not model human preferences and keep a narrow margin for subjectivity (cf. [41]). This way of prioritizing the modeling of a rational way of thought neglects the importance of reflecting decision-maker's subjectivity. To stretch bridges between decision analysis and Artificial Intelligence continues being an important issue. Mathematical methods of decision analysis keep centering on the modeling of preferences and risk attitude without emphasis on representing the experience and human reasoning. Functional and relational approaches (including beside those of the Classic Decision Theory, the Analytical Hierarchic Process, MACBETH, TOPSIS, ELECTRE, PROMETHEE and derivations) are kept as the principal manners of constructing the model of decision-maker multicriteria preferences. Although secondary in popularity and applications with regard to the first ones, the Artificial Intelligence close-related paradigm of learning from examples (e.g. [25, 26, 42]) has turned into the third relevant approach for supporting decision-making.

The knowledge of the decision-maker preferences (his/her decision policy) is represented by decisions of assigning objects of the universe to certain pre-existing categories (classification). Later, this knowledge is exploited to produce decisions about new objects. Dominance-based Rough Set approach (cf. [25]) is probably the best representative of this approach. The decision policy (that is implicit in a decision table) is transformed into rules of the type IF…THEN… that allow classifying or arranging new objects. Slowinski et al. [49] proved the equivalence of this paradigm with those that rest on the multi attribute value and on relational methods. Nevertheless, the rough set approach is more flexible because it does not need the fulfillment of preference independence conditions, which are needed by the other approaches. In fact, the decision-maker (*DM*) is obliged to satisfy a single condition: his/her aptitude to evaluate/ classify a core of reference objects. There is an explicit resignation to model rationality, even to model the reasoning within the decision process. In a certain form, the reasoning remains implicit in the reference objects and in the decision rules extracted from them. As decision tool, in practice, the quality of this approach depends on the wealth of the information system – decision table that represents the knowledge about the decision policy. In [48], Slowinski argues that it is easier for the *DM* to exert his/her capacity of making decisions than to make explicit his/her decision policy by constructing a parametric model. In our view, this is a partial truth. The cognitive human limitations make handling of conflicting attributes very difficult when the number of attributes increases beyond a few ones. In fact, many human agents

already have difficulty to treat 5 conflicting criteria (cf. [38]). This can limit severely the wealth of the set for learning from examples. On the other hand, the rejection to explain the decision policy disagrees with one of the aims of decision analysis: to encourage *DM* to think about his/her preferences and values, and in this frame to improve the consistency of his/her judgments, and finally of his/her decision policy (cf. [21]). Regarding this issue the methods based on functional or relational approaches have advantages, since they create an explicit model of preferences which facilitates the process of reflection and increasing consistency.

With the limitations that we have indicated, Dominance-based Rough-Set approach is, at the moment, the most successful bridge between Artificial Intelligence and multicriteria decision analysis. But the knowledge can be represented explicitly. As alternative to the paradigm of learning from examples, let us accept that frequently the *DM* is capable of thinking about his/her preferences with the help of a decision analyst and to explain them verbally making explicit the heuristic which he/she uses to make-decisions that concern the classification, evaluation or sorting of objects of his/her expert domain. Such transformation act of the content of his mind implies possession, discovery of a form of *knowledge* that goes beyond the typical preferential information with the multicriteria method work; it will be called *preferential knowledge*. Making decisions on the basis of this *knowledge* is a way of reasoning that has been called *preferential reasoning* (cf. [54]). The challenge is to use Artificial Intelligence related technologies to represent the preferential knowledge of a *DM*, including the way in which the *DM* aggregates the multicriteria information, and later, "to reason" on the basis of this knowledge in order to arrive to an evaluation or a final decision. If Logic is the best model of the human reasoning, and if making decisions is a way of reasoning on the preferential body, then Logic can be a tool for decision-making. Such idea was firstly proposed in the framework of Deontic Logics, with emphasis on the semantic of the dynamic preferential reasoning (e.g. [29, 30, 36, 55]). From MultiCriteria Decision Aid's perspective multivalent logics have been used as models of the preferential reasoning (cf. [52, 53]). In [5, 11, 20] a wide panorama of the axiomatic approaches to decision making (including the treatment of incomplete knowledge and fuzzy preference modeling) is shown. Nevertheless, a general axiomatic logical approach to decision making, which deals with evaluation and aggregation in a more compatible way with the vagueness of the language and the approximate reasoning, could be a welcome addition to that view.

Expert Systems are pioneering in the idea of obtaining models departing from verbal expressions, so that the human agents can apply their essential experience to concrete problems. The logic-based knowledge representation is here a central issue. More recently, a new discipline called Soft-Computing or Computational Intelligence has been developed (e.g. [56]), having Fuzzy Logic (e.g. [14, 37, 58]) in its foundations. Vagueness and uncertainty are modeled by Fuzzy Logic, which has allowed advances in the modeling of knowledge and decision-making based on verbal expressions (cf. [12, 31, 33, 34]).

The principal advantage of an approach to representation of the preferential knowledge based on Fuzzy Logic, would be exactly the opportunity to use the

language as an element of communication and modeling in the analysis of the decision, creating an explicit model of the preferential knowledge; and later to use the capacity of inference of the logical frame to propose decisions that reflect better the decision policy of the human agent.

That logic for decision-making would be, ultimately, a functional approach that is explicit in its predicates, but preference relations can be modeled like logical predicates too. The axioms that constitute the base of this logic must gather actual characteristics of the decision-making processes, and the way of reasoning of people who intervene in them. Its affinity with the approaches that adopt a descriptive position in supporting decision-making should be natural. But the axiomatic logical body must also contain elements of a rational thought. In this sense, it is possible to consider a logical approach for decision-making as a third position that combines normative and descriptive components. The multivalent logics, with their aptitude to treat the imprecision and the approximate reasoning, allow modeling properties that, though reasonable, lack general validity and therefore cannot be considered to be axioms.

This paper is done with a critical analysis of Decision-Making from the perspective of the Fuzzy Logic. It proposes a new axiomatic-based approach to obey for the Multivalent Fuzzy Logics, looking for a system compatible with the preferential reasoning that characterizes the real processes of decision-making. We have named this new proposal Compensatory Fuzzy Logic (*CFL*). This multivalent approach aims to relate, in a natural way, the deductive thought and human preferences. With a specific choice of its operators, *CFL* becomes in the Geometric Mean Based Compensatory Logic. In the aim to model rational approximate reasoning, the proposed multivalued system achieves a more complete compatibility with the Boolean Propositional Calculus.

This paper is organized as follows: the main motivations for a new decision support-oriented multivalent logic system are given in Section 2, followed by its axioms (Section 3) and a description of its operators. Its compatibility with Boolean Logic is analyzed in Section 4. Its compatibility with the order is reviewed in Section 5, and its usefulness is tested by solving a real decision problem (Section 6). Finally, some concluding remarks are discussed in the last section.

2 Fuzzy Logic and Multicriteria Decision-Making

One way of applying the "Gradualism Principle" – essential property of Fuzzy Logic- is the definition of logic where predicates are functions of the universe within the interval [0,1], and conjunction, disjunction, negation and implication operations are defined in such a way that their restriction to domain {0,1} results in Boolean logic. Different ways to define the operations and their properties determine different multivalent logics that are part of Fuzzy Logic Paradigm (cf. [14]). More recently, a new discipline has emerged from Fuzzy Logic, called narrow Fuzzy Logic or Mathematical Fuzzy Logic (cf. [16, 27, 28]).

In this framework the modeling of linguistic information is an important issue. The Calculus with Words proposed by Zadeh and Kacprzyk [59] departs from the so-called linguistic variables. Several approaches to decision making have used those variables for modeling linguistic information, combined with fuzzy preference relations and matrices (cf. [31, 32, 33, 59]). There are important advances in modeling of preferences through linguistic information (e.g. [12, 31, 33, 34]).

In the modeling of linguistic information the use of isolated aggregation operators has often supplanted the employment of groups of operators as a logical system (cf. [13, 15]). In consequence, there is a risk of losing the systemic conception of reasoning. The use of the language as an element of communication between an analyst and a decision-maker (the same as between a knowledge engineer and an expert in developing Expert Systems) needs more use of a system of operators (as in multivalent logic systems). The combined operators as a logic system can facilitate the modeling from judgments expressed in natural language, in the same way the logical formulae are used to represent reasoning in the mathematical logic. Here, a system formed by four fundamental operators will be considered in order to define a multivalent logic capable of reflecting and inferring preferential knowledge. These operators are conjunction, disjunction, negation and strict order. The properties of these operators must correspond to those of the way of reasoning to be modeled. To make this clearer, let's consider multicriteria decision in its three fundamental problems: classification, selection and ranking. In the three problems judgments are made on objects or actions $\boldsymbol{a} = (a_1, a_2,\ldots a_N)$, where a_i represents the condition of the i-th attribute. In classification the judgments are absolute; it is a question of determining to what class the object belongs, in agreement with its similarity to representative objects of each class.

Let $G_k = (g_1, g_2,\ldots g_N)$ be the pattern that identifies the k-th class. If x_i is the truth value that a_i is similar to g_i, the conjunction c of $x_1, x_2,\ldots x_N$ gives the truth value that **a** is similar to G_k. Comparison of similarity of truth values with the different classes should allow a prescription.

In the problems of selection, when a set A of potential actions is given, it is a question of finding the smallest $B \subset A$ possible so that non-consideration of all actions in A-B could be justified. Let us suppose that x_i is the truth value of the proposition "the level a_i of the i-th attribute is convenient for the *DM*´s objectives". Then, values provided by the conjunction of $x_1, x_2,\ldots x_N$ (or by a combination of logical operators) reflect the aggregated convenience of the action $\boldsymbol{a}$ as a solution of decision problem, that is, its convenience as a member of B. The truth values for the entire $\boldsymbol{y} \in A$ offer reasons to select B and discard A-B. In ranking problems, it is a question of performing a partial order of A, discovering and ordering equivalence classes. Values provided by conjunctions (or by a combination of logical operators) can be used to determine whether $\boldsymbol{a}$ and $\boldsymbol{b}$ belong to the same equivalence class or if one of them should be ranked better than the other. In all cases, the strict-order operator can be used to distinguish whether c($\boldsymbol{a}$) is clearly different from c($\boldsymbol{b}$) or if an indifference can be accepted.

Since ultimately a logic system is a functional model based on operators, a decision- making oriented logic should reflect properties from the functional approach to decision problems. In the following, this issue is illustrated by considering a very popular functional model. Let us suppose that the *DM*'s system of preferences is modeled by

$$U = w_1u_1 + w_2u_2 + \cdots + w_Nu_N \quad (1)$$

which is defined on a decision set A, with $u_i \in [0,1]$, $w_i>0$ and $w_1 + w_2 + \cdots + w_N = 1$. This is a widely used model under preferential independence conditions. In order to make a connection with logic approaches, we can accept that for each particular element of the decision set, the *i*-th cardinal function value measures how convenient the associated attribute is for the *DM*. *U* measures how globally convenient that element is in terms of the convenience of its attributes.

The following properties are held:

1. ui is a cardinal function on the domain of the respective i-th attribute; for each particular dimension, a strict order is established by >;

U represents a weak order on A; besides, if (a,b) ∈ A×A, U(a) > U(b) ⇔ "a is strictly preferred to b";

Let a be an element of A. Let us denote by $umin(a)$ $(umax(a))$ the minimum (maximum) value of $ui(a)$ $(i = 1, \dots N)$. Note that $umin(a) < U(a) < umax(a)$;

For each i, *U* is strictly increasing function of ui;

If for each $ui(a) = usame$, then $U = usame$;

If $wj = 1$ were permitted with $uj = umin(a)$, then U would be equal to $umin(a)$. With $uj = umax(a)$, U would be equal to $umax(a)$. These are the extreme Max-min and Max-max approaches, respectively.

The above properties are not limited to simple models. In more complex models for *DM*'s preferences as:

$$U = f_1(u_2, u_3 \dots u_N)u_1 + \cdots f_i(u_1, \dots u_{i-1}, u_{i+1} \dots u_N)u_i + \cdots f_N(u_1, \dots u_{N-1})u_N \quad (2)$$

with $f_1(u_2, u_3 \dots u_N) + \dots f_i(u_1, \dots u_{i-1}, u_{i+1} \dots u_N) + \dots f_N(u_1, \dots u_{N-1}) = 1$ and

$$U = ({u_1}^{w1}.{u_2}^{w2} \dots {u_N}^{wN})^{1/N} \quad (2')$$

with $w_1 + w_2 + \cdots + w_N = N$

Properties 1-6 are still satisfied.

Although different axiomatic initiatives have been recently proposed (see, for instance, [24] for the concept of uninorm), most of the multivalent systems use t-norm and c-norm axiomatic structure for conjunction and disjunction operators (cf. [27]). In our view, there are two properties of t-norm operators that are hardly compatible with preferential reasoning. These are the following:

A. If *c* is a t-norm, and x_1 and x_2 are truth values of two predicates, then $c(x_1, x_2) \leq \min(x_1, x_2)$;
B. If *c* is a t-norm , *c* is an associative operator.

Property A does not match the above Properties 3, 5 and 6. Dyson ([17]) proved that the conjunction "min" operator is equivalent to the maxmin approach for multicriteria decision making. As much as $c(x_1, x_2) \leq \min(x_1, x_2)$, it corresponds to a way of reasoning even more "pessimistic" than the maxmin approach. On the other hand, if the associative property is fulfilled, hierarchical trees of objectives representing different preferences produce equal truth values of their compound predicates. Under this property, the two trees in figure 1 would represent the same preferences, something inappropriate in a decision-making model. It is obvious, for example, that the objective *x* is of greater relevance in the right-side tree than in the left-side one.

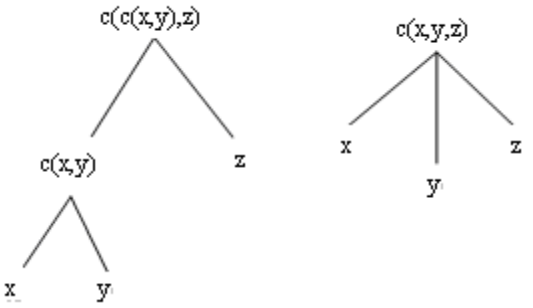


Fig. 1 The advantage of the non-associativeness of the CFL

From the above arguments, unlike most multivalent logic systems, a new logic for decision making should not be based on t-norms and c-norms.

The complete transitivity and comparability of functional models have been extensively criticized by the MultiCriteria Decision Aid (*MCDA*) European school (e.g. [45]). In [44], Roy described situations in which a real *DM* or decision actor cannot (or does not want to) make a decision. These hesitations may come from:

- the *DM* is a vaguely defined entity, or even a well precised entity with poorly defined preference rules [44];
- the existence in the *DM*'s mind (if the *DM* is a real person) of certain "zones" of uncertainty, imprecise beliefs, conflicts and competing aspirations [44];
- imprecise attribute values.

MCDA has introduced the notions of veto and incomparability. They are important issues for a logic system which pretends to reflect imprecise knowledge and reasoning.

The above Properties 1-6 together with veto and incomparability are desirable characteristics for a decision-making oriented logic system. Here we present a multivalent compensatory logic which satisfies the above requirements. This appears as a combined model of decision-making and approximate reasoning.

Preference systems more complex than those represented by (1) can be modeled by this approach, but using a combination of logical operators which comes from a linguistic expression of decision-maker's preferences. The *MCDA* point of view can be incorporated by i) modeling veto conditions and ii) modeling the above *DM*'s hesitations by fuzzy orders, which treats comparisons in a non-deterministic way. This process may be considered a way of making Knowledge Engineering on Preferences.

The aim of combining Logic and Decision-Making on a linguistic framework should be supported by a system of axioms which allows: a) to model some kind of rationality compatible with approximate reasoning, and descriptive (behavioral) effects of veto; b) to handle linguistic preference information from the *DM*; c) to aggregate preference information and handling compromises and trade-offs from the *DM*. This axiomatic system is presented in the next Section.

3 Compensatory Fuzzy Logic

Let n be a negation operator from [0,1] to [0,1], or a strictly decreasing operator fulfilling n(n(x)=x, n(0)=1 and n(1)=0. [14].

Let from now on $x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_n)$, $z = (z_1, z_2, \dots, z_n)$ be any element of the Cartesian product $[0,1]^n$.

A quartet of continuous operators (*c, d, o, n*), *c* and *d* from $[0,1]^n$ to [0,1], the operator *o* from $[0,1]^2$ to [0,1] and *n* a negation operator, constitute a Compensatory Fuzzy Logic (*CFL*) if the following group of axioms is satisfied:

I. Compensation Axiom:
$$\min(x_1, x_2, \dots, x_n) \leq c(x_1, x_2, \dots, x_n) \leq \max(x_1, x_2, \dots, x_n)$$
II. Commutativity or Symmetry Axiom:
$$c(x_1, x_2, \dots, x_i, \dots, x_j, \dots, x_n) = c(x_1, x_2, \dots, x_j, \dots, x_i, \dots, x_n)$$
III. Strict Growth Axiom: If $x_1 = y_1, x_2 = y_2, x_{i-1} = y_{i-1}$, $x_{i+1} = y_{i+1}, \dots, x_n = y_n$ are unequal to zero, and $x_i > y_i$ then $c(x_1, x_2, \dots, x_n) > c(y_1, y_2, \dots, y_n)$
IV. Veto Axiom: If $x_i = 0$ for an *i* then $c(x) = 0$.
V. Fuzzy Reciprocity Axiom: $o(x, y) = n[o(y, x)]$
VI. Fuzzy Transitivity Axiom: If $o(x, y) \geq 0.5$ and $o(y, z) \geq 0.5$, then $o(x, z) \geq \max(o(x, y), o(y, z))$
VII. Morgan's Laws:
$$n(c(x_1, x_2, \dots, x_n)) = d(n(x_1), n(x_2), \dots, n(x_n))$$
$$n(d(x_1, x_2, \dots, x_n)) = c(n(x_1), n(x_2), \dots, n(x_n))$$

The operators *c* and *d* are called conjunction and disjunction, respectively. The operator *o* is called fuzzy strict order.

The Compensation Axiom gives name to the proposed structure; the property it reflects is usually employed in the literature on fuzzy operators to define the concept of compensatory operator (cf. [13]). Note that for the particular case of two

components, the fact that operator's value is between the minimum and maximum can be interpreted like the second value compensates the value of the first in the truthfulness of a conjunction. The idea is generalized for the case of n components. This axiom is consistent with Properties 3 and 6 of Section 2. Note that as $c(x_1, x_2, \ldots, x_i, \ldots, x_j, \ldots, x_n) > min(x_1, x_2, \ldots, x_i, \ldots, x_j, \ldots, x_n)$, the *DM*'s reasoning is less "pessimistic" than the maxmin way of thinking.

The Symmetry Axiom is desirable for it is natural that the conjunction result is independent from the order in which basic predicates are taken.

The introduction of Strict Growth Axiom provides sensitivity to the system that makes any variation in the values of basic predicates modify the truth value of a compound predicate, provided that none of the basic predicates has a value of zero. An additional consequence of this axiom is the desired property of non-associativeness, since there are no strictly increasing associative compensatory operators (cf. [15]). This axiom is consistent with Property 4 of Section 2.

The Veto Axiom is inspired by the *MCDA* approach as was discussed in Section 2. This property provides any conjunction of a basic predicate with the capacity of vetoing, i.e., capacity of preventing any form of compensation when its value is equal to zero.

In a "fuzzy framework", Axioms V-VI match with the properties of the strict preference relation from functional decision models (Property 1-2 of Section 2). The fuzzy order allows associating a truth value with the strict preference statement; it is a way of modeling impreciseness that frequently leads to incomparability.

As from now, given a negation operator and in accordance with axioms V and VI, a strict order is a predicate o: $U^2 \rightarrow [0,1]$ that meets both of the following conditions:

A1. $o(x,y) = n[o(y,x)]$ (generalized fuzzy reciprocity)

B1. If $o(x,y) \geq 0.5$ and $o(y,z) \geq 0.5$, then $o(x,z) \geq \max(o(x,y), o(y,z))$ (max-max fuzzy transitivity or strong stochastic transitivity)

The concept of fuzzy strict order is approached in different ways in the literature (cf. [4, 8, 9, 10, 22, 23, 50, 51]). The property of anti-symmetry $p(x,y) > 0 \Longrightarrow p(y,x) = 0$ employed by other authors to define strict order is not compatible with the desired sensitive behavior in view of the changes in basic predicates (Strict Growth Axiom). According to Switalski [50, 51] and, García-Lapresta and Meneses-Poncio [22], the selection of the strong property, max-max fuzzy transitivity, in the presence of fuzzy reciprocity allows satisfying a group of desirable properties that provide greater meaning to the strict order. In this proposal fuzzy predicates are playing the role of utilities as was suggested by Switalski in [51]. Preference relations associated to the proposed strict fuzzy order are compatible with rational utility criteria (cf. [51]). This is a relevant fact for the aim of combining elements from the normative and descriptive approaches to decision-making.

The property of reciprocity is obtained in Axiom V selecting $n(x) = 1 - x$. Strong stochastic transitivity is compatible with reciprocity. The known alternative definitions of transitivity, weak, and moderate stochastic transitivity, can be

proved from the strong order in conditions of reciprocity (cf. [51]). T-norms-based transitivities studied by Switalski [51] are not compatible with the present axiomatic proposal because the norm concept is not playing any role in this context.

Note that transitive property of the models (1) and (2) is fulfilled in the deterministic case: If $o(x,y)=1$ and $o(y,z)=1$, then $o(x,z)=1$.

Multiplicative Transitivity has been proposed recently by Chiclana et al. (cf. [6]) for the analysis of transitivity of reciprocal preferences. Being compatible with reciprocity and having some other good properties, Multiplicative Transitivity could be used as alternative to the max-max fuzzy transitivity in the above Point $B1$.

With the definition provided by above $A1$ and $B1$, the function $o(x,y) = 0.5[c(x)-c(y)]+0.5$.

With $n(x)=1-x$ is a strict order over the universe of the predicate C (see [14], p. 12), which has been already used successfully [19]. The predicate $o(x,y)$ then allows measuring "how much better is $\boldsymbol{x}$ than $\mathbf{y}$" if C measures the convenience of *x, y* alternatives for the decision maker. If $o(x,y)=0.5$, then $\boldsymbol{x}, \boldsymbol{y}$ would be considered indifferent. Furthermore, this logic order can be used more generally to compare truthfulness of statements modelled through predicates. It is an instrument to establish a relationship between decision-maker's preferences and truthfulness attributed to his/her knowledge. These elements appear separately in most of the theoretical models proposed previously.

Morgan's laws are essential properties in the behavior that in natural and universally accepted manner relate conjunction (c) and disjunction (d) operators. After the above selection of o and n operators, their introduction allows easy confirmation of a behavior similar to that of the conjunctive operator expressed in the following properties:

1. Compensation Property:
   $\min(x_1, x_2, \dots, x_n) \le d(x_1, x_2, \dots, x_n) \le \max(x_1, x_2, \dots, x_n)$
2. Symmetry Property:
   $$d(x_1, x_2, \dots, x_i, \dots, x_j, \dots, x_n) = d(x_1, x_2, \dots, x_j, \dots, x_i, \dots, x_n)$$
3. Strict Growth Property:
   If $x_1 = y_1, x_2 = y_2, x_{i-1} = y_{i-1}, x_{i+1} = y_{i+1}, \dots, x_n = y_n$ are unequal to zero, and $x_i > y_i$ then $\mathrm{d}(x_1, x_2, \dots, x_n) > d(y_1, y_2, \dots, y_n)$
4. If $x_i = 1$ for an i then $d(x) = 1$

For a vector $(a, a, \dots, a)$ at any $a \in [0,1]$, from the compensation axiom follows that $a = \min(a, a, \dots, a) \le c(a, a, \dots, a) \le \max(a, a, \dots, a) = a$. This result alows the conclusion by means of one of Morgan's laws that the disjunction satisfies the same inequality. Therefore, the following Idempotency Property is met:

5. $c(a, a, \dots, a) = a, d(a, a, \dots, a) = a$

This property is consistent with Property 5 of Section 2. It is a consequence from Axiom I. Note that $c(1,1,\dots,1)=1$. With Veto Axiom this result generalizes the Boolean conjunction. Note also that non-compensatory operators such that $c(x_1, x_2) < min(x_1, x_2)$ do not satisfy idempotency, thus not matching with Property 5 of Section 2.

There are many possibilities to define the implication in accordance with advances found on the matter in the literature (cf. [1, 2, 16]); in this work, we rather start from definitions that use conjunction, disjunction and negation operators to explore the effect that this class of operators has on the implication behavior.

In general, the implication can be defined as $i_1(x,y) = d(n(x),y)$ (1) or $i_2(x,y) = d(n(x),c(x,y))$ (2), thus generalizing truth tables of Boolean logic in two different ways.

The equivalence is consequently defined from the implication operator as $e(x,y) = i(x,y) \wedge i(y,x) = c(i(x,y),i(y,x))$.

The universal and existential quantifiers must be introduced naturally from the selected conjunction and disjunction operators; when these are introduced, at any fuzzy predicate p over the universe U, universal and existential propositions are defined respectively as:

$$\underset{x\in U}{\forall} p(x) = \underset{x\in U}{\wedge} p(x) \tag{3}$$

$$\underset{x\in U}{\exists} p(x) = \underset{x\in U}{\vee} p(x) \tag{4}$$

Operators that satisfy Axiom 1 are called Compensatory Operators (cf. [13]). The symmetric compensatory operators found in the literature are the following [13]:

1. The maximum and minimum operators; the first one does not satisfy Strict Growth and Veto Axioms; the minimum operator does not satisfy the Strict Growth Axiom.

The k-order statistics, that include median operator; they do not satisfy the Strict Growth and Veto Axioms.

Combinations of norm and co-norm like exponential compensatory operators (that include the called Zimmerman operator) and convex linear compensatory operators; they do not satisfy the Veto Axiom.

The arithmetic mean, which does not satisfy the Veto Axiom.

The quasi-arithmetic means, which include for example the geometric mean. They are operators of the form $M_f(x_1,x_2,\ldots,x_n) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right)$, where f is a strictly monotone continuous function which is extended to non-defined points by using the corresponding limit. These operators satisfy Axioms I-III. If we have in addition that for all $i \in \{1,2,\ldots n\}, \lim_{x_i\to 0} M_f(x_1,x_2,\ldots,x_n) = 0$, we will be having axiom IV too. Then, taking $d(x_1,x_2,\ldots,x_n) = 1 - f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(1-x_i)\right)$ $n(x) = 1 - x$ and $o(x,y) = 0.5[c(x) - c(y)] + 0.5$ we have a class of Compensatory Logics that we can call Quasi Arithmetic Mean Based Compensatory Logic (QAMBCL).

Within the class of *QAMBCL* we should select a specific $M_f(x_1, x_2, \dots, x_n)$with additional desirable properties. In the next Section we discuss a particular multi-valued system based on the geometric mean operator.

For the case of delimited sets of $\mathfrak{R}^n$ universal and existential quantifiers in *QAMBCL* are defined naturally from conjunction and disjunction concepts, respectively, passing to the continuous case through integral calculus from formulas 3 and 4:

$$\forall x\, p(x) = \begin{cases} f^{-1}\left(\dfrac{\int_X f(p(x))dx}{\int_X dx}\right) & if\ p(x) > 0 \text{ for entire } x \in X \\ 0 & \text{In any other case} \end{cases} \tag{5}$$

$$\forall x\, p(x) = \begin{cases} 1 - f^{-1}\left(\dfrac{\int_X f(p(x))dx}{\int_X dx}\right) & if\ p(x) > 0 \text{ for entire } x \in X \\ 1 & \text{In any other case} \end{cases} \tag{6}$$

4 Geometric Mean Based Compensatory Logic (GMBCL)

The geometric mean belongs to the class of quasi arithmetic means. From the above Point 5, the geometric mean can be obtained making $f(x) = \ln x$. This choice satisfies $\lim_{x_i \to 0} M_f(x_1, x_2, \dots, x_n) = 0$ and consequently axioms I-IV.

The geometric mean is one of the most studied and applied quasi-arithmetic means, especially in the context of Decision Making. It is a particular case of the Ordered Weighted Geometric Operator (*OWG*) which has very good properties in the context of Fuzzy Decision Making. *OWG* keeps the reciprocal property of order relations after aggregation, and guarantees consistency of the obtained fuzzy order (cf. [32, 46, 47, 57]). The geometric mean is also simpler than other quasi-arithmetic means. So, we propose to take the geometric mean

$c(x_1, x_2, \dots, x_n) = (x_1, x_2, \dots, x_n)^{1/n}$ as conjunction operator and study a Compensatory Logic based on it, analyzing its properties related to the Boolean Logic and the compatibility with the order.

Consistently with Morgan's Laws, the corresponding disjunction would be: $d(x_1, x_2, \dots, x_n) = 1 - ((1 - x_1)(1 - x_2) \dots (1 - x_n))^{1/n}$

From the above statements, the quartet of operators formed by the geometric mean and its dual as conjunctive and disjunctive operators, together with the order $o(x, y) = 0.5[c(x) - c(y)] + 0.5$ and the negation $n(x) = 1 - x$ constitute a Compensatory Logic that will be named *Geometric Mean Based Compensatory Logic* (*GMBCL*).

For *GMBCL*, definitions of universal and existential quantifiers would be respectively:

$$\underset{x\in U}{\forall} p(x) = \underset{x\in U}{\wedge} p(x) = \sqrt[n]{\prod_{x\in U} p(x)}$$
$$= \begin{cases} \exp\left(\frac{1}{n}\sum_{x\in U}\ln\left(p(x)\right)\right) & if\ p(x) \neq 0, \text{for entire } x \in U \\ 0 & if \text{ for some } x, p(x) = 0 \end{cases} \quad (3')$$

$$\underset{x\in U}{\exists} p(x) = \underset{x\in U}{\vee} p(x) = 1 - \sqrt[n]{\prod_{x\in U}\left(1 - p(x)\right)}$$
$$= \begin{cases} 1 - \exp\left(\frac{1}{n}\sum_{x\in U}\ln\left(1 - p(x)\right)\right) & if\ p(x) \neq 0, \text{for entire } x \in U \\ 0 & if \text{ for some } x, p(x) = 0 \end{cases} \quad (4')$$

For the case of delimited sets of $\Re^n$, universal and existential quantifiers in *GMBCL* are defined naturally from conjunction and disjunction concepts, respectively, passing to the continuous case through integral calculus:

$$\forall x\, p(x) = \begin{cases} e^{\frac{\int_X \ln(p(x))dx}{\int_X dx}} & if\ p(x) > 0 \text{ for entire } x \in X \\ 0 & \text{In any other case} \end{cases} \quad (5')$$

$$\exists x\, p(x) = \begin{cases} 1 - e^{\frac{\int_X \ln(1-p(x))dx}{\int_X dx}} & if\ p(x) > 0 \text{ for entire } x \in X \\ 1 & \text{In any other case} \end{cases} \quad (6')$$

5 An Illustrative Example: Company Competitiveness

The following model sorts a group of companies on a competitive market using *GMBCL*. Expert specialists from *BIOMUNDI* consulting firm, which has achieved a huge development in the supply of competitive intelligence services in Cuba, took part in its construction. The model corresponds to a consulting work about the market of tissue adhesives in several geographic regions.

Below are verbal expressions and their translation to the language of Predicate Calculus:

A company is competitive in a line of products for a given market if 1*) the economy of the company is sound and* 2) *has a leading-edge technology position* 3) *is very strong in a product line on the reference market.*

1. A company is economically sound if it has a good financial situation and good sales. If the financial situation were little bad, it should be compensated by very good sales.
2. A company has a leading-edge technology position if its present technology is good and in addition, is patent owner or has products under Research & Development, or destines significant amounts of money for this activity. If its technology is little behind, then it should have many patents, or many products under Research & Development, or destine significant amounts of money for this activity.

A company is strong in a product line if it has strength on the market, has a varied line of products and is independent from the supplier

The model is the following compound predicate:

$$C(x) = s(x) \wedge T(x) \wedge l^2(x)$$

where:

$$s(x) = f(x) \wedge v(x) \wedge (\neg(f(x))^{0.5} \rightarrow v^2(x))$$

$$T(x) = t(x) \wedge (p(x) \vee i(x) \vee d(x)) \wedge (\neg t^{0.5}(x) \rightarrow (p^2(x) \vee i^2(x) \vee d^2(x)))$$

and

$$l(x) = m(x) \wedge vl(x) \wedge ip(x)$$

The predicates have the following meanings:

C(x): The company *x* is competitive
s(x): The company *x* has a sound economy
T(x): The company *x* has a leading-edge technology position
l(x): The company *x* is strong in the product line
f(x): The company *x* has a good financial situation
v(x): The company *x* has good sales
t(x):The company *x* has a good technology at present
p(x):The company *x* is owner of patents
i(x):The company *x* has products under research and development
d(x):The company *x* destines significant amounts of money to research & development
m(x):The company *x* has strength on the market
vl(x):The company *x* has a varied line of products
ip(x):The company *x* is independent from the supplier

The study and use of functions of the form $f(x) = x^a$ where a is a real number, or of other forms, as ways of modifying the function of associated belonging

to a predicate in order to achieve a new meaning that modifies the previous by either enhancing or attenuating it, is a usual practice in fuzzy logic applications (cf. [15, 40]). Note that in this case, exponents 0.5 and 2 are used to model words a bit (more or less) and much (very) respectively, which is a usual practice (cf. [39]).

Figures 2-5 illustrate the model through a logic tree. Figure 2 illustrates the conjunctive predicate *C(x)* for evaluating competitiveness. Figures 3, 4 and 5 include, step by step, tree predicates which define Economic Soundness ($s(x)$), Leading-Edge Technology Position ($T(x)$) and Strength in Product Line ($l(x)$).

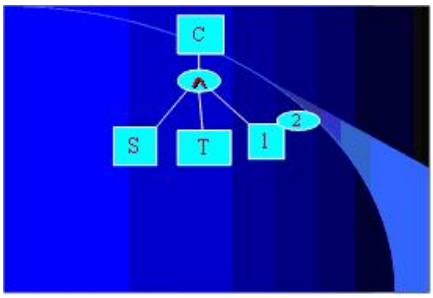


Fig. 2 Conjunctive predicate for evaluating competitiviness

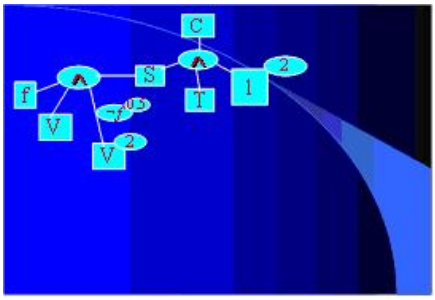


Fig. 3 Inclusion of the Economic Soundness ($S(x)$)

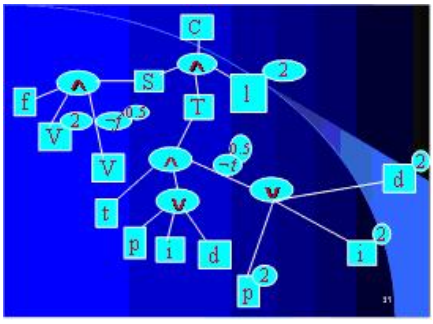


Fig. 4 Inclusion of the Leading-Edge Technology Position ($T(x)$)

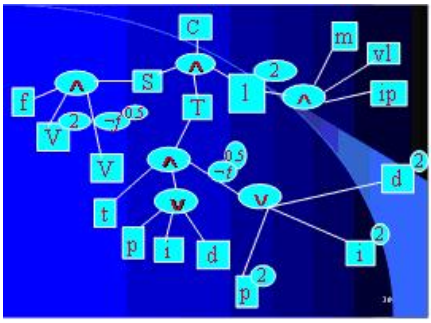


Fig. 5 Inclusion of the Strength in Product Line ($l(x)$)

Conditional expressions in the tree are expressed placing the predicate corresponding to premise over the arch, and the thesis as an end of the same arch.

Note that:

1. The preference model constructed by means of language can be very rich and varied.
2. In this case, the use of conditionals allowed increasing the demands of an attribute departing from the state of another one. This could describe, though it is not a case herein, preference-dependent situations.

The evaluation is performed as described by the tree using as attributes truthfulnesses associated with basic predicates f, V, t, p, i, d, ip, vl and m. In the example illustrated in Table 1, truthfulnesses were obtained either directly by evaluation of duly informed experts, or by using belonging functions over numerical data in the predicates where this was possible, as in the case of predicates V, p, i, d, and vl, obtained respectively from sales data, number of patents, number of products under Research & Development, amount invested in R&D, and number of products that make up the line.

Sigmoid membership functions were used, which in the case of increasing or decreasing functions are recommended in the literature because of theoretical considerations (cf. [15]). This is illustrated by figures 6 and 7 for the case of predicates V and i, dependent upon the sales data and the number of products under R&D. Some parameters of these functions are determined by setting pre-images of two values, as shown in the figures mentioned above. It defines expressions corresponding to predicates V and i from the meaning of said expressions using the data. The pre-image of 0.5 is established so that, based on this data, the statement contained in the predicate is considered acceptably true. The pre-image of 0.1 establishes a value for which data makes almost unacceptable the corresponding statement.

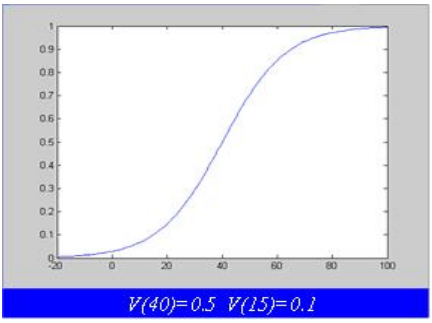


Fig. 6 Membership function for the case of predicate V

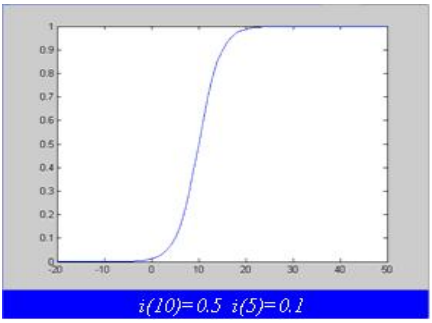


Fig. 7 Membership function for the case of predicate i

Table 1 gathers the evaluations of 4 companies on a tissue adhesive market of a Latin-American country. The implication used in calculus was i_2, for which the authors deem to have a better behavior (cf. [18]).

Table 1 Results of the example

Company x	$f(x)$	$v(x)$	$t(x)$	$p(x)$	$i(x)$	$d(x)$	$ip(x)$	$vl(x)$	$m(x)$	$S(x)$	$T(x)$	$l(x)$	$l^2(x)$	$C(x)$
A	0.5	0.47	0.3	0.93	0.81	0.61	0.6	0.23	0.1	0.5	0.516	0.234	0.058	0.246
B	0.6	0.63	0.5	0.41	1	0.95	0.8	0.77	0.4	0.611	0.682	0.627	0.393	0.545
C	0.9	0.75	0.7	0.62	0.55	0	1	0.92	0.8	0.812	0.584	0.903	0.815	0.728
D	0	0.99	0.8	0.81	0.79	0.7	0.5	0.39	1	0	0.763	0.58	0.336	0

From the predicate truth value $C(x)$ (Table 1) it is inferred that i) it is quite certain that the company C is competitive on the studied market; *ii*) it is acceptably true that B is also competitive; *iii*) it is false that A and D are competitive.

Note also that in Table 1:

1. All values of the predicate C are between the minimum and maximum truth values of the predicates S, T and l^2 in keeping with the compensation axiom fulfillment.

2. The value of the predicate C for the company D is zero due to the value of the predicate S, which itself is obtained from the basic predicate f, in both cases owing to veto axiom. In this case, no compensation is produced despite that, for instance, T has a high truthfulness value.
3. However, although the predicate d in the company C is 0, the final result is not such, because d is part of a disjunction within the predicate T. Precisely because the contribution of d is disjunctive, despite the fact that d has the value of 0, it is possible for the company C to be better evaluated through the predicate constructed to express the preferences.
4. The predicate l^2 expresses an increase in the demand in relation to l due to the modifier introduced. This exponent, as the entire tree structure, determines the contribution of each basic predicate or attribute to the evaluation of Competitiveness.
5. Truth values of the predicate over company competitiveness are different enough as to suggest the competitiveness ranking C-B-A-D. Note that the application of strict order definition of GMBCL results in o(0.246;0)=0.6230, o(0.545;0.246)=0.6495 and o(0.728;0.545)=0.5918. If we apply implication of the weak order used, it would result in o(0.246;0)=1, o(0.545; 0.246)=0.6057 and o(0.728;0.545)=0.5510.

6 Conclusions

This work presents a new approach for multivalent systems, called Compensatory Fuzzy Logic that besides contributing a formal system with logic properties of remarkable interest, represents a bridge between Logic and Decision-Making in the framework of linguistic preference information. The *CFL* enters to make part of the arsenal of methods for multicriteria evaluation, adapting itself especially to those situations in which *DM* can verbally describe, often in an ambiguous way, the heuristic he/she uses when executing actions of multicriteria evaluation/classification. However, the consistency of the logic platform provides this proposal with a capacity for formalization of reasoning that goes beyond descriptive approaches of decision-making process. It is an opportunity to use the language as key element of communication in the construction of semantic models that make easier the evaluation, decision-making and knowledge discovery. The *CFL* can be an important step to bring the scientific community closer to the objective to create a calculus with words, proposed by Zadeh and Kacprzyc in [59].

This approach makes emphasis in using language for aggregation through the combination of different logical operators and not only one operator, like other methods. However, in practical applications, this method could be combined with some existing results of obtaining and modeling the linguistic preference information about each attribute, as in [31, 33].

The study of fuzzy logic operators lead to a complete class of Compensatory Logics called Quasi-Arithmetic Means Based Compensatory Logic (QAMBCL) using Quasi-Arithmetic Means as conjunctions and their duals as disjunctions. The study of the properties of this class is relevant from the point of view of operators' selection. Their relation with bivalent logic, its use as inference logic

systems according mathematical logic, and properties of compatibility with the order according measurement theory, are desirable too [35,43].

Geometric Mean Based Compensatory Logic (*GMBCL*) is a particular case of QAMBCL. The illustrative case using *GMBCL* illustrates properties of CFL and its usefulness of modeling through language.

Acknowledgments. We acknowledge support from CYTED project 507RT0325 and CONACYT project no. 57255.

References

1. Baczyński, M., Jayaram, B.: Q-L implications, some properties and intersections. Fuzzy Sets and Systems 161(2), 158–188 (2010)
2. Baczyński, M., Jayaram, B.: (S,N)- and R-implications: A state of the art survey. Fuzzy Sets and Systems 159(14), 1836–1859 (2008)
3. Bellman, R., Giertz, M.: On the Analytic Formalism of the Theory of Fuzzy Sets. Information Sciences 5, 149–156 (1973)
4. Bodenhorf, B., Demircy, M.: Strict fuzzy orderings with a given context of similarity. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 16(2), 147–178 (2008)
5. Boyssou, D., Marchant, T., Pirlot, M., Tsoukias, A., Vincke, P.: Evaluation and Decision Models with Multiple Criteria. Springer, Berlin (2006)
6. Chiclana, F., Herrera-Viedma, E., Alonso, S., Herrera, F.: Cardinal Consistency of Reciprocal Preferce Relations: A Characterization of Multiplicative Transitivity. IEEE Transactions on Fuzzy Systems 17(1), 14–23 (2009)
7. Chiclana, F., Herrera, F., Herrera-Viedma, E.: A Study on the Rationality of Induced Ordered Weighted Operators Based on the Reliability of the Information Sources for Aggregation for Group Decision-Making. Kybernetika 40(1), 121–142 (2004)
8. Chiclana, F., Herrera, F., Herrera-Viedma, E., Martinez, L.: A note on the reciprocity in the aggregation of fuzzy preference relations using OWA operators. Fuzzy Sets and System 137, 71–83 (2003)
9. Dasgupta, M., Deb, R.: Transitivity and Fuzzy Preferences. Social Choice and Welfare 13, 305–318 (1996)
10. Dasgupta, M., Deb, R.: Factoring Fuzzy Transitivity. Fuzzy Set and System 118, 489–502 (2001)
11. De Baets, B., Fodor, J., Perny, P.: Preferences and Decisions under Incomplete Knowledge. Physica-Verlag, Berlin (2000)
12. Delgado, M., Herrera, F., Herrera-Viedma, E., Martínez, L.: Combining numerical and linguistic information in group decision making. Information Sciences 107(1-4), 177–194 (1998)
    Detiniecky, M.: Mathematical Aggregations operators and its application to Video Querying. Berkeley University (2000), `http://www.lip6.fr/reports/index-eng.html`
13. Dubois, D., Prade, H.: Fuzzy Sets and Systems: Theory and Applications. Academic Press Inc. (1980)
14. Dubois, D., Prade, H.: A review of fuzzy set aggregation connectives. Information Sciences 36, 85–121 (1985)
15. Dubois, D., Esteva, F., Godo, L., Prade, H.: Fuzzy-Set Based Logics. A history oriented presentation of their main development. In: Gabbay, D., Woods, J. (eds.) The Many Valued and Non-monotonic Turn Logic, Handbook of the History of Logic, vol. 8, pp. 325–449. Elsevier (2007)

16. Dyson, R.G.: Maximin Programming, Fuzzy Linear Programming and Multi-Criteria Decision Making. The Journal of the Operational Research Society 31, 263–267 (1980)
17. Espin, R., Fernández, E., Mazcorro, G., Marx-Gómez, J., Lecich, M.I.: Compensatory Logic: A fuzzy normative model for decision making. Investigation Operativa 27(2), 188–197 (2006)
18. Espin, R., Fernandez, E., Mazcorro, G.: A fuzzy approach to cooperative n-person games. European Journal of Operational Research 176, 1735–1751 (2007)
19. Fodor, J., Roubens, M.: Fuzzy Preference Modeling and Multicriteria Decision Support. Kluwer, Dordrecht (1994)
20. French, S.: Decision Theory: An Introduction to the Mathematics of Rationality. Halsted Press, NY (1986)
21. García-Lapresta, J.L., Meneses-Poncio, L.C.: An empirical analysis of fuzzy transitivities in Decision Making. In: VIII SIGEF Congress Proceedings, pp. 129–133 (2001)
22. García-Lapresta, J.L., Marques, R.A.: Constructing reciprocal and stable aggregation operators. In: Proceedings AGOP, Alcala de Henares, pp. 73–78 (2003)
23. Gavvav, D., Metcalfe, G.: Fuzzy logics based on (0, 1)-continuous uninorms. Springer (2006)
24. Greco, S., Matarazzo, B., Slowinski, R.: Rough sets theory for multicriteria decision analysis. European Journal of Operational Research 129, 1–47 (2001)
25. Greco, S., Matarazzo, B., Slowinski, R.: Rough sets methodology for sorting problems in presence of multiple attributes and criteria. European Journal of Operational Research 138, 247–259 (2002)
26. Hajek, P.: Methamathematics of Fuzzy Logic. Kluwer Academic Publishers, Dordrecht (1998)
27. Hajek, P.: What is Mathematical Fuzzy Logic? Fuzzy Sets and Systems 157, 597–603 (2006)
28. Hansson, B.: An analysis of some deontic logics. Nous 3, 373–398 (1969)
29. Hansson, S.O.: Preference-based deontic logic. Journal of Philosophical Logic 19, 75–93 (1990)
30. Herrera, F., Herrera-Viedma, E., Martinez, L.: A Fuzzy Linguistic Methodology To Deal With Unbalanced Linguistic Term Sets. IEEE Transactions on Fuzzy Systems 16(2), 354–370 (2008)
31. Herrera, F., Herrera-Viedma, E., Chiclana, F.: A Study of the Origin and Uses of the Ordered Weighted Geometric Operator in Multicriteria Decision Making. International Journal of Intelligent Systems 18(6), 689–707 (2003)
32. Herrera, F., Herrera-Viedma, E.: Linguistic Decision Analysis: Steps for Solving Decision Problems under Linguistic Information. Fuzzy Sets and Systems 115, 67–82 (2000)
33. Herrera, F., Herrera-Viedma, E., Verdegay, J.L.: A sequential selection process in Group decision making linguistic assessment approach. Information Sciences 85(4), 223–239 (1995)
34. Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A.: Foundations of measurement. Additive and polynomial representations, vol. 1. Academic Press, New York (1971)
35. Lang, J., Van der Torre, L.: From belief change to preference change. In: Proceedings of the 18th European Conference on Artificial Intelligence ECAI (2008)
36. Lindley, D.: In: Wright G., and P. Ayton (eds.) Subjective Probability, pp. 1–37. Wiley & Sons, England (1994)

37. Marakas, G.: Decision Support Systems in the 21st Century. Prentice Hall (2002)
38. Novak, V.: Fuzzy Sets and their applications. Adam Hilger, Bristol (1989)
39. Novak, V., Perfilieva, I.: Evaluating linguistic expressions and functional fuzzy theories in Fuzzy Logic. In: Zadeh, L.A., Kapricz, J. (eds.) Computing with words in Information/ Intelligent Systems 1. (Foundations), pp. 383–606. Physica-Verlag (1999)
40. Pomerol, C.: Artificial Intelligence and Human Decision Making. In: Slowinski, R. (ed.) OR: Toward Intelligent Decision Support, 14th European Conference on Operational Research, pp. 169–196 (1995)
41. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Los Altos (1993)
42. Roberts, F.S.: Measurement theory, with applications to Decision Making, Utility and the Social Sciences. Addison-Wesley, Boston (1979)
43. Roy, B.: The outranking approach and the foundations of ELECTRE methods. In: Bana e Costa, C.A. (ed.) Reading in Multiple Criteria Decision Aid, pp. 155–183. Springer, Berlin (1990)
44. Roy, B., Vanderpooten, D.: The European School of MCDA: A Historical Review. In: Slowinski, R. (ed.) OR: Toward Intelligent Decision Support, 14th European Conference on Operational Research, pp. 39–65 (1995)
45. Chen, S.-J., Chen, S.-M.: Fuzzy information retrieval based on geometric-mean averaging operators. Computers & Mathematics with Applications 49(7-8), 1213–1231 (2005)
46. Silvert, W.: Ecological impact classification with fuzzy sets. Ecological Modeling 96, 1–10 (1997)
47. Slowinski, R.: Rough Set approach to Decision Analysis. AI Expert, 19–25 (March 1995)
48. Slowinski, R., Greco, S.: Axiomatic Basis of Aggregation Functions: Utility Function, Associative Operator, Sugeno Integral, Max-Min Weighted Sum, Decision Rules. In: Invited Lecture in XVI MCDM World Conference, Semmering, Austria (2002)
49. Switalski, Z.: Transitivity of Fuzzy Preferences relations: An empirical study. Fuzzy Sets and Systems 118, 503–508 (2001)
50. Switalski, Z.: General Transitivity conditions for fuzzy reciprocal preference matrices. Fuzzy Sets and Systems 137(1), 85–100 (2003)
51. Tsoukias, A., Vincke, P.: Extended preference structures in MCDA. In: Climaco, J. (ed.) Multicriteria Analysis, pp. 37–50. Springer, Berlin (1997)
52. Tsoukias, A., Vincke, P.: A new axiomatic foundation of partial comparability. Theory and Decision 39, 79–114 (1995)
53. Tsoukias, A., Vincke, P.: A survey of non conventional preference modelling. Ricerca Operativa 61, 5–49 (1992)
54. Velazquez-Quesada, F.R.: Inference and update. Synthese 169, 283–300 (2009)
55. Verdegay, J.L.: A revision of Soft Computing Methodologies. In: Verdegay, J.L. (ed.) Acts from the Simposium on Fuzzy Logic and Soft Computing LFSC (EUSFLAT), pp. 151–156 (2005) (in Spanish)
56. Wang, Y.-M., Chin, K.-S., Poon, G.K.K., Yang, J.-B.: Risk evaluation in failure mode and effects analysis using fuzzy weighted geometric mean. Expert Systems with Applications 36, 1195–1207 (2009)
57. Zadeh, L.A.: Fuzzy Sets. Inf. Control 8, 338–353 (1965)
58. Zadeh, L.A., Kacpryzyk, J.: Computing with words in Information/Intelligent Systems 2 (Applications). STUDFUZZ, vol. 34. Phisyca-Verlag (1999)

Compensatory Fuzzy Logic Inference

Rafael Alejandro Espín Andrade, Erick González, Eduardo Fernández,
and Marlies Martinez Alonso

Abstract. This chapter deals with specific way of inferences possible in the framework of *Compensatory Fuzzy Logic* (CFL). It introduces new and useful inference systems based on CFL. They will be called *Compensatory Inference Systems* (CIS). It is a generalization of the deduction like in mathematical logic, with implication operators found in the literature regarding fuzzy logic. CIS is a combination of one Compensatory Logic and an implication operator. Every CIS is a logically rigorous inference system with easy application. In addition CIS may be coherently associated with the methods of deduction of the mathematical logic. The theoretical basis of this association is proved in this chapter. Valid formulas and right deductive structures according CFL are the new concepts introduced here. The formulas of the propositional calculus are valid in the bivalent logic if and only if they are valid according CFL. The same result is introduced for deductive right structures. The relevance of these results for approximate reasoning and knowledge discovery are illustrated. Further more, probabilistic properties expressed in a theorem allow applying statistical inference in the framework of CFL. This theorem expresses that the universal proposition over a sample can be a statistic estimator of the corresponding universal proposition over the entire universe. Thus CFL joins logical and statistical inferences and gives logical models of automated learning with properties of a statistical estimator.

1 Introduction

There are many approaches to the inference in fuzzy logic. They can be found in both Fuzzy Logic in the Narrow Sense and Fuzzy Logic in the Broad Sense.

Rafael A. Espín Andrade · Erick González · Marlies Martinez Alonso
"Jose Antonio Echeverria" Higher Technical Institute, Cuba
e-mail: rafaelespin@yahoo.com, erickgc@cemat.cujae.edu.cu,
marlies.martinez@gmail.com

Eduardo Fernández
"Autonomous University of Sinaloa", Sinaloa, Mexico
e-mail: eddyf171051@gmail.com

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_2,

The *Fuzzy logic in the narrow/technical sense* refers to syntax, semantic, axiomatization, completeness and other formalizations, proper for every many-valued logic. *Fuzzy logic in the broad* or *wide sense* is a singularity of this many-valued logic and refers to concepts like linguistic variable, fuzzy if-then rule, fuzzy quantification and defuzzification, truth qualification, the extension principle, the compositional rule of inference and interpolative reasoning, etc. [3].

The *fuzzy if-then rules* are a very appreciated tool of fuzzy logic; it is the basis of the *fuzzy reasoning* and the *fuzzy inference systems*. *Fuzzy if-then rules* are conditional statements of the form "if x_1 is A_1 and …and x_n is A_n, then y is B", where A_i and B are fuzzy sets. The inference is realized by using the *Generalized Modus Ponens*, which is basically the classical modus ponens, but consisting in particular of: "x is A^*" and "if x is A then y is B" we infer B^*. The fuzzy set B^* has membership function described in detail from Jang et al. [6]. For example, from the statement "the tomato is more or less red" and the rule "if the tomato is red, then it is ripe", it can be inferred that according to the generalized modus ponens "the tomato is more or less ripe". The Generalized Modus Ponens is also known as *Fuzzy Reasoning* or *Approximate Reasoning*.

The *Fuzzy Inference System*s are successful frameworks for inference calculation. The *Mamdani Fuzzy Models*, the *Sugeno Fuzzy Models* and the *Tsukamoto Fuzzy Models* are three of the most recurrent models in the literature. The basic scheme of the *Fuzzy Inference System*s is described as: from a set of crisp or fuzzy input data, other fuzzy sets are obtained as a consequence of every fuzzy if-then rule. The next step is the use of an aggregator, which is a unique function, representing the results of the conjoint of rules. This function needs to be defuzzified, in order to convert the fuzzy results into a single crisp output value.

The success in the application of the above tools is guaranteed, but they need to use some extra-logical methods, including the defuzzification. Besides, the functions defined as aggregators in the models, aren't well justified from the point of view of the classical logics.

The inference upon the Fuzzy Logic in the *narrow sense* provides some interesting results that are more related to the mathematical logic. The concept of *lattice*, which is defined in algebra and set theory, is the point of departure. A *lattice* is a particular case of *partially ordered set*. It includes a binary relation of order over every pair of elements of the set. The logic systems found in the literature are based on t-norm and t-conorm.

The axiomatic of the *Propositional Fuzzy Logic,* also known as *Basic Fuzzy Logic* or *Basic Fuzzy Propositional Logic*, was introduced by Hájek in 1998. It is a Hilbert-style deduction system with the classical modus ponens as the unique rule of inference [3]. A natural extension of this axiomatic is obtained when including some axioms, where the universal and existential quantifiers appear; it is the *Basic Fuzzy Predicate Logic*. In Hilbert-style deduction systems the axiomatic is formed by formulas with the implication operator. The Basic fuzzy logic takes the operators of implication and conjunction, and the constant 0 (false) for defining the negation operator and the disjunction operator.

The *Fuzzy Logic with Evaluated Syntax* includes the notion of lattice. Every antecedent formula is directly evaluated with its truth value, and the result of the deduction obtained with the application of the modus ponens is a formula computed by the order of the lattice with a truth value as well. The axioms, which are not necessarily fully true are equally accepted; see Nóvak and Dvorák [10]. The deduction in the Fuzzy logic with evaluated syntax uses the theory of proof of classical logic; hence, it is a natural generalization of the mathematical logic. Nonetheless, [11] agreed that these calculi in the narrow sense haven't enough variety of applications. Other approaches and tendencies according the inference can be found in [3, 10, 11].

The apparent contradiction between the use of mathematically rigorous calculi not widely applied and the use of pragmatic tools for inference with many successful applications, is perhaps due to the use of the t-norm and t-conorm paradigm as the principal rational option to define fuzzy logic systems. However, the compensatory operators seem to be more adequate to model the human thinking, according to some experiments [9]. In addition, some pragmatic calculations like defuzzification are compensatory because they are essentially a mean.

The unique paradigm of fuzzy logic system found in the literature, based on compensatory operators, and not exclusively on single isolated operators [2], is the *Compensatory Fuzzy Logic* (CFL) [4]. A CFL system is a quartet of: a conjunction operator, a disjunction operator, a negation operator and a strict fuzzy order operator. They must satisfy an axiomatic which belongs to the logic and the decision theory.

CFL is a formal development of a logic system, from the *Narrow Fuzzy Logic* point of view, but with properties allowing its application in solving all the problems like in Fuzzy Logic in broad sense. CFL axiomatic can be considered as an extension of the mathematical logic, furthermore it can be successfully associated with the methodology for *Experts Systems*, called *Knowledge Engineering* [1] and the notion of *Soft Computing* [13].

The CFL makes possible the idea of computation by words, as proposed by Zadeh [14]; it upgrades the use of solely simple linguistic variables, through implementation of complex phrases expressed in natural language. Thus, CFL allows to model problems expressed in natural language, using sentences provided by experts in the theme, following the methodology of the Expert Systems.

This chapter aims to introduce new and useful inference systems based on CFL. They will be called *Compensatory Inference Systems* (CIS). They are generalizations of the deduction, like in mathematical logic, with one implication operator found in the literature.

The chapter is organized as follows: in the second epigraph – Preliminaries – some basic concepts and operators of CFL are explained, and a one-parameter family system of logic systems is introduced. Likewise, proposed operators for CIS, are briefly described. Epigraph 3 defines the statistical properties of the CFL. The relation between CFL and Bivalent logic is treated in epigraph 4. Compensatory Inference Systems - as fifth epigraph - exposes the criteria and the results according the CIS. Epigraph 6 portrays some experiments with the CIS. Possibilities of CIS for Approximate Reasoning and Knowledge Discovery are discussed in the conclusion of this chapter.

2 Preliminaries

A Compensatory Logic (CL) is a quartet of operators (*c*, *d*, *o*, *n*), where *c* is conjunction, *d* is disjunction, *o* is fuzzy strict order and *n* is negation [4].

c and *d* map vectors of $[0,1]^n$ into [0,1], *o* is a mapping from $[0,1]^2$ into [0,1], and *n* is a unary operator of [0,1] into [0,1]. Some axiomatic must to be satisfied for the operators of conjunction and disjunction, like e.g. Compensation Axiom¸ Symmetry Axiom, Veto axiom and others [4].

A family of CL may be obtained from the quasi-arithmetic means, with the following formula [8]:

$$M_f(x_1, x_2, \dots, x_n) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \tag{1}$$

Where $f(x)$ is a continuous and strictly monotonic function of one real variable for $x \in (0,1]$.

Then CFL has probabilistic properties, which join logic and statistical inferences. It facilitates models for automated learning, which have properties of statistical estimator.

A particular one-parameter family can be introduced using the formula below:

$$M_f(x_1, x_2, \dots, x_n) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{1/p} \tag{2}$$

Where $p \in (-\infty, 0]$.

It satisfies the axiom of compensation, if the conjunction is defined as in (2). More details about the CL and formulas of family (2) can be found in [4].

Parameter *p* measures the degree of "compensation" in the correspondent logic. Since formula (2) is the minimum operator, when $p = -\infty$[2,8], $p = 0$ is the most "compensatory" of all the compensatory systems in the one-parameter family exposed in (2).

The universal proposition over the set *x*, in CFL, which is the generalization of formula (1), is defined as following:

$$\underset{x \in X}{\forall} x\, p(x) = \begin{cases} f^{-1}\left(\dfrac{\int_X f(p(x))dx}{\int_X dx}\right) & if\ p(x) > 0 \text{ for entire } x \in X \\ 0 & \text{In any other case} \end{cases} \tag{3}$$

Implication operators are mostly defined like conjunction, disjunction and negation of t-norms and t-conorms. In this chapter these concepts will be extended to compensatory logics. Other implication operators used here are associated with specific t-norm and t-conorms, but they are used in combination with compensatory logics to get a CIS.

The criteria for selecting implication operators for our purposes are the following:

1. The operator satisfies the truth value table of the bivalent classical logic.
The operator must be a continuous function with regard to both arguments or have a finite number of removable discontinuities.

Some implication operators suggested in the literature, which satisfy the two conditions expressed above are:

- Reichenbach implication: $x \to y = 1 - x + xy$
- Klir-Yuan implication: $x \to y = 1 - x + x^2 y$
- Natural implication, see [4]: $x \to y = d(n(x), y)$
- Generalized Zadeh implication: $x \to y = d(n(x), c(x, y))$
- Yager implication: $x \to y = y^x$

Other classifications can be found in [7].

3 Statistical Inference from CFL

The following theorem can be obtained directly from the Central Limit Theorem:

Theorem 1: If $p(x) > 0$ for any $\boldsymbol{x}$ in a universe *X*, then the universal proposition $\forall_{x \in M}\, p(x) = f^{-1}(\frac{1}{n} \sum_{x \in M} f(p(x)))$ of the predicate $p(x)$, is a statistical estimator of the truth value of $\underset{x \in X}{\forall} p(x)$, the universal proposition over *X*.

Proof

Since $\frac{1}{n}\sum_{x \in M} f(p(x))$ is distributed as $N\left(u, \frac{\sigma^2}{n}\right)$, where σ^2 is the variance of $f(p(x))$ and *u* is the mean of $f(p(x))$, then $\frac{1}{n} \sum_{x \in M} f(p(x))$ is an estimator of u because of the Central Limit theorem. Accordingly $\forall_{x \in M}\, p(x) = f^{-1}(\frac{1}{n} \sum_{x \in M} f(p(x)))$ is an estimator of $\underset{x \in X}{\forall} p(x)$. Consequently, the truth value of the universal proposition over a probabilistic sample *M* is an estimator of the truth value of the universal proposition over the universe *X*.

4 The CFL and the Bivalent Logic

Definition 1

Let p(**x**) be a formula of the propositional calculus in CFL.

A formula is *valid according CFL* if whatever could be Δ, $0<\Delta<1$, exist a set S of neighborhoods of elements belonging to $\{0,1\}^n$ such that:

$$\underset{x \in S}{\forall} x\, p(x) > \Delta \tag{4}$$

can be obtained.

Theorem 2: Let φ be formula of the propositional calculus, then φ is valid in the bivalent logic, if and only if, it is valid according CFL.

Proof

If φ is valid in the CFL, according to condition (4), every truth value of φ can not be 0 for any x∈S, including all $x \in \{0,1\}^n$. Hence, φ is valid in the bivalent logic, because its possible truth values for $x \in \{0,1\}^n$ are either 0 or 1, and 0 was excluded as we proved.

If φ is a tautology, then due to the continuity (or at least the finite number of removable discontinuities) of the operators used to define it and the mean value theorem, exists a set S of neighborhoods of elements belonging to $\{0,1\}^n$, such that $\forall_{x \in S}\, x\, p(x) > \Delta$, then φ is valid according CFL.

Remark 1

Let us note that if we select a tautology, for every Δ, independently how great value for it will be selected, 'sufficiently small sized' neighborhoods V_i exist, such that the truth value of a tautology is greater than Δ. Therefore, it is always possible to set beforehand a threshold Δ, with the property that the tautology is greater than Δ, for some sufficiently small sized set *S*, defined like in definition 1.

5 Compensatory Inference Systems

A CIS is the combination of a Compensatory Logic and an implication operator.

The axiomatic of Kleene is a well-known Hilbert-style deduction system, therefore the implication is essential part of its logical axioms, which represent tautologies of the bivalent logic. They are defined as shown below [3].

$AX1: A \to (B \to A)$
$AX2: (A \to B) \to ((\mathrm{A} \to (B \to C)) \to (A \to C))$
$AX3: A \to (B \to A \wedge B)$
$AX4: A \wedge B \to \mathrm{A} \ldots \mathrm{A} \wedge \mathrm{B} \to \mathrm{B}$
$AX5: A \to A \vee B \ldots B \to A \vee B$
$AX6: (A \to C) \to ((B \to C) \to (A \vee B \to C))$
$AX7: (A \to B) \to ((A \to \neg B) \to \neg A)$
$AX8: \neg(\neg A) \to A$

Kleene's axiomatic has been used as the basis for constructing bivalent logic in the framework of the well-known *proof theory*. This construction is based on the principle of *deductive right structures*. Such principle is called *Deduction Theorem* [5], and is described in definition 2.

Definition 2

$\alpha_1, \alpha_2, \ldots, \alpha_{n-1} \vdash \varphi$ is a *deductive right structure,* if the formula $\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_{n-1} \to \varphi$ is valid in the propositional calculus of the bivalent logic. $\alpha_1, \alpha_2, \ldots, \alpha_{n-1}$ are the *hypotheses* and ⊢ is the symbol of inference.

These classical concepts of deductive systems and proofs can be naturally extended to fuzzy logic, using the CFL approach.

Some right deductive structures are the following:

a) $\neg Q, P \to Q \vdash \neg P$ (Modus Tollens)
b) $\neg(P \wedge Q) \vdash \neg P \vee \neg Q$ (D' Morgan)
c) $P \vdash \neg(\neg P)$ (Double negation)
d) $P \vee Q, \neg P \vdash Q$ (Disjunctive syllogism)
e) $P \vdash P \vee Q$ (Disjunction)
f) $P, Q \vdash P \wedge Q$ (Conjunction)

Definition 3

$\alpha_1, \alpha_2, \ldots, \alpha_{n-1} \vdash \varphi$ is a *right deductive structure* according CFL, if the formula $\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_{n-1} \to \varphi$ is *valid* according CFL.

The truth value of the formula (5) below is considered the truth value of the demonstration of φ.

$$\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_{n-1} \to \varphi \tag{5}$$

According definition 3, we can obtain the following immediate result:

Corollary 1:

$\alpha_1, \alpha_2, \ldots, \alpha_{n-1} \vdash \varphi$ is a right structure according bivalent logic if and only if $\alpha_1, \alpha_2, \ldots, \alpha_{n-1} \vdash \varphi$ is a *right deductive* structure according CFL, independently of the CIS selected.

For example, all the right deductive structures used as examples of definition 2, are right deductive structures according CFL because of *Corollary 1*.

This means that the following formulas are valid according CFL:

a) $\neg Q, P \to Q \to \neg P$ (Modus Tollens)
b) $\neg(P \wedge Q) \to \neg P \vee \neg Q$ (D' Morgan)
c) $P \to \neg(\neg P)$ (Double negation)
d) $P \vee Q, \neg P \to Q$ (Disjunctive syllogism)
e) $P \to P \vee Q$ (Disjunction)
f) $P, Q \to P \wedge Q$ (Conjunction)

6 Experiment with CIS

All the valid formulas of Bivalent Logic can be deducted using Kleene's Axioms as premises in definition 2.

Then according Corollary 1, $(\wedge_{i=1}^{8} Ax_i) \to \varphi$ is *valid* for all φ.

If we fix the set *S* of neighborhoods, the greater the truth values of the Kleene's axioms are, according the CIS, the greater the number Δ of the valid condition according CFL is for each φ. Therefore the Kleene axioms' conjunction truth value over all set of interpretations is an indicator of how great the truth value of valid formulas of bivalent logic are according CFL (4).

As next tables with Kleene axioms values for different CIS will be presented.

The definite integrals, simple, double and triple, depending on which axiom is applied, are calculated using MATLAB. The values of the formula in expression

(4), for some fixed exponent p (see (2)) and every implication operator are calculated. The results are presented in seven tables summarizing the calculus of the truth values of the Kleene's axioms obtained for seven fixed p, where p is the parameter used in (2).

Table 1 bases its calculus on $p = 0$, where (2) results in Geometric Mean. Each column represents one of the implications' operators: Natural, Generalized Zadeh, Yager, Reichenbach and Klir-Yuan. Each row represents one of the eight Kleene's axioms. The last two rows calculate the values of the Universal Quantifier and the Minimum by implication operator, respectively.

Table 1 Results of the evaluation of each Klenee's axiom for the formula in condition (4), where $p = 0$ (Geometric mean)

	Natural Geometric	Generalized Zadeh	Yager	Reichenbach	Klir-Yuan
Ax1	0.5859	0.5685	0.8825	0.9143	0.7433
Ax2	0.5122	0.5073	0.8425	0.8709	0.6745
Ax3	0.5556	0.5669	0.8825	0.9088	0.7416
Ax4	0.5859	0.5661	0.7436	0.8160	0.7148
Ax5	0.5859	0.5859	0.7774	0.8160	0.7217
Ax6	0.5026	0.5038	0.8772	0.8911	0.6617
Ax7	0.5315	0.5137	0.7574	0.7882	0.6690
Ax8	0.5981	0.5981	0.7788	0.8301	0.7413
Universal Quantifier	0.5561	0.5502	0.8158	0.8532	0.7077
Minimum	0.50258	0.5038	0.74357	0.78820	0.66172

Table 2 Results of the evaluation of each Klenee's axiom for the formula in condition (4), where $p = -1$ (Harmonic mean)

	Natural Harmonic	Generalized Zadeh	Yager	Reichenbach	Klir-Yuan
Ax1	0.6369	0.6111	0.8784	0.9119	0.7369
Ax2	0.5611	0.5449	0.8344	0.8668	0.6698
Ax3	0.5996	0.6096	0.8686	0.9018	0.7335
Ax4	0.6369	0.6161	0.7649	0.8295	0.7285
Ax5	0.6369	0.6298	0.7905	0.8295	0.7265
Ax6	0.5440	0.5331	0.8671	0.8823	0.6585
Ax7	0.5753	0.5560	0.7411	0.7849	0.6640
Ax8	0.6366	0.6366	0.7744	0.8270	0.7330
Universal Quantifier	0.60116	0.58962	0.81181	0.85225	0.70475
Minimum	0.544	0.5331	0.7411	0.78493	0.6585

Table 3 Results of the evaluation of each Klenee's axiom for the formula in condition (4), where $p = -2$

	Natural	Generalized Zadeh	Yager	Reichenbach	Klir-Yuan
Ax1	0.6527	0.6254	0.8743	0.9094	0.7308
Ax2	0.5879	0.5677	0.826	0.8626	0.6656
Ax3	0.6173	0.6242	0.858	0.8961	0.726
Ax4	0.6527	0.6332	0.7715	0.8343	0.7311
Ax5	0.6527	0.6413	0.7939	0.8343	0.7251
Ax6	0.57	0.5565	0.86	0.8763	0.6554
Ax7	0.5947	0.5782	0.7192	0.7819	0.6595
Ax8	0.6446	0.6446	0.7703	0.824	0.7252
Universal Quantifier	0.6191	0.6061	0.8042	0.8496	0.69996
Minimum	0.57	0.5565	0.7192	0.7819	0.6554

Table 4 Results of the evaluation of each Klenee's axiom for the formula in condition (4), where $p = -3$

	Natural	Generalized Zadeh	Yager	Reichenbach	Klir-Yuan
Ax1	0.6552	0.6278	0.87	0.9068	0.7251
Ax2	0.5991	0.5782	0.8174	0.8585	0.6619
Ax3	0.6223	0.6268	0.8497	0.8913	0.719
Ax4	0.6552	0.6367	0.7728	0.8355	0.7295
Ax5	0.6552	0.6416	0.7936	0.8355	0.7215
Ax6	0.5819	0.5688	0.8544	0.8717	0.6527
Ax7	0.6014	0.5872	0.7033	0.779	0.6555
Ax8	0.643	0.643	0.7665	0.8211	0.718
Universal Quantifier	0.6242	0.61101	0.79627	0.84638	0.69482
Minimum	0.5819	0.5688	0.7033	0.779	0.6527

Table 5 Results of the evaluation of each Klenee's axiom for the formula in condition (4), where $p = -4$

	Natural	Generalized Zadeh	Yager	Reichenbach	Klir-Yuan
Ax1	0.6521	0.6252	0.8656	0.9042	0.7198
Ax2	0.6026	0.5821	0.8085	0.8543	0.6586
Ax3	0.6213	0.6243	0.8428	0.8871	0.7125
Ax4	0.6521	0.6345	0.7719	0.835	0.7262
Ax5	0.6521	0.6376	0.7915	0.835	0.7173
Ax6	0.5864	0.5741	0.8495	0.8679	0.6503
Ax7	0.6021	0.5899	0.6331	0.7763	0.6521
Ax8	0.6379	0.6379	0.7629	0.8184	0.7114
Universal Quantifier	0.6233	0.6106	0.77164	0.84291	0.6899
Minimum	0.5864	0.5741	0.6331	0.7763	0.6503

Table 6 Results of the evaluation of each Klenee's axiom for the formula in condition (4), where $p = -10$

	Natural	Generalized Zadeh	Yager	Reichenbach	Klir-Yuan
Ax1	0.6142	0.5941	0.8395	0.8876	0.6947
Ax2	0.5859	0.5711	0.7525	0.8309	0.6449
Ax3	0.5935	0.5936	0.8149	0.8673	0.6833
Ax4	0.6142	0.6018	0.7565	0.8231	0.7016
Ax5	0.6142	0.6022	0.7715	0.8231	0.6929
Ax6	0.5751	0.5665	0.8272	0.8519	0.6401
Ax7	0.5817	0.5755	0.3451	0.7635	0.6378
Ax8	0.601	0.601	0.7462	0.8048	0.683
Universal Quantifier	0.5954	0.58637	0.4248	0.8225	0.66699
Minimum	0.5751	0.5665	0.3451	0.7635	0.6378

Table 7 Results of the evaluation of each Klenee's axiom for the formula in condition (4), where $p = -300$

	Natural	Generalized Zadeh	Yager	Reichenbach	Klir-Yuan
Ax1	0.5094	0.5085	0.7089	0.7677	0.6183
Ax2	0.5084	0.5077	0	0.7228	0.6025
Ax3	0.5085	0.5085	0	0.7666	0.6036
Ax4	0.5094	0.5089	0.6983	0.7572	0.6222
Ax5	0.5094	0.5089	0.6991	0.7572	0.6216
Ax6	0.5079	0.5076	0.6925	0.7502	0.6024
Ax7	0.5081	0.5079	0	0.713	0.5993
Ax8	0.5088	0.5088	0.6978	0.7561	0.6209
Universal Quantifier	0.5086	0.5083	0	0.7179	0.6026
Minimum	0.5079	0.5076	0	0.713	0.5993

The combination of a CL, defined by a parameter p according to (2), and an implication operator for a specific Kleene's axiom, is a specific formula of the propositional calculus of the CFL. Its truth value is obtained by using condition (4), which essentially consists in a multiple integral calculus for $S = [0,1]^n$. MATLAB was utilized for the computation.

The penultimate and ultimate rows of the tables summarize the conjunction and the minimum of the results. Basically, they are aggregation operators that allow ordering the implication operators by their truth values. For example, in table 1 the maximum truth values are obtained for the Reichenbach implication with $p = 0$, and they are 0.8532 and 0.78820, (see the intersection of the ninth and tenth rows with the fourth column of table 1).

The tables show that the Reichenbach implication has the best results for all of the fixed p. Axiom 7 has the worst truth values. The fourth and fifth axioms increase their truth values, if p decreases, but they tend to decrease when certain p is reached. However, in general it can be assumed that the truth values decrease, if p decreases.

Since the best results were obtained with Reichenbach, Yager and Klir-Yuan implications, an optimization problem using these implications was developed. The parameters to be estimated are p and the implication operators. Eight objective functions were maximized according the eight Kleene's axioms and evaluated according condition (4). Genetic algorithms of MATLAB were used for the optimization. The results are shown in table 8.

Table 8 Results of maximizing the validation function evaluated in every Kleene axiom, by the three selected implication operators

Implication/Axiom	Reichenbach Implication		Yager implication		Klir-Yuan Implication	
	p estimated	Truth value	*p* estimated	Truth value	*p* estimated	Truth value
Ax1	0	0.914	-1.907e-6	0.883	0	0.743
Ax2	0	0.871	-1.9073e-6	0.855	0	0.674
Ax3	-1.9073e-6	0.909	0	0.882	0	0.742
Ax4	-3.13379	0.835	-2.95215	0.773	-1.97656	0.731
Ax5	-3.13379	0.835	-2.33789	0.7941	-1.09863	0.727
Ax6	0	0.892	0	0,8772	0	0,670
Ax7	0	0.788	0	0.7574	-1.9073e-6	0.669
Ax8	-1.9073e-6	0.830	-1.9073e-6	0.7790	0	0.741

It shall be remarked that the procedures to compute the truth values in table 8 remain equal to those applied in the calculations of values in the previous tables. The difference in this case is that p is particularly estimated for each Kleene's axiom and not a fixed parameter for all of them. The genetic algorithm was used as the method of optimization.

Table 8 confirms preliminary results in the tables 1-7; Reichenbach implication generates the highest truth values for each Kleene's axiom. For all the axioms the best estimated p were $p = 0$ and $p = -1.9073e\text{-}6$, except for the fourth and fifth axioms, where p is between -1 and -4. The seventh axiom is the worst in all cases, except in the Klir-Yuan implication.

Example 1

Let us study the CIS, which consists of the Geometric Mean Based Compensatory Logic and the Reichenbach implication.

The tautology of the propositional calculus, which represents the proof of the well-known rule named *modus ponens*, $x \to ((x \to y) \to y)$, have truth value 0.8572 for $S = [0,1]^2$.

Δ and Δ' exist, $0<\Delta,\Delta'<1$, such that if the formula $(\wedge_{i=1}^{8} Ax_i) \to \varphi$ is valid, where Ax_i are the Kleene's axioms, then φ is valid. In this case, $\Delta' = 0.8572$ and $\Delta = 0.8799$.

Let us suppose the truth values of the data we are studying are greater than 0.7 or below 0.3, that means they are 0.3 sized, then $S = V_1 \cup V_2 \cup V_3 \cup V_4$, where $V_1 = [0,0.3] \times [0,0.3]$, $V_2 = [0,0.3] \times [0.7,1]$, $V_3 = [0.7,1] \times [0,0.3]$ and $V_4 = [0.7,1] \times [0.7,1]$; consequently, modus ponens have a truth value great than 0.8911, which is a greater (a better) truth value than 0.8572.

It can be exemplary concluded that if the neighborhood is $[0,1]^2$ then the truth value of the first Kleene's axiom is 0.9143, see table 1, and for S, its truth value is 0.9382.

7 Approximate Reasoning and Knowledge Discovery by CIS

Because CFL is compatible with the bivalent logic, CIS can be applied in Approximate Reasoning and Knowledge Discovery by the use of right deductive structures of bivalent mathematical logic. This will be illustrated as next.

A Knowledge Discovery method is compounded by three basic components: (1) model representation, (2) model evaluation and (3) search.

A general method of Knowledge Discovery by logical predicates has been introduced in literature [12]. It uses logic predicates like way of representation, multivalued fuzzy logic for evaluation and any meta-heuristic or optimization method as the way to search.

Searching is made applying as objective function the universal proposition truth value, over the set of instances of the used Data Base.

Searching can be made while fixing specific structure of the predicates, and changing the parameters' value with the purpose to solve 1) specific problem of Data Mining or adjusting a hypothesis to improve the truth value; 2) modifying the structure of predicates, looking into the space of all possible predicates.

The first kind of problems is better answered by optimization methods, finding local maximums. The second finds solution, using any meta-heuristic for obtaining all predicates with good truth values.

The way of knowledge discovery from hypotheses shall be illustrated on an example of Mexican economy.

Example 2:

The following hypotheses can be enunciated:

a) If the time t after the moment t_0 is short, Gross Domestic Product (GDP) is big, the value of Mexican Peso is good, and also inflation, then inflation at the moment $t_0 + t$ will be good (Enough Condition for future inflation).
b) If the time t after the moment t_0 is short, GDP is big, the value of Mexican Peso is good, and inflation too, then the value of Mexican Peso at the moment $t_0 + t$ will be big. (Enough Condition for future value of Mexican Peso).
c) If the time t after the moment t_0 is short GDP is big, the value of Mexican peso is good, and inflation too, then GDP in moment $t_0 + t$ will be good. (Enough Condition for future GDP).

Hypotheses were modeled using CFL with basic predicates as sigmoid functions.

Table 9 resumes the truth value of the three hypotheses, where 1', 2' y 3' are the necessary conditions corresponding to 1, 2 and 3. Its first half expresses the truth value for each hypothesis using the sample. The first row in the first half uses fifty cases, and the second row is based on another six additional cases. The second half expresses the results after doing a continuous search in the parameters space to achieve a better necessary condition for the future inflation. A greater proximity between the truth values of hypothesis' 1' estimations for both samples is observed after optimization, due to the variance reduction associated to a truth

value closer to 1. Then, getting truer value means getting less uncertain one from the probabilistic point of view too.

Table 9 Truth value of the universal propositions of the two samples, before and after the parameters searching

Hypothesis 1	Hypothesis 2	Hypothesis 3	Hypothesis 1'	Hypothesis 2'	Hypothesis 3'
0.439202131	0.574390855	0.604586702	0.846168613	0.281018624	0.395033281
0.519845804	0.657303193	0.754040445	0.890357379	0.330713346	0.578926238
Hypothesis 1	Hypothesis 2	Hypothesis 3	Hypothesis 1'	Hypothesis 2'	Hypothesis 3'
0.15293434	0.493060755	0.476469006	0.992585058	0.361083333	0.353165336
0.064133476	0.727586495	0.58911645	0.999644446	0.552232295	0.534562806

Table 10 represents the value of sigmoid membership functions parameters, Gamma and Beta: the initial ones and the obtained from searching. They are the inverse image of 0.5 and 0.1. Consequently, results can be interpreted as a substitution of the states in each fuzzy variable with new states, represented by those functions, or in some cases an adjustment of the parameters value for expressing better the linguistic variable state.

Table 10 Parameters, before and after the search by optimization

	Gamma	Beta	Gamma	Beta
Inflation	11	5	10.3022482	5.30220976
GDP	2	0	2.70067599	0.12747186
Money Value	7	12	6.8657769	12.0650321
Future Inflation	11	5	6.39146712	6.35080391
Future GDP	2	0	2	0
Future Money Value	7	12	7	12
Time	2	4	1.19916547	4.14284971

Corollary 1 is fundamental for the use of deductive reasoning as search way for new 'good predicates', using the knowledge base of predicates, discovered previously.

According to *Corollary 1*, using CIS gives the possibility to use right deductive structures of Classical Bivalent Logic from discovered good predicates, in order to obtain new ones and selecting from the 'deduced predicates' new ones with higher truth value.

To summarize, if the premises $\alpha_1, \alpha_2, \ldots, \alpha_{n-1}$ are approximately true according the Data Base, and the deductive structure $\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_{n-1} \vdash \varphi$ is approximately right, over a set S of interpretations containing approximately the truth values of $\alpha_1, \alpha_2, \ldots, \alpha_{n-1}$, then, the predicate φ will be approximately true.

Such approximate reasoning utilizing bivalent logic enables heuristic obtaining of predicates, which can have high truth value, according the Data Knowledge from the previously obtained predicates.

Example 3:

A Data Base B has the following seven variables. They describe patients and are presented in detail in one chapter, which will appear in the second part of this book:

a) Age
b) Race
c) Hypertension
d) Body Mass Index (BMI)
e) Cardiovascular and/or Cerebral Vascular Accident (CVA) antecedents (both known for the expression: "Antecedents")
f) Sex
g) Classification of diabetes (Diabetes)

Combination of Geometric Mean Compensatory Logic with Generalized Zadeh implication was used as CIS.

The seven variables above define fuzzy variables. Many new predicates were discovered by using the following two predicates below, which have high truth values for every patient:

- (((Race=white) ∧ (Age=advanced)) → (Classification =diabetes)) (premise)
  Truth value of the universal proposition over B: 0.8632
- (((Antecedents=true) ∧ (Age=advanced)) → (Classification =diabetes))
  Truth value of the universal proposition over B: 0.8574

If we consider these predicates like premises, valid for all the instances (patients), we can make the following sequential demonstration. The right deductive structures used, were used as examples before, in section 5:

Hypothesis

1. $Age = advanced$
2. $\neg(Classification = diabetes)$

Premises

3. $((Race = white) \wedge (Age = advanced)) \rightarrow (Classification = diabetes)$
4. $((Antecedents = true) \wedge (Age = advanced)) \rightarrow (Classification = diabetes)$

5. $\neg(Race = white \wedge Age = advanced)$ Modus Tollens (2,3)
6. $\neg(Antecedents = true \wedge Age = advanced)$ Modus Tollens (2,4)
7. $\neg(Race = white) \vee \neg(Age = advanced)$ D' Morgan (5)
8. $\neg(Antecedents = true) \vee \neg(Age = advanced)$ D' Morgan (6)
9. $\neg(\neg(Age = advanced))$ Double negation (1)
10. $\neg(Race = white)$ Disjunctive Syllogism (7,9)
11. $\neg(Antecedents = true)$ Disjunctive Syllogism (8,9)

12. $\neg(Race = white) \vee \neg(Antecedents = true)$ Addition (10)
13. $\neg(Race = white) \wedge \neg(Antecedents = true)$ Product (10, 11)

According the sequential demonstration, the following deductive structures are approximately right.

$(Age = advanced), \neg(Classification = diabetes) \vdash \neg(Race = white)$
$(Age = advanced), \neg(Classification = diabetes) \vdash \neg(Antecedents = true)$
$(Age = advanced), \neg(Classification = diabetes) \vdash \neg(Race = white) \vee \neg(Antecedents = true)$
$(Age = advanced), \neg(Classification = diabetes) \vdash \neg(Race = white) \wedge \neg(Antecedents = true)$

Then the following predicates are probably approximately truth

$(Age = advanced) \wedge \neg(Classification = diabetes) \rightarrow \neg(Race = white)$
$(Age = advanced) \wedge \neg(Classification = diabetes) \rightarrow \neg(Antecedents = true)$
$(Age = advanced) \wedge \neg(Classification = diabetes) \rightarrow \neg(Race = white) \vee \neg(Antecedents = true)$
$(Age = advanced) \wedge \neg(Classification = diabetes) \rightarrow \neg(Race = white) \wedge \neg(Antecedents = true)$

Because the predicates we used as premises are approximately true, we should expect that some of the predicates obtained through reasoning from them, since we use deductive structures approximately true, have good truth values as well.

The truth values of these predicates illustrate that among predicates obtained by Boolean reasoning, there are obtained predicates with good truth values:

$(Age = advanced) \wedge \neg(Classification = diabetes) \rightarrow \neg(Race = white)$
Truth value of the universal proposition over B: 0.6479

$(Age = advanced) \wedge \neg(Classification = diabetes) \rightarrow \neg(Antecedents = true)$
Truth value of the universal proposition over B: 0.8348

$(Age = advanced) \wedge \neg(Classification = diabetes) \rightarrow \neg(Race = white) \vee \neg(Antecedents = true)$
Truth value of the universal proposition over B: 0.7419

$(Age = advanced) \wedge \neg(Classification = diabetes) \rightarrow \neg(Race = white) \wedge \neg(Antecedents = true)$
Truth value of the universal proposition over B: 0.66021

Based on example 3, it the conclusion of the experiments can be made as follows:

According tables 1-7, the best CIS for using any deductive right structure as a way of searching, is the combination of Reichenbach with the conjunction obtained by $p = 0$ (Geometric mean).

As observed in table 8, for the case of Kleene axioms, it is possible to find more convenient CIS for obtaining the best value of each formula. Consequently, it is possible to select the best CIS for using a specific right deductive structure like way of searching.

8 Concluding Remarks

This chapter introduced new fuzzy inference systems based on Compensatory Fuzzy Logic: a quartet of operators, where *c*, *d*, *o* and *n* are the operators of the Compensatory Logic and a fuzzy implication operator.

Some properties of CFL recommend the study of inference in its framework; see [4]:

- Their logic operators seem to be more adequate to model human thinking, according to some experiments [9].
- The truth values of the composed predicates can be interpreted semantically by themselves, because their operators of conjunction and disjunction are idempotent and continuous. This is an advantage over the systems based on norm and conorm.
- Usually, fuzzy systems model the vagueness of the natural language by means of one simple linguistic variable. It doesn't use only simple linguistic variables, but complex phrases expressed in natural language. Thus, CFL allows to model problems expressed in natural language, using sentences provided by experts in the filed, following the methodology of the Expert Systems. Hence, CFL is more adequate than other logic systems to compute with words, according to the idea proposed by Zadeh [14].
- Theorem 2, establishes a very important result: the equivalence between the valid formulas of CFL and bivalent logic. *Corollary 1* expresses the same result between right deductive structures of CFL and Boolean Logic.

As consequence of this corollary, a very important result illustrated through the examples in this chapter is the possibility of applying the classical Boolean reasoning as the search component of Knowledge Discovery by using logical predicates. This is a very important advance in the ways of Approximate Reasoning and Knowledge Discovery. The here demonstrated properties built a very coherent relation between approximate reasoning and Boolean Logic.

Probabilistic properties expressed in theorem 1 allow applying statistical inference in the framework of CFL. That theorem expresses that the universal proposition over a sample can be a statistic estimator of the corresponding universal

proposition over the entire universe. Then CFL joins logical and statistical inferences, it gives logical models of automated learning with properties of a statistical estimator too.

The study of other implications as components of CISs is further needed. The experimentation in the use of these new ways of inference and the development of tools is very important. Some chapters included in this book are beginning studies of Knowledge Discovery by CFL predicates, but studies according searching by reasoning and statistical inference by it, should be studied like promising ways of inference. The use of Kleene axioms is a good approach to evaluate the behavior of CIS systems, but studies of sensitivity and robustness of different CIS operators can offer more light over their selection. The realization of the experiments of Knowledge Discovery using reasoning by bivalent logic as way to search in the space of fuzzy predicates with high truth value according different CIS, different deductive right structures, and looking for the solutions of different knowledge discovery problems are proved to be important, according the results shown in this chapter.

References

1. Buchanan, B.G., Shortliffe, E.H.: Knowledge Engineering. Rule-Based Expert Systems –The MYCIN Experiments of the Stanford Heuristic Programming Project, pp. 147–158. Addison-Wesley, Massachusetts (1984)
2. Detyniecki, M.: Mathematical aggregation operators and their application to video querying. Paris, University of Paris VI. Ph.D. Thesis (2001)
3. Dubois, D., et al.: Fuzzy-Set Based Logic- An History -Oriented Presentation of their Main Developments. In: Gabbay, D.M., Woods, J. (eds.) Handbook of the History of Logic, pp. 325–449. North-Holland, Elsevier BV (2007)
4. Espin, R., et al.: Un sistema lógico para el razonamiento y la toma de decisio-nes: la Lógica Difusa Compensatoria Basada en la Media Geométrica (A logic system for reasoning and decision making: Compensatory Fuzzy Logic on Geometric Mean). Revista Investigación Operacional 32, 230–245 (2011) (in Spanish)
5. Gallier, J.H.: Logic for Computer Science: Foundations of Automatic Theorem Proving. Harper & Row Publishers (1986)
6. Jang, J.-S., et al.: Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Prentice-Hall, Inc., Upper Saddle River (1997)
7. Jayaram, B.: On the Law of Importation in Fuzzy Logic. IEEE Transactions on Fuzzy Systems 16, 130–144 (2008)
8. Mitrinovic, D.S., et al.: Classical and New Inequalities in Analysis. Kluwer Academic Publishers, Dordrecht (1993)
9. Mizumoto, M.: Pictorial Representations of Fuzzy Connectives: Part II. Cases of Compensatory Operators and Self-Dual Operators. Fuzzy Sets and Systems 32, 45–79 (1989)
10. Nóvak, V., Dvorák, A.: Research Report No.125: Fuzzy Logic: A powerful Tool for Modeling of Vagueness. Ostrava, University of Ostrava (2008)
11. Novák, V.: Reasoning about mathematical fuzzy logic and its future. Fuzzy Sets and Systems 192, 25–44 (2012)

12. Rosete, A., et al.: A General Method for Knowledge Discovery Using Compensatory Fuzzy Logic and Metaheuristics. In: Espin, R., Marx, J., Racet, A. (eds.) Toward a Trans-Disciplinary Technology for Business Intelligence, pp. 240–268. Shaker Verlag (2011)
13. Zadeh, L.: Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Computing 2, 23–25 (1998)
14. Zadeh, L.: From computing with numbers to computing with words –from manipulation of measurements to manipulation of perceptions. International Journal of Applied Mathematics and Computational Sciences 3, 307–324 (2002)

A Fuzzy Approach to Prospect Theory

Rafael Alejandro Espín Andrade, Erick González,
Eduardo Fernández, and Salvador Muñoz Gutiérrez

Abstract. The aim of this chapter is to revisit an experiment of Kahneman and Tversky to arrive at conclusions about Prospect theory and the ways of human thinking, but using a fuzzy approach, especially the compensatory one. New results shall be proved and others well-known shall be changed or confirmed. The study comprises the examination of logical predicates like those expressed by the following sentences: "if a scenario is probable then it is *convenient*", "there exist probable and *convenient* scenarios" and "all the scenarios are probable and *convenient*". According to the empirical results, the Reichenbach implication and the Geometric Mean are closest to the people's way of thinking.

1 Introduction

Prospect theory has been well accepted by Decision Theory community. This success is due to its right and simple answer to the question: actually how human beings make decisions under uncertainty? [8]. The expected utility theory, another classic, can't deal with situations where the subjectivity of persons is relevant and, hence, objectivity is not the only factor to be taken into account [6].

Prospect theory is a consequence of many experiments carried out by Kahneman and Tversky about the attitude of human being under uncertainty situations. They maintained the concept of *lottery*, used for computing expected utility functions, which consists of a set of premiums often representing money quantities,

Rafael A. Espín Andrade · Erick González · Salvador Muñoz Gutiérrez
"José Antonio Echeverría" Higher Technical Institute,
Havana, Cuba
e-mail: rafaelespin@yahoo.com,
erickgc@cemat.cujae.edu.cu,
salvador@ind.cujae.edu.cu

Eduardo Fernández
"Autonomous University of Sinaloa", Sinaloa, Mexico
e-mail: eddyf171051@gmail.com

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_3,

positive if they are gains or negative if they are losses, while being associated with the probability of occurrence, such that the probabilities of all the premiums sum one. They studied the shape, slope and other characteristics of a function, named *value function* that measures the risk attitude and preferences of persons. In this context, lotteries are called *prospects*.

On the other hand, Fuzzy logic is a multi-valued logic, with a wide range of applications [4]. Some of their essential properties are their facilities to model the "vagueness" proper to the natural language and the uncertainty. These properties are arguments to justify the relevance of searching for nexuses between Fuzzy logic and Prospect theory. Also, fuzzy logic has been a useful tool for modelling preferences.

The notion of t-norm and t-conorm doesn't seem to be adequate to solve problems in decision making; however, it is the most extended approach of all, even though empirical studies prove that some compensatory operators are closest to represent real human thinking than any t-norm or t-conorm system [10].

The insufficient study of compensatory operators in fuzzy literature [2], usually provokes that the concept of operator prevails over the concept of integrated operators' system. Maybe, the only exception in the literature is *Compensatory Fuzzy Logic* (CFL) [5]. The CFL consists of a set of axioms, some of them inspired in logic and others in Decision theory, which are grouped in a coherent way. It is a quartet of continuous operators (c, d, o, n) of, respectively, a conjunction operator, a disjunction operator, a fuzzy strict order operator and a negation operator.

The conjunction operator of the CFL could be defined with formulas of the quasi-arithmetic means and the disjunction operator could be their duals. CFL is a recommendable tool to be used in *Soft-computing*, which is the classification given by Zadeh [14] to all the branches of Artificial Intelligence opposites to *hard-computing*, such that a good or approximate solution is accepted, even if it is not optimal, and fuzzy logic is one of their bases.

CFL is designed to calculate using complex sentences expressed in natural language, and not the so usually exclusive employment of simple linguistic variables. The conception of this new tool is to reaffirm the Zadeh's idea to compute with words rather than with numbers [15]. This characteristic can be used to link CFL with Artificial Intelligence branches like *Knowledge Engineering*, the *Expert System*'s methodology [1].

The aim of this chapter is to revisit an experiment of Kahneman and Tversky [8] to arrive at conclusions about Prospect theory and the ways of human thinking, but using a fuzzy approach, especially the compensatory one. New results shall be proved and others well-known shall be changed or confirmed. The study comprises the examination of logical predicates like those expressed by the following sentences: "if a scenario is probable then it is *convenient*", "there exist probable and *convenient* scenarios" and "all the scenarios are probable and *convenient*".

In this chapter, a scenario is a premium, which is associated with a probability. An implication operator upon a set of five and a one-parameter family of compensatory systems will be selected for representing these predicates.

The chapter is structured as follows: next section, called *Preliminaries*, is divided in two parts, the first of them explains the basic concepts of Prospect theory and the second one exposes some notions about CFL, including the introduction of a compensatory one-parameter family. The third section describes the experiment of Kahneman and Tversky that shall be used in the chapter; some other notions like implication operators that will be useful are included. This section finishes with the description of a fuzzy approach to Prospect theory. The fourth section describes the analysis of the results.

2 Preliminaries

A *prospect* in Prospect theory, as a *lottery* in Utility theory, is represented by $L = (x_1, p_1, x_2, p_2, \dots, x_n, p_n)$, where p_i is the probability to obtain the potential outcome or premium x_i and $\sum_{i=1}^{n} p_i = 1$.

The detailed manner to measure the prospects can be found in [8], it is basically $V(L) = \sum_{i=1}^{n} \pi(p_i) v(x_i)$.

$\pi(p)$ is called the *weighting function* or *decision weight*, which maps over the probabilities and $v(x)$ is called the *value function*, which maps over the outcomes or premiums.

Let us note that probabilities aren't used directly in the final valorisation of the prospect, because they don't influence objectively the result, but subjectively, according to a function $\pi(p)$ defined by the decision maker. Usually, $\pi(p)$ is assumed by individual decision makers as non-linear weights, which are concave over certain interval [0, b] and convex over the interval [b, 1], where 0<b<1.

The value function has the characteristics summarized below, according to empirical results:

1. There exists a *reference point* that is valued as indifferent by people. The other points are assumed like deviations from this point; therefore, people think in terms of gains and losses.
2. The function is concave over gains and convex over losses. That is to say, it is an s-shaped or sigmoidal function.
3. It is steeper for losses than for gains. This is because people experience losses more intensively than gains.

Hypothetical figures of a value function and a decision weight are represented in figures 1 and 2, respectively.

In brief, people are risk-averse for gains and risk-seeking for losses.

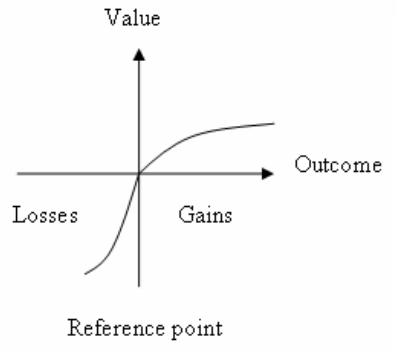


Fig. 1 A hypothetical Value Function

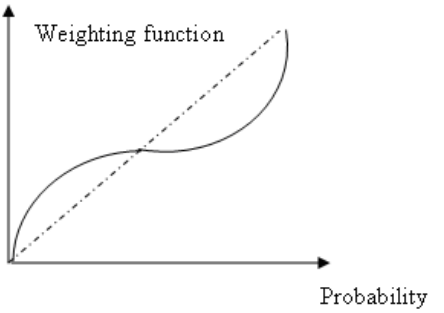


Fig. 2 A hypothetical weighting function

A *Compensatory Fuzzy Logic* (CFL) system is a quartet (*c*,*d*,*o*,*n*) of operators of conjunction, disjunction, fuzzy strict order and negation, respectively [5].

c and *d* map vectors of $[0,1]^n$ into [0,1], *o* is a mapping from $[0,1]^2$ into [0,1], and *n* is a unary operator of [0,1] into [0,1]. Some axiomatic must to be satisfied for the operators of conjunction and disjunction, like for example, Compensation Axiom¸ Symmetry Axiom and others [5].

A family of CFL systems may be obtained from the quasi-arithmetic means, with the following formula below [9]:

$$M_f(x_1, x_2, \dots, x_n) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \tag{1}$$

Where f(x) is a continuous and strictly monotonic function of one real variable.

In this chapter the one-parameter family with formula:

$$M_f(x_1, x_2, \dots, x_n) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{1/p} \tag{2}$$

Where $p \in (-\infty, 0]$ satisfies the axiom of compensation, if the conjunction is defined as in (2). More details about the CFL and formulas of family (2) can be found in (Espin et al. 2011).

Therefore, conjunction is defined as follows:

$$c(x_1, x_2, \dots, x_n) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{1/p} \tag{3}$$

The disjunction is defined as the dual of the conjunction, that is to say:

$$d(x_1, x_2, \dots, x_n) = \left(\frac{1}{n}\sum_{i=1}^{n} (1 - x_i)^p\right)^{1/p} \tag{4}$$

The fuzzy negation is:

$$n(x) = 1 - x \tag{5}$$

The fuzzy strict order is:

$$o(x, y) = 0.5[c(x) - c(y)] + 0.5 \tag{6}$$

$p(\boldsymbol{x})$ is a formula of the propositional calculus in CFL.

This formula is valid in the CFL if it satisfies the condition (7), below:

$$f^{-1}\left(\frac{\int_{[0,1]^n} f(p(x))dx}{\int_{[0,1]^n} dx}\right) > \frac{1}{2} \tag{7}$$

3 The Experiments

This section begins with a useful resume of fuzzy implications.

In fuzzy literature the classification of implication operators is usually defined using other operators, like conjunction, disjunction and negation, but they are always based on t-norm and t-conorm paradigm. In this chapter, these concepts will be extended to any fuzzy system, including the compensatory ones. Here, when it would be necessary, the operators will preserve their exact definition, even if they don't correspond to any classification and taking into account that often the definition of an implication operator is associated with a specific t-norm and t-conorm.

The criteria for selecting implication operators for our purposes are the following:

1. The operator satisfies the truth-value table of the bivalent classical logic, when the truth-values calculus is restricted only to the set $\{0, 1\}$. Briefly, the truth-value of the formula $x \to y$ is 1 if $x = 0$ or $x = y = 1$, and is 0 if $x = 1$ and $y = 0$.
2. The operator must be a continuous function with regard to both arguments or it has a finite number of removable discontinuities.

The reason for imposing condition 1 is that this must be a natural extension of the mathematical logic. Whereas condition 2 guarantees the "sensitiveness" of the composed predicates, that is to say, any change in the simple predicates will be reflected in the final results of their corresponding composed predicates.

Some classifications definitions appeared in the literature are:

- S-implication [4]: $I_s(x, y) = d(n(x), y)$, where d and n are the disjunction and negation operators, respectively.
- R-implication [4]: $I_R(x, y) = sup\{z \in [0,1] : c(x, z) \leq y\}$, where c is the conjunction operator.
- QM-implication [11], which is also known as QL-implication [4]: $I_{QL}(x, y) = d(n(x), c(x, y))$

- A-implication [12]: The operator satisfies a group of axioms, which implicitly associate it with the conjunction, disjunction and negation operators. For example, the Law of Importation $(x \wedge y \rightarrow z) \leftrightarrow (x \rightarrow (y \rightarrow z))$ is one of its axioms, where the symbol $\leftrightarrow$ is the logic equivalence.

The implication operators that have appeared in the literature satisfy the two conditions expressed above, and their classifications are:

- Reichenbach implication (S-implication): $x \rightarrow y = 1 - x + xy$
- Klir-Yuan implication (a variation of the above case without a classification): $x \rightarrow y = 1 - x + x^2 y$
- Natural implication (S-implication), see [5]: $x \rightarrow y = d(n(x), y)$
- Zadeh implication (QL-implication): $x \rightarrow y = d(n(x), c(x, y))$
- Yager implication (A-implication): $x \rightarrow y = y^x$

The formula of the equivalence is defined as: $x \leftrightarrow y = (x \rightarrow y) \wedge (y \rightarrow x)$. It is valid for any implication operator and any conjunction operator.

Other classifications can be found in [7].

This chapter shall revisit an experiment of Tversky and Kahneman appeared in [13]. The results are summarized in the table 1:

Table 1 Results of an experiment of Tversky and Kahneman

Premium 1	Premium 2	Probability 1	Probability 2	Equivalent
0	50	0.9	0.1	9
0	50	0.5	0.5	21
0	50	0.1	0.9	37
0	-50	0.9	0.1	-8
0	-50	0.5	0.5	-21
0	-50	0.1	0.9	-37
0	100	0.95	0.05	14
0	100	0.75	0.25	25
0	100	0.5	0.5	36
0	100	0.25	0.75	52
0	100	0.05	0.95	78
0	100	0.95	0.05	-8
0	100	0.75	0.25	-23.5
0	100	0.5	0.5	-42
0	100	0.25	0.75	-63
0	100	0.05	0.95	-84
0	200	0.99	0.01	10
0	200	0.9	0.1	20
0	200	0.5	0.5	76
0	200	0.1	0.9	131
0	200	0.01	0.99	188
0	-200	0.99	0.01	-3
0	-200	0.9	0.1	-23
0	-200	0.5	0.5	-89

Table 1 (*continued*)

Premium 1	Premium 2	Probability 1	Probability 2	Equivalent
0	-200	0.1	0.9	-155
0	-200	0.01	0.99	-190
0	400	0.99	0.01	12
0	400	0.01	0.99	377
0	-400	0.99	0.01	-14
0	-400	0.01	0.99	-380
50	100	0.9	0.1	59
50	100	0.5	0.5	71
50	100	0.1	0.9	83
-50	-100	0.9	0.1	-59
-50	-100	0.5	0.5	-71
-50	-100	0.1	0.9	-85
50	150	0.95	0.05	64
50	150	0.75	0.25	72.5
50	150	0.5	0.5	86
50	150	0.25	0.75	102
50	150	0.05	0.95	128
-50	-150	0.95	0.05	-60
-50	-150	0.75	0.25	-71
-50	-150	0.5	0.5	-92
-50	-150	0.25	0.75	-113
-50	-150	0.05	0.95	-132
100	200	0.95	0.05	118
100	200	0.75	0.25	130
100	200	0.5	0.5	141
100	200	0.25	0.75	162
100	200	0.05	0.95	178
-100	-200	0.95	0.05	-112
-100	-200	0.75	0.25	-121
-100	-200	0.5	0.5	-142
-100	-200	0.25	0.75	-158
-100	-200	0.05	0.95	-179

Columns 1, 2, 3 and 4 of table 1 represent prospects of two alternatives and the ultimate column summarizes equivalent values of their acceptance.

The data in table 1 will be interpreted with fuzzy models. Sigmoidal is the membership function that will be used, according to the recommendation appeared in [3].

The sigmoidal function formula is:

$$sigm(x, \alpha, y) = \frac{1}{1 + e^{-\alpha(x-y)}} \tag{8}$$

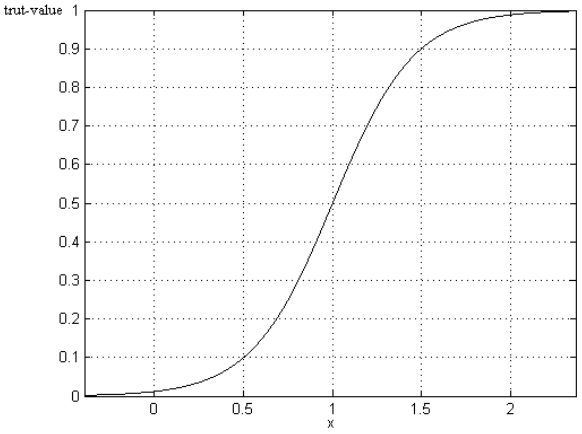


Fig. 3 A generic sigmoidal function with parameters $\gamma = 1$ and $\beta = 0.5$

Figure 3 is the graphic of a generic sigmoidal membership function, where $\gamma = 1$ and $\beta = 0.5$. $\alpha = 4.3944$ was calculated by the following formula below:

$$\alpha = \frac{\ln(0.9) - \ln(0.1)}{\gamma - \beta}$$

Let us note that $sigm(\gamma, \alpha, \gamma) = 0.5$, $sigm(\beta, \alpha, \gamma) = 0.1$ and it is s-shaped, different from function in figure 2 and equal to figure 1. $x = \gamma$ is an "indifferent" value and $x = \beta$ is "almost false" in formula (8).

Here a "scenario" is a premium associated with a probability and it will be classified with the term "convenient".

Three predicates will be calculated using fuzzy variables:

1. "If the scenario is probable then it is convenient".

"All the scenarios are probable and convenient". This statement measures the risk-aversion tendency by the decision makers.

"There exist probable and convenient scenarios". This statement measures the risk-seeking tendency by the decision makers.

It is converted in an optimization (maximization) problem which will be detailed below in order of apparition:

1. The first proposition is divided in the following two: "If all the scenarios are probable then they are convenient" and "If there are probable scenarios then they are convenient".

The maximization problems are respectively:

1.1. Maximize $P_{11}(x)$ such that $P_{11}(x)$ is:

$$\wedge_{i=1}^{56} ((u_p(p_i) \rightarrow u_x(x_i)) \wedge (u_p(1-p_i) \rightarrow u_x(y_i))) \leftrightarrow sigm(eq_i, \alpha_{eq}, \gamma_{eq}) \quad (9)$$

Where u_p and u_x are the sigmoidal functions sigm(p,α_p,γ_p) and sigm(x,α_x,γ_x), representing respectively the predicates "the scenario is probable" and "the scenario is convenient". sigm(eq_i,α_{eq},γ_{eq}) is the sigmoidal function of the equivalent values.

1.2. Besides, the second problem consists in maximizing $P_{12}(x)$ such that $P_{12}(x)$ is:

$$max \wedge_{i=1}^{56} ((u_p(p_i) \rightarrow u_x(x_i)) \vee (u_p(1-p_i) \rightarrow u_x(y_i))) \leftrightarrow sigm(eq_i, \alpha_{eq}, \gamma_{eq}) \quad (10)$$

2. The maximization problem is to find the maximum of $P_2(x)$ such that $P_2(x)$ is:

$$max \wedge_{i=1}^{56} ((u_p(p_i) \wedge u_x(x_i)) \wedge (u_p(1-p_i) \wedge u_x(y_i))) \leftrightarrow sigm(eq_i, \alpha_{eq}, \gamma_{eq}) \quad (11)$$

3. The maximization problem consists in maximizing $P_3(x)$ such that $P_3(x)$ is:

$$max \wedge_{i=1}^{56} ((u_p(p_i) \wedge u_x(x_i)) \vee (u_p(1-p_i) \wedge u_x(y_i))) \leftrightarrow sigm(eq_i, \alpha_{eq}, \gamma_{eq}) \quad (12)$$

Formulas 9-12 are aggregation operators for all the lotteries, the conjunctions $\wedge_{i=1}^{56}$ were defined on the set of the 56 lotteries, see table 1. u_p and u_x are modelled by using sigmoidal membership functions, the first of them represents the subjective perception of probability by people and the second one is the value function.

Because each lottery in table 1 consists in two scenarios with two probabilities, there are two evaluations for u_p and u_x in the lottery, first for p_i and $1-p_i$, see third and fourth columns in table 1, and secondly for x_i and y_i, see the two first columns. The last column represents an equivalent valorisation of the lottery in the experiment. It is also modelled with a sigmoidal function which depends on two parameters, α_{eq} and γ_{eq}.

The search of the three sigmoidal functions u_p, u_x and $sigm(eq_i, \alpha_{eq}, \gamma_{eq})$ by each problem is reduced to the optimization on the space of the six parameters $\gamma_p, \alpha_x, \gamma_x, \alpha_{eq}$ and γ_{eq}, where the objective functions are those represented in formulas 9-12. Other unknown in formulas above are the implication operator $\rightarrow$ and hence, the equivalence $\leftrightarrow$, therefore, the Reichenbach, Yager, Klir-Yuan, Natural

and Zadeh are tested. CFL, depending on parameter p in formula (2) are tested too, and the search actually depends on eight parameters, if p and $\rightarrow$ are included.

The optimization problems, 1.1, 1.2, 2 and 3, are reduced to estimate the maximum truth-values of formulas 9-12 respectively, with a fixed $\rightarrow$ and varying the other seven parameters that were detailed in the paragraph above.

Every formula 9-12 is equivalent to a linguistic problem. For example, in formula 9, $((u_p(p_i) \rightarrow u_x(x_i)) \wedge (u_p(1-p_i) \rightarrow u_x(y_i)))$ means for the lottery i, "if the first scenario is probable then it is convenient and if the second scenario is probable then it is convenient". On the other hand, the logical equivalence $\leftrightarrow$ emulates the experimental equivalence summarized in the last column of table 1. This reasoning can be generalized to the other predicates which represent the other problems.

To sum up, each optimization problem depends on a CFL system. The one-parameter family of formulas 3, 4, 5, 6 will be one of the parameter to be estimated. Also, each problem derives in five cases, where the implication operator is applied from the five proposed in the beginning of the section.

Some heuristic restrictions of the alphas and gammas that will be applied are:

1. All the alphas are strictly equal to 0. This condition guarantees that sigmoidal is an increasing function and not a constant one, such as the case where it is equal 0.

The values of gammas are between the minimum and the maximum data in table 1, where they do not represent the equivalent values.

In case of the equivalent values of the last column in table 1, the gamma will be restricted between 0 and 76. As a result of the Prospect theory, it is well-known that people don't accept non-positive values with indifference; taking into account that gamma is the value which represents indifference (0.5). 76 is 20% of the absolute value of the maximum number in the last column in table 1, which has been selected heuristically.

The optimization will be based on the genetic algorithm coded in MATLAB.

4 Results

Tables 2-5 summarize the results for every optimization problem exposed above. Table 2, for example, may be read as following: The maximum truth-value of the objective function of formula (9) in the case of Reichenbach implication is 0.93791284, see second column and ultimate file. This is the biggest value by column in this table; hence, Reichenbach implication is the best of all implication operators for problem 1.1, which linguistically represents the predicate: "If all the scenarios are probable then they are convenient".

The values which maximize the problem 1.1 are: α_x=64, γ_x=1, α_p=11.0376854, γ_p=0.02615738, α_{eq}=45 and γ_{eq}=56. The last parameter estimated is $p = 0$, which corresponds to the Geometric Mean in formula (3).

Table 2 Estimated parameters for problem 1.1. "If all the scenarios are probable then they are convenient"

Estimated Parameters	Reichenbach	Yager	Klir-Yuan	Natural	Zadeh
α_x	64	0.88085938	64	57	128
γ_x	1	30.6367188	0	1	129
α_p	11.0376854	2.12890625	230	19.8595638	21.6115036
γ_p	0.02615738	1	0	0.10683823	0.4031105
α_{eq}	45	0.09375	65	32	74.8601074
γ_{eq}	56	60.7246094	1	57	73
P	0	0	0	0	0
Maximum truth-value	0.93791284	0.85109059	0.87150398	0.88978479	0.79259532

Every pair of parameters represents a sigmoidal membership function and hence, a fuzzy selection pattern by people. In case of the Table 2 they are plotted in figure 4.

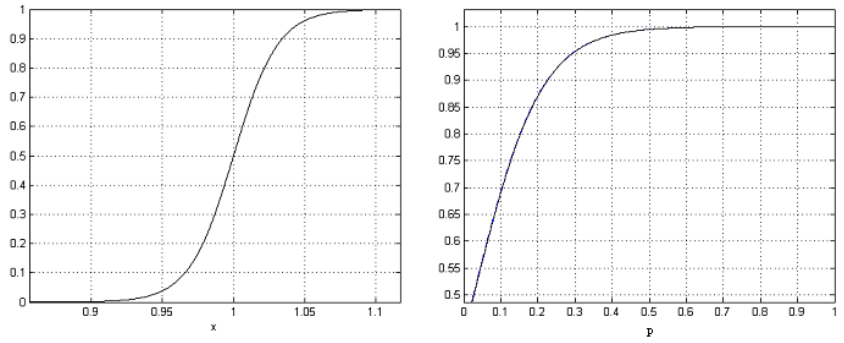


Fig. 4 Membership functions of the predicates: "The scenario is convenient" (left) and "the scenario is probable" (right), for the problem 1.1, with Reichenbach implication. See table 2

According to the meaning of each parameter, in problem 1.1, representing the predicate: "if the scenario is probable then it is convenient", people is indifferent when the function value is 1 and when the probability is 0.02615738, because $\gamma_x = 1$ and $\gamma_p = 0.02615738$, respectively.

Parameter $\alpha_x = 64$ corresponds to $\beta_x = 0.9657$, according to the formula of α, appeared above. Therefore, people consider "almost false" a value function equalling 0.9657.

A negative value of p suggests a "pessimistic" tendency in the people's behaviour. Let us note that p=0 or $p \approx 0$ for all the cases; therefore, people actually have a neutral's behaviour.

The reasoning above for the problem 1.1 can be extended to the other three problems, which their corresponding results are summarized in tables 3-5. Table 3,

4 and 5 summarize the values of the optimization problems with objective functions showed in equations (10), (11) and (12), respectively. Their corresponding figures are 5, 6 and 7.

Table 3 Estimated parameters for problem 1.2. "If there are probable scenarios then they are convenient"

Estimated parameters	Reichenbach	Yager	Klir-Yuan	Natural	Zadeh
α_x	32	0.0703125	64	31	128
γ_x	1.5	317.71875	1	1	1
α_p	230	7.5625	6.76686478	97	17.2717075
γ_p	0	1	0	1	0
α_{eq}	0.03500748	0	25	0	65
γ_{eq}	0	1	16	1	17
P	0	0	0	0	-1.9073E-06
Maximum truth-value	0.85970284	0.7147067	0.8335026	0.5411961	0.76028923

Table 4 Estimated parameters for problem 2. "All the scenarios are probable and convenient"

Estimated parameters	Reichenbach	Yager	Klir-Yuan	Natural	Zadeh
α_x	5.52869034	0.734375	6.02235603	7.0930481	12.0120811
γ_x	1	14.8125	1	1	1
α_p	230	230	230	230	230
γ_p	0	0	0	0	0
α_{eq}	62	0.09375	30	24	24
γ_{eq}	53	58.9453125	57	57	57
P	0	0	0	0	0
Maximum truth-value	0.90420119	0.81944258	0.89626517	0.83563393	0.87025163

Table 5 Estimated parameters for problem 3. "There exist probable and convenient scenarios"

Estimated parameters	Reichenbach	Yager	Klir-Yuan	Natural	Zadeh
α_x	97	0.3203	97	128	97
γ_x	1	18.6406	1	65	1
α_p	8.4261	6.7734	8.17059708	15.8297119	14.9041805
γ_p	0.354	0	0.24069786	0.58897972	0.82479858
α_{eq}	229.2813	0.0781	48	129	74.8599014
γ_{eq}	20.375	0	17	73	73
P	0	0	0	0	0
Maximum truth-value	0.9046	0.7803	0.87730427	0.83112339	0.76983507

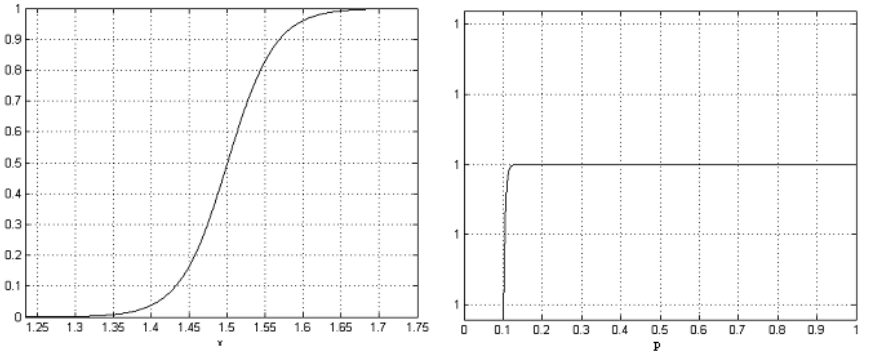


Fig. 5 Membership functions of the predicates: "The scenario is convenient" (left) and "the scenario is probable" (right), for the problem 1.2, with Reichenbach implication. See table 3.

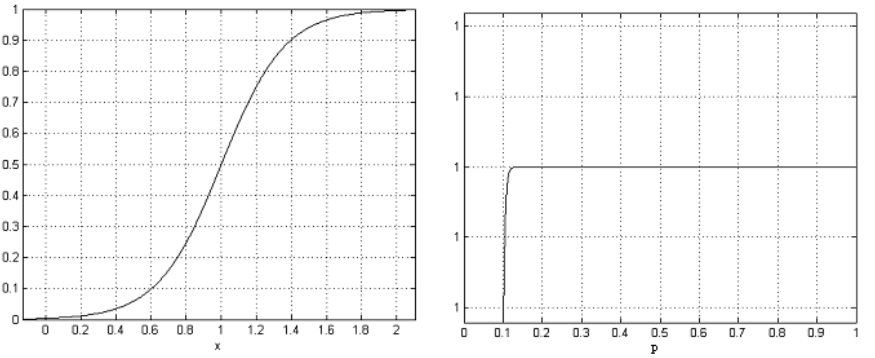


Fig. 6 Membership functions of the predicates: "The scenario is convenient" (left) and "the scenario is probable" (right), for the problem 2, with Reichenbach implication. See table 4.

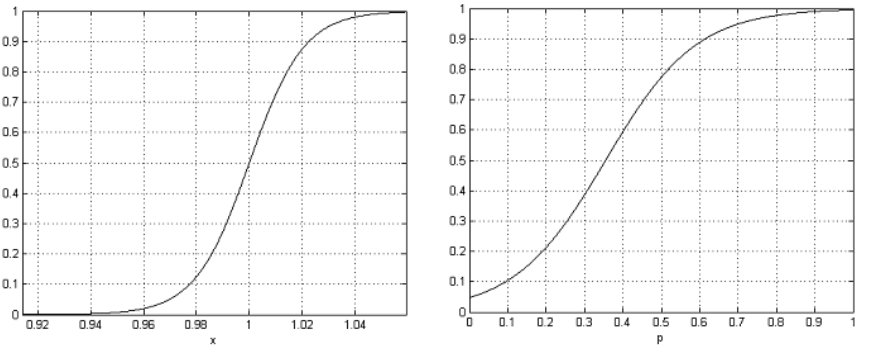


Fig. 7 Membership functions of the predicates: "The scenario is convenient" (left) and "the scenario is probable" (right), for the problem 3, with Reichenbach implication. See table 5.

These results allow arriving to some conclusions:

- All the predicates show better results with the Reichenbach implication.
- People measure preferences with Geometric Mean ($p = 0$).
- With Reichenbach implication, the values of indifference for the scenarios are equal or slightly bigger than 1.
- The probabilities are measured with small slopes and γ_p>0, for problems: "If all the scenarios are probable then they are convenient" and "There exist probable and convenient scenarios" (risk-seeking), see tables 2 and 5. Besides, the probabilities for: "If there are probable scenarios then they are convenient" and "There exist probable and convenient scenarios" (risk-aversion), have big slopes and the minimum of their values is 0.5 for the probability 0, see tables 3 and 4.

Other experiments made by authors, show that the shape of the membership function, like in figure 2, doesn't contribute to better results of the truth-values in the maximizations.

Tables below indicate the application of precedent results in the experiment summarized in table 1.

Table 6 Predicates "If all the scenarios are probable then they are convenient" and "If there are probable scenarios then they are convenient", respectively in columns 1 and 2 for Reichenbach, Yager and Klir-Yuan implications applied to the experiment of table 1. The next-to-last and last columns for every implication represent the conjunction and disjunction of the two predicates, respectively.

	Reichenbach				Yager				Klir-Yuan			
Lottery	∀	∃	∧	∨	∀	∃	∧	∨	∀	∃	∧	∨
1	0.01	1.00	0.09	1.00	0.00	0.86	0.05	0.62	0.71	0.53	0.84	1.00
2	0.07	1.00	0.27	1.00	0.03	0.63	0.14	0.40	0.71	0.82	0.84	1.00
3	0.55	1.00	0.74	1.00	0.18	0.84	0.39	0.64	0.71	0.96	0.84	1.00
4	0.00	0.00	0.00	0.00	0.00	0.83	0.00	0.59	0.00	0.19	0.00	0.16
5	0.01	0.00	0.00	0.00	0.00	0.59	0.00	0.36	0.00	0.03	0.00	0.16
6	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.60	0.00	0.19	0.00	0.16
7	0.01	1.00	0.08	1.00	0.00	0.89	0.04	0.67	0.71	0.51	0.84	0.97
8	0.02	1.00	0.14	1.00	0.01	0.78	0.07	0.53	0.71	0.64	0.84	1.00
9	0.07	1.00	0.27	1.00	0.03	0.66	0.14	0.43	0.71	0.82	0.84	1.00
10	0.28	1.00	0.53	1.00	0.10	0.75	0.28	0.52	0.71	0.93	0.84	1.00
11	0.66	1.00	0.81	1.00	0.21	0.87	0.42	0.68	0.71	0.97	0.84	1.00
12	0.01	1.00	0.08	1.00	0.00	0.89	0.04	0.67	0.71	0.51	0.84	0.97
13	0.02	1.00	0.14	1.00	0.01	0.78	0.07	0.53	0.71	0.64	0.84	1.00
14	0.07	1.00	0.27	1.00	0.03	0.66	0.14	0.43	0.71	0.82	0.84	1.00
15	0.28	1.00	0.53	1.00	0.10	0.75	0.28	0.52	0.71	0.93	0.84	1.00
16	0.66	1.00	0.81	1.00	0.21	0.87	0.42	0.68	0.71	0.97	0.84	1.00

Table 6 (*continued*)

17	0.00	1.00	0.07	1.00	0.00	0.93	0.03	0.74	0.68	0.50	0.73	0.74
18	0.01	1.00	0.09	1.00	0.00	0.90	0.05	0.69	0.71	0.53	0.84	1.00
19	0.07	1.00	0.27	1.00	0.03	0.74	0.15	0.50	0.71	0.82	0.84	1.00
20	0.55	1.00	0.74	1.00	0.18	0.85	0.39	0.65	0.71	0.96	0.84	1.00
21	0.74	1.00	0.86	1.00	0.23	0.89	0.45	0.71	0.71	0.97	0.84	1.00
22	0.00	0.05	0.01	0.03	0.00	0.86	0.00	0.62	0.21	0.28	0.26	0.27
23	0.00	0.00	0.00	0.00	0.00	0.80	0.00	0.55	0.00	0.19	0.00	0.16
24	0.01	0.00	0.00	0.00	0.00	0.53	0.00	0.32	0.00	0.03	0.00	0.16
25	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.60	0.00	0.19	0.00	0.16
26	0.00	0.05	0.01	0.03	0.00	0.89	0.00	0.67	0.00	0.28	0.00	0.16
27	0.00	1.00	0.07	1.00	0.00	1.00	0.04	0.96	0.68	0.50	0.73	0.74
28	0.74	1.00	0.86	1.00	0.23	1.00	0.48	0.94	0.71	0.97	0.84	1.00
29	0.00	0.05	0.01	0.03	0.00	0.83	0.00	0.59	0.21	0.28	0.26	0.27
30	0.00	0.05	0.01	0.03	0.00	0.89	0.00	0.67	0.00	0.28	0.00	0.16
31	1.00	1.00	1.00	1.00	1.00	0.87	0.93	1.00	1.00	0.98	1.00	1.00
32	1.00	1.00	1.00	1.00	1.00	0.69	0.83	1.00	1.00	0.97	1.00	1.00
33	1.00	1.00	1.00	1.00	1.00	0.86	0.93	1.00	1.00	0.98	1.00	1.00
34	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.58	0.00	0.19	0.00	0.00
35	0.01	0.00	0.00	0.00	0.00	0.54	0.00	0.32	0.00	0.03	0.00	0.00
36	0.00	0.00	0.00	0.00	0.00	0.83	0.00	0.59	0.00	0.19	0.00	0.00
37	1.00	1.00	1.00	1.00	1.00	0.91	0.95	1.00	1.00	0.98	1.00	1.00
38	1.00	1.00	1.00	1.00	1.00	0.81	0.90	1.00	1.00	0.97	1.00	1.00
39	1.00	1.00	1.00	1.00	1.00	0.72	0.85	1.00	1.00	0.97	1.00	1.00
40	1.00	1.00	1.00	1.00	1.00	0.78	0.88	1.00	1.00	0.97	1.00	1.00
41	1.00	1.00	1.00	1.00	1.00	0.88	0.94	1.00	1.00	0.98	1.00	1.00
42	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.60	0.00	0.24	0.00	0.00
43	0.01	0.00	0.00	0.00	0.00	0.68	0.00	0.43	0.00	0.08	0.00	0.00
44	0.01	0.00	0.00	0.00	0.00	0.52	0.00	0.31	0.00	0.03	0.00	0.00
45	0.01	0.00	0.00	0.00	0.00	0.71	0.00	0.46	0.00	0.08	0.00	0.00
46	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.63	0.00	0.24	0.00	0.00
47	1.00	1.00	1.00	1.00	1.00	0.92	0.96	1.00	1.00	0.98	1.00	1.00
48	1.00	1.00	1.00	1.00	1.00	0.84	0.92	1.00	1.00	0.97	1.00	1.00
49	1.00	1.00	1.00	1.00	1.00	0.78	0.88	1.00	1.00	0.97	1.00	1.00
50	1.00	1.00	1.00	1.00	1.00	0.82	0.90	1.00	1.00	0.97	1.00	1.00
51	1.00	1.00	1.00	1.00	1.00	0.89	0.95	1.00	1.00	0.98	1.00	1.00
52	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.59	0.00	0.24	0.00	0.00
53	0.01	0.00	0.00	0.00	0.00	0.66	0.00	0.42	0.00	0.08	0.00	0.00
54	0.01	0.00	0.00	0.00	0.00	0.48	0.00	0.28	0.00	0.03	0.00	0.00
55	0.01	0.00	0.00	0.00	0.00	0.69	0.00	0.44	0.00	0.08	0.00	0.00
56	0.00	0.00	0.00	0.00	0.00	0.85	0.00	0.61	0.00	0.24	0.00	0.00

Table 7 Predicates "If all the scenarios are probable then they are convenient" and "If there are probable scenarios then they are convenient" in columns 1 and 2 for Natural and Zadeh implications, respectively, applied to the experiment of table 1. The next-to-last and last columns for every implication represent the conjunction and disjunction of the two predicates, respectively.

	Natural				Zadeh			
Lottery	$\forall$	$\exists$	$\wedge$	$\vee$	$\forall$	$\exists$	$\wedge$	$\vee$
1	0.00	1.00	0.02	1.00	0.00	0.49	0.05	0.56
2	0.01	1.00	0.12	1.00	0.06	0.90	0.06	0.06
3	0.56	1.00	0.75	1.00	0.00	0.98	0.05	0.56
4	0.00	1.00	0.01	0.09	0.00	0.04	0.05	0.56
5	0.00	1.00	0.00	0.00	0.06	0.00	0.06	0.06
6	0.00	1.00	0.01	0.09	0.00	0.04	0.05	0.56
7	0.00	1.00	0.01	1.00	0.00	0.42	0.04	0.62
8	0.00	1.00	0.03	1.00	0.02	0.72	0.09	0.35
9	0.01	1.00	0.12	1.00	0.06	0.90	0.06	0.06
10	0.17	1.00	0.41	1.00	0.02	0.97	0.09	0.35
11	0.71	1.00	0.84	1.00	0.00	0.99	0.04	0.62
12	0.00	1.00	0.01	1.00	0.00	0.42	0.04	0.62
13	0.00	1.00	0.03	1.00	0.02	0.72	0.09	0.35
14	0.01	1.00	0.12	1.00	0.06	0.90	0.06	0.06
15	0.17	1.00	0.41	1.00	0.02	0.97	0.09	0.35
16	0.71	1.00	0.84	1.00	0.00	0.99	0.04	0.62
17	0.00	1.00	0.01	1.00	0.00	0.39	0.03	0.66
18	0.00	1.00	0.02	1.00	0.00	0.49	0.05	0.56
19	0.01	1.00	0.12	1.00	0.21	0.90	0.34	0.40
20	0.56	1.00	0.75	1.00	0.98	0.98	0.98	0.98
21	0.80	1.00	0.90	1.00	0.99	0.99	0.99	0.99
22	0.00	1.00	0.01	0.23	0.00	0.14	0.03	0.65
23	0.00	1.00	0.01	0.09	0.00	0.04	0.05	0.56
24	0.00	1.00	0.00	0.00	0.06	0.00	0.06	0.06
25	0.00	1.00	0.01	0.09	0.00	0.04	0.05	0.56
26	0.00	1.00	0.01	0.23	0.00	0.14	0.03	0.65
27	0.00	1.00	0.01	1.00	0.00	0.39	0.03	0.66
28	0.80	1.00	0.90	1.00	0.99	0.99	0.99	0.99
29	0.00	1.00	0.01	0.23	0.00	0.14	0.03	0.65
30	0.00	1.00	0.01	0.23	0.00	0.14	0.03	0.65
31	1.00	1.00	1.00	1.00	0.00	0.99	0.05	0.56
32	1.00	1.00	1.00	1.00	0.06	0.99	0.06	0.06
33	1.00	1.00	1.00	1.00	0.00	0.99	0.05	0.56
34	0.00	1.00	0.01	0.09	0.00	0.04	0.05	0.56

Table 7 (*continued*)

35	0.00	1.00	0.00	0.00	0.06	0.00	0.06	0.06
36	0.00	1.00	0.01	0.09	0.00	0.04	0.05	0.56
37	1.00	1.00	1.00	1.00	0.00	0.99	0.04	0.62
38	1.00	1.00	1.00	1.00	0.02	0.99	0.09	0.36
39	1.00	1.00	1.00	1.00	0.21	0.99	0.34	0.40
40	1.00	1.00	1.00	1.00	0.89	0.99	0.92	0.92
41	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99
42	0.00	1.00	0.01	0.16	0.00	0.08	0.04	0.62
43	0.00	1.00	0.00	0.01	0.02	0.00	0.09	0.35
44	0.00	1.00	0.00	0.00	0.06	0.00	0.06	0.06
45	0.00	1.00	0.00	0.01	0.02	0.00	0.09	0.35
46	0.00	1.00	0.01	0.16	0.00	0.08	0.04	0.62
47	1.00	1.00	1.00	1.00	0.00	0.99	0.04	0.62
48	1.00	1.00	1.00	1.00	0.02	0.99	0.09	0.36
49	1.00	1.00	1.00	1.00	0.21	0.99	0.34	0.40
50	1.00	1.00	1.00	1.00	0.89	0.99	0.92	0.92
51	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99
52	0.00	1.00	0.01	0.16	0.00	0.08	0.04	0.62
53	0.00	1.00	0.00	0.01	0.02	0.00	0.09	0.35
54	0.00	1.00	0.00	0.00	0.06	0.00	0.06	0.06
55	0.00	1.00	0.00	0.01	0.02	0.00	0.09	0.35
56	0.00	1.00	0.01	0.16	0.00	0.08	0.04	0.62

Table 8 Predicate "All the scenarios are probable and convenient"

Lottery	Reichenbach	Yager	Klir-Yuan	Natural	Zadeh
1	0.17806275	0.58009323	0.28618914	0	0.00225796
2	0.65314345	0.8747217	0.76511623	0	0.04529298
3	0.9293946	0.96685524	0.9522479	0	0.63724983
4	0	0.02553434	0	0	0
5	0	0.02513789	0	0	0
6	0	0.02076962	0	0	0
7	0.14424611	0.52671956	0.23647781	0.0070434	0.00155506
8	0.32336441	0.72274194	0.47123815	0.03470226	0.00692078
9	0.65314345	0.87480273	0.76511623	0.25384054	0.04529298
10	0.8684096	0.94566937	0.9122294	0.80782289	0.29075092
11	0.94273215	0.97224114	0.96104794	0.95945049	0.73654208
12	0.14424611	0.52671956	0.23647781	0.0070434	0.00155506
13	0.32336441	0.72274194	0.47123815	0.03470226	0.00692078
14	0.65314345	0.87480273	0.76511623	0.25384054	0.04529298

Table 8 (*continued*)

15	0.8684096	0.94566937	0.9122294	0.80782289	0.29075092
16	0.94273215	0.97224114	0.96104794	0.95945049	0.73654208
17	0.12166376	0.48342306	0.20192342	0.00512731	0.001154
18	0.17806275	0.58011323	0.28618914	0.0104797	0.00225796
19	0.65314345	0.87480273	0.76511623	0.25384054	0.04529298
20	0.9293946	0.96717048	0.9522479	0.93985357	0.63724983
21	0.95157961	0.97572583	0.96690981	0.97043746	0.79980276
22	0	0.02554093	0	0	0
23	0	0.02552766	0	0	0
24	0	0.02512983	0	0	0
25	0	0.02076139	0	0	0
26	0	0.01830749	0	0	0
27	0.12166376	0.48342306	0.20192342	0.00512731	0.001154
28	0.95157961	0.97572583	0.96690981	0.97043746	0.79980276
29	0	0.02554093	0	0	0
30	0	0.01830749	0	0	0
31	0.94196679	0.98541558	0.96591403	0.0104797	0.63806891
32	0.87969053	0.98349646	0.94482962	0.25384054	0.0885345
33	0.94196679	0.98555361	0.96591403	0.93985357	0.63806891
34	0	8.4041E-06	0	0	0
35	0	8.2753E-06	0	0	0
36	0	6.8524E-06	0	0	0
37	0.95099281	0.98606545	0.97025924	0.0070434	0.73695177
38	0.91096126	0.98411595	0.95359026	0.03470226	0.29565947
39	0.87969053	0.98349646	0.94482962	0.25384054	0.0885345
40	0.91096126	0.98416925	0.95359026	0.80782289	0.29565947
41	0.95099281	0.98624983	0.97025924	0.95945049	0.73695177
42	0	8.4046E-06	0	0	0
43	0	8.3853E-06	0	0	0
44	0	8.2726E-06	0	0	0
45	0	7.7305E-06	0	0	0
46	0	6.4272E-06	0	0	0
47	0.95099281	0.98625032	0.97025924	0.95973609	0.73695177
48	0.91096126	0.9841712	0.95359026	0.81449187	0.29565947
49	0.87969053	0.98350713	0.94482962	0.44324606	0.0885345
50	0.91096126	0.9841712	0.95359026	0.81449187	0.29565947
51	0.95099281	0.98625032	0.97025924	0.95973609	0.73695177
52	0	2.7983E-09	0	0	0
53	0	2.7919E-09	0	0	0
54	0	2.7544E-09	0	0	0
55	0	2.5739E-09	0	0	0
56	0	2.14E-09	0	0	0

Table 9 Predicate "There exist probable and convenient scenarios"

Lottery	Reichenbach	Yager	Klir-Yuan	Natural	Zadeh
1	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
2	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
3	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
4	6.1029E-32	4.4792E-07	9.9667E-35	8.9904E-41	1.52E-68
5	6.1029E-32	4.4792E-07	9.9667E-35	8.9904E-41	1.52E-68
6	6.1029E-32	4.4792E-07	9.9667E-35	8.9904E-41	1.52E-68
7	0.25078333	0.06590838	0.22175174	0.16974259	0.04963672
8	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
9	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
10	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
11	0.25078333	0.06590838	0.22175174	0.16974259	0.04963672
12	0.25078333	0.06590838	0.22175174	0.16974259	0.04963672
13	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
14	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
15	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
16	0.25078333	0.06590838	0.22175174	0.16974259	0.04963672
17	0.24486462	0.06435289	0.21651821	0.16573651	0.04846525
18	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
19	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
20	0.25078396	0.06590855	0.22175231	0.16974302	0.04963685
21	0.24486462	0.06435289	0.21651821	0.16573651	0.04846525
22	0	4.7947E-19	0	0	0
23	0	4.9106E-19	0	0	0
24	0	4.9106E-19	0	0	0
25	0	4.9106E-19	0	0	0
26	0	4.7947E-19	0	0	0
27	0.24486462	0.06435289	0.21651821	0.16573651	0.04846525
28	0.24486462	0.06435289	0.21651821	0.16573651	0.04846525
29	0	5.4201E-35	0	0	0
30	0	5.4201E-35	0	0	0
31	1	1	1	1	1
32	1	1	1	1	1
33	1	1	1	1	1
34	5.7408E-92	4.7624E-15	4.092E-100	0	0
35	5.7408E-92	4.7624E-15	4.092E-100	0	0
36	5.7408E-92	4.7624E-15	4.092E-100	0	0
37	0.99999747	0.99999747	0.99999747	0.99999747	0.99999747
38	1	1	1	1	1

Table 9 (*continued*)

39	1	1	1	1	1
40	1	1	1	1	1
41	0.99999747	0.99999747	0.99999747	0.99999747	0.99999747
42	0	4.9106E-19	0	0	0
43	0	4.9106E-19	0	0	0
44	0	4.9106E-19	0	0	0
45	0	4.9106E-19	0	0	0
46	0	4.9106E-19	0	0	0
47	0.99999747	0.99999747	0.99999747	0.99999747	0.99999747
48	1	1	1	1	1
49	1	1	1	1	1
50	1	1	1	1	1
51	0.99999747	0.99999747	0.99999747	0.99999747	0.99999747
52	0	5.2211E-27	0	0	0
53	0	5.2211E-27	0	0	0
54	0	5.2211E-27	0	0	0
55	0	5.2211E-27	0	0	0
56	0	5.2211E-27	0	0	0

Tables 6 and 7 summarize the calculus of the predicates: “If all the scenarios are probable then they are convenient” and “If there are probable scenarios then they are convenient”, corresponding to problem 1 by each of the 56 lotteries shown in table 1. The results were separated taking into account the five implication operators: Reichenbach, Yager and Klir-Yuan in table 6, Natural and Zadeh in table 7.

Every implication operator has associated four columns, the first of them, represented by symbol $\forall$, is the value of the predicate: “If all the scenarios are probable then they are convenient”. Sigmoidal functions of figure 4 were used.

The second column with symbol $\exists$ corresponds to results of the predicate: “If there are probable scenarios then they are convenient”, which uses the membership functions of figure 5. The third and fourth columns are conjunction (symbol $\wedge$) and disjunction (symbol $\vee$), respectively, of the two first columns.

The first column represents risk-aversion by people and the second one represents risk-seeking. The other two columns compute its aggregation using the conjunction and the disjunction.

For instance, with Reichenbach implication the first lottery in table 1 (0, 0.9; 50, 0.1), has truth-value 0.01 for the predicate “if all the scenarios are probable then they are convenient”, see second column and third row in table 6, and truth-value 1 for “If there are probable scenarios then they are convenient”. The conjunction and disjunction of these two truth-values may be found in the two next columns; they are 0.09 and 1, respectively. The computation is based on the sigmoidal functions obtained from the optimization problems 1.1 and 1.2, see tables 2 and 3.

Table 8 summarizes the results by every lottery in table 1 of the predicate: "All the scenarios are probable and convenient", specifying the implication operator used in the calculation. Table 9 is structured as table 8, but here the predicate "There exist probable and convenient scenarios" is computed.

The membership functions appeared in table 4 and figure 6 were used to calculate the values of table 8. On the other hand, the elements of table 9 were calculated with the aid of the results in table 5 and figure 7.

5 Concluding Remarks

This chapter makes a fuzzy approach to Prospect theory by using the compensatory fuzzy logic. A one-parameter family of Compensatory Fuzzy Logic and five implication operators selected are used to obtain the maximization of four objective functions with the genetic algorithm coded in MATLAB. This approach is a revisit to a 1992 experiment of Kahneman and Tversky.

The family of CFL depends on a parameter p, equal to or less than 0 and they are based on the formula of the quasi-arithmetic mean. On the other hand, Reichenbach implication, Yager implication, Klir-Yuan implication, Natural implication and Zadeh implication are selected because they generalize the truth table of the bivalent logic, when they are restricted to values 0 or 1. Also, they are continuous or they have at most a finite number of removable discontinuities.

According to the empirical results, the Reichenbach implication and the Geometric Mean are closest to the people's way of thinking. The sigmoidal membership functions of some predicates, like "the scenario is convenient" or "the scenario is probable" are found to be included in the composed predicates like "If the scenario is probable then it is convenient", "All the scenarios are probable and convenient" or "There exist probable and convenient scenarios".

1 or 1.5 are the values of indifference for the premiums, according to Reichenbach implication results. The membership function for probabilities changes for each predicate.

The sigmoidal functions were used for modelling every predicate, including those related with probabilities, even though its shape differs from the function illustrated in figure 2.

References

1. Buchanan, B.G., Shortliffe, E.H.: Knowledge Engineering. Rule-Based Expert Systems –The MYCIN Experiments of the Stanford Heuristic Programming Project, pp. 147–158. Addison-Wesley, Massachusetts (1984)
2. Detyniecki, M.: Mathematical aggregation operators and their application to video querying. Paris, University of Paris VI. Ph.D. Thesis (2001)
3. Dubois, D., Prade, H.: Criteria Aggregation and Ranking of alternatives in the framework of Fuzzy Sets Theory. In: Zimmermann, H.J., Gaines, B.R., Zadeh, L.A. (eds.) Fuzzy Set and Decision Analysis, pp. 209–240. North Holland, Amsterdam (1984)

4. Dubois, D. et al.: Fuzzy-Set Based Logic- An History -Oriented Presentation of their Main Developments Handbook of the History of Logic. In: Gabbay, D.M., Woods, J. (eds.), pp. 325–449. North-Holland, Elsevier BV (2007)
5. Espin, R., et al.: Un sistema lógico para el razonamiento y la toma de decisiones: la Lógica Difusa Compensatoria Basada en la Media Geométrica (A logic system for reasoning and decision making: Compensatory Fuzzy Logic on Geometric Mean). Revista Investigación Operacional 32, 230–245 (2011) (in Spanish)
6. French, S.: Decision Theory: An Introduction to the Mathematics of Rationality. Halsted Press, New York (1986)
7. Jayaram, B.: On the Law of Importation in Fuzzy Logic. IEEE Transactions on Fuzzy Systems 16, 130–144 (2008)
8. Kahneman, D., Tversky, A.: Prospect Theory: An Analysis of Decision under Risk. Econometrica 47, 263–291 (1979)
9. Mitrinovic, D.S., et al.: Classical and New Inequalities in Analysis (1993)
10. Dordrecht/Boston/London. Kluwer Academic Publishers
11. Mizumoto, M.: Pictorial Representations of Fuzzy Connectives: Part II. Cases of Compensatory Operators and Self-Dual Operators. Fuzzy Sets and Systems 32, 45–79 (1989)
12. Trillas, E., et al.: When QM-operators are implication functions and conditional fuzzy relations. International Journal of Intelligent Systems 15, 647–655 (2000)
13. Turksen, I.B., et al.: A new class of Fuzzy Implications: Axioms of Fuzzy Implication revisited. Fuzzy Sets and Systems 100, 267–272 (1998)
14. Tversky, A., Kahneman, D.: Advances in Prospect Theory: Cumulative Representation of Uncertainty. Journal of Risk and Uncertainty 5, 297–323 (1992)
15. Zadeh, L.: Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Computing 2, 23–25 (1998)
16. Zadeh, L.: From computing with numbers to computing with words –from manipulation of measurements to manipulation of perceptions. International Journal of Applied Mathematics and Computational Sciences 3, 307–324 (2002)

Probabilistic Approaches to the Rough Set Theory and Their Applications in Decision-Making

Rafael Bello Pérez and Maria M. Garcia

Abstract. Rough sets were presented by Professor Zdzislaw Pawlak in a seminal paper published in 1982. Rough Sets Theory (RST) has evolved into a methodology for dealing with different types of problems, such as the uncertainty produced by inconsistencies in data. RST is the best tool for modeling uncertainty when it shows up as inconsistency, according to several analyses. This is the main reason for which the RST has been included in the family of Soft Computing techniques. The classical RST is defined by using an equivalence relation as an indiscernibility relation. This is very restrictive in different domains, so several extensions of the theory have been formulated. One of these alternatives is based on a probabilistic approach, where several variants have been proposed such as the Variable Precision Rough Sets model, Rough Bayesian model, and Parameterized Rough Set model. Here is presented an analysis about the evolution of the RST in order to enrich the applicability to solve real problems by means of the probabilistic approaches of rough sets and its application to knowledge discovering and decision making, two main activities in Business Intelligence.

1 Introduction

The aim of the information systems is to model the real world, but the uncertainty pervades our understanding of the real world; for this reason, it is necessary to consider and properly handle the uncertainty to implement computational systems and solve real problems [22]. The applications in Business Informatics are not the exceptions.

Logical experts and philosophers have considered the problems of vagueness and uncertainty for many years. More recently, computer scientists, particularly

Rafael Bello Pérez · Maria M. Garcia
Central University of Las Villas, Cuba
e-mail: {rbellop,mmgarcia}@uclv.edu.cu

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_4,

researchers related with Artificial Intelligence have worked out novel approaches on this research field, being the Fuzzy Sets Theory [40] one of the most outstanding and representative approaches.

Bonnisone and Tong [6] classify knowledge faultiness in: *uncertainty*, *imprecision* and *incomplete information*. This latter term is used to indicate the lack of a value, whereas "*imprecision*" denotes the existence of a value which cannot be measured with the suitable accuracy. Finally, the term "*uncertainty*" stands for the fact that an agent has formulated a subjective opinion on a veracity of a fact which is not known for sure. Bosc and Prade [8] coin two new terms: *vagueness* and *inconsistency*. The vagueness is a new category modeled by means of fuzzy sets but that essentially falls under the classification of imprecision, just like the aforementioned concept. The "inconsistency" describes a situation in which there are two or more conflicting values to be assigned to a variable [6] and [22].

This is the background where in Soft Computing is developed as an emergent approach of Computer Science whose goal is the remarkable ability of the human mind to reason and learn in an uncertain environment [7] and [13]. The essence of Soft Computing is to consider the pervading imprecision of the real world in the computational systems. For that reason, the ruling principle of Soft Computing is exploiting the tolerance, imprecision, uncertainty and partial truths in order to get flexibility, robustness, low solution costs and a better harmony with reality. Soft Computing is not a single methodology; on the contrary, it is an umbrella of approaches whose key members are fuzzy logic, neurocomputing and probabilistic reasoning, in addition to genetic algorithms, chaotic systems, belief networks and some elements of the learning theory (Lotfi A. Zadeh in [13]). The common denominator of these technologies is that they are not based in the classical reasoning and modeling approaches usually relying on Boolean logic, analytical models, hard classifications and determinist search. In a similar way to Zadeh, more recently Verdegay, Yager and Bonissone [29] provided their own definition of Soft Computing, in which the metaheuristics are emphasized.

It is quite common that a given combination of observations is associated with two or more different outcomes. According to [11], RST is the best tool for handling uncertainty when this is provoked by inconsistency. Li et al [16] argue that fuzzy logic, neural networks, and probabilistic reasoning were the initial components of Soft Computing, and that subsequently other components were added, such as the rough sets. In [2] and [3], the authors present an analysis of the relationship between rough sets and other components of Soft Computing.

The RST provides ways to directly model the uncertainty caused by inconsistencies in the information, and it also to takes into account the granularity of information. However, the classical approach based on a relation of inseparability is an equivalence relation; it is very strict to model real problems. In many real word applications, the assumption of exact data is not fulfilled and some objects are misclassified or condition attribute values are corrupted.

On the other hand, the definition of lower approximation is also very strict; it is enough for two objects in a universe of one million objects that they were inseparable and belong to different classes for the system to result inconsistent, and consequently it will affect the lower approximation of those classes. The strict definition of the approximations has consequences in all the techniques that are built from rough sets. One of the most important aspects is the selection of features based on the concept of reduct (a reduct is generally defined as a minimal subset of attributes that can classify the same domain of objects as unambiguously as the original set of attributes).

The RST can be more flexible by using a weaker indiscernibility relation or relaxing the lower and upper definitions, the two basic concepts of the theory. The Pawlak's Rough Set Model may be extended by using an arbitrary binary relation instead of equivalent relations, [24]; by considering any binary relations on attribute values, instead of the trivial equality relation (=); an object x is related to another object y, based on an attribute a, if their values on a are related, with respect to a subset A of attributes, x is related to y if their values are related for every attribute in A. When all relations R_a are chosen to be =, the proposed definition is reduced to the definition in the Pawlak's Rough Set Model.

In this chapter the second alternativeis discussed, presenting an analysis of ways to make flexible the RST using a probabilistic approach. The main objective in this chapter is to provide an analysis and review of probabilistic approaches to rough sets. Several probabilistic extensions of the rough set model have been proposed to make the approach more applicable to real life data analysis problems, in which the information may be incomplete, with noises or uncertainties; important reviews about this subject are presented in [33] and [44]. This alternative approaches to the classical rough set theory that can be achieved by decreasing the classification precision in the knowledge obtained through rough set's analysis. Also, there is included the use of them to achieve decision-making troubles, which result to be of great interest in Business Informatics, especially in Business Intelligence.

Business Intelligence (BI) mainly refers to computer-based techniques used in identifying, extracting and analyzing business data. BI is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. Business Intelligence -understood broadly- can include the subset of competitive intelligence. Using data that has been stored, software applications are able to use this data to report past business information as well as predict future business information, including trends, threats, opportunities and patterns; to do this, BI includes techniques for different activities, among them decision making.

2 Rough Set Theory

The whole knowledge about the domain is contained in the set of objects, called Information System. A *decision system*$(U, A \cup \{d\}$, *where* $d \notin A)$ is obtained

when a new attribute *d*, called decision attribute, is added for each object in *U*. A simple idea of rough sets is the following: objects having exactly the same values of condition attributes are indiscernible by using these attributes. This indiscernibility relation is the mathematical basis of RST. Such relation induces a partition of the universe U in equivalence classes; that is, a set of indiscernible objects according to the relation.

Any subset $X \subseteq U$ can be expressed exactly or approximately in terms of these sets by using two crisp sets called lower approximation and upper approximation. The lower approximation ($B_*(X)$) of a set of objects (concerning those attributes) is a collection of objects whose equivalence classes are fully contained into the set of objects we want to approximate; while the upper approximation ($B^*(X)$) of the same set of objects is a collection of objects whose equivalence classes are at least partially contained into the set of objects it is wanted to be approximated.

$$B_*(X) = \{x \in U \mid B(x) \subseteq X\} \tag{1}$$

$$B^*(X) = \{x \in U \mid B(x) \cap X \neq \emptyset\} \tag{2}$$

$$BN_B(X) = B^*(X) - B_*(X) \tag{3}$$

If the boundary region (BN_B) is empty ($BN_B(X) = \emptyset$ then X is *crisp* according to B; otherwise X is said to be *rough*. Objects members of the boundary region have a membership status that cannot be classified with certainty as members of the underlying concept. Using this approximations the positive, negative, and boundary regions of X can be defined: the positive region, $POS(X) = B_*(X)$, consists of all objects that are definitely contained in the set X, the negative region, $NEG(X) = U - B^*(X)$, consists of all objects that are definitely not contained in the set X, and the boundary region, defined by (3), consists of all objects that may be contained in X.

RST provides several measures to characterize a given set. These measures are very useful in order to evaluate the quality of the results computed via rough-set-based methods, for instance, the strength of the decision-making processes and the certainty of the discovered knowledge. Pawlak defines a measure to evaluate the quality of classification [21], this measure (γ_B) is used to calculate the degree of consistency of a decision system: If $\gamma_B(Y) = 1$, the decision system is consistent; otherwise it is inconsistent [28].

In this classical rough set model, the lower and upper approximations are defined based on the two extreme cases (full inclusion or non-empty overlap) regarding the relationships between an equivalence class and a target set, this requirement limits unnecessarily the applications of rough sets in practical problems; in the next sections other approaches are analyzed, these are based on considering the degree of set overlap in the rough set formulation.

3 Rough Sets Based on Probabilistic Approaches

Based on the notion of rough membership functions, different approaches for the construction of a probabilistic rough set model have been developed. Yao et al. [38] put into groups the existing rough set models into two major classes, the algebraic and probabilistic rough set models, depending on whether statistical information is used; some of them are analyzed in this section. While non-probabilistic studies of rough sets focus on algebraic and qualitative properties of the theory, probabilistic approaches are more practical and capture quantitative properties of the theory [32]. Algebraic rough set approximations may be considered as qualitative approximations of a set, in this case the extent of overlap between a set and an equivalence class is not considered; by incorporating the overlap, probabilistic rough set approximations have been introduced. Probabilistic rough set approximations can be formulated based on the notions of rough membership functions and rough inclusion. The probabilistic approaches expand the positive and negative regions (defined from the lower and upper approximations) by providing probabilities that define boundary regions; since the boundary region introduces uncertainty into the discernibility of objects, the major challenge in data analysis by using rough sets is to minimize the size of this region, this is done by relaxing the definitions of the *POS* and *NEG* regions to include objects that would otherwise not have been previously included.

An attempt to use probabilistic information for approximations was suggested by Pawlak et al. [20]. Their model is based essentially on the majority rule. Yao and Wong [37] introduced a more general probabilistic approximation in the decision-theoretic rough set model (DTRSM).

$$B_*(X) - \alpha = \{x \in U \mid P(A|[x]) \geq \alpha\} \tag{4a}$$

$$B^*(X) - \beta = \{x \in U \mid P(A|[x]) > \beta\} \tag{4b}$$

Where $0 \leq \beta < \alpha \leq 1$. If α =1 and β =0, the classical lower and upper approximations are obtained. Based on Bayesian decision procedure, DTRSM provides systematic methods for deriving the required thresholds on probabilities for defining the three regions: positive region, boundary region and negative region. A review on decision-theoretic rough sets is presented in [9].

Liu et. al. [9] introduce three-way decision-theoretic rough sets and answer "why" and "how" to use DTRSM to solve practical problems. It divides the universe into three regions, which lead the generalized three-way decision rules. The probabilistic positive rules express that an object or object sets belong to one decision class when the threshold is more than α, which enable us to make decisions of acceptance; the probabilistic boundary rule express that an object or sets of objects belong to one decision class when the thresholds are between α and β, which lead to doubt about the decision; the probabilistic negative rules express that an object or sets of objects do not belong to one decision class when the threshold is less than β, which enable to make decisions of rejection. A great chal-

lenge for the probabilistic rough set models is the acquirement of a pair of thresholds. Unfortunately, the thresholds are usually given by expert's experience in most of the probabilistic rough sets.

3.1 Rough Sets Based on Rough Membership Function

By definition, elements in the same equivalent class have the same degree of membership. The rough membership may be interpreted as the probability of x belonging to X given that x belongs to an equivalence class, this interpretation leads to probabilistic rough sets; the probabilistic rough set models may be interpreted based on rough membership functions [38]. The rough membership function is defined by (5), this measure in the interval [0, 1].

$$\mu_X^B(x) = \frac{|X \cap B(x)|}{|B(x)|} \tag{5}$$

$B(x)$ denotes the equivalence class of object x according to the relation B. By definition, elements in the same equivalent class have the same degree of membership. This value may be interpreted analogously to conditional probability (as a frequency-based judgment of conditional probability). This interpretation leads to probabilistic rough sets [30] and [20]. Using the rough membership function, the lower and upper approximations are defined by (6) and (7).

$$B_*(X) = \{x \in U \mid \mu_X^B(x) = 1\} \tag{6}$$

$$B^*(X) = \{x \in U \mid \mu_X^B(x) > 0\} \tag{7}$$

The former definitions of lower and upper approximations can be made more general by using an arbitrary precision threshold "τ", expression (8) and (9):

$$B_*\tau(X) = \{x \in U \mid \mu_X^B(x) \geq \tau\} \tag{8}$$

$$B^*\tau(X) = \{x \in U \mid \mu_X^B(x) > 1 - \tau\} \tag{9}$$

3.2 Variable Precision Rough Sets Model

The Variable Precision Rough Sets (VPRS) model is a generalized version of the conventional rough set approach which inherits all of its fundamental mathematical properties and aims at handling vague information. This model was introduced in [43]. The VPRS model defines the positive region as an area where, on the basis of available data, the rough membership of objects to the given set is certain to some degree.

The VPRS model allows for a controlled degree of misclassification. Any partially incorrect classification rule provides valuable trend information about future test cases if most of the available data which are applied to such a rule can

be correctly classified. The target of this model is to loose the former definition of lower and upper approximations introduced in the classical rough set methodology.

This model deals with this type of information by introducing a precision parameterβ, the value of this parameter represents a bound on the conditional probability of a proportion of objects in a condition class, which are classified to the same decision class [27]. Ziarko in [43] considers the parameter β as an admissible level of classification error defined in the domain $\beta \in [0,0.5)$. Other alternative presented in [1] and [5] considered the parameter β to denote the proportion of correct classifications, in which case the appropriate range is $(0.5,1]$.

The concepts of β-lower approximation and β-upper approximations are defined as follows, where $D_0(Y/X)$ is an inclusion degree is defined by (12) and $\beta \in (0.5,1]$,

$$B_*^{\beta}(X) = \{x \in U : D_0(X/[x]_B) \geq \beta\} = \cup \{[x]_B : D_0(X/[x]_B) \geq \beta\} \tag{10}$$

$$\begin{aligned} B^{*\beta}(X) &= \{x \in U : D_0(X/[x]_B) > 1-\beta\} \\ &= \cup \{[x]_B : D_0(X/[x]_B) > 1-\beta\} \end{aligned} \tag{11}$$

Let U be a finite set, $F = \{X : X \subseteq U\}$ and $\subseteq$ a partial order relation on F. For all $X, Y \in F$, D_0 is computed as follows:

$$D_0(Y/X) = \begin{cases} \dfrac{|Y \cap X|}{|X|} & if\ X \neq \emptyset \\ 1 & \text{otherwise} \end{cases} \tag{12}$$

This means that an elementary class belongs to the lower approximation of X if and only if a 100% * β of its elements belong to X; in a similar way, an elementary class is excluded from the upper approximation of X if and only if a 100% * β of its elements does not belong to X. The grade of looseness allowed in our model is fixed in advance by properly setting the value of the parameter β. Due to the existence of β, the VPRS can resist data noise or remove data errors.

According to [27], the VPRS model lacks a feasible method to determine the value of the parameter β. Ziarko [43] proposed the β value to be specified by the user, Beynon proposed the allowable β value range to be an interval [5], and for the case of reduct calculation proposed two methods of selecting a β-reduct without determining a β value [4]. Other method to determine the precision parameter value in the context of reduct calculation is introduced in [27]. In a similar way, authors in [14] analyze the decision-theoretic rough set model from an optimization viewpoint.

3.3 Rough Bayesian Model

Sleezak [25] proposed an alternative parameterized rough set model, called Rough Bayesian model, in which the lower and upper approximations of X are defined as follows:

$$B_*(X, \varepsilon t) = \left\{ x \in U : \frac{|[x]_B \cap X|}{|X|} \geq \varepsilon t * \frac{|[x]_B \cap (U - X)|}{|(U - X)|} \right\} \tag{13}$$

$$B^*(X, \varepsilon q) = \left\{ x \in U : \frac{|[x]_B \cap X|}{|X|} \succ \varepsilon q * \frac{|[x]_B \cap (U - X)|}{|(U - X)|} \right\} \tag{14}$$

Where $\varepsilon t, \varepsilon q \in [1, +\infty]$, such that $\varepsilon t > \varepsilon q$.

3.4 Parameterized Rough Set Model

In [10], a generalization of the original definition of rough sets and variable precision rough sets is introduced, this generalization is based on the concept of absolute and relative rough membership, it is called Parameterized Rough Set model; according to the authors, the classical rough set model, the VPRS model, and the Rough Bayesian model are special cases of this.

The generalized VPRS model proposed in [10] assumes that, in order to include an object x in the positive region of set *X*, it is not sufficient to have a minimum percentage of objects from *X* in $[x]_R$, but it is also necessary that the percentage of objects from *X* in $[x]$ R is sufficiently greater than the percentage of objects from *X* outside $[x]_R$. In other words, it is necessary that both, the absolute and the relative memberships of *x* in *X* are not smaller than the given thresholds *t* and *a*, respectively.

This model is defined as follows: Let α and β, $\alpha \geq \beta$, be two real values in the range of variation of each relative rough membership $c(x, X)$ and $0 \leq q \leq t \leq 1$. The parameterized lower and upper approximations of *X* in *U* with respect to relative rough membership $c(x, X)$ are defined, respectively, by (15) and (16):

$$B_*(X, t, \alpha) = \left\{ x \in U : \frac{|[x]_B \cap X|}{|[x]_B|} \geq t \; and \; c(x, X) \geq \alpha \right\} \tag{15}$$

$$B^*(X, q, \beta) = \left\{ x \in U : \frac{|[x]_B \cap X|}{|[x]_B|} \succ q \; or \; c(x, X) \succ \beta \right\} \tag{16}$$

Where $c(x, X)$ is a relative rough membership measure; in [10] several expressions for $c(x, X)$ are proposed, for example (17):

$$c(x, X) = \frac{|[x]_B \cap X|}{|[x]_B|} - \frac{|X|}{|U|} \tag{17}$$

The Parameterized Rough Set model is the most general since it involves both, absolute and relative rough memberships; moreover, it can be generalized further by considering more than one relative rough membership.

3.5 Applications in Decision-Making

Decision-Making is a chosen strategy in order to achieve some purposes. Decision theory considers how is the best way to make decisions in the light of uncertainty about data. The basic approach to make decisions with a rough set model is to analyze a dataset in order to acquire lower and upper approximations. Immediate decisions (Unambiguous) can be formulated from the positive and negative regions, while Delayed decisions (Ambiguous) are based on the boundary region.

According to [38] the probabilistic rough set models are justified based on the framework of the decision theory; the results given in that work suggest that both algebraic rough set and probabilistic rough set models can be viewed as a special case of the decision theoretic framework. Several decision rules are derived using the probabilistic approach based on the membership function. The VPRSM has been used in many areas to support decision making [12]; for instance, a multi-attribute decision making method based on the concept of extended dominance relation and variable precision rough sets in this paper is proposed in [18], other example is presented in [15]. The use of a probabilistic approach for Decision-Making in Incomplete Information is analyzed in [39].

The concept of three-way decisions plays an important role in many real world Decision-Making problems; usually the Decision-Making is based on available information and evidence, when the evidence is insufficient or weak, it might be impossible to make either a positive or a negative decision. Yao [34], [35] and [36] propose to formulate decision rules according to three categories of decisions; this kinds of rules are derived from the three regions. As it was explained before, rules generated by the three regions form three-way decision rules: the positive rules generated by the positive region make decisions of acceptance; the negative rules generated by the negative region make decisions of rejection; and the boundary rules generated by the boundary region make deferred or non-committed decisions. Using this three-way decision approach, a solution to multi-category decision-making is proposed in [42]; other application is presented in [17], from the idea of three-way decisions of a new discriminate analysis approach by combining decision-theoretic rough sets and binary logistic regression is proponed. A multi-view decision method based on decision-theoretic rough set model is proposed in [41], in which optimistic decision, pessimistic decision, and indifferent decision are provided according to the cost of misclassification.

In real life, many important decision problems are not determined by a single decision-maker but by a group of them. In group Decision-Making, the members usually make judgments on the same decision problem independently. Due to the difference among them, there could be great disagreements on the same decision problem. Therefore, how to effectively integrate the evaluation of the decision-maker is an interesting problem. In [31], a study about how to use the variable precision rough set model as a tool to support group decision-making in credit risk management is presented. This technique is able to remove errors or inconsistency in a set of decision.

In group decision-making, the individual importance of the decision makers is introduced by using different weights for them; the analytical hierarchy process (AHP) technique is used to obtain members' weight, and aggregate group preference [23]. AHP is the multicriteria decision technique that can combine qualitative and quantitative factors for prioritizing, ranking and evaluating different decision alternatives. In [31], the VPRS and AHP are combined to obtain the weight of condition attribute sets decided by each decision maker, three VPRS-based models to obtain Integrated Risk Exposure (IRE) are discussed; the effectiveness of the methods is evaluated in an application in credit risk management, credit risk is represented by IRE.

One of the challenges a decision maker faces in using rough sets is to choose a suitable rough set model for data analysis; authors in [12] have observed that the availability of information regarding the analysis data is crucial for selecting a suitable rough set approach, and they present a list of decision types corresponding to the available information and user's needs.

3.6 An Example of Three-Way Decisions

Suppose a bank has to decide on the orders it receives from companies whether to grant a loan or not. The purpose of the credits can be to make investments, to cover unexpected expenses, to address problems of lack of financial capacity, etc. In order to do this, a criterion is established for helping the bank's management to decide on the granting of the credit.

The past experience of the bank can be used to set the criterion, on which credits have been effective or not. The available information is shown in Table 1, in which the applicant companies are described by a set of attributes; in this analysis is only considered the following information: the company's sector, business productivity, the company's production market, the company's finances state. Furthermore, it is known if the credit granted had a positive effect or not.

Table 1 Previous cases met by the bank

Company	Sector	Productivity	Market	Finances	Effectiveness
E1	Agricultural	low	limited	low	no
E2	Industry	high	wide	low	yes
E3	Industry	average	wide	average	no
E4	Agricultural	average	average	high	yes
E5	Industry	average	wide	average	yes
E6	Services	high	average	high	yes

From the information contained in Table 1, there may be formulated rules such as "three-way decisions", as follows:

R1: $Des([x]) \rightarrow_P Des(C), for\ [x] \subseteq POS(C)$, with all certainty the decision is C.

R2: $Des([x]) \rightarrow_B Des(C), for\ [x] \subseteq BND(C)$, uncertainty over the decision C.

R3: $Des([x]) \rightarrow_N Des(C), for\ [x] \subseteq NEG(C)$, with allcertainty the decision is NOT C.

After applying the definitions of RST the positive region of the decision Effectiveness class = "if" is:

POS(Effectiveness=”if”)= {E2, E4, E6}

BND (Effectiveness=”if”)= {E3, E5}

NEG (Effectiveness =”if”)= {E1}

Rule1 means that if the description of a company applying for a loan is inseparable from companies E2, E4, or E6, then it should be given the credit for sure. By rule 3, if it is inseparable of E1, with all certainty it must not be given the credit. According to the second rule, if it is inseparable from E3 or E5, there will be a doubt about granting the credit or not.

Obviously in a real situation where the available information is hundreds or thousands of cases rather than the 6 in Table 1, there can be inconsistencies on the information (such as between the cases of E3 and E5 in Table 1) but that because of the given amount of information available is not significant. In this case it would be necessary to use probabilistic approaches, and therefore the rules to use would be:

R1: $Des([x]) \rightarrow_P Des(C), for\ [x] \subseteq POS_{\alpha,\beta}(C)$

R2: $Des([x]) \rightarrow_B Des(C), for\ [x] \subseteq BND_{\alpha,\beta}(C)$

R3: $Des([x]) \rightarrow_N Des(C), for\ [x] \subseteq NEG_{\alpha,\beta}(C)$

Where

$POS_{\alpha,\beta}(C) = \{x \in U : Pr(C|[x]) \geq \alpha\}$

$BND_{\alpha,\beta}(C) = \{x \in U : \beta < Pr(C|[x]) < \alpha\}$

$NEG_{\alpha,\beta}(C) = \{x \in U : Pr(C|[x]) \leq \beta\}$

And

$Pr(C|[x]) = D0(C|[x]) = |C \cap [X]|\ /\ |[X]|$, from expression (12).

References

1. An, A., et al.: Discovering rules for water demand prediction: an enhanced rough-set approach. Engineering Applications in Artificial Intelligence 9(6), 645–653 (1996)
2. Bello Pérez, R., Verdegay Galdeano, J.L.: On Hybridizations in Soft Computing: Rough Sets and Metaheuristics. In: Proceeding of World Conference on SoftComputing, San Francisco, USA, May 23-26 (2011)
3. Bello Pérez, R., Verdegay Galdeano, J.L.: Rough sets in the Soft Computing environment. Information Sciences 212, 1–14 (2012)
4. Beynon, M.J.: An investigation of β-reduct selection within the variable precision rough sets model. In: Ziarko, W.P., Yao, Y. (eds.) RSCTC 2000. LNCS (LNAI), vol. 2005, pp. 114–122. Springer, Heidelberg (2001)
5. Beynon, M.: Reducts within the variable precision rough sets models: a further investigation. European Journal of Operational Research 134, 592–605 (2001)
6. Bonnissone, P.P., Tong, R.M.: Editorial: Reasoning with uncertainty in expert systems. Int. J. Man Machine Studies 22, 241–250 (1985)
7. Bonissone, P.P.: Soft Computing: the Convergence of Emerging Reasoning Technologies. Journal of Soft Computing 1(1), 6–18 (1997)
8. Bosc, P., Prade, H.: An introduction to fuzzy set and possibility theory based approaches to the treatment of uncertainty and imprecision in database management systems. In: Proc. of Second Workshop Uncertainty management in Information Systems: from Needs to Solutions, California (1993)
9. Dun, L., Huaxiong, L., Xianzhong, Z.: Two Decades' Research on Decision-theoretic Rough Sets. In: Proceedings of the 9th IEEE International Conference on Cognitive Informatics, ICCI 2010, article number 5599770, pp. 968–973 (2010)
10. Greco, S., Matarazzo, B., Slowinski, R.: Parameterized rough set model using rough membership and Bayesian confirmation measures. International Journal of Approximate Reasoning 49, 285–300 (2008)
11. Grzymala-Busse, J.W.: Managing uncertainty in machine learning from examples. In: Proceedings of the Workshop Intelligent Information Systems III, Poland, June 6-10, pp. 6–10 (1994)
12. Herbert, J., Yao, J.: Criteria for choosing a rough set model. Computers and Mathematics with Applications 57, 908–918 (2009)
13. Jang, J.R., et al.: Neuro-Fuzzy and Soft Computing: a computational approach to learning and machine intelligence. Prentice Hall (1997)
14. Jia, X., Li, W., Shang, L., Chen, J.: An Optimization Viewpoint of Decision-Theoretic Rough Set Model. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 457–465. Springer, Heidelberg (2011)
15. Jun-hua, H., Xiao-hong, C.: Multi-criteria decision making method based on dominance relation and variable precision rough set. Systems Engineering and Electronics 32(4), 759–763 (2010)
16. Li, R., Zhao, Y., Zhang, F., Song, L.: Rough Sets in Hybrid Soft Computing Systems. In: Alhajj, R., Gao, H., Li, X., Li, J., Zaïane, O.R. (eds.) ADMA 2007. LNCS (LNAI), vol. 4632, pp. 35–44. Springer, Heidelberg (2007)
17. Liu, D., Li, T., Liang, D.: A New Discriminant Analysis Approach under Decision-Theoretic Rough Sets. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 476–485. Springer, Heidelberg (2011)

18. Ming-li, H., Fei-fei, S., Ya-huan, C.: A multi-attribute decision analysis method based on rough sets dealing with uncertain information. In: Proceedings of 2011 IEEE International Conference on Grey Systems and Intelligent Services, GSIS 2011, art. no. 6043984, pp. 576–581 (2011)
19. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
20. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. International Journal of Man-Machine Studies 29, 81–95 (1988)
21. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Boston (1991)
22. Parsons, S.: Current approaches to handling imperfect information in data and knowledge bases. IEEE Trans. on Knowledge and Data Engineering 8(3) (June 1996)
23. Ramanathan, G.: Group preference aggregation methods employed in AHP: an evaluation and an intrinsic process for deriving members' weightages. European Journal on Operation Research 79, 249–265 (1994)
24. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae 27(2-3), 245–253 (1996)
25. Ślęzak, D.: Rough sets and Bayes factor. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets III. LNCS, vol. 3400, pp. 202–229. Springer, Heidelberg (2005)
26. Slowinski, R., Vanderpooten, D.: Similarity relation as a basis for rough approximations. Advances in Machine Intelligence & Soft-Computing IV, 17–33
27. Su, C.T., Hsu, J.H.: Precision parameter in the variable precision rough set model: an application. Omega 34, 149–157 (2006)
28. Tay, F.E., Shen, L.: Economic and financial prediction using rough set model. European Journal of Operational Research 141, 641–659 (2002)
29. Verdegay, J.L., Yager, R.R., Bonissone, P.P.: On heuristics as a fundamental constituent of soft computing. Fuzzy Sets and Systems 159, 846–855 (2008)
30. Wong, S.K.M., Ziarko, W.: Comparison of the probabilistic approximate classification and the fuzzy set model. Fuzzy Sets and Systems 21, 357–362 (1987)
31. Xie, G., Zhang, J., Lai, K.K., Yu, L.: Variable precision rough set for group decision-making: An application. International Journal of Approximate Reasoning 49, 331–343 (2008)
32. Yao, Y.Y.: Probabilistic Approaches to Rough Sets. Expert Systems 20(5), 287–297 (2003)
33. Yao, Y.: Probabilistic rough set approximations. International Journal of Approximate Reasoning 49, 255–271 (2008)
34. Yao, Y.: Three-way decision: an interpretation of rules in rough set theory. In: Wen, P., Li, Y., Polkowski, L., Yao, Y., Tsumoto, S., Wang, G. (eds.) RSKT 2009. LNCS (LNAI), vol. 5589, pp. 642–649. Springer, Heidelberg (2009)
35. Yao, Y.Y.: Three-way decisions with probabilistic rough sets. Information Sciences 180, 341–353 (2010)
36. Yao, Y.Y.: The superiority of three-way decision in probabilistic rough set models. Information Sciences 181, 1080–1096 (2011)
37. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximating concepts. International Journal of Man–Machine Studies 37, 793–809 (1992)
38. Yao, Y.Y., Wong, S.K.M., Lin, T.Y.: A review of rough set models. In: Lin, T.Y., Cercone, N. (eds.) Rough Sets and Data Mining: Analysis for Imprecise Data, pp. 47–75. Kluwer Academic Publishers, Boston (1997)

39. Yang, X., Song, H., Li, T.-J.: Decision Making in Incomplete Information System Based on Decision-Theoretic Rough Sets. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 495–503. Springer, Heidelberg (2011)
40. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
41. Zhou, X., Li, H.: A Multi-View Decision Model Based on Decision-Theoretic Rough Set. In: Wen, P., Li, Y., Polkowski, L., Yao, Y., Tsumoto, S., Wang, G. (eds.) RSKT 2009. LNCS, vol. 5589, pp. 650–657. Springer, Heidelberg (2009)
42. Zhou, B.: A New Formulation of Multi-category Decision-Theoretic Rough Sets. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 514–522. Springer, Heidelberg (2011)
43. Ziarko, W.: Variable precision Rough sets model. Journal of Computer and System Science 46(1), 39–59 (1993)
44. Ziarko, W.: Probabilistic approach to rough sets. International Journal of Approximate Reasoning 49, 272–284 (2008)

New Rough Sets by Influence Zones Morphological Concept

Juan I. Pastore, Agustina Bouchet, and Virginia L. Ballarin

Abstract. Rough Sets and Mathematical Morphology theories are both defined through dual operators sharing similar properties. This allows to establish equivalences between the basic morphological operators and rough sets. The concept of Influence Zones has been widely studied and used successfully in applications that are solved by Mathematical Morphology techniques. In this work we define the Rough Sets by Influence Zones based in morphological concept. To the best of our knowledge, this approach has not been explored until now.

1 Introduction

Rough Sets (RS) theory is an important tool to solve the problem of obtaining useful knowledge that is not evident in information stored in a database. This information may be affected by imprecision, uncertainty and even it may be incomplete. RS theory is based on the "discernibility and approach" concepts. To discern means "to distinguish one thing from another, through the senses or human reason," the idea is to find all objects that produce different types of information. When it says "approach", it refers to the existence of vagueness, ie inaccurate information of a set of objects. From these concepts all the mathematical structure of RS is built. The main idea of this methodology is the approximation of a set with two others, called lower and upper approximation.

The Mathematical Morphology (MM) is a theory for image processing based on concepts of geometry, algebra, topology and sets theory, created originally with

Juan I. Pastore · Agustina Bouchet · Virginia L. Ballarin
Digital Image Processing Group, School of Engineering,
National University of Mar del Plata. J.B.Justo 4302, Mar del Plata, Argentina
e-mail: {juan.pastore,agustina.bouchet,vballari}@gmail.com

Juan I. Pastore · Agustina Bouchet
Consejo Nacional de Investigaciones Científicas y Técnicas,
CONICET, Argentina

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_5, © Springer-Verlag Berlin Heidelberg 2014

the goal to characterize physical and structural properties of different materials. Currently the MM has become a solid mathematical theory providing powerful tools for Digital Image Processing (DIP). The central idea of this theory is to analyze the geometric structures in an image by probing it with small patterns, called structuring elements (SE). The MM allows enhancing areas, edge detection, analyzing structure and segment regions, among others. One advantage of MM is its simplicity of implementation, and its pillars are the two basic operations, erosion and dilation. From these two operators it is possible to construct new operators by composition, which differentiates it from other image processing techniques. The shape and size of the SE are chosen depending on the type of analysis to be performed and the shape of the objects that make up the images. The SE is moved over the image, so that it covers the whole image pixel by pixel, performing a comparison between the SE and the image. The result is a new binary image containing the result of the comparison.

Both theories, defined from the upper and lower approximation sets in the case of RS and dilation and erosion operations in the case of MM, share similar properties. The definitions of upper approximation set and dilation are mathematically equivalent, while the definitions of lower approximation set and erosion are also equivalent.

The upper approximation set is the union of the elementary sets having nonempty intersection with the concept. That is, these sets that contain all the objects based on knowledge of attributes that can not be classified as not belonging to the concept. Dilation of a binary image by a SE consists on all the pixels for which the translation of the SE intercepts the image. The application of the dilation adds all the bottom pixels that touch the edge of an object, ie fill contrasts that do not fit the SE, like small holes and bays.

The lower approximation of a set is the union of the elementary sets contained in the concept. That is, contains all objects, based on knowledge of attributes that can be classified with certainty that they belongs to the concept. The erosion of a binary image by a SE produces a thinning of the image taking as reference the SE. The application of erosion removes groups of pixels where the SE does not "fit", ie, it removes groups of pixels smaller than the SE, like small islands and protrusions.

This work aims to analyze deeply the relationship between the two theories and to explore a novel characterizations of RS which we will call Rougt Sets by Influence Zones (*RS-IZ*) based in morphological concepts.

2 Basic Concepts of Rough Sets

RS theory was proposed by Pawlak [13-15] as a new mathematical model for knowledge representation and treatment of uncertainty. This theory has evolved into a methodology to address a wide variety of issues, including the uncertainty in the information. The philosophy of this new theory is based on the assumption that each object in the universe can be associated with some information [12].

A major advantage of RS theory applied to data analysis is that it requires no preliminary or additional information, but is based solely on the internal structure of the original data to model knowledge. Therefore, it is not necessary any assumptions about the data, and it used to analyze both qualitative and quantitative features. This theory is useful when the classes are imprecise, but nevertheless can be approximated by precise sets [10]. It has been used successfully in various applications, such as analysis of the information, data analysis and data mining, and knowledge discovery, ie those applications where the need arises for intelligent decision support.

The indiscernible relationship -objects characterized by the same information are not discernible- is the key concept of the RS theory. Inaccurate information is the cause of not discernibility of objects in terms of available data, therefore, their allocation needs a set. Rough could be understood as "vague, imprecise," therefore, a RS is a set of objects that, in general, can not be accurately characterized in terms of available information. This theory assumes that a RS can be replaced or represented by a pair of precise sets called lower and upper approximation. The lower approximation contains of all objects which surely belong to the set and upper approximation contains objects that possibly belong to the set. The boundary (or region of doubt) is the set of elements that can not be with certainty classified using the set of attributes.

The following section presents some basic notions related to information systems and RS.

3 Information Systems

Information systems (IS) are the basic information structure in the theory of RS [11]. An IS comprises a quadruple $S = (U, A, V, f)$ where U is a nonempty finite set called the universe; $A = \{a_1, a_2, \dots, a_n\}$ is a nonempty finite set of attributes describing the objects; $V = \bigcup_{i=1}^{n} V_i$ where V_i represents the domain associated with each attribute a_i; and $f: U \times A \to V$ such that $f(x, a_i) \in V_i$ for each $a_i \in A$ and $x \in U$, is the total decision function called the *information function* [9]. Each information system $S = (U, A, V, f)$ has associated a multidimensional binary array $M \in \{0,1\}^{c(U)\times c(A)\times \lambda}$ where each position determines the value of the category that identifies each atributte, being $c(\cdot)$the cardinality of a set and λ the maximum number of categories between the attributes.

Table 1 shows an example of IS [17], the universe consists of six objects, which are described by the set of attributes A given by:

$$A = \{(a_1)\ headache, (a_2)\ muscle\ pain, (a_3)\ temperature, (a_4)\ flu\}$$

Figure 1 shows the binary multidimensional array $M \in \{0,1\}^{6\times 3\times 3}$ associated to the former IS. The first column of M represents the headache values, where one (1) represents the presence of headache and zero (0) the absence. The second column of M represents the muscle pain values, where one (1) indicates the presence

of muscle pain and zero (0) represents the absense. Finally, the third dimension of M indicates the temperature defined as a category as follows:

$$M(i,j,1) = \begin{cases} 0 & \text{not normal temperature} \\ 1 & \text{normal temperature} \end{cases}$$
$$M(i,j,2) = \begin{cases} 0 & \text{not high temperature} \\ 1 & \text{high temperature} \end{cases}$$
$$M(i,j,3) = \begin{cases} 0 & \text{not very high temperature} \\ 1 & \text{very high temperature} \end{cases}$$

Table 1 Information System

Patient	Headache	Muscle pain	Temperature	Flu
P_1	No	Yes	High	Yes
P_2	Yes	No	High	Yes
P_3	Yes	Yes	Very high	Yes
P_4	No	Yes	Normal	No
P_5	Yes	No	High	No
P_6	No	Yes	Very high	Yes

$$M(:,:,1) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad M(:,:,2) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad M(:,:,3) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Fig. 1 Binary multidimensional array associated to the example given in Table 1

3.1 Indiscernible Relation

In an information system, for each subset of attributes, a binary relation called *indiscernible relationship* can defined.

Being $S = (U, A, V, f)$ an IS, it is said that the objects x and y are *inseparable* for a subset of attributes $B \subseteq A$, if and only if $f(x, a_i) = f(y, a_i)$ for $\forall a_i \in B$ [11]. This relationship is denoted by $\langle x, y \rangle_B$. Otherwise, are said to be separable or distinguishable with respect to the subset B.

The relationship of inseparability $\Re = \langle x, y \rangle_B$ defined above is an equivalence relation. An equivalence relation induces a partition of the universe U. The equivalence class of an item $x \in U$ is the set of all elements $y \in U$ such that $x \Re y$, i.e. the elements of U that are similar to x considering the attributes contained within B. By $B(x)$ we denote the set of all $y \in U$ such that $\langle x, y \rangle_B$. Various relations of separability will lead to different variants of approximate sets.

3.2 Decision System

Being $S = (U, A, V, f)$ an IS, if to each element of U is assigned a new attribute $d \notin A$ called *decision* indicating the decision in that state or situation, then a *decision system*$(U, A \cup \{d\}, V, f)$ is defined [11]. The decision attribute d induces a partition of the universe U. That is to say, $U = \bigcup_{i=1}^{m} X_i$, where $X_i = \{x \in U | d(x) = i\}$ are called *decision classes.* These decisions classes are discussed as RS and often are called *concepts*.

3.3 Rough Sets

Being $S = (U, A, V, f)$ an IS, $B \subseteq A$ a subset of attributes and $X \subset U$, the set X can be approximated using only the information contained in B by building the *lower and upper approximations* of X, named $B_*(X)$ and $B^*(X)$, respectively, defined by [3, 20]:

$$B_*(X) = \{x \in U \mid B(x) \subseteq X\} \tag{1}$$

$$B^*(X) = \{x \in U \mid B(x) \cap X \neq \emptyset\} \tag{2}$$

From equations (1) and (2) follows that the lower approximation of a set with respect to a set of attributes is defined as the collection of objects whose equivalence classes are fully contained in the set, while the upper approximation is defined as the collection of objects whose equivalence classes are at least partially contained in the set.

Being $S = (U, A, V, f)$ an IS, the sets defined above satisfy the following properties [1]:

Property 1: $B_*(X) \subseteq X \subseteq B^*(X) \subseteq U$
Property 2: $B_*(\emptyset) = B^*(\emptyset) = \emptyset$
Property 3: $B_*(\mathrm{U}) = B^*(\mathrm{U}) = \mathrm{U}$
Property 4: $B_*(\mathrm{X} \cup \mathrm{Y}) = B^*(\mathrm{X}) \cup B^*(\mathrm{Y})$
Property 5: $B_*(\mathrm{X} \cap \mathrm{Y}) = B_*(\mathrm{X}) \cap B_*(\mathrm{Y})$
Property 6: $X \subseteq Y \Rightarrow B_*(\mathrm{X}) \subseteq B_*(\mathrm{Y}) \wedge B^*(\mathrm{X}) \subseteq B^*(\mathrm{Y})$
Property 7: $B_*(X) = \mathrm{U} - B^*(U - X)$

A *Rough Set* is defined as the pair $(B_*(X), B^*(X))$. Therefore, a RS is a pair of lower and upper approximations of a set in terms of objects not discernible. In other words, a RS is a collection of objects which, in general, can not accurately be classified in terms of the values of the set of attributes, while the upper and lower approximations can. Consequently, each RS has border cases, i.e. objects that can not be classified with certainty as members of the set or its complement and, therefore, can be replaced or represented by a pair of precise sets, their approaches.

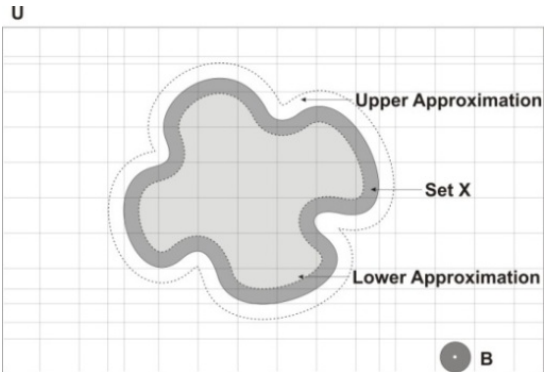


Fig. 2 Lower and upper aproximation

4 Basic Concepts of Mathematical Morphology

The word Morphology originates from two Greek words morphe (form) and logos (science). From a scientific perspective, this word refers to the study of forms and structures that matter can take. In digital image processing, Mathematical Morphology (MM) is the name of a specific methodology for analyzing the geometric structures in an image [18,19].

MM, whose early works are due to the German scientist Hermann Minkowski (1897 and 1901) and was subsequently studied by H. Hadwiger (1957 and 1959), is based on sets theory. The continuation of such research with some reformulations was conducted by two researchers at the Centre of Morphologie Mathematique (CMM) from l'Ecole des Mines de Paris in Fontainebleau, Georges Matheron and Jean Serra, who in the sixties worked on mineralogy and petrography problems. Its main objective was to characterize physical properties of certain materials, such as the permeability of porous media, examining its geometric structure. That was how the MM theory started, based on concepts of geometry, algebra, topology and set theory [16, 18]. The MM can process images for purposes of enhancement, filtering, restoration, segmentation, edge, detection, skeletonization, filling regions, thickening, structural analysis, etc.

In particular, the MM studies the geometrical structures of the components present in the images. Characteristics related to the geometry and topology of the components of the images are extracted by nonlinear operations. This theory allows to analyze the shape, size, orientation and overlapping of objects in digital images. The key of this techniques lies on the "structuring element" (SE), a set completely defined and of known geometry, that is traslated and compared with

the image pixel by pixel. The shape and size of SE allow to test and quantify how the item "is or is not contained" in the image [5].

One advantage of MM is its simplicity of implementation and its pillars are the two basic operations, erosion and dilation. From them, by composition, it is possible to construct new operators, a property that differentiates it from other image processing techniques [2].

Binary MM is defined from two basic operations called dilation and erosion. These operations compare the subsets within the binary image with a SE, which is a two dimensional set. The shape and size of the SE is chosen depending on the type of analysis to perform, and the shape of the objects within the images. The SE is moved, so that performs a pixel by pixel comparison within the full image. The result is a new binary image containing the result of the comparison. The morphological dilation and erosion operators are the basis of the MM, dilation is equivalent to Minkowski Sum while erosion is equivalent to Minkowski subtraction [7, 18].

Below is the definition of these operators.

Let A and B be two subsets of $\mathbb{R}^2$. *Binary dilation* of A by a SE B, denoted by $\delta_B(A)$, is the set of points $x \in \mathbb{R}^2$ such that the set $B_x = \{b + x | b \in B\}$ has non-empty intersection with A [6, 18]:

$$\delta_B(A) = \{x \in \mathbb{R}^2 \,|B_x \cap A \neq \emptyset\} \tag{3}$$

Dilate the image A by the SE B consists in the removal of all the points x for which the set B_x is not included, or equivalently to assign to the dilated image all points x such that B_x intercepts the image. The dilation adds all the points that touch the edge of an object, i.e., fill contrasts that do not fit the SE (eg, small holes and bays).

Binary erosion of A by a SE B, denoted by $\varepsilon_B(A)$, is the set of points $x \in \mathbb{R}^2$ such that the set $B_x = \{b + x | b \in B\}$ is contained in the set A [6, 18]:

$$\delta_B(A) = \{x \in \mathbb{R}^2 \,|B_x \cap A\} \tag{4}$$

Erode the image A by the SE B consists in decreasing the set A through a process of removing elements, taken as reference the SE B. The final size and shape of the eroded set depend heavily on the size and shape of the SE [4]. This is because the application of erosion removes groups of pixels where the SE does not "fit", i.e. removes groups of pixels, like small islands and protrusions, smaller than the size of the SE.

Figure 3 shows an example of the dilation and erosion of an image with a circular SE.

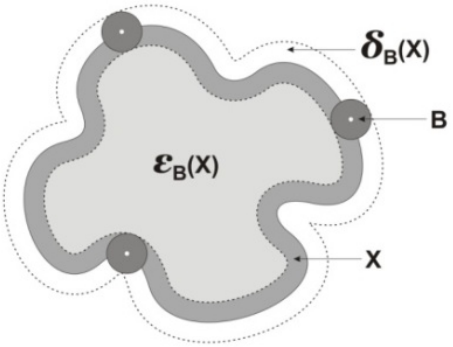


Fig. 3 Erosion and dilation

Noting the examples of erosion and dilation shown in Figure 3, it is clear that the erosion operator "shrinks" the set, while the dilatation as "expand".

The basic binary operators have the following properties [18]:

Property 1: Dilation and erosion are increasing operators, that is to say, preserve the inclusion relation between sets. Being $A, A', B \subseteq \mathbb{R}^2$:

$$A \subseteq A' \Rightarrow \delta_B(A) \subseteq \delta_B(A')$$
$$A \subseteq A' \Rightarrow \varepsilon_B(A) \subseteq \varepsilon_B(A')$$

Property 2: The expansion is increasing while erosion is decreasing for the SE B, i.e., being $A, B, B' \subseteq \mathbb{R}^2$:

$$B \subseteq B' \Rightarrow \delta_B(A) \subseteq \delta_{B'}(A)$$
$$B \subseteq B' \Rightarrow \varepsilon_{B'}(A) \subseteq \varepsilon_B(A)$$

Property 3: Erosion and dilation are not idempotent operators:

$$\delta_B(\delta_B(A)) \neq \delta_B(A) \quad \forall A \neq \emptyset, B \neq \{0\}$$
$$\varepsilon_B(\varepsilon_B(A)) \neq \varepsilon_B(A) \quad \forall A \neq \emptyset, B \neq \{0\}$$

Property 4: The dilatation is commutative, while erosion is not:

$$\delta_B(A) = \delta_B(B) \quad \forall A, B \subseteq \mathbb{R}^2$$

Usually, $\varepsilon_B(A) \neq \varepsilon_A(B)\ with\ A, B \subseteq \mathbb{R}^2, A \neq B$.

Property 5: The dilatation is associative, while erosion is not, ie, whether $A, B, C \subseteq \mathbb{R}^2$:

$$\delta_C(\delta_B(A)) = \delta_{\delta_C(B)}(A)$$

Usually, $\varepsilon_C(\varepsilon_B(A)) \neq \varepsilon_{\varepsilon_C(B)}(A)$.

Property 6: Dilatation and erosion are dual operators, ie $\forall A, B \subseteq \mathbb{R}^2$:

$$\delta_B(A^C) = [(\varepsilon_B(A)]^C\ \ with\ \varepsilon_B(A^C) = [(\delta_B(A)]^C$$

Property 7: When the SE has its center O, the dilation by B is extensive and erosion by B is no extensive, that is to say:

$$if\ O \in B \Rightarrow A \subseteq \delta_B(A) \quad \forall A \subseteq \mathbb{R}^2$$
$$if\ O \in B \Rightarrow \varepsilon_B(A) \subseteq A \quad \forall A \subseteq \mathbb{R}^2$$

Property 8: $\delta_{\delta_{B_2}(B_1)}(A) = \delta_{B_1}\left(\delta_{B_2}(A)\right) = \delta_{B_2}\left(\delta_{B_1}(A)\right) \ \forall A, B_1, B_2 \subseteq \mathbb{R}^2$

Property 9: $\varepsilon_{\delta_{B_2}(B_1)}(A) = \varepsilon_{B_1}\left(\varepsilon_{B_2}(A)\right) = \varepsilon_{B_2}\left(\varepsilon_{B_1}(A)\right) \ \forall A, B_1, B_2 \subseteq \mathbb{R}^2$

Property 10:

$$\delta_{B_1 \cup B_2}(A) = \delta_{B_1}(A) \cup \delta_{B_2}(A) = \delta_{B_2}(A) \cup \delta_{B_1}(A) \ \forall A, B_1, B_2 \subseteq \mathbb{R}^2$$

Property 11:

$$\varepsilon_{B_1 \cup B_2}(A) = \varepsilon_{B_1}(A) \cup \varepsilon_{B_2}(A) = \varepsilon_{B_2}(A) \cup \varepsilon_{B_1}(A) \ \forall A, B_1, B_2 \subseteq \mathbb{R}^2$$

Property 6 is particularly interesting in practice, since it says that erode the object is equivalent to dilate its background.

From property 7 it follows that if $O \in B : \ \varepsilon_B(A) \leq A \leq \delta_B(A)$.

Properties 8 and 9 are also of particular interest. Given a complex SE B, if you can find a decomposition as a union of simple elements, then you can perform a dilation (or erosion) by B as a combination of dilation (or erosion) for simple forms.

Dilation and erosion are not inverse operations. Applying a dilation or erosion to dilation of a given set by the same SE, in general, do not obtain the original set. Therefore, typically: $\delta_B(\varepsilon_B(A)) \neq \varepsilon_B(\delta_B(A))$.

The combination of erosion and dilation operators produces in new morphological operators [18].

4.1 Relationship between Both Theories

The lower and upper approximations may be obtained from the erosion and dilation [8]. For a given SE the equivalence relation is defined by:

$$x \Re y \ \Leftrightarrow y \in B_x \tag{5}$$

For the equivalence relation $\Re$, the class equivalence is obtained by $r(x) = \{y \in U | y \in B_x\} = B_x \ \ \forall x \in U$.

It is assumed that the origin of U belongs to the SE B. This means that: $\forall x \in U, x \in B_x$ and therefore $\forall x \in U, x \Re x$, then $\Re$ is reflexive. If B is symmetric $\left(B = \check{B}\right)$ then:

$$\forall (x, y) \in U^2, x \Re y \Leftrightarrow y \in B_x \Leftrightarrow y - x \in B \Leftrightarrow x - y \in \check{B} \Leftrightarrow x \in B_y \Leftrightarrow y \Re x$$

which proves $\Re$ is symmetric.

From this relation we see that the lower approximation and erosion coincide:

$$\forall X \subset U, B_*(X) = \{x \in U | r(x) \subset X\} = \{x \in U | B_x \subset X\} = \varepsilon_B(X)$$

Similarly, it can be observed that the upper and dilation approximation coincide:

$$\forall X \subset U, B^*(X) = \{x \in U | r(x) \cap X \neq \emptyset\} = \{x \in U | B_x \cap X \neq \emptyset\} = \delta_B(X)$$

5 Influence Zones

The concept of Influence Zones (IZ) has been widely studied and used successfully in applications that are resolved by MM techniques. This concept is defined from the basic operations, erosion and dilation. Equivalences discussed in the previous section, between the lower and upper RS and basic morphological operators is of great interest to define the equivalent in the theory of RS, which until now has not been explored.

Let $(\Im, d)$ be a metric space and let S a subset de $\Im$ such that $S = \bigcup_{i \in I} X_i$ where $\{X_i\}_{i \in I}$ is a disjoint family of connected components $\Im$. The *Influence Zones* (IZ) of X_i, denoted by $IZ(X_i)$, represents the set of all points that are closest to the component X_i than any other component $X_j, i \neq j$. Formally [19]:

$$IZ(X_i) = \{x \in X | d(x, X_i) < d(x, S - X_i)\} \tag{6}$$

From the definitions of morphological dilation and erosion, the $IZ(X_i)$ can be expressed equivalently [5]:

$$\begin{aligned} IZ(X_i) &= \bigcup_{\lambda=1}^{n} \left(\delta_{n-\lambda}(X_i) \cap \varepsilon_\lambda \left(\left(\bigcup_{j \neq i} X_i \right)^c \right) \right) \\ &= \bigcup_{\lambda=1}^{n} \left(\delta_\lambda(X_i) \backslash \delta_\lambda \left(\bigcup_{j \neq i} X_j \right) \right) \end{aligned} \tag{7}$$

where δ_λ represents the dilation of a ball of radius λ and ε_λ the erosion of a ball of radius λ.

Given an image, the *IZ* operator creates another image with the influences zones of the connected components according to the connectivity defined by the SE B. Figure 4 shows an example of this operator.

The expression given above in terms of morphological dilations and erosions is the basis for the definition of RS by influence zones proposed in this work.

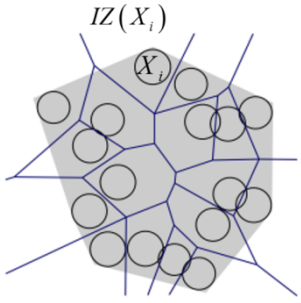


Fig. 4 Influence Zones

5.1 Rough Sets by Influence Zones

If does not exist a decision attribute d to determine a partition of the universe U, the Rough Sets by Influence Zones (RS-IZ) will allow to process a family of data sets, identifying influence zones, thus obtaining their separation.

Let $S = (U, A, V, f)$ an IS such that $U = \bigcup_{i \in I} X_i$ where the $\{X_i\}_{i \in I}$ is a family of disjoint sets. The Rough Sets by Influence Zones (*RS-IZ*), denoted by $RS - IZ(X_i)$, are defined as follows:

$$RS - IZ(X_i) = \bigcup_{\lambda} \left[B_\lambda^*(X_i) \backslash B_\lambda^* \left(\bigcup_{j \neq i} X_j \right) \right] \tag{8}$$

where λ decreases and denotes the cardinal of the set of attributes $B \subseteq A$.

5.2 Relationship between IZ and RS-IZ

When defining the *IZ* in MM, the value of λ, in the union, is growing, since δ_λ represents the dilation of a ball of radius λ, and Property 2 indicates that: $\lambda_1 < \lambda_2 \Rightarrow \delta_{\lambda_1} \subseteq \delta_{\lambda_2}$.

However, in the RS theory this inclusion relationship is not satisfied. For upper approximation the inclusion is verified as the cardinal of the set of attributes decreases, ie: $\lambda_1 > \lambda_2 \Rightarrow B_{\lambda_1}^* \subseteq B_{\lambda_2}^*$.

Therefore, in the definition of *RS-IZ* λ values are decreasing.

6 Application Example

Below an example of an IS is shown. Its universe is composed of six items, which are described by the set of attributes A given by:

$$A = \{(a_1)\ headache, (a_2)\ muscle\ pain, (a_3)\ temperature, (a_4)\ flu\}$$

Table 1 defined the IS to be used in this example. Being $S = \{P_1, P_2, P_3, P_4, P_5, P_6\}$ the universe consists of the six patients. First we provide a disjoint partition of the universe as $S = \bigcup_{i \in I} X_i$ where the $X_i$sets must be disjoint.

Example 1: Being $B_{\lambda=2} = \{headache, muscle\ pain\}$ the subset of attributes to be considered first, with λ the number of attributes in the set B. In this case, the classes are determined as follows:

$B(P_1) = B(P_4) = B(P_6) = \{P_1, P_4, P_6\}$; $B(P_2) = B(P_5) = \{P_2, P_5\}$ and $B(P_3) = \{P_3\}$

These classes induce a partition of the universe as follows:

$$X_1 = \{P_1, P_4, P_6\};\ X_2 = \{P_2, P_5\} \text{ and } X_3 = \{P_3\}$$

Being $B_{\lambda=1} = \{headache\}$. Therefore, the classes are determined as follows:

$$B(P_1) = B(P_4) = B(P_6) = \{P_1, P_4, P_6\} \text{ and}$$
$$B(P_2) = B(P_3) = B(P_5) = \{P_2, P_3, P_5\}$$

We compute the $RS - IZ(X_1)$:

- To $\lambda = 2$:

$$\left.\begin{array}{l} B^*(X_1) = \{P_1, P_4, P_6\} \\ B^*(X_2 \cup X_3) = \{P_2, P_3, P_5\} \end{array}\right\} RS - IZ_{\lambda=2}(X_1) = \{P_1, P_4, P_6\}$$

- To $\lambda = 1$:

$$\left.\begin{array}{l} B^*(X_1) = \{P_1, P_4, P_6\} \\ B^*(X_2 \cup X_3) = \{P_2, P_3, P_5\} \end{array}\right\} RS - IZ_{\lambda=1}(X_1) = \{P_1, P_4, P_6\}$$

Then: $RS - IZ(X_1) = \{P_1, P_4, P_6\}$

We compute the $RS - IZ(X_2)$:

- To $\lambda = 2$:

$$\left.\begin{array}{l} B^*(X_2) = \{P_2, P_5\} \\ B^*(X_1 \cup X_3) = \{P_1, P_3, P_4, P_6\} \end{array}\right\} RS - IZ_{\lambda=2}(X_2) = \{P_2, P_5\}$$

- To $\lambda = 1$:

$$\left.\begin{array}{l} B^*(X_2) = \{P_2, P_3, P_5\} \\ B^*(X_1 \cup X_3) = \{P_1, P_2, P_3, P_4, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=1}(X_2) = \emptyset$$

Then: $RS - IZ(X_2) = \{P_2, P_5\}$

We compute the $RS - IZ(X_3)$:

- To $\lambda = 2$:

$$\left.\begin{array}{l} B^*(X_3) = \{P_3\} \\ B^*(X_1 \cup X_2) = \{P_1, P_2, P_4, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=2}(X_3) = \{P_3\}$$

- To $\lambda = 1$:

$$\left.\begin{array}{l} B^*(X_3) = \{P_2, P_3, P_5\} \\ B^*(X_1 \cup X_2) = \{P_1, P_2, P_3, P_4, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=1}(X_3) = \emptyset$$

Then: $RS - IZ(X_3) = \{P_3\}$

Figure 5 shows the composition of each of the RS-IZ of the previous example.

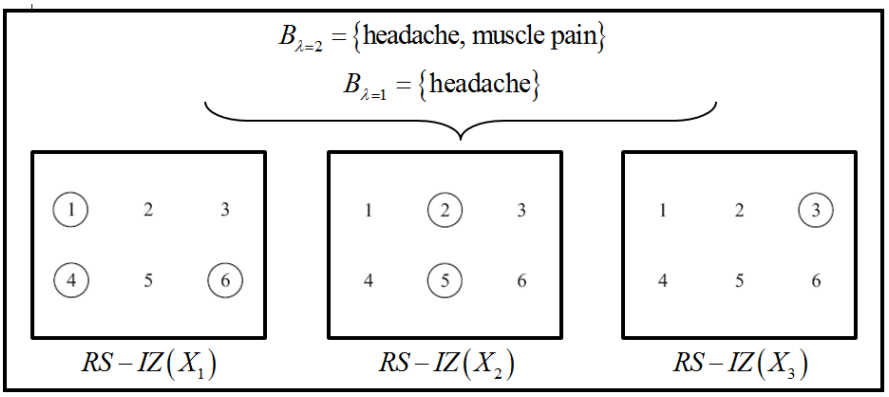


Fig. 5 Rough Sets by Influence Zones of the example 1

Example 2: Here we analyze an example where the partition of the universe S is defined a priori, unlike the previous example where the partition is induced from the classes defined by the set B.

Being $X_1 = \{P_1, P_5, P_6\}$, $X_2 = \{P_2, P_3\}$ and $X_3 = \{P_4\}$ disjoint partition of space $S = \{P_1, P_2, P_3, P_4, P_5, P_6\}$.

Being $B_{\lambda=3} = \{headache, muscle\ pain, temperature\}$.

Therefore, the classes are determined as follows:

$B(P_1) = \{P_1\}$; $B(P_2) = B(P_5) = \{P_2, P_5\}$; $B(P_3) = \{P_3\}$; $B(P_4) = \{P_4\}$ and $B(P_6) = \{P_6\}$

Being $B_{\lambda=2} = \{headache, temperature\}$. Therefore, the classes are determined as follows:

$B(P_1) = \{P_1\}$; $B(P_2) = B(P_5) = \{P_2, P_5\}$; $B(P_3) = B(P_6) = \{P_3, P_6\}$ and $B(P_4) = \{P_4\}$

Being $B_{\lambda=1} = \{temperature\}$. Therefore, the classes are determined as follows:

$B(P_1) = B(P_2) = B(P_5) = \{P_1, P_2, P_5\}$; $B(P_3) = B(P_6) = \{P_3, P_6\}$ and $B(P_4) = \{P_4\}$

We compute the $RS - IZ(X_1)$:

- To $\lambda = 3$:

$$\left.\begin{array}{l} B^*(X_1) = \{P_1, P_2, P_5, P_6\} \\ B^*(X_2 \cup X_3) = \{P_2, P_3, P_4, P_5\} \end{array}\right\} RS - IZ_{\lambda=3}(X_1) = \{P_1, P_6\}$$

- To $\lambda = 2$:

$$\left.\begin{array}{l} B^*(X_1) = \{P_1, P_2, P_3, P_5, P_6\} \\ B^*(X_2 \cup X_3) = \{P_2, P_3, P_4, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=2}(X_1) = \{P_1\}$$

- To $\lambda = 1$:

$$\left.\begin{array}{l} B^*(X_1) = \{P_1, P_2, P_3, P_5, P_6\} \\ B^*(X_2 \cup X_3) = \{P_1, P_2, P_3, P_4, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=1}(X_1) = \emptyset$$

Then: $RS - IZ(X_1) = \{P_1, P_6\}$

We compute the $RS - IZ(X_2)$:

- To $\lambda = 3$:

$$\left.\begin{array}{l} B^*(X_2) = \{P_2, P_3, P_5\} \\ B^*(X_1 \cup X_3) = \{P_1, P_2, P_4, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=3}(X_2) = \{P_3\}$$

- To $\lambda = 2$:

$$\left.\begin{array}{l} B^*(X_2) = \{P_2, P_3, P_5, P_6\} \\ B^*(X_1 \cup X_3) = \{P_1, P_2, P_3, P_4, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=2}(X_2) = \emptyset$$

- To $\lambda = 1$:

$$\left.\begin{array}{l} B^*(X_2) = \{P_1, P_2, P_3, P_5, P_6\} \\ B^*(X_1 \cup X_3) = \{P_1, P_2, P_3, P_4, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=1}(X_2) = \emptyset$$

Then: $RS - IZ(X_2) = \{P_3\}$

We compute the $RS - IZ(X_3)$:

- To $\lambda = 3$:

$$\left.\begin{array}{l} B^*(X_3) = \{P_4\} \\ B^*(X_1 \cup X_2) = \{P_1, P_2, P_3, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=3}(X_3) = \{P_4\}$$

- To $\lambda = 2$:

$$\left.\begin{array}{l} B^*(X_3) = \{P_4\} \\ B^*(X_1 \cup X_2) = \{P_1, P_2, P_3, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=2}(X_3) = \{P_4\}$$

- To $\lambda = 1$:

$$\left.\begin{array}{l} B^*(X_3) = \{P_4\} \\ B^*(X_1 \cup X_2) = \{P_1, P_2, P_3, P_5, P_6\} \end{array}\right\} RS - IZ_{\lambda=1}(X_3) = \{P_4\}$$

Then: $RS - IZ(X_3) = \{P_4\}$

Figure 6 shows the composition of each of the RS-IZ of the previous example.

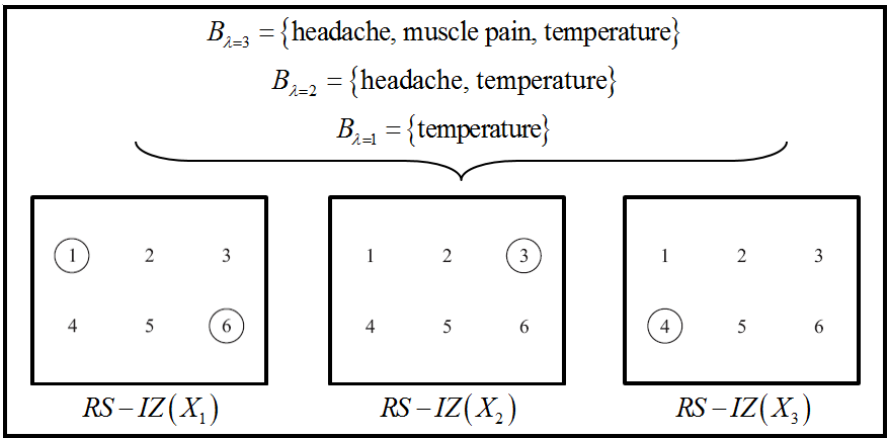


Fig. 6 Rough Sets by Influence Zones of Example 2

7 Conclusion

In this work we analized deeply the relationship between the two theories, Rough Sets and Mathematical Morphology, and we defined a novel characterization of Rough Sets which we called Rough Sets by Influence Zones based in morphological concepts.

When it does not exist a decision attribute or it is not available, this novel definition of Rough Sets by Influence Zones, allows determining a new partition of the universe generating influence zones.

Finally we presented an example, with various partitions of the space obtaining different Rough Sets by Influence Zones.

References

1. Abraham, A., Falcón, R., Bello, R.: Rough Set Theory. SCI, vol. 174. Springer, Heidelberg (2009)
2. Bangham, J.A., Marshall, S.: Image and Signal processing with mathematical morphology. IEE Electronics & Communication Engineering Journal 10, 117–128 (1998)
3. Caballero, Y., Bello, R., Salgado, Y., García, M.M.: A Method to Edit Training Set Based on Rough Sets. International Journal of Computational Intelligence Research 3, 219–229 (2007)
4. Dougherty, E.R.: Mathematical morphology in image processing. M. Dekker (1993)
5. Facon, J.: Morfología Matemática. Teoría e ejemplos. Editora Universitária Champagnat da Pontífica Universidade Católica do Paraná, Curitiba, Brasil (1996)
6. Glasbey, C.A., Horgan, G.W.: Image Analysis for the Biological Sciences. Jonh Wiley and Sons, England (1995)

7. Haralick, R.M., Sternberg, S.R., Zhuang, X.: Image Analysis Using Mathematical Morphology. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-9, 532–550 (1987)
8. Bloch, I.: On links between mathematical morphology and rough sets. Pattern Recognition 33, 1487–1496 (2000)
9. Muir, A., Düntsch, I., Gediga, G.: Rough set data representation using binary decision diagrams. RACSAM Rev. R. Acad. Cien. Serie A. Mat. 98, 197–211 (2004)
10. Nurmi, H., Kacprzyk, J., Fedrizzi, M.: Probabilistic, fuzzy and rough concepts in social choice. European Journal of Operational Research 95, 264–277 (1996)
11. Pal, S.K., Mitra, P.: Case Generation Using Rough Sets with Fuzzy Representation. IEEE Transactions on Knowledge and Data Engineering 16, 292–300 (2004)
12. Pawlak, Z.: Decision rules, Bayes' rule and Rough Sets. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) RSFDGrC 1999. LNCS (LNAI), vol. 1711, pp. 1–9. Springer, Heidelberg (1999)
13. Pawlak, Z.: Hard and soft sets. In: Rough Sets, Fuzzy Sets and Knowledge Discovery, pp. 130–135. Springer, London (1994)
14. Pawlak, Z.: Rough Sets -Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
15. Pawlak, Z.: Rough sets. Internat. J. Comput. Inform. Sci. 11, 341–356 (1982)
16. Ronse, C., Heijmans, H.J.A.M.: The algebraic basis of mathematical morphology: II. Openings and closings. Computer Vision, Graphics, and Image Processing: Image Understanding 54, 74–97 (1991)
17. Salama, A.S., Abu-Donia, H.M.: New Approaches for Data Reduction. International Journal of Pure and Applied Mathematics 27, 39–53 (2006)
18. Serra, J.: Image analysis and mathematical morphology. Academic Press, London (1982)
19. Serra, J.: Image analysis and mathematical morphology. Academic Press, London (1988)
20. Yao, Y.Y.: Constructive and Algebraic Methods of the Theory of Rough Sets. Information Sciences 109, 21–47 (1998)

Part II

Business Intelligence and Knowledge Discovery

Fuzzy Tree Studio: A Tool for the Design of the Scorecard for the Management of Protected Areas

Gustavo J. Meschino, Marcela Nabte, Sebastián Gesualdo,
Adrián Monjeau, and Lucía I. Passoni

Abstract. This paper presents a Scorecard, which associated with a geographic information system (GIS), will provide a management tool to assess vulnerability within a protected area. To accomplish this task a novel framework is presented, which enables the design of logical predicates evaluated with fuzzy logic. This tool may guide decisions and investment priorities in protected areas. We have taken the Valdes Peninsula Protected Natural Area as a case study, which has been declared a World Heritage Site by UNESCO. In this area we have released an intense amount of variables related to natural resources, as well as human uses of land and territory and the effectiveness of the management plan and management area. To evaluate the vulnerability values of different parcels, according to a set of field collected variables is proposed a framework that manages logic predicates using fuzzy logic. Several ecologists evaluated this framework satisfactorily due to the easy-to-use interface and that the shown results are highly understandable for those who need to make decisions on environmental care.

Gustavo J. Meschino · Lucía I. Passoni
Laboratorio de Bioingeniería, Facultad de Ingeniería,
Universidad Nacional de Mar del Plata, Argentina
e-mail: {gustavo.meschino,isabel.passoni}@gmail.com

Marcela Nabte · Adrián Monjeau
Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Argentina
e-mail: mjnabte@yahoo.com.ar, amonjeau@gmail.com

Sebastián Gesualdo
Universidad CAECE, Mar del Plata, Argentina
e-mail: sebastián.gesualdo@gmail.com

Adrián Monjeau
Universidad Atlántida Argentina, Mar del Plata, Argentina

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_6,

1 Introduction

When ecologists look at the need to generate predictive knowledge, their descriptions must be converted into quantitative data. A quantitative approach requires precise numerical data in order to gain the benefits of mathematics and statistics. However, in ecology often the precise number is inaccessible because the imprecision is a constitutive part of reality. Most of the ecological patterns and processes can be predicted only beyond a certain "noise" [2]. Where random reigns, as the deep root of the vagueness, is where our tools are useless to make predictions [12]. This drawback has led scientists to ridiculous situations where all the statistical and mathematical paraphernalia was blind to what was obvious to the senses [9], all in the pursuit of accurate data to apply a biological reality that refuses to be tamed by the accuracy intended. Fuzzy logic has come to address this problem, lavishing powerful tools to predict situations from elusive data using logical predicates and logical syntaxes [14] connected together in a cascade of reasoning that comes from experience and common sense [4].

In this case study we propose to apply Fuzzy Logic to assess the vulnerability within a protected area. Protected areas are needed to prevent the extinction of its main elements [6, 7, 13]. These elements (e.g., biological species) can be valued according to their likelihood of extinction [5]. The level of threat to these species and the effectiveness of protected area management are key elements in the analyzed system. To analyze this issue using fuzzy logic it is necessary to define, with as little ambiguity as possible, logical predicates that permit the formalization and evaluation of its truth value.

Some knowledge describing the status of protected area, from the experience based on multiple studies [10, 11], can be formalized as:

1. The vulnerability increases when the value of their conservation targets increases,

The value of the conservation targets increases when increasing the possibility of extinction.
The vulnerability increases when the conflict increases.
The conflict increases as threats increase.
The conflict increases when management inefficiency increases.

This paper presents a Scorecard, which associated with a geographic information system (GIS), will provide a management tool to measure vulnerability. This tool may guide decisions and investment priorities in protected areas. We have taken the Valdes Peninsula Protected Natural Area as a case study, which has been declared a World Heritage Site by UNESCO. In this area we have released an intense amount of variables related to natural resources, as well as human uses of land and territory and the effectiveness of the management plan and management area.

2 Materials and Methods

2.1 Study Site

The Valdes Peninsula lies northeast of the province of Chubut, Argentina, between 42 ° and 42 ° 45 'S and 63 ° 35' and 65 ° 17 'W. Its area is 400 000 ha. It was declared a Natural Heritage Site by UNESCO in 1999. The province of Chubut created a natural reserve by the Provincial Law Nº 4722, called Protected Natural Area Valdés Peninsula. According to the classification of conservation units of the World Conservation Union (IUCN), the area has been placed in category VI (Managed Resource Protected Area). Fig. 1 shows a map of Valdes Peninsula, including the ten lots where data was collected clearly pointed.

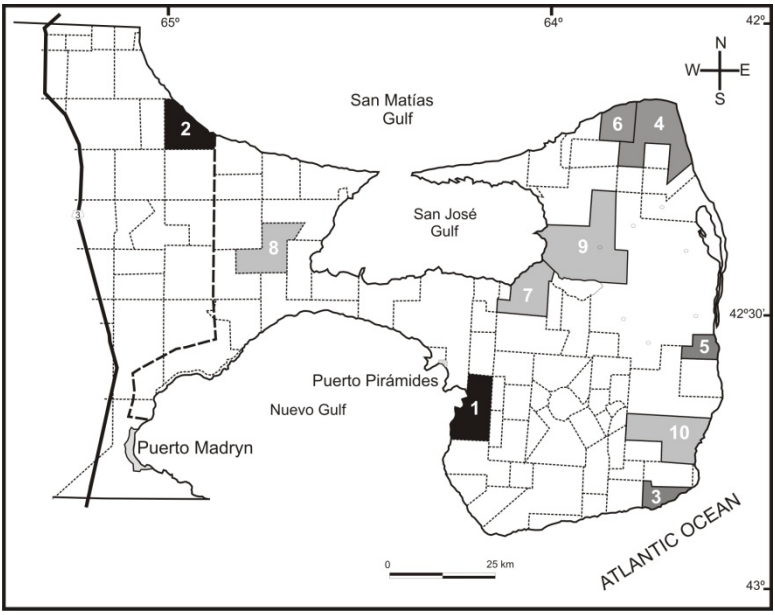


Fig. 1 Map of Valdés Peninsula, showing the ten lots where data was collected clearly pointed

The Valdés Peninsula has worldwide relevance for its spectacular marine mammals (whales, dolphins, elephants and sea lions). It is in these coastal areas, attractive for tourism, where management efforts are concentrated. Masked by these charismatic megafauna, Valdes Peninsula has a rich and varied terrestrial fauna in the interior [11], which represents 98% of World Heritage area. This area is completely occupied by private ranches devoted to sheep farming [10]. In this sector the protected area management is not only poor [8] but wildlife is combated as a threat to sheep or competition with resources [3, 8], with some exceptions due to circumstantial conservationist owners [10].

2.2 Data Acquisition

Information on natural resources, threats and management effectiveness is based on five sources:

- Survey of published information (See [10] and its references).
- Materials collected: Marcela Nabte campaigned in 2005, 2006 and mid-2007, in which materials were collected mainly from skeletal elements of specimens run over, killed or found dead in the field [10]. The collected material is temporarily deposited in the Laboratory Animals of the National Center Patagonia (Puerto Madryn, Chubut, Argentina).
- Direct observations: the most detailed observations were made by the authors of this paper (MN, AM) or people with experience in the field;
- Identification of skins: during the fieldwork surveys were conducted on skins of animals hunted presumably by rural people, which in some cases were documented photographically [10].
- Interviews with rural people and park guards.

2.3 Species Account

For this case study we have consider only the most conspicuous elements of the terrestrial mammal fauna of the Peninsula Valdes, excluding small mammals and bats from the analysis. The species included in this study are:

- Guanaco (Lama guanicoe),
- Patagonian Mara (Dolichotis patagonum),
- Large Hairy Armadillo (Chaetophractus villosus),
- Patagonian Pichi (Zaedyus pichiy),
- Grey Fox (Pseudalopex griseus),
- Geoffroy's Cat (Leopardus geoffroyi),
- Pampas´Cat (Leopardus colocolo pajeros),
- Puma (Puma concolor),
- Patagonian Hog-nosed Skunk (Conepatus humboldtii),
- Lesser Grison (Galictis cuja).

2.4 Fuzzy Logic Predicates Concepts

Definition #1. A fuzzy predicate p is a linguistic expression (a proposition) with degree of truth μ_p into [0, 1] interval. It applies the "principle of gradualism" which states that a proposition may be both true and false, having some degree of truth (or falsehood) assigned.

Definition #2. A simple fuzzy predicate sp is a fuzzy predicate whose degree of truth μ_{sp} can be obtained by some of the next alternatives:

- The application of a membership function associated with a fuzzy term, to a quantitative variable. E.g. *sp* = "Intensity is high", is associated with the variable "intensity" which is measured in meters and the concept "high" is defined by a membership function over the magnitude of the intensity.
- The association of discrete values into the interval [0, 1] to language labels (generally adjectives) of a variable. For example: variable "intensity", and its labels "high": $\mu_{sp} = 0.9$; "medium": $\mu_{sp} = 0.5$; "low": $\mu_{sp} = 0.1$.
- Determination of real value into the [0, 1] interval by an expert. It is normally required in situations of some subjectivity where there is a variable that cannot be quantified by using one of the two previous cases, e.g. "Infrastructure is adequate".

Definition #3. A compound predicate *cp* is a fuzzy predicate obtained by combination of simple fuzzy predicates or other compound fuzzy predicates, joined by logical connectives and operators (and, or, not, implication, double-implication). For example:

$$cp = cp1\ and\ (cp2\ or\ sp1)\ and\ sp2$$

Definition #4. Compound predicates can be represented as a tree structure, having its nodes associated by logical connectives (and, or, not, implication, double-implication) and the successive branches related to lower hierarchical level predicates (simple or compound). Of course, the root of the tree corresponds to a main compound predicate and the leaves will be simple predicates.

It is needed defining logic systems based on a quadruple of continuous operators: conjunction, disjunction, order and negation, over a set of truth values for predicates, into the real interval [0, 1], such that when the truth values are restricted to $\{0, 1\}$, these operations become classic Boolean predicates [1].

Some logic systems are quite simple, for example using minimum and maximum operations for conjunction and disjunction respectively. Some others quit some axioms in order to achieve a sensitive and idempotent multi-valued system, as compensatory logics systems. These systems are sensible to the value of truth of the predicates involved and they have been widely used to represent knowledge as a predicates system.

In Table 1 we present some systems for conjunction and disjunction (“and” and “or”) operators, their operations and references. The “not” operator is usually implemented as though there are another proposals.

2.5 Fuzzy Tree Studio Product Perspective

Fuzzy Tree Studio is a software product conceived as a Decision Support System. It tackles Fuzzy Logic concepts by means of a friendly graphical user interface. The main objective is giving some help in data analysis by applying previous expert knowledge, allowing evaluation and comparison of alternatives.

The Fuzzy Tree Studio user interface provides easy access to the various tools and features to browse, view, edit, manage and query Fuzzy Predicates Trees.

Table 1 Some systems for conjunction and disjunction operators

	Conjunction and Disjunction	Reference
Max-Min	$C(x_1, x_2, \ldots, x_n) = min(x_1, x_2, \ldots, x_n)$ $D(x_1, x_2, \ldots, x_n) = max(x_1, x_2, \ldots, x_n)$	Dubois & Prade, 1985
Probabilistic	$C(x_1, x_2) = x_1 x_2$ $D(x_1, x_2) = x_1 + x_2 - x_1 x_2$	Dubois & Prade, 1985
GMBCL (Geometric Mean Based Compensatory Logic)	$C(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_n)^{\frac{1}{n}}$ $D(x_1, x_2, \ldots, x_n) = 1 - [(1 - x_1)(1 - x_2) \ldots (1 - x_n)]^{\frac{1}{n}}$	Espin Andrade, Marx Gómez, Mazcorro Téllez, & Fernández González, 2003
AMBCL (Arithmetic Mean Based Compensatory Logic)	$C(x_1, x_2, \ldots, x_n) = \left[\min(x_1, x_2, \ldots, x_n) \frac{1}{n} \sum_{i=1}^{n} x_i\right]^{\frac{1}{2}}$ $D(x_1, x_2, \ldots, x_n) =$ $1 - \left[\min(1 - x_1, 1 - x_2, \ldots, 1 - x_n) \frac{1}{n} \sum_{i=1}^{n} (1 - x_i)\right]^{\frac{1}{2}}$	Bouchet, Pastore, Espin Andrade, Brun, & Ballarin, 2010

Queries are based on objective data coming from different cases to be considered as alternatives. A degree of truth of a main Predicate will be delivered for each case.

Fuzzy Tree Studio consists of three main parts: the Tree designer and viewer, the Data generator and importer, and the Query analyzer.

The pipeline to work will contain the next three stages:

- Tree design: user should provide a predicates tree structure. The root is the main predicate to evaluate, and leaves are the simple predicates that will be evaluated considering data values.
- Data generation: user is able to import data from files or write them down in a table. Data may be numerical or keywords, choices from list of options. Data must be coherent with respect to the designed tree.
- Tree evaluation: based on a data set, the degree of truth of the root will be given for each case. But additionally, useful graphical and numerical information is delivered, to allow the analysis of the particular cases.

Fuzzy Tree Studio was thought to be scalable. In future versions it is planned to include some Evolutionary Algorithms to improve the tree design and the parameters of the membership functions included in the model.

Fuzzy Tree Studio has proven to be helpful as a tool for model generation and analysis from automated control, psychology, ecology, economics, finances, biology, medicine, social sciences, management, etc.

Fuzzy Tree Studio is intended to be used by people having some experience in Fuzzy Logic and Predicates Logic. Its use is simple for any user familiarized with typical Windows applications.

2.5.1 Some other General Specifications

FTS allows working with more than one project at the same time. Project data can be stored in XML format to be compatible with other or future systems.

During the tree design, the user can see errors or omissions causing problems for the future evaluation stage. There are functions for undo, redo, copy, cut, paste as usual. Additionally, information about the tree is given at design time: weight, number of leaves nodes, number of composed predicates, depth, and grade.

When a valid design is achieved, a linguistic expression for the main predicate is shown. It is formed taking the description of the nodes.

Simple predicates (leaves nodes) are intended to be evaluated by data. Their degree of truth can be defined by:

- Membership functions (triangular, trapezoidal, Gaussian, sigmoid, S shaped, Z shaped). In this case, a value from the dataset will be taken and evaluated using the membership function.
- User-defined Labels, related with different degrees of truth. In this case, the dataset should contain the description of the label (for example, “big”, “enough”, “small”) and it will be associated with a value of true previously defined.
- User-values. In this case the value is directly taken from the dataset. It is supposed that some expert gave the value according to his expertise. For example, used in cases such “The quality of the soil is good”, which could not be quantified by using a numerical variable.

Composed predicates are characterized by a logical operator (and, or, not, implication, double implication) and associated to one or more simple predicates.

The membership functions can be changed by varying their parameters. Besides the shape of the function can be visualized as the parameters are changed, the user can modify the function interactively with the mouse by displacing some points.

2.5.2 Tree Evaluation

Since there are a variety of options to compute the fuzzy operations between degrees of truth, in FTS the user can choose between of them which was presented in Table 1: Max-Min, Probabilistic, GMBCL (Geometric Mean Based Compensatory Logic) or AMBCL (Arithmetic Mean Based Compensatory Logic).

Results of the degree of truth for the main predicate are given as a table for each case in the dataset, showing the results of the different systems in columns.

Given a particular case in the table, a graphical representation of the tree for this case may be asked. In this representation, the partial degree of truth across the tree is shown using a color scale. So the user will be able to analyze which branches of the tree (partial composed predicates) are strongly true or weakly true.

2.6 Data Quantification

We have considered three kinds of indicators for the case study, a) the conflictivity of the protected area, b) threats, c) efficiency / inefficiency of the area management.

a) *Conflictivity* is a function of the *conservation value* of each species. This conservation value of target species has been characterized by three numerical indicators related with the possibility of extinction (logical predicates R(i), C(i) and I(i)) based on previous studies [8, 10]: rarity, criticality, irreplaceability:

- *Rarity*: was estimated from the relative abundance assigned by sample (parcelas) obtained from the sources mentioned above.
- *Criticality*: was estimated from population trend assessments made from surveys and historical records [11]. A 0 (zero) is assigned to a growing population, and to declining populations increasing values up to 1.
- *Irreplaceability*: was estimated from considering the ecological role within the area and its taxonomic uniqueness. For example, a species of large carnivore, unique and the only predator of large herbivores is irreplaceable in the area because no other species is capable of filling that role, and therefore the value is 1; but a small herbivore that shares its diet with other species is more "replaceable" in its ecosystem function, and their absence may be covered by other elements of their trophic guild.

b) *Conflictivity* depends on the level of *Anthropic pressure* to the conservation values. The *Anthropic pressure* has been characterized by three numerical indicators (*Area*, *Time*, *and Intensity*):

- *Area*: the total relative area occupied by the disturbance. The maximum value (1) is whether it occurs throughout the area and the minimum (0) if the disturbance does not occur in the plot.
- *Time*: the length of time in which the disturbance occurs. The maximum value (1) is if the disturbance happens all the time, intermediate values are estimates of the frequency with which the action takes place, from occasional to frequent, and the minimum value (0) if the disturbance never happens.
- *Intensity*: It is the power to modify the existing natural conditions, i.e., number of hunters or amount of poison, or grazing load per square kilometer.

c) *Vulnerability* is also a function of the *Inefficiency of the management* (predicate EI). The Ineffective management indicator is estimated from studies of the management plan revision [8] after considering time and money spent on control and surveillance in various sectors of the protected area.

2.7 Fuzzy Predicates Definition

We reformulated the available knowledge as logical predicates in a tree scheme, for each Anthropic Pressure.

Vulnerability is high when Conflictivity is high and the Environmental Management is inefficient.

Conflictivity is high when relative conservative value of the SIG area is high and at the same time, the prevalent anthropic pressure is high.

Relative conservative value of the SIG area is high when Specie 1 is rare or *Specie 1 is critical* or *Specie 1 is irreplaceable;* or *when Specie 2 is rare* or *Specie 2 is critical* or *Specie 2 is irreplaceable; … ;* or *when Specie n is rare* or *Specie n is critical* or *Specie n is irreplaceable.*

Prevalent anthropic pressure is high when its time is long, its area is big and its intensity is strong.

Being n the number of species considered.
Symbolically:

$$VULH(ap) := CH \wedge EI$$

$$CH := VH \wedge PH$$

$$VH := [R(1) \vee C(1) \vee I(1)] \vee [R(2) \vee C(2) \vee I(2)] \vee \ldots \vee [R(n) \vee C(n) \vee I(n)]$$

$$PH := TL \wedge AH \wedge IS$$

Where:

VULH(ap) = "Vulnerability for the anthropic pressure considered is high."
CH = "Conflictivity is high."
EI = "Environmental Management is inefficient."
VH = "Relative conservative value of the SIG area is high."
PH = "Prevalent anthropic pressure is high."
R(i) = "Specie i is rare."
C(i) = "Specie i is critical."
I(i) = "Specie i is irreplaceable."
TL = "Prevalent anthropic pressure time is long."
AB = "Prevalent anthropic pressure area is big."
IS = "Prevalent anthropic pressure intensity is strong."

In this way, we can compute the degree of truth of the main predicate *VULH* by operating with the degree of truth of lower level predicates (*R(i), C(i), I(i), TL, AB, IS*).

After computing the degree of truth of the main predicate for each considered anthropic pressure, we compute a general value of Vulnerability, considering.

$$AreaVulnerability := VULH(hunting) \vee VULH(poisoning) \vee VULH(sheepGrazing)$$

3 Results

3.1 *Application of Fuzzy Tree Studio*

In Figure 2, a general visualization of the Fuzzy Tree Studio Environment is shown. On the left is a bar with tools to create both compound and simple predicates nodes. On the right, a project browser shows the available elements (diagrams and datasets) and there are two settings windows.

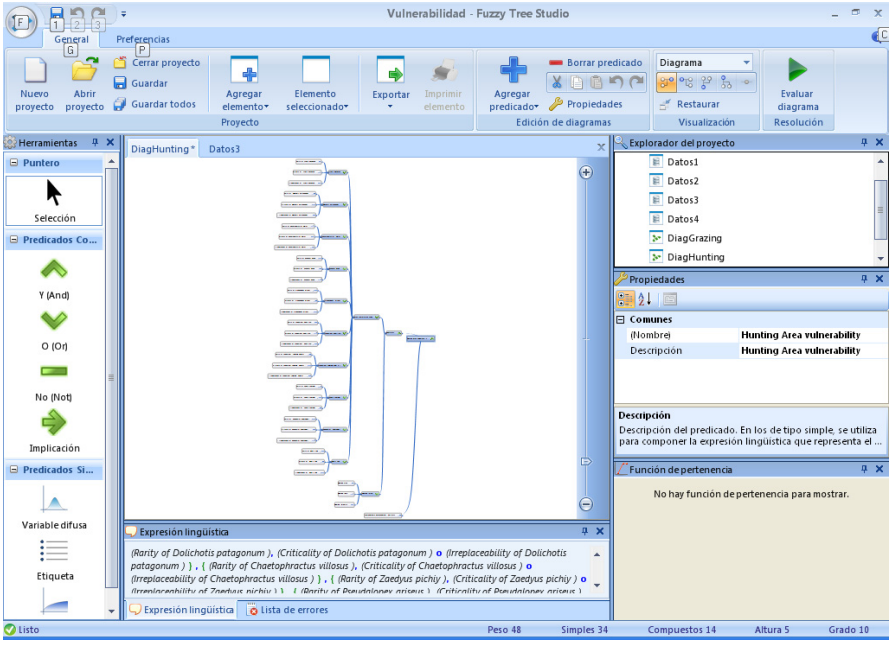


Fig. 2 Fuzzy Tree Studio Environment

When a tree is about to be evaluated using a dataset, the window in Figure 3 allows selection of the logic models that will be used when degrees of truth will be computed.

A results table is delivered after computing the degree of truth of the main predicate for each dataset register, which is shown partially in Figure 4. Also, these results are shown for some areas in Table 2.

In order to deep into the analysis, if user clicks on a result cell for a specific register, a detailed colored tree is made. It can be seen on Figure 5.

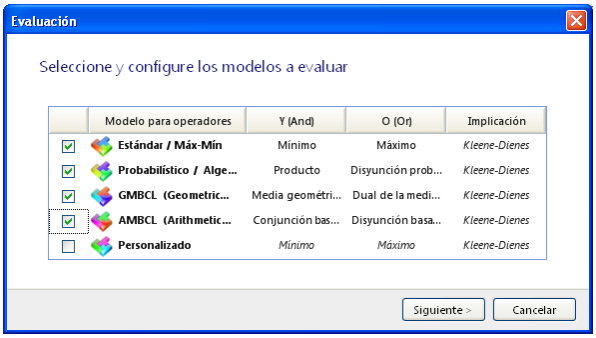


Fig. 3 Logical models selections for evaluate a predicates tree

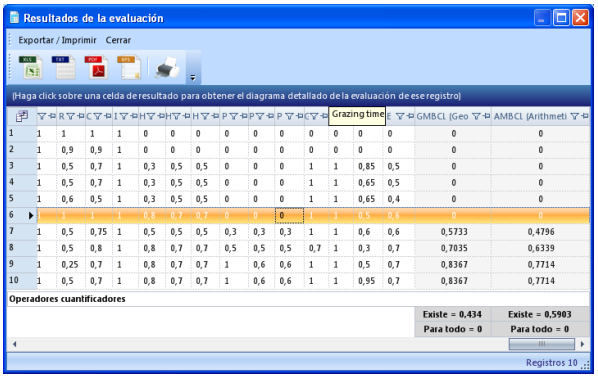


Fig. 4 Results table after computing the degree of truth of the main predicate for each data-set register. There is a column for each logic system (in this example, GMBCL and AMBCL)

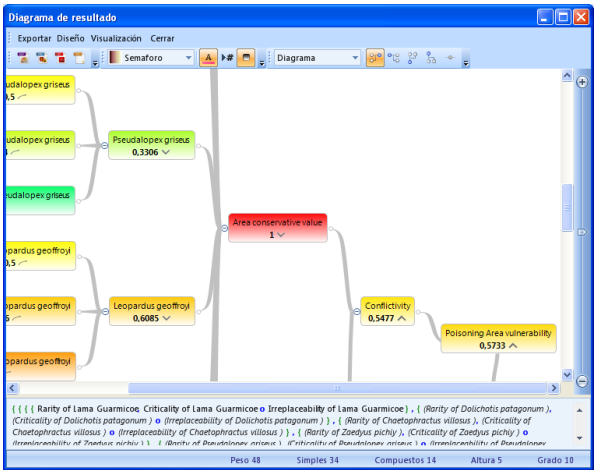


Fig. 5 Detailed colored tree for a specific register. Red means high degree of truth, while green means low degrees of truth

Table 2 Results for a partial set of parcels, using Min-Max (MM), Geometric Mean Based Compensatory Logic (GM) and Arithmetic Mean Based Compensatory Logic (AM) operators

	Poisoning Vulnerability			Hunting Vulnerability			Grazing Vulnerability			Total threat		
Parcel	MM	GM	AM	MM	GM	AM	MM	GM	AM	MM	GM	AM
#1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
#2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
#3	0.00	0.00	0.00	0.50	0.57	0.52	0.50	0.70	0.61	0.50	0.50	0.51
#4	0.30	0.57	0.48	0.50	0.65	0.60	0.60	0.77	0.69	0.60	0.67	0.65
#5	0.00	0.00	0.00	0.50	0.57	0.52	0.50	0.70	0.61	0.50	0.50	0.51
#6	0.00	0.00	0.00	0.40	0.52	0.44	0.40	0.63	0.52	0.40	0.43	0.44
#7	0.00	0.00	0.00	0.60	0.72	0.65	0.60	0.77	0.69	0.60	0.60	0.59
#8	0.70	0.84	0.77	0.70	0.78	0.73	0.70	0.83	0.77	0.70	0.82	0.76
#9	0.50	0.70	0.64	0.70	0.78	0.73	0.70	0.83	0.77	0.70	0.78	0.74
#10	0.70	0.84	0.77	0.70	0.78	0.73	0.70	0.83	0.77	0.70	0.82	0.76

Using results of Table 2, SIG maps of vulnerability can be obtained, coloring with a semaphore-color bar the different parcels according with their vulnerability values (red=1, green=0). The Kruskal-Wallis test was performed on each set of Vulnerability source (Poisoning, Hunting and Grazing) and also over the Total threat to compare results from each logical operator: Max-Min (MM), Geometric Mean Based Compensatory Logic (GM) and Arithmetic Mean Based Compensatory Logic (AM). The multiple comparison test ($\alpha = 0.05$) showed no mean ranks significantly different among any of the groups ($p > 0.05$) within all the Vulnerability sources and also on the Total Threat data.

4 Discussion and Conclusions

Analyzing results in Table 2, we can consider that the first two parcels show null values of vulnerability. When these records were inspected using the colored tree (Figure 5), null inefficiency values of the Environmental Management were shown, forcing to get null vulnerability values, despite the existence of high conservative values of the present species. Other parcels show a wide spread of vulnerability values, showing the use of compensatory logics a wider range than those obtained with the standard max-min operators.

The Fuzzy Tree Studio framework into a making decision process oriented to the environmental care, has proved to have an easy-to-use interface and helpful for those who need to prioritize policies.

When results of this paper are contrasted against the viewpoint of a person with deep experience in Peninsula Valdes (e.g. a park guard), they are logical and expected. Consequently we can consider that the system implements the expert knowledge successfully. What is the contribution of this paper to that person?

Probably, it is only the simplification of the main problems of the area in a color map. However, these results are significant to those who are not near of the troublesome territory and those who need making decisions that involves the area. Indeed, one of the contributions of the fuzzy logic based system is being a surrogate for experience [9].

For this particular application was discarded the use of fuzzy logic systems with a knowledge base of rules (Mamdani type), since the design contained several systems with cascade stages. The parameterization of such a design was not simple; given that the logical operators and also the defuzzification method must be selected according to improve the efficacy. The performance obtained with this kind of design did not improve the results achieved with the Fuzzy Tree Studio; and also this approach did not provide how visualize the results, as does the FTS framework, (Figure 5).

Another contribution, perhaps the most important, is to provide a solution that synergizes the SIG and the fuzzy data processing technologies, integrating geographical information with data from surveys and countless opinions, individual experiences, subjectivities, biases, and good or bad luck of observers and samplers. A data source like this, subject to a classical statistical package, would not yield results showing differences between plots. When results say there are no statistical significant differences, the work of the scientist finds its border. Of course there are patterns beyond the limits of statistics. Field ecologists are well aware of this tension between differences that are seen with the naked eye and the difficulty of statistics to detect them [9]. Fuzzy logic provides a tool that enables slant a step beyond the boundaries that the mathematics of accurate data has demarcated between models and reality. Fuzzy logic provides another way of thinking that can treat this type of data and model situations where the change of state variables can predict behavior in the dependent variables, not linked by correlation, regression or deterministic models, but by a cascade of logical predicates and the logical correspondence with warning color matching, that guide the decision making process.

References

1. Bouchet, A., Pastore, J.I., Andrade, R.E., Brun, M., Ballarin, V.: Arithmetic Mean Based Compensatory Fuzzy Logic. International Journal of Computational Intelligence and Ap-plications 10, 231–243 (2011)
2. Brown, J.H.: Macroecology. University of Chicago Press, Chicago (1995)
3. Flueck, W.T., Smith-Flueck, J.M., Monjeau, J.A.: Protected areas and extensive produc-tions systems: a phosphorus challenge beyond human food. BioScience 61 (2011)
4. Meschino, G.J., Espin Andrade, R.A., Ballarin, V.L.: A framework for tissue discrimina-tion in Magnetic Resonance brain images based on Predicates Analysis and Compensatory Fuzzy Logic. International Journal of Intelligent Computing in Medical Sciences and Image Processing 2, 1–16 (2008)

5. Monjeau, J.A., Marquez, J., Zuleta, G.: Aproximación metodológica para establecer los principales conflictos y priorización de la toma de decisiones en áreas protegidas. In: Ediciones Universidad Atlántida Argentina, pp. 240 + illustrations. Ediciones Universidad Atlántida Argentina, Mar del Plata (2006)
6. Monjeau, J.A., Solari, H.: Conservacionismo. In: Biagini, Roig, eds. (2009)
7. Monjeau, J.A.: Conservation crossroads and the role of hierarchy in the decision-making process. Natureza & Conservação 8, 1–8 (2010)
8. Monjeau, J.A.: Evaluación y actualización del plan de manejo del área natural protegida península Valdés, patrimonio natural de la humanidad (2011)
9. Monjeau, J.A.: La sombra de la experiencia. Dendron 1 (1989)
10. Nabte, M.: Criterios ecológicos para la toma de decisiones para la conservación de mamí-feros de la Península Valdés. Ph.D. Mar del Plata, p. 214. Universidad Nacional de Mar del Plata (2010)
11. Nabte, M., Saba, S.I., Monjeau, J.A.: Mamíferos terrestres de la Península Valdés: lista sistemática comentada. Mastozoología Neotropical 16 (2009)
12. Popper, K.R.: Búsqueda sin término. Una autobiografía intelectual. Editorial Tecnos, Madrid, España (1985)
13. Terborgh, J.: Requiem for Nature. Island Press, Washington (2004)
14. Wittgenstein, L.: Investigaciones filosóficas. Editorial Altaza, España (1999)

Framework for the Alignment of Business Goals with Technological Infrastructure

Roberto Pérez López de Castro, Pablo M. Marin Ortega,
and Patricia Pérez Lorences

Abstract. There is not a full integration between business and technological domains in organizations; it creates problems with the availability of the necessary information for the decision-making process, and the under use and exploit of installed Information Technologies (IT) capabilities. In the present investigation is proposed a framework to solve this problem. The framework consists of an enterprise architecture, several specific procedures, from which we proposed two global indicators; the first one for management control based on Compensatory Fuzzy Logic (CFL), which measures the strategy performance from the compensation of the indicators defined in a Balanced Scorecard (BSC); the other one is an indicator to evaluate the IT Management (IGTI) based on the assessment of process maturity expressing a single comprehensive measure. The framework considers the alignment among business requirements, business processes and IT resources taking into account the risks management and benefits.

1 Introduction

A survey of 385 finance and IT executives, by CFO Research Services, asked them to identify the drivers for poor information quality (IQ). Nearly half of them pointed (45 percent) the non-integration of IT systems and the variability of business processes as an acute problem that constrains management's ability to work effectively and focus on high-value activities. Approximately the same number agrees that finance and business units alike spend too much time developing supplemental reports and analysis. Other disappointing and productivity sapping by products of poor information quality include that "multiple versions of the truth," misguides incentive programs, and leads to unrealistic plans and budgets [11].

In fact, 61 percent of respondents say they could still do a better job of just making sure the financial information they generate accurately reflects the performance of their businesses. The business impact of this poor IQ, say respondents, includes widespread decision-making problems that are often tied to inaccurate, untimely, and irrelevant information.

Roberto Pérez López de Castro · Pablo M. Marin Ortega · Patricia Pérez Lorences
Central University of Las Villas, Cuba
e-mail: {robertop,pablomo,patriciapl}@uclv.edu.cu

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_7,

Recently, studies showed how managers of companies had begun using Enterprise Architectures (EA) for various reasons. A study in 2007 conducted by the Society for Information Management's (SIM) Enterprise Architecture Working Group (EAWG) to better understand the state of the practices of the EA in organizations and assessed the status of the IT capabilities of the organizations to develop an EA [9], the main results shows that 85% of respondents were in full agreement with the EA facilitate systemic change, 80% think that EA is a tool to align the business goals and IT initiatives, 90% are strongly agreed that the EA provide a guide to the business, data, applications and technologies, nearly 90% agreed that EA is a tool for business planning , 84% strongly agreed that the EA is a tool for decision making, more than 83% strongly agreed that the purpose / function of the EA is to align business goals and IT investments, over 87% agreed that improves the interoperability of information systems.

The main goal of this research is to develop a framework that facilitates the integration of both: business and technology domains, as a tool that contributes to the improvement of necessary information availability for the decision making process. The framework considers the alignment among business requirements, business processes and IT resources taking into account the risks management and benefits.

2 Framework Description

The framework proposal is based structurally on the matrix proposed by Zachman (Figure 1), considering their first four rows, in which are defined: the business strategy, the business model, the system model and the technology infrastructure.

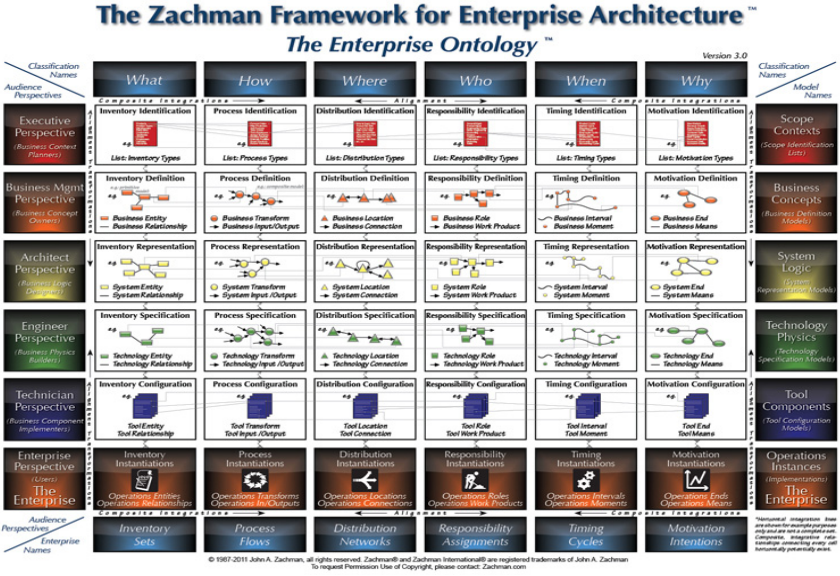


Fig. 1 Zachman Framework [14]

In addition we considered necessary to include an interface between business strategy and business model, in order to translate into operational terms the enterprise strategy. The general idea would be to construct a pyramid (architecture) that incorporates in each of its levels (rows) the specifications needed to support the upper level (Figure 2).

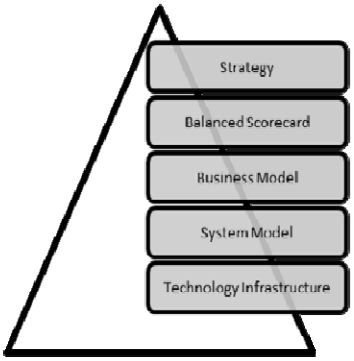


Fig. 2 Proposed Architecture

For each row were defined a set of tools, best practices, input and output elements. This document will only show the new elements; an indicator to measure the performance of the strategic plan, a semantic model to support the integration of business processes and an indicator to evaluate the level of IT Governance (I_{GTI}). These elements will be explained in details in the following sections.

3 Indicator to Measure the Performance of the Strategic Plan

In the literature we find different methods for the design and implementation of a Balanced Scorecard [2, 5, 7, 8, 10, 12, 13]. After their analysis we could concluded that there are a set of common elements in any process involving the application of this tool:

- Strategic plan review
- Identification of key success factors
- Identification and design of indicators
- Design of the strategic map
- Formulation of strategic projects through feedback

It is important to recognize the strategic map as the most important conceptual contribution of the balanced scorecard because it helps to appreciate the importance of each strategic goal, because it is linked to the Balanced Scorecard perspectives; which are arranged according to the cause - effect criterion.

In this paper we only focus in creating a global indicator to measure the strategy performance. One of the most important steps in the development of any strategic control tool is the development of strategic plans, through feedbacks, where the indicators capable of monitoring the management control, play an important role. Being it, one of the principal difficulties found in the literature used for this research.

In this sense, we proposed an indicator to measure of management control using compensatory fuzzy logic; it is capable to evaluate based on the behavior of the indicators defined in the BSC (Figure 3).

Step 1: Built the matrix $GI_{(m,k)}$

To prepare a matrix where all the indicators and the strategic goals appear defined for the organization. The prepared matrix, must be presented to a group of chosen experts who have to answer the question: How true is it, that the indicator "*k*" is an important element in the measurement of the fulfillment of the strategic goal "*m*"? The scale to be used would be a continuous scale between 0 and 1; where 0 would be the most false value and 1 the most truthful one.

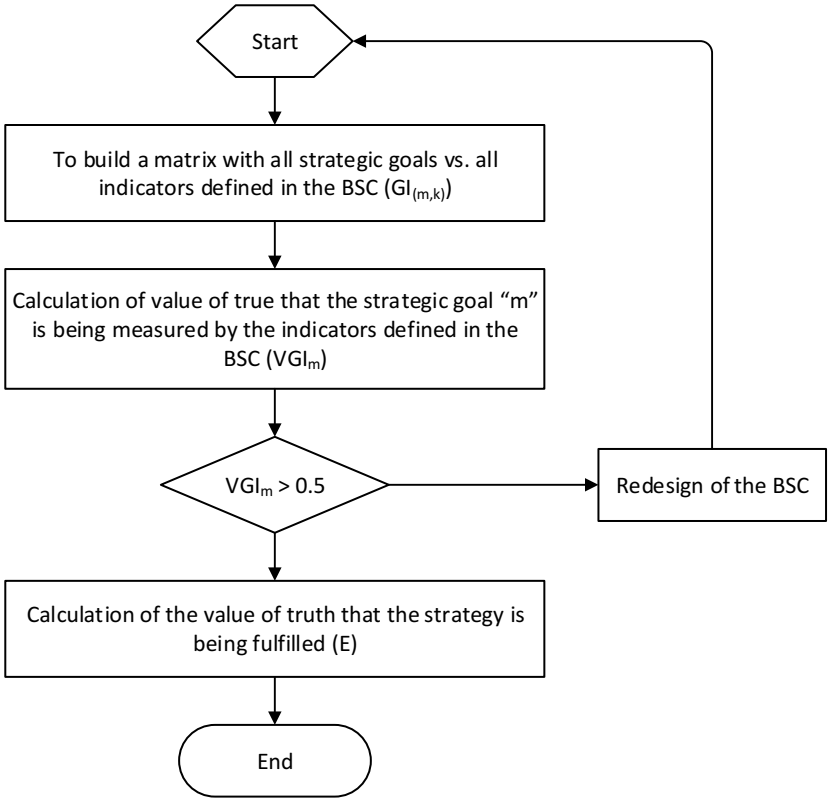


Fig. 3 Specific procedure for calculating the truth value of the strategic plan implementation

Step 2: Calculation of value of truth, that the strategic goal "m" is being measured by the indicators defined in the BSC (VGI_m)

With the previous information we can answer the question: How true it is, that the strategic goal "*m*" is being measured by the indicators defined in the BSC?

- A strategic goal is being measured if and only if exist indicators that measure it.

This can be expressed using compensatory fuzzy logic as:

$$VGI_m = \exists_k(VGI_{(m,k)}) \quad (1)$$

Where:

VGI_m: Value of truth that the strategic goal "*m*" is being measured by the indicators defined in the BSC.

$VGI_{(m,k)}$: Matrix with the truth value of the expert consensus that the presence of the strategic goal "*m*" is measuring by the indicator "*k*".

The people in charge of designing the BSC should be ensuring that the value obtained in $VGI_{(m)}$ for each goal, has a value greater than 0.5.The ideal value would be given by: maximizing $VGI_{(m)}$and minimize the number of indicators defined in the BSC.

Step 3: Calculation of the value of truth that the strategy is being fulfilled (E)

As a premise, must be ensured that the VGI_m value was more true than false for all strategic goals.

A strategy is being fulfilled if and only if all the important strategic goals are being met.

- A strategic goal is important if there are critical success factors that justify its approach.
- An important strategic goal is being met if and only if all the indicators defined in the BSC for its measurement are being met.

Based on the principles stated above and using compensatory fuzzy logic to compensate the indicator defined in the BSC, given as:

$$E = \forall_m(VG_m \rightarrow (\forall_k(VGI_m \rightarrow VI_k))) \quad (2)$$

Where:

E: Value of truth that the strategy is being fulfilled.

VG_m: Value of truth that the element "j" is a key success factor and in turn advises the strategic goal "m".

VGI_m: Value of truth that the strategic goal "m" is being measured by the indicators defined in the BSC.

VI_k: Value of truth of the criterion of measurement of indicator "k".

To calculate VI_k we propose to use the sigmoidal membership function.

Where:

$S = VI_k$: Value of truth of the criterion of measurement of indicator "k".

$X = I_k$: Calculated value of the indicator "k" according to the company.

Gamma (γ): Value acceptable. It would be equal to the value at which the indicator is considered acceptable.

Beta (β): Value almost unacceptable: It would be equal to the pre-image of a symmetric sigmoidal function for the optimal value defined for the indicator, or it would be the same β = (Value at which the indicator is acceptable - Value from which the indicator is optimal).

Alfa (α): Sigmoidal function parameter.

The use of compensatory fuzzy logic as a knowledge engineering tool, allows managers to better analyze the information needed to design a BSC, given that the concepts involved in strategic business, are essentially subjective and imprecise. The proposed indicator for measuring the strategy gives managers a tool for consistent feedback to the development of strategic projects to meet the dynamic changes in the environment. With the combination of metaheuristics which are able to determine from a broad set of indicators and goals, and compensatory fuzzy logic we would define the ideal balanced scorecard to maximize VGIm value and minimize the number of indicators defined in the balanced scorecard.

4 Semantic Model to Support the Integration of Business Processes

When it comes to Business Process Management (BPM) the main problem found is the non-existence of an integration solution that allows merging the semantically supported modeling and orchestration of business processes, with interoperability solutions at data level. Despite the fact that process modeling languages allow the combination of process definition (their structure) with Web services orchestration (as process execution structure), they are not able until now to define the integration mechanisms for heterogeneous data schemas. On the other hand, the solutions designed to achieve interoperability of the information systems don't achieve a complete integration since they don't include the integration at processes level.

Hepp [6] argues that BPM is "*the approach to manage the execution of IT-supported business operations from a business expert's process view rather than from a technical perspective*" [6]. In this definition Hepp points out the main problem in initial stages of BPM, the divorce between the Business Perspective and the IT Perspective; creating a need for a unified view on business processes in a machine-readable form that allows querying their process spaces by logical expressions, with the lack of such a machine-readable representation of their process space as a whole on a semantic level been a major obstacle towards mechanization of BPM [6]. As it turns out is quite difficult for business analysts to use the existing tools due to high complexity, and equally difficult to IT specialists to understand the business needs and what a process represents.

Another issue presents itself whenever two or more companies need to collaborate, or even applications within the same company, they invariably need to exchange information electronically and integrate the results in each system. In an ideal scenario, this information exchange and integration is performed in an automatic way allowing business partners to interoperate information seamlessly. However, because of the large number of diverse information systems the data format (syntax) of each exchange (message) usually differs from company to company, or sometimes even within the same company if more than one software product is used. The situation is similar to the Tower of Babel involving many different people that want to work together on a specific task without understanding each other.

This makes it very challenging to exchange information in an interoperable way. Interoperability in this context means "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" (IEEE Standard Computer Dictionary). Even when you have internal information integrated (said in a modern ERP system), to manage the information exchange with other companies in the Supply Chain, users either have to agree on a common model to express the data to be exchanged or they have to individually translate the data received from the business partner to the data format they own.

The traditional solution consists on providing interfaces that allows the access of providers and clients to the necessary data for the management. When the vertical integration degree in the chain is high, it could be viable the adoption or agreement of a common data model, but when this it is not the case, it is very complex the task of providing a different interface for each provider or client with a different schema. The use of a unique interface is a solution of compromise that is able to transfer the responsibility of translating the data schemas to those that you access, from providers to clients. This type of solutions is recognized under the common name of "integration projects".

The proposed model combines semantic Business Process Modeling, with support from Supply Chain Operations Reference model (SCOR) ontology; this will allow modeling the logistic integration processes in the Supply Chain based on a recognized standard. The main goal is to define the mechanisms to orchestrate the integration processes in the Supply Chain and provide the business analysts in charge of modeling the new processes common knowledge base to assist the proceeding.

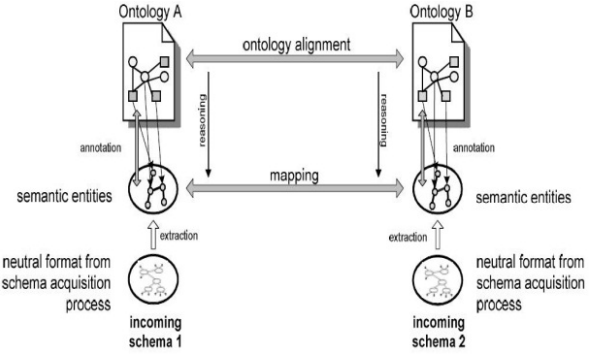


Fig. 4 Sematic Mapping in STASIS [3]

The neutral representation of incoming schemata provides the basis for the identification of the relevant semantic entities being the basis of the mapping process; based on the annotation made with respect to the ontologies and on the logic relations identified between these ontologies, reasoning can identify correspondences on the semantic entity level and support the mapping process. In Figure 4 we show an example implementation of how this is perceived in project STASIS [3].

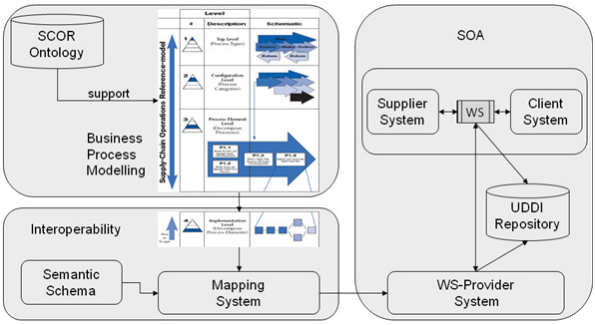


Fig. 5 Proposed Model for semantically supported Business Process Integration

The main contribution in this aspect will be the addition of the fourth level of the SCOR process model from each company to the mapping process; based on methodologies [1] to align, map and merge ontologies; the hierarchical structure of the SCOR model should prove to be useful to enhance the results of this procedure.

As result from the mapping system you get a transformation language (could be expressed in XSLT), it will be used to generate the web service which will be used as translator for the different schemas in the current information interchange. This way seamless interoperation at the process level is achieved including these translators in the executable process model of the workflow.

5 Indicator to Evaluate the Level of IT Governance (I_{GTI})

At this stage we assess IT governance in the organization to which we proposed an indicator to evaluate the level of IT Governance (I_{GTI}). The steps to develop this stage are:

Step 1: Determination of the relative importance of domains and control objectives

The first step of this stage is to define the domains and control objectives to diagnose. From the framework COBIT 4.1[4] was made a general proposal, which must be adapted by the team considering elements to be added or removed depending on the characteristics of the organization. Then we proceed with the collection, verification and analysis of information for which we designed a set of interview guides. Based on previous results, is passed to determine the maturity level of each control objective according the maturity models defined by COBIT.

Step 2: Assessment of the domains and control objectives

We propose the assessment of each control objective through the following expression:

$$EOC_{dg} = \frac{W_{dg} \times NM_{dg}}{5} \tag{3}$$

Where:

EOC_{dg}: Assessment of the control objective "d" of the domain "g"

W_{dg}: Weight (relative importance) of the control objective "d" of the domain "g"

NM_{dg}: Maturity level of the control objective "d" of the domain "g"

The sum of the assessments of the control objectives is the domain result

$$RD_g = \sum_{d=1}^{m_g} EOC_{dg} \tag{4}$$

Where:

RD_g: Result of the domain "g"

The evaluation of each domain is calculated using the following expression:

$$ED_g = w_g \times RD_g \times 100 \tag{5}$$

Where:

ED_g: Evaluation of the domain "g"

w_g: Weight of the domain "g"

Step 3: Determination of indicator I_{GTI}. Graphical representation of results

The Indicator to evaluate the level of IT Governance (I_{GTI}) is calculated as shown:

$$I_{GTI} = \sum_{g=1}^{4} ED_g \tag{6}$$

We define the scale for assessment of IT Governance from Non-existent level to Optimized, as shown in table 1. We propose a graphical representation of results, using control radars and Cause-Effect graphical.

Table 1 Scale for assessment of IT Governance

Intervals I_{GTI} (%)	IT Governance Assessment
($95 \leq I_{GTI} \leq 100$)	Level 5: Optimized
($75 \leq I_{GTI} < 95$)	Level 4: Managed
($55 \leq I_{GTI} < 75$)	Level 3: Defined
($35 \leq I_{GTI} < 55$)	Level 2: Repeatable
($15 \leq I_{GTI} < 35$)	Level 1: Initial/Ad Hoc
($I_{GTI} < 15$)	Level 0: Non-existent

Step 4: Preparation of evaluation report

From the results obtained in the previous stages, this step is required to produce a report which includes assessing: the analysis of IT resources and alignment to business objectives, analysis of IT risk management, the analysis of the characterization of worker satisfaction, and a list of domains and control objectives that reflected greater difficulty in evaluating management. Should be noted the main problems affecting the governance of IT in the organization.

For the calculation of all the expressions defined above can be used math packages such as Matlab, but it is also left well defined expressions for easy matrix calculation in Excel spreadsheet, if the user does not have access to any of those packages.

6 Conclusions

The proposed indicators contributed as new measures for management control and strategy performance from the compensation using CFL of the indicators defined in a Balanced Scorecard; and the assessment of process maturity expressing a single comprehensive measure in the global IT Governance indicator.

The lack of a commonly accepted schema is still a major handicap for Business Process Management. Competing standardization bodies have proposed numerous specifications and competing schemas that capture only parts of the business process life cycle. We proposed an integration model helping to merge the heterogeneous proposals for BPM in an attempt to bridge the gap between the IT and business domains.

The proposed framework, as well as all the tools that conform it, help to improve the availability of the necessary information for the decision making process, based on the integration of the business and technological domains.

References

1. Abels, S., Haak, L., Hahn, A.: Identifying ontology integration methods and their applicability in the context of product classification and knowledge integration tasks. WI-OL, Oldenburg (2005)
2. Amat, J.: Control de Gestión. Una Perspectiva de Dirección. Gestión 2000. S.A, Barcelona (2000)
3. Beneventano, D., Dahlem, N., El Haoum, S., Hahn, A., Montanari, D., Reinelt, M.: Ontology-driven Semantic Mapping. In: I-ESA 2008, Berlin (2008)
4. COBIT. Control Objectives for Information and related Technology (2007), from COBIT: `http://www.itgi.org/COBIT.htm` (retrieved April 17, 2012)
5. Consulting, S.M.: Guía práctica de implantación del Balanced Scorecard. Mapas Estratégicos. P. Hall (2000)
6. Hepp, M., Leymann, F., Domingue, J., Wahler, A., Fensel, D.: Semantic Business Process Management: A Vision Towards Using Semantic Web Services for Business Process Management. In: IEEE - ICEBE (2005)

7. Kaplan, R.N.: The Strategy Focused Organization: How Balanced Scorecard Companies Thrive in the New Business Environment. Harvard Business School, Boston (2001)
8. Kaplan, R.N.: Cómo utilizar el Cuadro de Mando Integral para implementar y gestionar su estrategia. Gestion 2000, Barcelona (2005)
9. Kappelman, L.: The SIM Guide to Enterprise Architecture. Broken Sound Parkway NW, Suite 300. Boca Raton, FL 33487-2742.: Taylor & Francis Gruop, an informabusiness (2010)
10. Marin Ortega, P.: Fuzzy method for Balanced Scorecard Design. In: Espin Andrade, R., Marx Gómez, J., Racet Valdés, A. (eds.) Towards a Trans-disciplinary Technology for Business Intelligence, pp. 73–89. Shaker Verlag, Oldengurg (2011)
11. Myers, R.: IT Executives Seek to Boost Information Quality (2005), Retrieved from CFO Research Services http://www.cfo.com/article.cfm/5545859/c_2984335/?f=archives
12. Prieto, D.: Procedimiento para el diseño del CMI en pequeños y medianos hoteles. Samuel Feijó, Santa Clara (2007)
13. Rivero, M.: Procedimiento para el diseño del CMI en la División Centro TRD Caribe. Samuel Feijó, Santa Clara (2006)
14. ZIFA. Zachman International Enterprise Architecture, from Zachman Website: http://www.zachman.com/ (retrieved April 17, 2012)

Time Series Classification with Motifs and Characteristics

André Gustavo Maletzke, Huei Diana Lee,
Gustavo Enrique Almeida Prado Alves Batista, Cláudio Saddy Rodrigues Coy,
João José Fagundes, and Wu Feng Chung

Abstract. In the last years, there is a huge increase of interest in application of time series. Virtually all human endeavors create time-oriented data, and the Data Mining community has proposed a large number of approaches to analyze such data. One of the most common tasks in Data Mining is classification, in which each time series should be associated to a class. Empirical evidence has shown that the nearest neighbor rule is very effective to classify time series data. However, the nearest neighbor classifier is unable to provide any form of explanation. In this chapter we describe a novel method to induce classifiers from time series data. Our approach uses standard Machine Learning classifiers using motifs and characteristics as features. We show that our approach can be very effective for classification, providing higher accuracy for most of the data sets used in an empirical evaluation. In addition, when used with symbolic models, such as decision trees, our approach provides very compact decision rules, leveraging knowledge discovery from time series. We also show two case studies with real world medical data.

André Gustavo Maletzke · Huei Diana Lee · Wu Feng Chung
Laboratory of Bioinformatics (LABI), Western Paraná State University (UNIOESTE), Foz do Iguaçu, Brazil
e-mail: {andregustavom,huei}@gmail.com

Gustavo Enrique · Almeida Prado Alves Batista · Wu Feng Chung
Laboratory of Computational Intelligence (LABIC), Institute of Mathematical and Computing Sciences (ICMC), University of São Paulo (USP) São Carlos, Brazil
e-mail: gbatista@icmc.usp.br

Cláudio Saddy Rodrigues Coy · João José Fagundes
Department of Surgery, Coloproctology Service, School of Medical Sciences, State University of Campinas (UNICAMP), Campinas, Brazil
e-mail: wufengchung@gmail.com

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_8, © Springer-Verlag Berlin Heidelberg 2014

1 Introduction

The Data Mining process has been applied to several areas with the objective of extracting relevant and interesting knowledge from large data sets, so that such knowledge can be used to support the decision-making process. Knowledge extraction from time series is a subject that has attracted the attention of researchers and experts in several application areas, since several of those areas generate time-oriented data. A few examples of application areas for time series are the analysis of stock market, manufacturing processes, amino acid sequences data, and the medical area with monitoring of chemical, physical and biological variables that describe a patient clinical state. Time series data consist of an ordered set of observations, about a determined phenomenon, measured along a time period, being the temporal characteristic the main interest aspect for the Data Mining process.

Machine Learning is one of the research areas that contribute with algorithms and methods used in the Data Mining process. However, dealing with temporal and sequential data is a challenge for most of the Machine Learning algorithms, since many of them assume the data are independent and identically distributed (i.i.d.). In contrast, time series data have a natural order, and consequently the probability of occurrence of an observation in a certain time instant usually depends of previously observed values.

This chapter presents an approach to mine temporal data using characteristic extraction and motif discovery. The proposed approach consists of two strategies, the extraction of global characteristics and the discovery of motifs. The first strategy is widely used in time series research, and describes the global data features using, for instance, descriptive statistics. However, in some application domains, the global data behavior may not evidence some important details and a more detailed analysis may be necessary. The second strategy uses motifs discovery and aims to provide a local view of the temporal data. The proposed approach has the objective of unifying the global view given by the general characteristics with the local view provided by the motifs identification.

We show examples of use of the proposed approach with experiments performed in time series data sets available in literature, specifically from the UCR *Time Series Classification/Clustering* archive. We also show two case studies with time series from the medical area for the classification of electrocardiograms and anorectal manometry exams.

2 Definitions and Notations

This section presents some definitions and terminologies used in this chapter.

Definition 1 (Time Series) [2] A time series Z is an ordered collection of real-valued observations of length m, that is, $Z = (z_1, z_2, \ldots, z_m)$ with $z_t \in R$, for $1 \leq t \leq m$.

Definition 1 states that a time series is a collection of real values ordered temporally. For some problems, it may be necessary to transform the real-valued

observations into symbolic values. In doing so, we avail a wide class of algorithms developed exclusively to work with symbolic sequences, for instance, hashing and SuffixTrees data structures as well as algorithms to string matching [6]. The definition of a symbolic time series is presented in Definition 2.

Definition 2 (Symbolic Time Series) [8] A time series Z of length m' is a collection of ordered values $Z = (z_1, z_2, \dots, z_m)$ with $z_{t'} \in \sum$, for $1 \leq t' \leq m'$ where $\sum$, is a finite alphabet of symbols.

Some methods rely on analyzing small portions of a time series with the objective of, for example, identifying local characteristics or reducing the search space. These small portions are named subsequence (Definition 3) and are extracted with a sliding window (Definition 4).

Definition 3 (Subsequence) [2] Given a time series Z of length m, a subsequence C of Z is a continuous sample of Z of length n, with $n << m$. Therefore, $C = (z_p, \dots, z_{p+n-1})$ for $1 \leq p \leq m - n + 1$.

Definition 4 (Sliding Window) consists of extracting all subsequences of length n of a time series Z of length m, resulting *in* subsequences $(z_1, \dots, z_n)$, $(z_2, \dots, z_{n+1}), \dots, (z_i, \dots, z_{n+i-1})$, for $1 \leq i \leq m - n + 1$.

As previously mentioned, the proposed method jointly applies two strategies to construct the input for Machine Learning algorithms. The characteristics[1] extraction and motifs identification are used to construct an attribute-value representation. Features are typically related to descriptive statistics such as average, standard deviation and maximums and minimums which supply information on the global behavior of a time series. In contrast, motifs provide information about the existence of local behaviors. Motifs can be understood as a frequent subsequence present in a time series that have morphological similarity.

In the following definitions, concepts related to motifs are formalized, starting with the match concept, necessary to determine the similarity between two subsequences.

Definition 5 (Match) [2] Given a positive real number r and a time series Z containing a subsequence C beginning at position p and another M in position q, in which the distance between the two objects is denoted by D, if $D(C, M) \leq r$, then M is similar to C.

Although the match definition is simple and intuitive, in some situations two subsequences can be considered similar due to the fact that they share a large number of observations. This fact contributes to generate what is called a false motif. In these cases, it is necessary to apply the concept of trivial match formalized in the Definition 6.

Definition 6 (Trivial Match) [2] Given a time series Z and subsequences C and M beginning at positions p and q, trivial match between M and C occurs if $p = q$ or if there is not a subsequence M' beginning at q' such that $D(C, M') > r$, for $q < q' < p$ or $p < q < q$.

[1] In this chapter, the words characteristics, attributes and features are used indistinctly.

From the presented concepts, the idea of motifs can be formalized according to Definition 7.

Definition 7 (k-Motifs) [7] Given a time series Z and a threshold r, the most significant motif, called Motif-1, located in Z is the C_1 subsequence which has the largest amount of non-trivial matches. Consequently, the most significant k^{th} motif present in Z is the subsequence C_k having the k^{th} greatest amount of non-trivial matches and satisfying the condition $D(C_k, C_i) > 2r$, for all $1 \leq i < k$.

It is important to observe that the $D(C_k, C_i) > 2r$ is necessary to guarantee that the subsequences are mutually exclusives. This idea is graphically illustrated in Figure 1 (a) considering a minimum distance of just one r and in Figure 1 (b) considering a minimum distance of $2r$.

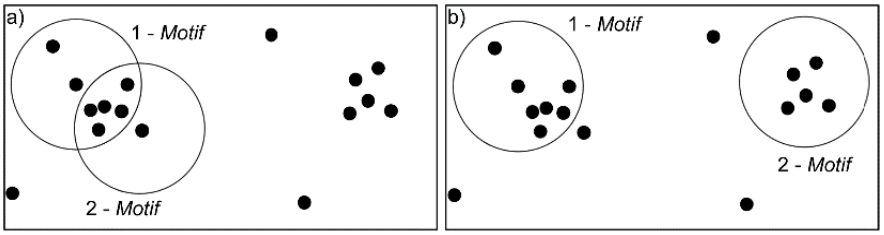


Fig. 1 Schematic representation of the k-motifs concept [7]

3 Mining Time Series Data

Our proposed method allows applying the Data Mining process to time series data, using standard Machine Learning techniques. The method consists of three main phases. The first phase preprocesses the data aiming to solve some common problems found on temporal data, such as differences in scale and time intervals. The second phase, performed after the data preprocessing, applies the two mentioned strategies, i.e. the extraction of global characteristics[2] and motifs identification. The third phase uses Machine Learning techniques to learn from the attribute-value table generated in the previous phase. Next, we present each phase in greater detail.

3.1 First Phase – Time Series Preprocessing

Data preprocessing is one of the most important tasks in the Data Mining process, since the data quality directly affects the quality of the induced knowledge [11]. Figure 2 presents some of the most common problems found in a time series data.

[2] For simplicity, we use the term extraction of characteristics instead of extraction of global characteristics in the remaining of this text.

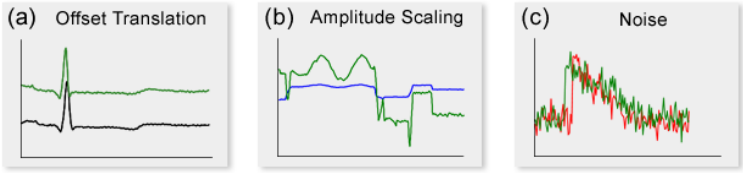


Fig. 2 Examples of problems that can be found in time series

The problems illustrated in Figure 2 are time series measured with different offsets (a), with different amplitudes (b) and with noise (c). Such problems can make the data analysis more difficult hindering the true behavior of the phenomena being measured. Another problem not illustrated in Figure 2 is the presence of missing values, a very common issue in real world data.

3.2 *Second Phase – Characteristics Extraction and Motif Identification*

The objective of this phase is the identification of attributes through characteristics and motifs from time series data. This phase is divided in two independent stages, presented next.

Stage 1: Characteristics Extraction

This stage defines the characteristics that will be extracted from the time series data set. As previously noted, one of the objectives of the proposed method is to provide global and local descriptions of the time series. The extracted characteristics can be descriptive statistics such as mean values, maximum and minimum, and variance as well as domain dependent characteristics obtained, for instance, by interviews with experts. Figure 3 presents a schematic representation of this stage. In this example, two time series (T_1, and T_2) are represented in an attribute-value table using five characteristics Ca_1, Ca_2, Ca_3, Ca_4 and Ca_5. The values occurring in the table, Vca_{ij} with i=1...2 and j=1...5, are the characteristics measured over the two time series.

Fig. 3 Schematic representation of the characteristic extraction stage [8]

Notice that the relevance of the characteristics to be used as features in the classification problem is data dependent. Therefore, choosing the right characteristics is a task that requires knowledge about the data as well as the application domain.

Stage 2: Motifs Extraction

The naïve algorithm for motif identification evaluates all possible matchings and, therefore, has time complexity of $O(m^2)$, being m the size of the time series. Such time complexity can be considered very high and will result in large execution times for most classification problems. Alternative approaches have been proposed aiming to reduce the time complexity [3, 14]. In this chapter we use the approach proposed by [1], namely *Random Projections* in conjunction with the proposal of [9]. The second stage, motifs identification, is divided in three steps, as described next:

Step 1 – Subsequence matrix building: the process of building a Subsequence Matrix (SM) consists of extracting all subsequences of length n from the time series, using a sliding window. Each subsequence is transformed in a string using the *Symbolic Aggregate aproXimation* (SAX) method [7]. The SAX discretization uses an alphabet, which size should be provided by the user. Figure 4 (a) presents an example with an alphabet of three symbols (a, b, c) in which subsequences of length 16 are extracted with a sliding window, and each subsequence is discretized using SAX giving origin to string of length n_{sax}=4.

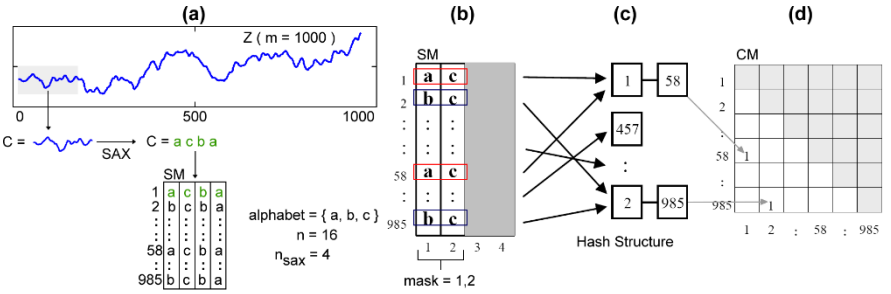


Fig. 4 Representation of the motif identification process, adapted from [2]

Step 2 – Collision matrix building: the Collision Matrix (CM) is used to identify subsequences that are likely motifs. Such matrix, initially null, has number of rows and columns equal to the number of rows of matrix SM. CM is filled in using an iterative process. In each iteration, a different randomly chosen mask is used to indicate which columns of SM are currently active. The active columns form a hash key, and the locations of the subsequences are inserted in a hash table according to those keys – Figure 4 (c). For instance, in Figure 4 (b) the current mask is (1, 2); therefore, the subsequences in the rows (1 and 58) and (2 and 985) of SM will collide since they have the same hash key, considering only the values of columns 1 and 2. At the end of each iteration, MC is updated by counting the number of subsequences that collided – Figure 4 (d). The process is repeated for a determined number of times. Each iteration requires that the hash structure is cleaned and filled in again.

Step 3 – Collision matrix analysis: a large value in a CM position is an indicative, although it is not a guarantee, of the existence of a motif. In order to identify a motif, we should verify in the CM matrix the location of the subsequences that resulted in the largest number of collisions. The distance between those subsequences is calculated over the original (real-valued) data. If two subsequences are inside a radius *r*, they are considered motifs. Other subsequences may also be inside the same radius and need to be also identified as motifs [2]. In general, a sequential search is performed using the subsequence defined as motif over the entire time series.

This strategy to identify motifs is an interactive and probabilistic process. Since this strategy does not explore the entire search space, it is more efficient than other approaches such as the naïve brute-force algorithm [2]. In contrast, the probabilistic strategy may lead to false negatives, i.e., it may not identify all existing motifs. An empirical comparison between the probabilistic and the brute force algorithm was performed in [9]. In that experiment, the probabilistic approach was significantly faster than the brute-force and both methods identified the same motifs.

Figure 5 illustrates schematically the process of identifying motifs and mapping them to an attribute-value table. In this example, four motifs, Mo_1, Mo_2, Mo_3 e Mo_4, are identified in two time series T_1 and T_2. The attribute-value representation indicates the occurrence (1) or not (0) of each motif in the time series.

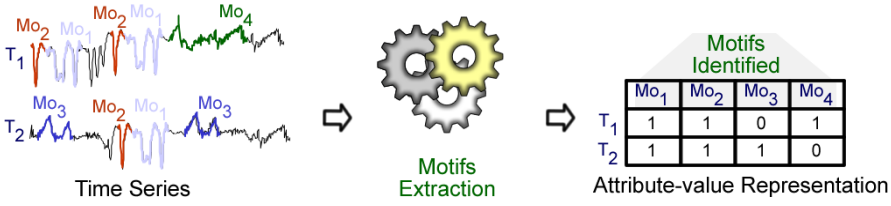


Fig. 5 Schematic representation of motif identification [8]

Motifs allied with extracted characteristics can be represented together in a final attribute-value table. In such table, each attribute is associated with an extracted characteristic or with a presence (0 or 1), frequency or location of an identified motif. Figure 6 illustrates this process.

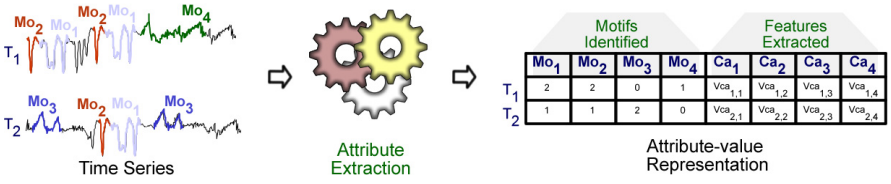


Fig. 6 Attribute-value representation obtained with characteristics and motifs

3.3 Third Phase– Knowledge Extraction from Time Series Databases

This phase defines the Machine Learning algorithms that will be used to build a prediction model or to explore and understand the data. Most of the previous work applied non-symbolic algorithms due to the temporal nature of the data. Our experiments show we can obtain competitive classification performance even when symbolic algorithms are used. In addition, our representation using motifs and characteristics usually lead to very simple models that can be easily interpretable by non-experts.

4 Case Study – Benchmark Datasets

This section presents a case study of the proposed method applied to six temporal benchmark datasets: *FaceFour*, *Coffee*, *Beef*, *Trace*, *Wafer* and *Gun-Point*, available at the UCR *Times Series Classification/Clustering* data base [5]. Table 1 presents a summarized description of these datasets.

Table 1 Description of the benchmarks datasets

Dataset	#Ex.	Time Series Length	Number of Classes	%Class	Trivial classifier Error
FaceFour	112	350	1	19.6%	69.6%
			2	30.4%	
			3	25.9%	
			4	24.1%	
Coffee	56	286	1	48.2%	51.8%
			2	51.8%	
Beef	60	470	1	20.0%	80.0%
			2	20.0%	
			3	20.0%	
			4	20.0%	
			5	20.0%	
Trace	200	275	1	25.0%	75.0%
			2	25.0%	
			3	25.0%	
			4	25.0%	
Wafer	7164	128	1	10.6%	10.6%
			2	89.4%	
Gun-Point	200	150	1	50.0%	50.0%
			2	50.0%	

Our evaluation compares the proposed approach to a frequently used strategy to classify time series, in which data are directly given to a Machine Learning algorithm. This approach is widely used in the area, and is able to provide excellent results in terms of classification error. In particular, the *k*-nearest neighbors (*k*NN), with $k = 1$, is widely used in classification of time series, frequently providing results that are very difficult to beat [4].

In this case study, models were induced using WEKA[3] Data Mining software [15] with *J48* and *k*NN algorithms with their default settings.

We use the average error rate as the main metric to assess our results. We chose this measure due its simple interpretability and because of its frequent use in time series classification papers; allowing direct comparison with other results in the literature. The average error was estimated with 2×5 fold cross-validation since most of the selected datasets have a limited number of samples, each sample corresponding to a time series. In order to facilitate the reproduction of results and comparison with methods proposed by other researchers, we selected publicly available datasets widely used in the area.

The comparison of the results was performed using the *t*-Student statistical test, using the mathematical and statistical environment R[4].

4.1 Results and Discussion

Table 2 presents the mean error rates and their standard deviations for the induced classifiers for each dataset. The lowest rates are shown in bold and those that present statistical significant difference (**s.s.d.**) are marked at the **s.s.d.** column.

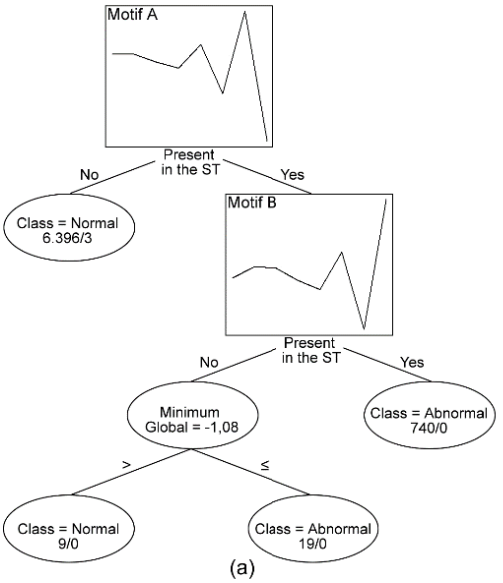
In this table, it is possible to notice that the symbolic models induced by *J48* over the attribute-value table created by the proposed method presented a higher accuracy than the traditional approach, in which *J48* induces a classifier over the raw data, with **s.s.d** in four of the six datasets. Similar results were obtained with the *k*NN algorithm.

Table 2 Mean errors and standard deviations of the induced models

	J48			kNN		
Datasets	Proposed method	Traditional approach	s.s.d.	Proposed method	Traditional approach	s.s.d.
FaceFour	13,4 (5,3)	19,8 (5,7)	†	3,7 (2,5)	7,0 (3,7)	†
Coffee	9,1 (6,0)	39,0 (10,6)	†	7,8 (4,6)	34,7 (10,3)	†
Beef	43,7 (14,4)	49,7 (9,6)		47,0 (9,7)	53,3 (11,8)	
Trace	1,1 (1,5)	22,7 (5,1)	†	0,0 (0,0)	13,8 (3,4)	†
Wafer	0,2 (0,2)	1,1 (0,4)	†	0,1 (0,1)	0,2 (0,1)	†
Gun-Point	11,1 (5,8)	12,2 (6,6)		7,2 (6,1)	6,6 (1,4)	

[3] http://www.cs.waikato.ac.nz/ml/weka/
[4] http://www.r-project.org/

The use of symbolic models enabled the joint representation of global characteristics, as well as motifs (local characteristics), which in real cases may provide relevant information to the domain experts. Figure 7 shows a tree built with the proposed method (a) and by the traditional approach (b) for the *Wafer* data set. Values between parentheses indicate the coverage of each branch till the leaf node.




```
TS[18] ≤ 1,91
|  TS[89] ≤ 0,96
|  |  TS[59] ≤ -1,07
|  |  |  TS[109] ≤ -0,41: 1 (25)
|  |  |  TS[109] > -0,41: -1 (121/1)
|  |  TS[59] > -1,07
        ⋮      ⋮
|  |  TS[108] > 0,38
|  |  |  TS[107] ≤ 0,40: -1 (4/0)
|  |  |  TS[107] > 0,40: 1 (12/0)
|  TS[129] > 0,56
|  |  TS[16] ≤ 1,06: 1 (15/0)
|  |  TS[16] > 1,06: -1 (5/0)
```

(b)

Fig. 7 Induced trees using the proposed method (a) and with the traditional ap-proach (b)

The knowledge represented by the tree in Figure 7 (a) is more understandable and intuitive in comparison with the knowledge represented in the tree shown in Figure 7 (b). Thus, the proposed method allowed the construction of models of easier interpretability, due to the use of low complexity characteristics as well as the use of motifs as attributes for the induction of models.

5 Case Study – Medical Area Datasets

This section presents a case study with medical data in which time series are derived from Electrocardiogram (ECG) and Anorectal Manometry (AM) examinations. The ECG is a test that aims to describe the electrical phenomena related to cardiac activity captured over time through electrodes pre-positioned in the body. This test, when performed in the full mode, is composed by twelve electrodes. However, many devices often use only a part of these electrodes [12]. The ECG data set used in this case study was obtained from the UCR repository *Times Series Classification/Clustering*. This data set is composed of 200 examinations represented by time series composed of 96 observations. Figure 8 (a) and (b) shows examples of two time series from the ECG dataset.

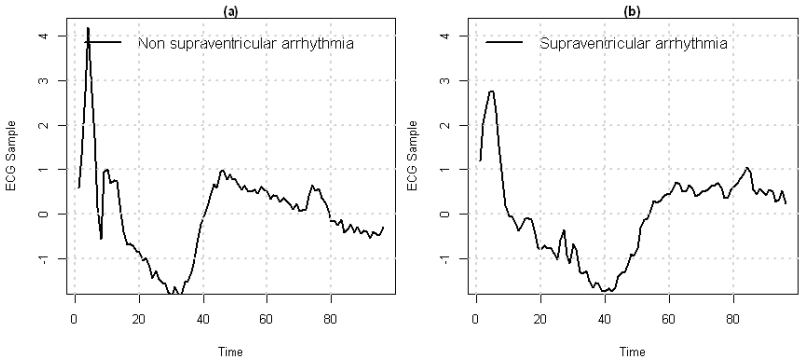


Fig. 8 Time series examples extracted from the ECG dataset

Two types of abnormal characteristics are described in this data set: patients with (66.5%) and without (33.5%) clinical signs of supraventricular tachycardia.

The second data set consists of an important exam for the diagnosis of patients with fecal incontinence. This condition, of varying degrees, is characterized by the patient's loss of the ability and capability to control the passage of feces and gases, in appropriate and socially acceptable time and place. The data from this exam were captured by eight sensors arranged radially in the anal sphincter [13].

This data set consists of 17 Anorectal Manometry exams, which were performed by the Coloproctology Service, Faculty of Medical Sciences at UNICAMP during the period of May/1995 to November/1996. Nine of these are normal patients and the other eight exams represent patients with abnormal fecal incontinence Grade III. The AM examination follows a protocol in which three sections of voluntary contraction are performed, in order to capture the work done by the anatomic anal region.

Due to the fact that each exam is composed of eight time series derived from each sensor, both global characteristics and motifs were extracted from these eight time series considering only the periods of voluntary contraction.

Figure 9 (a) shows a time series obtained from a sensor of a AM examination and Figure 9 (b) shows the bounded sections of the voluntary contraction periods with the help of experts.

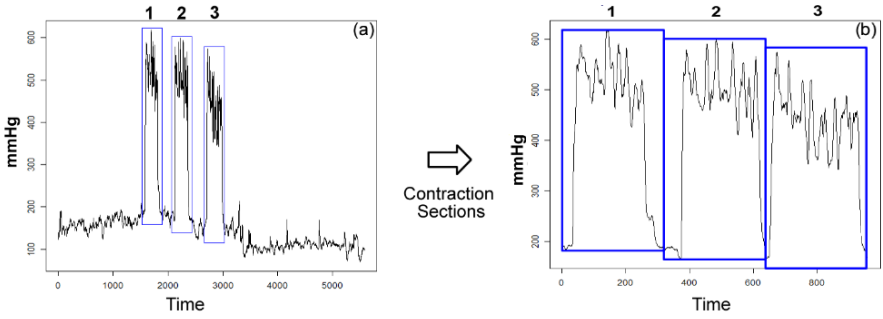


Fig. 9 Time series example obtained from the AM exam

Table 3 presents a summary of the characteristics of the ECG and AM datasets.

Table 3 Description of the ECG and AM datasets

Dataset	#Ex.	Time Series Length	Number of Classes	%Class	Trivial classifier Error
ECG	200	96	1	66,5%	33,5%
			2	33,5%	
AM	17	320	1	53,0%	47,0%
			2	47,0%	

In this case study the focus was concentrated at the task of classification by the induction of decision trees. Once again, induction of the models was also performed using the WEKA Data Mining software [15] using the *J48* algorithm with its default settings. Preliminary experiments were conducted using the holdout approach, in which the data set is divided into two subsets: a training data set and a testing data set. For the AM data set, 65% of the cases were used for training and 35% for the testing set. As for the ECG data set, both training and testing sets were composed by the same number of examples.

Sizes of motifs were determined according to previous work presented in Maletzke [8] and Maletzke et al. [10]. Thus, ten different values were used for the parameter size of motifs determined by the variation of 1% to 10% relative to the size of time series in each data set, in increments of 1%. As for the global characteristics, three of them were considered: mean value for all observations, as well as maximum and minimum values.

5.1 Results and Discussion

A major challenge of the presented approach is the definition of parameters such as the characteristics to be used and the size of motifs. These parameters are specific for each domain. Therefore, the results presented in this section refer to the use of motifs of different sizes for ECG and AM. Nevertheless, it may contribute as a guide for other domains, specially the medical ones.

As mentioned before, evaluation was performed considering the classification accuracy of symbolic models induced by the *J48* algorithm for the induction of decision trees.

Table 4 presents the error rates for the induced classifiers for each dataset. For the ECG data set, the proposed method presented an error value of one order of magnitude lower than the traditional approach. For the AM dataset, both approaches achieved the same performance. However, the proposed approach has the advantage that the generated models are significantly more understandable than the ones generated with the traditional approach.

Table 4 Comparative performance between the proposed and the traditional meth-ods for ECG and AM datasets

Dataset	Proposed method	Traditional approach
ECG	2.0	25.2
AM	16.7	16.7

In this case study, we opted for the induction of decision trees due to the fact that these models are more adequate when the objective is to interpret and understand of the details of the generated model, a characteristic that becomes quite complex when symbolic models are not used.

In a joint analysis among domain and computer science specialists, the conjunction of global information (characteristic) with local information (motifs) was considered to contribute to a more complete construction of models involving time series.

6 Conclusion

The chapter presented a new approach for mining time series by extracting global and local attributes. This approach can be applied in order to construct an attribute-value representation of temporal data, allowing the application of traditional Machine Learning methods and in particular symbolic methods such as decision trees. In contrast to the traditional analysis, that considers each temporal observation as an attribute, this approach identifies patterns (local and global) that enable to construct intelligible structures that simplify model analysis. In addition, the analysis of each component that conform this structure can reveal a novelty pattern in the data.

Two case studies of the proposed approach were presented in this chapter: one using benchmark datasets available in the scientific community and another using two real medical datasets. The results from both the case studies were considered satisfactory and with promising future research.

It is important to emphasize the contribution of the approach using motifs and global characteristics for the induction of symbolic models, especially in the medical field, as in these cases the analysis and the verification of the generated conclusions are fundamental issues to the research, diagnosis, treatment and prevention of diseases.

References

1. Buhler, J., Tompa, M.: Finding motifs using random projections. Journal of Computational Biology 9(2), 225–242 (2002)
2. Chiu, B., Keogh, E., Lonard, S.: Probabilistic discovery of time series motifs. In: Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining, New York, USA, pp. 493–498 (2003)
3. Ferreira, P.G., Azevedo, P.J., Silva, C.G., Brito, R.M.M.: Mining approximate motifs in time series. In: Todorovski, L., Lavrač, N., Jantke, K.P. (eds.) DS 2006. LNCS (LNAI), vol. 4265, pp. 89–101. Springer, Heidelberg (2006)
4. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. In: Proceedings of the VLDB Endowment, pp. 1542–1552 (2008)
5. Keogh, E., Zhu, Q., Hu, B., Hao, Y., Xi, X., Wei, L., Ratanamahatana, C.A.: The UCR Time Series Classification/Clustering (2011), http://www.cs.ucr.edu/~eamonn/time_series_data/ (accessed February 28, 2012)
6. Last, M., Kandel, A., Bunke, H.: Data Mining in Time Series Databases. Machine perception and artificial intelligence, vol. 57. World Scientific, Danvers (2004)
7. Lin, J., Keogh, E., Lonardi, S., Patel, P.: Finding motifs in time series. In: Proceedings of the Second Workshop on Temporal Data Mining at the Eighth Interntional Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta,Canada, pp. 53–68 (2002)
8. Maletzke, A.G.: Uma metodologia para a extração de conhecimento em séries temporais por meio da identificação de motifs e extração de características. Master Thesis. Universidade de São Paulo, São Paulo, Brazil (2009)
9. Maletzke, A.G., Batista, G.E., Lee, H.D.: Uma avaliação sobre a identificaçãode motifs em séries temporais. In: Anais do Congresso da Academia Trinacional de Ciências, Foz do Iguaçu, Paraná, Brazil, vol. 1, pp. 1–10 (2008)
10. Maletzke, A.G., Lee, H.D., Zalewski, W., Oliva, J.T., Machado, R.B., Coy, C.S.R., Fagundes, J.J., Wu, F.C.: Estudo do Parâmetro Tamanho de Motif para a Classificação de Séries Temporais de ECG. In: Congresso da Sociedade Brasileira de Computação, Workshop de Informática Médica, Natal, Rio Grande do Norte, pp. 1–10 (2011)
11. Michalski, R.S., Bratko, I., Kubat, M.: Machine learning and data mining. Wiley, Chichester (1998)
12. Olszewski, R.T.: Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data. PhD Thesis, Carnegie Mellon University, Pitts-burgh, PA (2001)
13. Saad, L.H.C.: Quantificação da função esfincteriana pela medida da capaci-dade de sustentação da pressão de contração voluntária do canal anal. PhD Thesis, Faculdade de Ciências Médicas da Universidade Estadual de Campi-nas, Campinas, SP (2002)
14. Tanaka, Y., Iwamoto, K., Uehara, K.: Discovery of time-series motif from multidimensional data based on mdl principle. Machine Learning 58(2-3), 269–300 (2005)
15. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques, 2nd edn. Elsevier, San Francisco (2005)

Solving Regression Analysis by Fuzzy Quadratic Programming

Ricardo Coelho Silva, Carlos Cruz Corona, and José Luis Verdegay Galdeano

Abstract. Regression analysis, which includes any techniques for modeling and analyzing several variables, is a statistical tool that focuses in finding a relationship between a dependent variable and one or more independent variables. When this relationship is found, some values of parameters are determined which help a function to best fit in a set of data observations. In regression analysis, it is also interesting to characterize the variation of the depend variable around the independent ones. A regression problem can be formulated as a mathematical programming problem, where the objective is to minimize the difference between the estimated values and the observed values. This proposal provides a fuzzy solution to the problem that involves all particular -punctual- solutions provided by other methods. To clarify the above developments, a numerical example about the price mechanism of prefabricated houses is analyzed.

1 Introduction

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. It is also used to understand, which among the independent variables are relates to the dependent one, and to explore the forms of these relationships. The performance of regression analysis in practice depends on the form of the data-generating process, and how it relates to the regression approach being used. However, it is more natural that the

Ricardo Coelho Silva
Institute of Science and Technology. Federal University of São Paulo,
Rua Talim, 330, Vila Nair,
12231-280, São José dos Campos, São Paulo, Brazil
e-mail: ricardo.coelho@unifesp.br

Carlos Cruz Corona · José Luis Verdegay Galdeano
Deparment of Computer Science and Artificial Intelligence, University of Granada, Spain
e-mail {carloscruz,verdegay}@decsai.ugr.es

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_9, © Springer-Verlag Berlin Heidelberg 2014

given data are imprecise and vague because they are observations that are subjective. In this context, some technique to deal with these imprecise data must be used, and one of these techniques is fuzzy regression analysis, which is more considered in systems where human estimations is influential.

Tanaka and Lee [5] classified the fuzzy regression analysis in two categories that are possibilistic regression analysis, which is based on possibility concepts, and least squares method, which minimizes errors between the given outputs and the estimated ones. The first work about possibilistic regression analysis was proposed by Tanaka et al. [8] in which is used a fuzzy linear system as a regression model. In [7] and [5], a interval regression analysis, which is a simplest version of possibilistic regression, based on quadratic programming is shown. It is an approach that unifies the possibility and necessity regression analyses. Other proposals based on the use of possibility concepts can be found in [3, 4, 6]. Another direction of fuzzy regression is fuzzy least squares approaches and some approaches can be found in [2, 12]. In [14], an approach that uses quadratic programming is shown and it integrates the two categories from fuzzy regression analysis, where it reconciles the minimization of estimated deviations of the central tendency with the minimization of estimated deviations in the spreads of membership functions.

A regression problem can be formulated as a mathematical programming problem, where the objective is to minimize the difference between the estimated values and the observed values. These data lack this kind of exact knowledge, and only approximate, vague and imprecise values are known. Moreover, these imprecise values can be dealt with fuzzy logic. In this case, the concept of fuzzy mathematical programming emerges when it is used. In this work, the input and output data are real numbers and the coefficients of these data are fuzzy numbers.

With this in mind, the goal of this paper is to apply a parametrical approach [1, 13] developed by authors based on fuzzy quadratic programming to solve this problem. This proposal provides a fuzzy solution to the problem that involves all particular -punctual- solutions provided by other methods. To clarify the above developments, a numerical example about the price mechanism of prefabricated houses is analyzed.

The paper is organized as follows: Section 2 briefly shows about fuzzy linear regression and how transforming a regression problem into mathematical programming problem with fuzzy costs. In this section, a novel approach is shown that solves quadratic programming problems with fuzzy costs where these are transformed into parametrical quadratic multi-objective programming problems. To clarify the above developments, a numerical example is analyzed in Section 3. Finally, conclusions are presented in Section 4.

2 Fuzzy Regression by Quadratic Programming

In this section, a brief introduction about fuzzy linear regression, which is used extensively in practical applications, is shown that has been based on linear programming when introduced by Tanaka et al. [8]. An approach that solves

quadratic programming problems with fuzzy costs, where these are transformed into parametrical quadratic multi-objective programming problems, is also described in this section.

2.1 Fuzzy Linear Regression

Fuzzy linear regression was the first type of regression analysis to be studied because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters.

Fuzzy linear regression models are often fitted using the least squares approach, which is used in this work, or possibilistic regression analysis.

In fuzzy linear regression, the model is a linear combination of the parameters, which are represented by fuzzy numbers, and independent variables that determines the dependent variable.

Table 1 Input-Ouput data

Samples	Output	Inputs
1	y_i	$x_{11}, \dots, x_{1n}$
.	.	.
.	.	.
.	.	.
m	y_m	$x_{m1}, \dots, x_{mn}$

In Table 1, the independent variable, x_{ij}, represents the jth input for the ith sample, when the dependent variable, $\hat{y}_i$, is an estimated value of the output, y_i, that is an observation for the ith sample.

The goal is to find the fuzzy parameters out that obtain the best estimation of the regression problem. Thus, the problem of fuzzy regression can be formulated by a straight line as:

$$\tilde{y}_i = \tilde{A}_0 + \sum_{i=1}^{m} \tilde{A}_i x_i \tag{1}$$

where $\tilde{A}_0$ is a fuzzy coefficient and $\tilde{A}_i (i = 1, \dots, m)$ is a vector of fuzzy coefficients.

Each fuzzy coefficient is characterized by a function membership defined by decision maker. The membership function can be defined as $\mu_{\tilde{A}_j}: R \to [0,1]$, $j \in \{0,1, \dots, m\}$.

In particular these membership functions will be supposed as:

$$\mu_{\tilde{A}_j}(y) = \begin{cases} 0 & if\ A_j^U \le y\ or\ y \le A_j^L \\ h_{\tilde{A}_j}(y) & if\ A_j^L \le y \le A_j \quad j \in J \\ g_{\tilde{A}_j}(y) & if\ A_j \le y \le A_j^U \end{cases} \tag{2}$$

The target is to minimize the difference between the observations and the estimations.

$$y_i - \tilde{y}_i = \epsilon_i \tag{3}$$

where ϵ_i is the residual, which is a random variable with zero mean. Thus, according to [8] the fuzzy regression analysis by using least square model can be transformed into a fuzzy quadratic programming problem as:

$$\begin{aligned} & min \sum_{i=1}^{m} \left(y_i - \tilde{A}_0 - \sum_{j=1}^{n} \tilde{A}_j x_{ij} \right)^2 \\ & s.t.\ A^t x_j + A^{U^t} |x_j| \ge y_j \\ & \quad A^t x_j - A^{L^t} |x_j| \le y_j \\ & \quad c_i \ge 0, \qquad i = 0,1,\dots,n \qquad j = 1,\dots,m \end{aligned} \tag{4}$$

2.2 *Fuzzy Quadratic Programming*

An optimization problem that is described with a quadratic objective function subject to linear constraints is called a "Quadratic Programming" problem. QP can be viewed both as a special case of the nonlinear programming and a generalization of the linear programming. In the real world problems have parameters that are seldom known exactly and have to be estimated by decision maker. Hence, the $n \times n$ symmetric matrix Q and the n vector $\boldsymbol{c}$ have fuzzy numbers in these components. A fuzzy quadratic programming problem can be formulated as:

$$\begin{aligned} & \min \tilde{c}^t\, x + \frac{1}{2} x^t \tilde{Q} x \\ & s.t.\ Ax \le b \\ & \quad x \ge 0 \end{aligned} \tag{5}$$

where the fuzzy numbers are characterized by membership functions that are defined by decision makers. The membership functions can be defined as $\mu_j, \mu_{ij}: R \to [0,1],\ \ i,j \in J = \{1,2,\dots,n\}$.

In particular these membership functions will be supposed as:

$$\mu_j(y) = \begin{cases} 0 & if\ c_j^U \le y\ or\ y \le c_j^L \\ h_j(y) & if\ c_j^L \le y \le c_j \quad j \in J \\ g_j(y) & if\ c_j \le y \le c_j^U \end{cases} \tag{6}$$

and

$$\mu_{ij}(y) = \begin{cases} 0 & if\ q_{ij}^{U} \le y\ or\ y \le q_{ij}^{L} \\ h_{ij}(y) & if\ q_{ij}^{L} \le y \le q_{ij} \quad i,j \in J \\ g_{ij}(y) & if\ q_{ij} \le y \le q_{ij}^{U} \end{cases} \tag{7}$$

where $h(\cdot)$ and $g(\cdot)$ are assumed to be strictly increasing and decreasing continuous functions, respectively, $h_j(c_j) = g_j(c_j) = 1,\ j \in J$ and $h_{ij}(q_{ij}) = g_{ij}(q_{ij}) = 1,\ i,j \in J$.

A multi-objective approach to solve a linear programming problem with imprecise costs is described in [10, 11]. As the linear problem is a particular case of quadratic problem, this approach can be extended to solve quadratic programming problems with fuzzy costs.

The quadratic objective function can be divided into two parts, where the first one is a linear term and the second one is a quadratic term. According to this, the fuzzy costs can only be in the first part or only the second part or in both.

The linear problem considered in [10] used trapezoid membership functions for the costs but here, for the sake of simplicity, they will be supposed to be like (6) and (7).

Then, by considering the $(1-\alpha)$-cut of every cost, $\alpha \in [0,1]$,

$$\begin{aligned} &\forall x \in R, \quad \mu_j(x) \ge 1-\alpha \Leftrightarrow h_j^{-1}(1-\alpha) \le x \le g_j^{-1}(1-\alpha), \\ &\forall x \in R, \quad \mu_{ij}(x) \ge 1-\alpha \Leftrightarrow h_{ij}^{-1}(1-\alpha) \le x \le g_{ij}^{-1}(1-\alpha), \end{aligned} \tag{8}$$

Where $i,j \in J = \{1,2,\dots,n\}$.

Thus, according to the parametric transformations shown above, a fuzzy solution to (5) may be found from the parametric solution of the multi-objective parametric quadratic programming problem

$$\begin{aligned} \min \Big[&(c^1)^t x + \frac{1}{2} x^t Q^1 x, (c^2)^t x + \frac{1}{2} x^t Q^1 x, \dots, \left(c^{2^n}\right)^t x + \frac{1}{2} x^t Q^1 x, (c^1)^t x \\ &+ \frac{1}{2} x^t Q^2 x, \dots, \left(c^{2^n}\right)^t x + \frac{1}{2} x^t Q^2 x, \dots, \left(c^{2^n}\right)^t x \\ &+ \frac{1}{2} x^t Q^{2^{n^2}} x \Big] \\ s.t.\ & Ax \le b, x \ge 0, \\ & c^k, Q^p \in E(1-\alpha), \alpha \in [0,1], \\ & k = 1,2,\dots,2^n \ and\ p = 1,2,\dots,2^{n^2} \end{aligned} \tag{9}$$

where $E(1-\alpha)$, for each $\alpha \in [0,1]$, is the set of vectors in R^n such that each of its components is either in the lower bound, $h_j^{-1}(1-\alpha)$, or in the upper bound, $g_j^{-1}(1-\alpha)$, of the respective $(1-\alpha)$-cut, that is, $\forall k = 1,2,\dots,2^n$, and $\forall i,j \in J$.

$$c^k = (c_1^k, c_2^k, \dots, c_n^k) \in E(1-\alpha) \Leftrightarrow c_j^k = \begin{cases} h_j^{-1}(1-\alpha) & or \\ g_j^{-1}(1-\alpha) \end{cases} \tag{10}$$

and

$$Q^k = (q_{11}^k, \ldots, q_{1n}^k, \ldots, q_{nn}^k) \in E(1-\alpha) \Leftrightarrow q_{ij}^k = \begin{cases} h_{ij}^{-1}(1-\alpha) & or \\ g_{ij}^{-1}(1-\alpha) \end{cases} \quad (11)$$

The obtained parametrical solutions for any of multi-objective models above, to the different α values, generate a set of solutions and then we use the Representation Theorem to integrate all of these particular alpha-solutions.

3 Numerical Example

In this section, a kind of data sets that represents the price mechanism of prefabricated houses is analyzed. The input and output data shown in Table 2 are obtained from the catalogue issued by some corporations, as described in [8]. The computational results and a comparative analysis of the linear and quadratic parametric approaches responses will be presented.

Table 2 Data related to prefabricated houses

No.	y_i			
1	606	1	38.09	36.43
2	710	1	62.10	26.50
3	808	1	63.76	44.71
4	826	1	74.52	38.09
5	865	1	75.38	41.10
6	852	2	52.99	26.49
7	917	2	62.93	26.49
8	1031	2	72.04	33.12
9	1092	2	76.12	43.06
10	1203	2	90.26	42.64
11	1394	3	85.70	31.33
12	1420	3	92.27	27.64
13	1601	3	105.98	27.64
14	1632	3	79.25	66.21
15	1699	3	120.50	32.25

The tests were all performed on a PC with 2.26GHZ Intel® Core™ 2 Duo processor, 4GB RAM running Ubuntu 9.10 operational system. All the problems presented in this work were resolved using NSGA-II evolutionary algorithm [9] which was implemented in MATLAB® 7.8.0 program. The evolutionary algorithm

parameters are: 100 generations and 100 individuals in the population, while the crossover and mutation index are 0.6 and 0.3, respectively.

Figure 1 shows that the observations, which are represented by diamonds, are inside of the interval formed by the obtained solutions for the linear formulation. The objective function of this formulation is described for $\min~\sum_{i=1}^{m}\left|y_{i}\tilde{A}_{0}\sum_{j=1}^{n}\tilde{A}_{j}x_{ij}\right|$ subject to the same constraints presented in Problem (4). The upper and lower box-and-whisker diagram, which shows the spread of the data, represent upper and lower values, respectively, of each one fuzzy number of the population of obtained optimal solutions. The box lies between the upper and lower quartiles, and the median can also be indicated by dividing the box into two. The whiskers are straight line extending from the end of the box to the maximum and minimum extreme values.

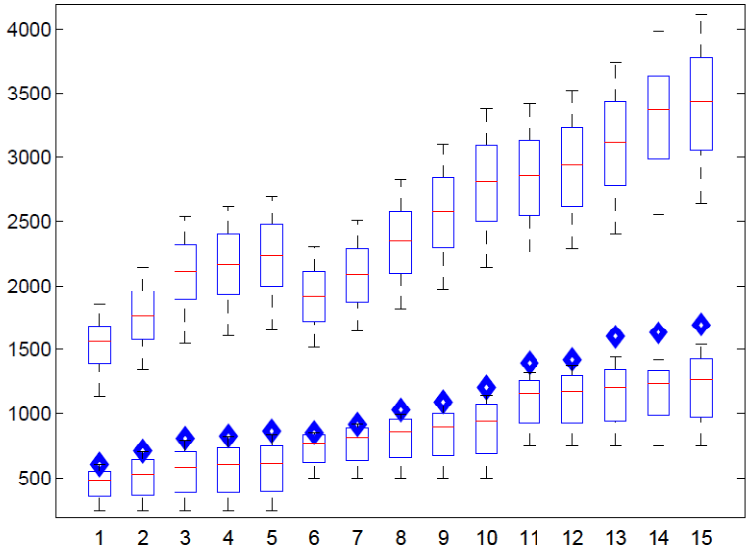


Fig. 1 Fuzzy regression model by linear programming for house price data

Figure 2 shows that the observations, which are represented by diamonds, are also inside of the interval formed by the obtained solutions for the quadratic formulation. It is easy to see that the quadratic formulation obtains an interval closer than the linear formulation. Another point is that the most of the lower values are more concentrate and some minimum extreme values are more distant for the lower quartile. Therefore, the quadratic formulation obtains parameters that fit better to the observed values.

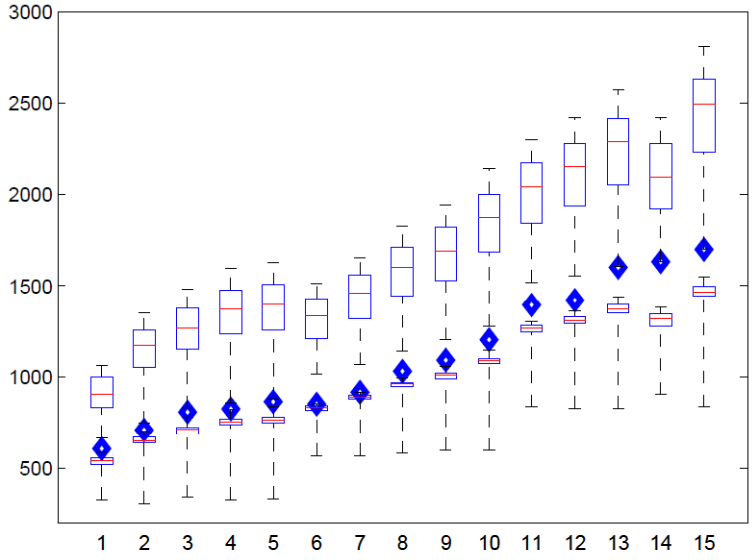


Fig. 2 Fuzzy regression model by quadratic programming for house price data

4 Conclusions

Quadratic Programming problems are very important in a variety of both theoretical and practical areas. When real-world applications are considered, the vagueness appears in a natural way, and hence it makes perfect sense to think in Fuzzy Quadratic Programming problems. In contrast to what happens with Fuzzy Linear Programming problems, unfortunately until now no solution method has been found for this important class of problems. In this context this paper has presented an operative and novel method for solving Fuzzy Quadratic Multi-Objective Programming problems which is carried out by performing two phases which finally provide the user with a fuzzy solution. The obtained solutions allow the authors to follow along this research line trying to solve real problems in practice, in such a way that oriented Decision Support Systems involving Fuzzy Quadratic Programming problems can be built.

An evolutionary algorithm called NSGA-II was used and it produces a sequence of points according to a prescribed set of instructions, together with a termination criterion. Usually we look for a sequence that converges to a set of efficient solutions, but in many cases however we have to be satisfied with less favorable solutions. Then the procedure may stop either 1) if a point belonging to a prefixed set (the solution set) is reached, or 2) if some prefixed condition for satisfaction is verified.

In any case, assuming that a solution set is prefixed, the algorithm would stop if a point in that solution set is reached. Frequently, however, the convergence to a point in the solution set is not easy because, for example, of the existence of local optimum points. Hence we must redefine some rules to finish the iterative procedure.

Hence the control rules of the algorithms solving convex programming problems could be associated to the solution set, and to the criteria for terminating the algorithm. As it is clear, fuzziness could be introduced in both points, not assuming it as inherent in the problem, but as help for obtaining, in a more effective way, some solution for satisfying the decision-maker's wishes. This mean that the decision maker might be more comfortable obtaining a solution expressed in terms of satisfaction instead of optimization, as it is the case when fuzzy control rules are applied to the processes.

Acknowledgments. The authors want to thank the financial support from the agency FAPESP (project number 2010/51069-2) and the Spanish projects CEI BioTic GENIL from the MICINN, as well as TIN2011-27696-C02-01, P11-TIC-8001, TIN2008-06872-C04-04, and TIN2008-01948.

References

1. Cruz, C., Silva, R.C., Verdegay, J.L.: Extending and relating different approaches for solving fuzzy quadratic problems. Fuzzy Optimization and Decision Making 10(3), 193–210 (2011)
2. Savic, D., Pedrycz, W.: Evaluation of fuzzy linear regression models. Fuzzy Sets and Systems 39, 51–63 (1991)
3. Redden, D.T., Woodall, W.H.: Properties of certain fuzzy linear regrssion methods. Fuzzy Sets and Systems 64, 361–375 (1994)
4. Peters, G.: Fuzzy linear regression with fuzzy interval. Fuzzy Sets and Systems 63, 45–55 (1994)
5. Tanaka, H., Lee, H.: Interval regression analysis by quadratic programming approach. IEEE Transactions on Fuzzy Systems 6(4), 473–481 (1998)
6. Tanaka, H., Watada, J.: Possibilistic linear systems and their application to the linear regression model. Fuzzy Sets and Systems 27, 275–289 (1988)
7. Tanaka, H., Koyama, K., Lee, H.: Interval regression analysis based on quadratic programming, pp. 325–329. IEEE (1996)
8. Tanaka, H., Uejima, S., Asai, K.: Linear regression analysis with fuzzy model. IEEE Transactions on Systems, Man and Cybernetics 12(6), 903–907 (1982)
9. Deb, K.: Multi-objective optimization using evolutionary algorithms. JohnWiley& Sons, Ltd., Chichester (2001)
10. Delgado, M., Verdegay, J.L., Vila, M.A.: Imprecise costs in mathematical programming problems. Control and Cybernetics 16(2), 113–121 (1987)
11. Delgado, M., Verdegay, J.L., Vila, M.A.: Relating different approaches to solve linear programming problems with imprecise costs. Fuzzy Sets and Systems 37, 33–42 (1990)
12. Diamond, P.: Fuzzy least squares. Information Sciences 46, 141–157 (1988)
13. Silva, R.C., Cruz, C., Yamakami, A.: A parametric method to solve quadratic programming problems with fuzzy costs. In: IFSA 2009 (2009)
14. Donoso, S., Marín, N., Vila, M.A.: Fuzzy regression with quadratic programming: An application to financial data. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 1304–1311. Springer, Heidelberg (2006)

Business Intelligence Taxonomy

Pablo M. Marin Ortega, Lourdes García Ávila, and Jorge Marx Gómez

Abstract. The tourism industry has been steadily rising in the world, creating new job opportunities in many countries. Today's information management solutions for the complex tasks of tourism industry are still at an early stage. This paper presents some preliminary results on the state of the art analysis on the existing tourism ontologies and enterprise architectures. First result presented in this paper identifies taxonomy for business intelligence in the tourism domain. We identify several tourism standards which are suitable as basis for business intelligence.

1 Introduction

The World Tourism Organization [21] vision forecasts that international arrivals are expected to reach nearly 1.6 billion by the year 2020. Of those worldwide arrivals in 2020, 1.2 billion will be intraregional and 378 million will be long-haul travelers. The total tourist arrivals by region shows that by 2020 the top three receiving regions will be Europe (717 million tourists), East Asia and the Pacific (397 million) and the Americas (282 million), followed by Africa, the Middle East and South Asia. In Cuba the tourism industry according to [2] represent the 7% of the Gross Domestic Product (PIB by its Spanish acronym). In the most recent 20 years, the revenues for this sector to the internal economy came to 30 billion dollars, and they covered 46% of the whole imports.

Tourism is viewed as an information intensive industry where information plays an important role for decision and action making and the Information Technology (IT) starts to play a challenging role in the tourism domain[17]. Business Intelligence (BI) provides the ability to analyze business information in order to

Pablo Marin Ortega · Lourdes García Ávila
Department of Industrial Engineering, Central University of Las Villas, Cuba
e-mail: {pablomo,lourdes}@uclv.edu.cu

Jorge Marx Gómez
Department of Computing Science, Business Information Systems I / VLBA,
Carl von Ossietzky University Oldenburg, Germany
e-mail:jorge.marx.gomez@wi-ol.de

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_10,

support and improve the decision making process across a broad range of business activities. But the BI tools alone facilitate "how" it is possible to achieve a solution from the IT point of view, but they do not assure "what" is the information that is really needed. Enterprise Architectures (EA) are a good starting point for the design of any BI solution, because EA captures a wide information variety, establishes relations between the technological domain and the business domain, and stores all the information joined on a single repository. Existing EA, however, lack of semantics for humans and systems to understand them exactly and commonly, this causes communication problems between humans or between systems or between human and system. These communication problems keep enterprises from implementing integration and collaboration with other companies.

This paper focuses on the state of the art analysis on the existing tourism ontologies, EA frameworks and BI components. We present several existing tourism ontologies and EA frameworks which are suitable to serve as a basis for the design of BI. Furthermore we propose taxonomy to design BI in the tourism industry that allows the mapping of tourism ontologies, an EA and the components of any BI solution.

This paper is organized as follows: Section 2 provides related work. In Section 3, we present important outlines about tourism ontologies. Section 4 we chose an EA. A new taxonomy to design BI is suggested in Section 5. Finally, conclusions and future work are discussed in Section 6.

2 Related Work

In tourism domain, there already exist different standard and ontologies which are designed and used internally by tourism agents to help them manage heterogeneous tourism data, in the Table 1, we summarize some work aimed to generate global standards to facilitate inter and intra tourism data exchange.

There are many enterprise architecture frameworks, and new ones are being added every day. Recent work shows that the most commonly used framework is the Zachman Framework, followed by the organization's own frameworks, followed by TOGAF, U.S. DoD (this covers about two-thirds of all enterprises)[1][6]. Frameworks provide guidance on how to describe architectures and based on them we can define taxonomy.

BI technologies are used to provide historical, current, and predictive views of business operations. Common functions of BI technologies are reporting, on-line analytical process (OLAP), analytics, data mining, business performance management, benchmarks, text mining, and predictive analytics[6].

[1] A press time survey showed the following statistics [IEA200501]: Zachman Framework: 25%; Organization's own: 22%; TOGAF: 11%; U.S., DoD Architecture Framework: 11%; E2AF: 9%; FEAF: 9%; IAF: 3%; TAFIM: 2%; TEAF: 0%; ISO/IEC 14252 (IEEE Std 1003.0): 0%; Other: 9%.

Table 1 Global standards to facilitate inter and intra tourism data exchange

Related Works	Description	Source
Mondecas tourism ontology	Used by Destination Marketing Organizations and distributors, it's maintained and manipulates comprehensive knowledge bases and catalogs from many sources.	[15]
WTO thesaurus	Includes information and definitions of the topic tourism and leisure activities. In addition to showing equivalent terms in the five official languages of the Organization (Arabic, English, French, Russian and Spanish), some records contain definitions, links to online references and other useful information.	[22]
e-Tourism Ontology	The goal of the ontology is to support tourism organizations with exchanging data and information without changing their local data structures and information systems.	[20]
Travel Itinerary Ontology	Simple ontology for representing a travel itinerary	[4]
General Geographic Ontology	Provides geographic location information for cities, airports, ports, and other facilities. Includes name, country, lat/long, etc.	[18]
TAGA Travel Ontology	Provides typical concepts of travelling combined with concepts describing typical tourism activities.	[26]
German Hotel Classification / Deutsche Hotelklassifizierung	Offers a German standardized classification system called "Deutsche Hotelklassifizierung" (German Hotel Classification).	[3]
NC 127:2001	Offers a Cuban's standard to the tourism industry, included a classification system.	[16]
ISO 18513:2003	Offers a standard that it defines terms used in the tourism industry in relation to the various types of tourism accommodation and other related services.	[9]

Generally, a BI system should have the following basic features[23]:

- Data Management: including data extraction, data cleaning, data integration, as well as efficient storage and maintenance of large amounts of data.
- Data Analysis: including information queries, report generation, and data visualization functions.
- Knowledge Discovery: extracting useful information (knowledge) from the rapidly growing volumes of digital data in databases.

Nowadays, there are many companies around the world considered as BI distributors, each introducing their own products as qualified in meeting all organizational needs, in Table 2. we present a study developed by [19] about the top 5 Worldwide BI tools revenue by vendor, 2007-2009. Nevertheless, these applications do not assure the totality of the necessary information in the decision making process is available. Most of the existing solutions are focused on the technological

capacities, and it answers to the question of, how achieving the solution? But not, what would be the necessary information? That the solution must support, according to the real needs for the business; these elements indicate that do not exist alignment between the business domain and technological domain.

Table 2 (Top 5) Worldwide Business Intelligence Tools Revenue by Vendor, 2007–2009

	Revenue ($M)			Share (%)			2007-2008	2008-2009
Company	2007	2008	2009	2007	2008	2009	Growth (%)	Growth (%)
SAP	1356.7	1,574.6	1,557.1	19.0	20.2	19.5	16.1	-1.1
IBM	1153.3	1,145.6	1,224.3	16.1	14.7	15.3	-0.7	6.9
SAS	785.4	870.5	909.5	11.0	11.1	11.4	10.8	4.5
Oracle	596.7	701.1	719.5	8.3	9.0	9.0	17.5	2.6
Microsoft	554.9	648.7	701.3	7.8	8.3	8.8	16.9	8.1
Other	2,706.2	2872.7	2893.6	37.8	36.8	36.1	6.2	0.7
Total	7,153.2	7,813.4	8,005.3	100.0	100.0	100.0	9.2	2.5

3 Tourism Ontologies

Several tourism ontologies were considered for reuse, one of the most popular is e-Tourism Ontology. It organizes tourism related information and concepts. The ontology will allow achieving interoperability through the use of a shared vocabulary and meanings for terms with respect to other terms. The Class Overview of the ontology is show below.

Class Overview[13]

- Accommodation
- Activity
- ContactData
- DateTime
  - OpeningHours
  - Period
    - DatePeriod
    - TimePeriod
  - Season
- Event
- Infrastructure
- Location
  - GPSCoordinates
  - PostalAddress
- Room
  - ConferenceRoom
  - Guestroom
- Ticket

With the above Class Overview of the e-Tourism Ontology we can get an idea about the main concepts needed today in the Tourism Industry, and based on these we will propose taxonomy for developing a BI solution for the Tourism Industry.

4 Enterprise Architecture

The purpose of enterprise architecture is to create a map of IT assets and business processes and a set of governance principles that drive an ongoing discussion about business strategy and how it can be expressed through IT. There are many different suggested frameworks to develop enterprise architecture. However, most frameworks contain four basic domains, as follows: (1) business architecture: documentation that outlines the company's most important business processes; (2) information architecture: identifies where important blocks of information, such as a customer record, are kept and how one typically accesses them; (3) application system architecture: a map of the relationships of software applications to one another; and (4) the infrastructure technology architecture: a blueprint for the gamut of hardware, storage systems, and networks. The business architecture is the most critical, but also the most difficult to implement, according to industry practitioners[14].

In order to define a taxonomy model, a framework to define EA is required. There are various EA frameworks, as explained above. Among them, the Zachman Framework [24, 25] is selected as a base EA framework to define the taxonomy. Although the Zachman framework lacks in modeling for detailed EA components and relationships among them and does not provide concrete implementing method, it is valuable in the point that it presents general framework which every enterprise can use to build its EA [10]. Beside "the Zachman Framework is an ontology – a theory of the existence of a structured set of essential components of an object for which explicit expression is necessary, and perhaps even mandatory for creating, operating, and changing the object (the object being an enterprise, a department, a value chain, a solution, a project, an airplane, a building, a product, a profession, or whatever)" [12]. In order to cope the taxonomy with the BI solution we considering only the first four rows of the framework, which are defined as: the strategy model, business model, system model and technology model. The structure of the framework is show in the Table 3.

Based in the structure of the framework we can define taxonomy of the enterprise, where the first hierarchical level is the "Scope content", followed by "Business Concept" and so on (see Figure 1), in each level is necessary to define the elements defined in each row of the enterprise architecture.

Table 3 The Zachman Enterprise Framework$^{2\,TM}$ [12]

	What	How	Where	Who	When	Why
Scope Contents	Inventory Identification.	Process Identification.	Network Identification.	Organization Identification.	Timing Identification.	Motivation Identification.
	Inventory types.	Process types.	Network types.	Organization types.	Timing types.	Motivation types.
Business Concepts	Inventory Definition.	Process Definition.	Network Definition.	Organization Definition.	Timing Definition.	Motivation Definition.
	Business Entity.	Business Transform.	Business Location.	Business Role.	Business Cycle.	Business End.
	Business Relationships.	Business Input.	Business Connection.	Business Work.	Business Moment.	Business Means.
System Logic	Inventory Representation.	Process Representation.	Network Representation.	Organization Representation.	Timing Representation.	Motivation Representation.
	System Entity.	System Transform.	System Location.	System Role.	System Cycle.	System End.
	System Relationships.	System Input.	System Connection.	System Work.	System Moment.	System Means.
Technology Physics	Inventory Specification.	Process Specification.	Network Specification.	Organization Specification.	Timing Specification.	Motivation Specification.
	Technology Entity.	Technology Transform.	Technology Location.	Technology Role.	Technology Cycle.	Technology End.
	Technology Relationship.	Technology Input.	Technology Connection.	Technology Work.	Technology Moment.	Technology Means.

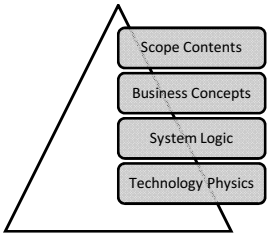


Fig. 1 Hierarchical Level of the enterprise taxonomy

5 Taxonomy for the Business Intelligence Solution in the Tourism Industry

In accordance with the previously expressed, the proposed taxonomy is comprised of EA components, a Balanced Scorecard (BSC), BI components and relationships between them as shown in Figure 2.

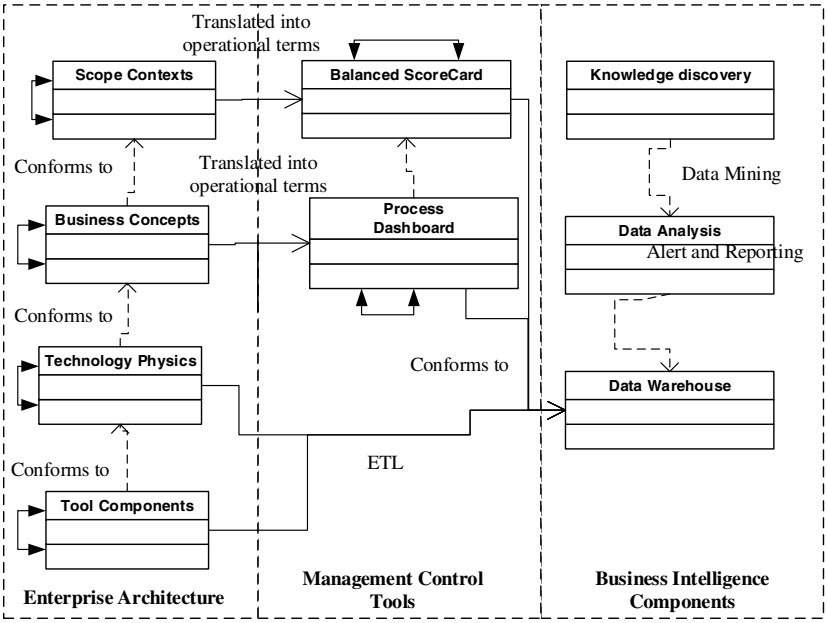


Fig. 2 Business Intelligence Taxonomy

The BSC [11] was included because it is a representative tool on business strategy alignments. The BSC suggests four perspectives: financial, customers, internal business processes, and learning and growth to measure enterprise performance. The four perspectives of the BSC offer a balance between short term and long term goals, between desired outcomes and performance drivers of those outcomes, and between hard goals measures and softer, more subjective measures. Using the BSC, enterprises can align strategies and make processes executed according to strategies[5], and this process help to know which is the necessary information for the decision making process.

Attending on the structure defined in the EA (Table 3), it must be completed by rows, given that each row represents a top level with respect to the one that follows in order, nevertheless there exist a big dependency among each of the elements for columns. In Table 4 appear the dependencies that we proposed between each cells. The order to fulfill each cell depends on the relationships among the cells.

In order to fulfill the goals of this paper is necessary to map the approach found in the Tourism Industry (see Table 1), we propose to use as vocabulary to describe the enterprise architecture the ISO 18513:2003, because ISO (International Organization for Standardization) is the world's largest developer of voluntary technical standards. ISO is a non-governmental organization with 163 members[8], included Cuba. The ISO 18513:2003 standard, "Tourism services Hotel and other types of tourism accommodation Terminology", defines terms used in the tourism industry

in relation to the various types of tourism accommodation and other related services. It has been published in January 2001 and is directly adopted from a standard by the European Committee for Standardization (CEN). The standard is designed to define a common vocabulary for a better understanding between the users and providers of tourism services.

Table 4 Fulfillment rules

	What	How	Where	Who	When	Why
Scope Contents	A1	B1	C1	D1	E1	F1
Business Concepts	A2 ← (A1)	B2← (B1+A2)	C2← (C1+B2)	D2← (D1+B2+C2)	E2← (E1+A2+C2)	F2← (F1+B2)
System Logic	A3← (A2+B2+F2)	B3← (B2+F2)	C3← (C2+A3+B3)	D3← (D2+F2+B3)	E3← (E2+B3+C3)	F3←(F2)
Technology Physics	A4← (A3)	B4← (B3+A4)	C4← (C3+A4+B4)	D4← (D3+A4+B4)	E4← (E3+D4)	F4←(F3)

Furthermore we proposed to used in the "Business Concepts" row the elements defined in e-Tourism Ontology, because this ontology cover common concepts of the tourism industry, besides it was developed taking into account existing standard initiatives and its main goal is to support tourism organizations with data and information exchange without changing their local data structures and information systems. In the Table 5 we proposed how to map the Overview Class of the ontology with each column of the Business Concepts row.

Table 5 Mapping between Business Concepts and Class Overview

	What	How	Where	Who	When	Why
Business Concepts	Infrastructure	Accommodation Room Ticket	Location	ContactData	DateTime Event	Activity

The BSC in the strategic level help to the manager to translate the strategy to operative terms, because it develops a strategic planning and management with the goal to align business activities with the organization strategy, improves internal and external communications, and monitors organization performance against strategic goals in order to continuously improve strategic performance and results[1].

Wherever is necessary to have a set of Key Performance Indicators (KPI) in the tactical level (Business Metric in the Figure 2) mapping with the BSC, so tactical staff knows which are they goals and what they should do; beside this linking is necessary for the knowledge discovery process. Using different algorithms is possible to find the relationships among KPIs and among BSC indicators and KPIs, these relationships help to explain some behaviors in the environment of the

enterprise. This process helps the manager in the decision making process and clarifies for them which is the right way for the organization to succeed.

The other part of the any BI solution is the "Data Management", the first step of the Data Management and sometime the most difficult and hard is the extract, transform, and load (ETL) process, because the information is distributed throughout the length and wide of organizational systems, with different technology and different models. To have an EA linking with a BSC to develop BI it helps to know:

1. What will be the necessary information to load in the data warehouse?

What will be the necessary changes to make on the data before loading them in the data warehouse?
Which are the technologies that support the necessary data to load in the data warehouse?
Where are distribute the necessary data to load in the data warehouse?
When will be necessary to load the information in the data warehouse?
Which will be the main roles of the BI?
Which will be the necessary reports?

To know the answer of these questions before developing BI is very important to reduce cost and time. Sometimes the cost and the time increase because the company staff does not know the answer of these questions and then they lost a lot time developing patches to fix errors previously committed. The problem is that to develop BI successfully is necessary first an alignment between business domain and technology domain, and is necessary too a common understanding among the business staff and the technological staff, because BI is: "An interactive process for exploring and analyzing structured, domain-specific information (often stored in data warehouses) to discern business trends or patterns, thereby deriving insights and drawing conclusions. The business intelligence process includes communicating findings and effecting change. Domains include customers, suppliers, products, services and competitors"[7].

6 Conclusion and Future Work

In this paper, we presented some early stage work on the taxonomy for BI in tourism industry, we also identify several tourism standards and ontologies and several problems in the development of BI. Based on this, we proposed taxonomy for BI to improve the availability of the necessary information for the decision making process, based on the integration of the business and technological domains. In the future we work to define based on the proposed taxonomy, an ontological model to help to improve the process of mapping among EA component, BSC and BI component as well as to design some tools, to measure the alignment degree between business and technological domains and the performance of BI.

References

1. Balanced Scorecard Institute. What is the Balanced Scorecard (2010), http://www.balancedscorecard.org/BSCResources/AbouttheBalancedScorecard/tabid/55/Default.aspx (cited August 04, 2010)
2. Campos Roberto, F.: El turismo empuja a la economía cubana, dice experto. In: Prensa Latina, La Habana, Prensa Latina (2010)
3. H.-u, D. G.a.: Deutsche Hotelklassifizierung, DEHOGA (1996), http://www.dehoga.de (cited July 28, 2010)
4. Dean, M.: (2001), http://www.daml.org/ontologies/178 (cited July 27, 2010)
5. Dongwoo Kang, J.L.: Kwangsoo Kim, Alignment of Business Enterprise Architectures using fact-based ontologies. ELSEVIER (2010)
6. FatemehMoghimi, C.Z.: Senior Lecturer, A decision-making model to choose Business Intelligence platforms for organizations. In: IEEE Third International Symposium on Intelligent Information Technology Application. IEEE (2009)
7. Group, G.: The Gartner Glossary of Information Technology Acronyms and Terms (2004), http://www.gartner.com/6_help/glossary/Gartner_IT_Glossary.pdf (cited)
8. ISO. ISO members S/F, http://www.iso.org/iso/about/iso_members.htm (cited 2010 August 04, 2010)
9. ISO. Tourism services - Hotel and other types of tourism accommodation - Terminology (2003), http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=31812 (cited July 28, 2010)
10. Kang Dongwoo, L.J.: Choi Sungchul, Kim, Kwangsoo, An ontology-based Enterprise Architecture. ELSEVIER (2009); (Expert Systems with Applications)
11. Kaplan, R., Norton, D.: The Strategy Focused Organization. In: Cómo utilizar el Cuadro de Mando Integral, Para Implementar y Gestionar su Estrategia. G. 2000, Barcelona (2001)
12. Kappelman, L. (ed.): The SIM Guide to Enterprise Architecture ed. C.P.T.F. Group, Taylor & Francis Gruop, an informa business. 6000 Broken Sound Parkway NW, Suite 300. Boca Raton, FL 33487-2742 (2010)
13. Katharina Siorpaes, K.P.: Daniel Bachlechner. Class Hierarchy for the e-Tourism Ontology (2004), http://e-tourism.deri.at/ont/etourism%20onto%2008.pdf (cited August 02, 2010)
14. Minoli, D.: Enterprise Architecture A to Z. Frameworks, Business Process Modeling, SOA, and Infrastructure Technology, T.F. Group, Editor, Auerbach Publications: Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742 (2008)
15. Mondeca. Mondecas tourism ontology (2010), http://www.mondeca.com (cited July 27, 2010)
16. Oficina Nacional de Normalización (ONN), NC 127:2001: Industria Turística - Requisitos para la clasificación por categorías de los establecimientos de alojamiento turístico, Oficina Nacional de Normalización (NC) La habana (2001)
17. Prantner Kathrin, D.Y., Michael, L., Zhixian, Y.: Tourism ontology and semantic management system: state-of-the-arts analysis. In: IADIS International Conference WWW/Internet 2007, Vienna, Austria (2007)
18. Rager, D.: GEOFILE (2001), http://www.daml.org/ontologies/156 (cited July 27, 2010)

19. Vesset, D.: Worldwide Business Intelligence Tools (2009), Vendor Shares (2010), http://www.idc.com/getdoc.jsp?sessionId=&containerId=222725&sessionId=7E6D9B56D86AA02D45CC71C6F280EC34 (cited July 29, 2010)
20. Wolfram Höpken, C.C.: HarmoNET Ontology RDFS Specification (2008), http://www.etourism-austria.at/harmonet/images/software/HTOv4002.rdfs (cited July 27, 2010)
21. World Tourism Organization (UNWTO/OMT). Tourism 2020, Vision (2010), http://www.unwto.org/facts/eng/vision.htm (cited July 27, 2010)
22. World Tourism Organization (UNWTO/OMT). TourisTerm (2010), http://www.unwto.org/WebTerm6/UI/index.xsl (cited July 27, 2010)
23. Ying Wang, Z.L.: Study on Port Business Intelligence System Combined with Business Performance Management. In: IEEE Second International Conference on Future Information Technology and Management Engineering. IEEE (2009)
24. Zachman, J.: Enterprise Architecture: The Issue of the Century, Zachman Institute for Framework Advanced, ZIFA (2010), http://www.cioindex.com/nm/articlefiles/63503-EAIssueForTheCenturyZachman.pdf (cited January 20, 2010)
25. ZIFA. Zachman Framework (2010), http://www.zachmaninternational.com/index.php/the-zachman-framework (cited January 20, 2010)
26. Zou, Y.: Travel business ontology. S/F, http://taga.sourceforge.net/owl/travel.owl (cited July 27, 2010)

Discovering Knowledge by Fuzzy Predicates in Compensatory Fuzzy Logic Using Metaheuristic Algorithms

Marlies Martínez Alonso, Rafael Alejandro Espín Andrade,
Vivian López Batista, and Alejandro Rosete Suárez

Abstract. Compensatory Fuzzy Logic (CFL) is a logical system that enables an optimal way of modeling knowledge. Its axiomatic character enables the work of natural language translation of logic, so it is used in knowledge discovery and decision-making.Obtaining LDC predicates with high values of truth is a general and flexible approach that can be used to discover patterns and new knowledge from data. This work proposes a method for knowledge discovery from obtaining LDC predicates, to obtain different structures of knowledge using a metaheuristic approach. A series of experiments and results descriptions of certains advantages for representing several patterns and tendencies from data is used to prove the proposed method.

1 Introduction

Multivalent Logics generalize Bivalent Logic (BL) to a whole range of intermediate values between true (1) and false (0). Thus, it is possible to work with sets that do not have perfectly defined limits, where the transition between membership and non-membership of a variable in a set is gradual [3].

Marlies Martínez Alonso · Rafael A. Espín Andrade
"Jose Antonio Echeverria" Higher Technical Institute, Cuba
e-mail: marlies.martinez@gmail.com, rafaelespin@yahoo.com, rosete@ceis.cujae.edu.cu

Vivian López Batista · Alejandro Rosete Suárez
Department of Computer Science and Automation, University of Salamanca, Spain
e-mail: vivian@usal.es

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_11,

Although the many advantages of multivalent systems to model ambiguous or vague knowledge have certain difficulties modeling knowledge in a natural way. For that reason is frequent in these applications the use of free operators and the extralogical resource called defuzzification [5].

The main difficulties of multivalent systems in the work of knowledge modeling are that not generalize completely all formulas of the Bivalent Logic. This difficulty sometimes causes bad behaviors for some interpretations of the logical variables. Another difficulty becomes evident when modeling decision-making problems, where not achieve the most appropriate behavior. In this case, the deficiency is given by the associative character of the operators of conjunction and disjunction, and the lack of sensibility to changes in the value of true of basic predicates when calculating the value of truth of compound predicates [7].

CFL is a multivalent system distinguished by its quality to generalize all formulas of BL. It has been demonstrated using Kleene´s axiomatic that the valid formulas of BL are exactly the formulas that have truth value greater than 0.5 in the context of the CFL. It has the characteristic of being a sensible system which ensures that any variation in the truth value of the compound predicates. CFL waiver of compliance with the associative classical properties of conjunction and disjunction to achieve a knowledge representation closer to human thinking. Its ability to formalize the reasoning makes possible its use in situations requiring multicriteria evaluations and ambiguous verbal descriptions of knowledge. Therefore, offers an opportunity to use language in the construction of semantic models that facilitate evaluation, decision making and knowledge discovery [7, 8].

The knowledge discovery by finding CFL predicates makes use of metaheuristics for exploring the space of possible predicates, identifying those that have high value of true. This is a way to generalize a part of the genetic programming dedicated to the collection of predicates that represent logical formulas; part that is intended mainly to circuit design [1, 6]. In this case, unlike genetic programming, the learning is not supervised; neither have we used the classical binary trees that have been used by genetic programming to prevent the deficiencies of the binary representation.

This work begins by giving some basic notions of the CFL. Then is describes some of the basic notions of the proposed approach. Finally, show some experimental results which show certain advantages of the proposed approach as a tool for knowledge discovery.

2 Basic Notions of CFL

The main difficulties of multivalent logical approaches in the modeling of knowledge are:

- The associative property of conjunction and disjunction operators used.
- The lack of sensitivity to changes in the values of truth of the basic predicates when calculating the truth value of the compound predicate.

- The total absence of compensation of truth values of predicates when calculating the truth value of the compound predicates using operators.

The association is a characteristic present in a great part of the operators used for aggregation. This characteristic is not good for data mining due to the existence of equality of hierarchies of objectives and preferences. Moreover, the sensitivity is the ability to react to the values of the predicates. This makes that different situations in the state of the systems generate different assessments in the veracity of knowledge and different models have different behaviors. Finally, the compensation is the capacities of the values of basic predicates are compensated to each other when calculating the truth value of the compound predicate. Classical approaches of Decision Theory include models such as additives, which accept compensation without limits, and the descriptive, which accept partially compensation, the latter being more akin to the reasoning of the actual agents [7].

In the CFL operation conjunction (∧ (and)) is given by the geometric mean of the truth value that takes the predicate of the analyzed variable. In equation 1 is represents the conjunction operator of the CFL. In this case c is the operator representing conjunction and X_n is the true value of the variable n.

$$c(x_1, x_2, x_3, \dots, x_n) = \sqrt[n]{x_1 + x_2 + x_3, \dots \dots, +x_1} \tag{1}$$

The disjunction (∨ (or)) is represented by the complement of the geometric mean of denials of the truth values of the variables. It is calculated according to the equation 2, where d is the disjunction operator and X_n is the truth value having the variable n.

$$d(x_1, x_2, \dots, x_n) = 1 - \sqrt[n]{(1 - x_1) * (1 - x_2) *, .., (1 - x_n)} \tag{2}$$

The negation (¬), as in the rest of the operators in the Fuzzy Logic, calculated using the complement of the value of the variable denied. In equation 2 shows how to calculate it. In this case, n represents the negation operator and X_i represents the value of the variable i.

$$n(x_i) = 1 - x_i \tag{3}$$

The implication (⟶) can be defined in either of two ways shown in equations 4 and 5. In these equations, i represents the implication operator and x, y represent any two variables.

$$i_1(x, y) = d(n(x), y) \tag{4}$$

or

$$i_2(x, y) = d(n(x), c(x, y)) \tag{5}$$

Where d is the disjunction operator, c is the conjunction operator and n is the negation operator mentioned above.

The equivalence or double implication ($\leftrightarrow$) remains as the conjunction of the implication and its reciprocal. For any two variables x and y can be defined equivalence as shown in equation 6.

$$e(x,y) = c(i(x,y), i(y,x)) \tag{6}$$

Where c and i are the conjunction and implication operators presented above.

The universal and existential quantifiers are calculated according to equations 7 and 8. For any fuzzy predicate p in the universe U, universal propositions and existential are respectively defined as:

$$\underset{x\in U}{\forall} p(x) = \underset{x\in U}{\wedge} p(x) \tag{7}$$

$$\underset{x\in U}{\exists} p(x) = \underset{x\in U}{\vee} p(x) \tag{8}$$

3 Search Method of CFL Predicates

The method proposed in this paper is based on providing flexibility for different knowledge structures, using different search algorithms. To achieve this flexibility we use a declarative approach, which consists in separating the mechanism to express requirements, of the mechanism used to satisfy them. The declarative approach is based on the use of optimization methods and general purpose searches such as: Genetic Algorithms, Evolutionary Algorithms, Simulated Annealing, the Search Tabu and the classical methods of Artificial Intelligence such as the Stochastic Hill Climbing (SHC) [1, 6]. To discover knowledge by obtaining CFL predicates, we need to find good predicates in the space of possible predicates. A predicate is considered good if it has a high truth value in the set of examples. Therefore, the problem is oriented to the use of a metaheuristic approach, which consists in optimizing a function in a space (objective function) [2, 6].

The metaheuristic approach to perform searches enables the separation of the evaluation of solutions of the search mechanism used. The separation of both mechanisms increases the flexibility of the method for possible changes to the selected search algorithm and it is related to the function evaluates predicates (function to be optimized).

The proposed search algorithm is composed of three key elements:

- knowledge representation in the form of predicates
- evaluation of predicates using the operators of the CFL
- metaheuristic approach to perform searches

Representation:

For the representation of predicates we use general trees. A general tree is defined as non-empty finite set T, of elements called nodes, such as:

- T contains an element distinguished *R*, called the root of *T*.
- The remaining elements of *T* form an ordered collection of zero or more disjoint trees *T1*, *T2*, .., *Tn*.

In this case, the terminal nodes of the tree are related variables with the problem and the internal nodes of the tree are the operators (negation, ¬), (conjunction, ∧), (disjunction, ∨), (implication, →) , (double implication or equivalence (↔)) of the CFL.

Evaluation:

To evaluate predicates one of the fundamental elements is the truth value that takes the predicate in the set of data studied. In the Predicate Logic, universal and existential quantifiers are frequently used. The universal quantifier determines whether a formula (predicate) is true for all values within the domain. The existential quantifier indicates that a formula is true for any values within the domain. In this paper we use the universal quantifier CFL (see equation 7) to calculate the true value that acquires a predicate in the dataset.

The proposed method takes into account other important characteristics when evaluating a predicate. They are:

1. Not to repeat (to avoid obvious predicates as $p \vee p \rightarrow p$).
2. To have a specific structure (if one wishes to obtain rules (implies) as the root node).
3. Involving as many variables in the examples used.
4. To be small.

The first feature is important because it allows avoiding finding predicates evident, which acquire high truth value but provide little knowledge new. To achieve the goal we define to penalize with 0 those predicates that have repeated variables.

The second characteristic is used for association rules. In this case is assessed that the root node of the tree is an implication (implies) and are penalized with *0* to those that are not rules. This makes the search is directed to obtain only rules.

Feature number three is defined in order to obtain more relations between the variables analyzed and therefore more knowledge. The four features is a natural interest from the point of view of knowledge engineering, because small predicates are easier to interpret than large and complex predicates.

In equation 9 shows the function that evaluates the quality of a predicate:

$$E = \frac{T * DV}{SC * (SC - SSC + 1)} \tag{9}$$

Where:

- Evaluation (*E*): evaluation of the predicate.
- Value of true (*T*): truth value of the predicate in the set of examples.
- Number of different variables (*DV*): number of different variables present in the tree.
- Size with constant (*SC*): number of nodes (terminal and internal) having the tree counting constants.
- Size without constant (*SSC*): number of nodes (terminal and internal) that has the tree without the constants.

Equation 9 represents the objective function to be optimized for search predicates. As shown, this feature aims to achieve to find predicates with high values of truth, small, and with the most variables involved in the examples. In this equation, the evaluation is directly proportional to the value of truth and the number of variables to be used, and inversely proportional to the length.

4 Metaheuristic Approach for Searching

The metaheuristic approach uses two basic mechanisms: the evaluation mechanism and the search mechanism. Both mechanisms are implemented separately and independently. This facilitates scalability and flexibility to changes in the requirements of the predicates to obtain and in the algorithm to use.

The method of obtaining predicates proposed in this paper uses three fundamental and independent components:

- Mutations (generate new solutions from others).
- Evaluation (seen in the previous section).
- Search algorithm.

To mutations defined a set of mutation operators more general and more specific according to the structure of knowledge that will be obtained. The general mutation operators are as follows:

Operators 1

- A terminal node is replaced by another terminal node, taken at random.
- An internal node is replaced by another internal node, taken at random.
- A subtree is replaced by a terminal node, both selected at random.
- A terminal node is replaced by a subtree, both selected at random.

To obtain different knowledge structures we define the following more specific operators:

Operators 2

Operators to obtain classification rules:

1. Set the implication operator (→) as the root node.
2. Set the variable representing the class as a consequent of the rule.
3. Use the operators conjunction (∧) in the antecedent of the rule.
4. Only mutate the antecedent of the rule using the Operators 1, except mutation number 2, since no change of logic connective, always is conjuntion (∧).

Operators for supervised learning:

1. Set the double implication operator (↔) as the root node.
2. Set one of the ends of the equivalence the variable representing the class.
3. Use the conjunction (∧) as logical connective between variables.

4. Mutate the extreme of equivalence in that does not appear the variable representing the class using the Operators 1, except mutation number 2, since no change of logic connective, always is conjuntion (∧).

Operators to obtain cluster:

1. Set the disjunction operator (∨) as the root node of the tree.
2. Set as subtrees clauses in conjunction (∧).
3. Only mutate the clauses in conjunction using Operators 1, except mutation number 2, since no change of logic connective, always is conjunction (∧).

The search of CFL predicates begins to generate initials solutions. These solutions are generated depending on the knowledge one wishes to obtain. If we wish to obtain classification rules, cluster, etc. Then used mutation operators allowing obtaining such a structure directly.

Otherwise, the initial solution is generated entirely at random, from an initial tree representing a predicate with only one node: the node terminal 0.5. Subsequently, 10 mutations are applied to the starting node to obtain a tree, using mutation operators Operator 1.

In both cases, at each iteration a new solution is generated from the previous solutions comparing the evaluation of the current tree (current solution) with the mutated tree (candidate solution) and select the tree with better evaluation. The search ends when it finds a tree with the desired truth value or when finishes running the predefined number of iterations.

5 Experiments

To prove the proposed method it was necessary to implement an experimental tool that would enable experiments using real data. For this implementation was defined using Visual Prolog programming language. This is a logic programming language objects oriented based on Prolog, which like Prolog has a high capacity of deduction to find relationships between the objects created, variables and lists [9].

Experiments were designed to obtain association rules, general predicates, classification rules and supervised learning. We used 800 iterations of SHC to search each of the predicates. The main metric used to assess the quality of the results is the truth value that acquires the predicates in the data set.

The diabetes database is derived from a study in a group of persons identified as diabetics or at risk for it. This database works with actual data taken from the results of a survey conducted by the Health Center of Jaruco, Mayabeque province, Cuba.

It is noteworthy that the data used in the experiments were to be processed, since the method works with linguistic labels and truth value and not the actual values of the variables. It was therefore necessary assigning degrees of membership in each of the variables, with respect to joint fuzzy previously defined.

The diabetes database contains 7 variables that describe a set of characteristics associated with patients. These variables are:

1. Age
2. Race
3. Hypertension
4. Body Mass Index (BMI)
5. Cardiovascular and/or Cerebral Vascular Accident (CVA) antecedents (both known for the expression: "Antecedents")
6. Sex
7. Classification of diabetes (Diabetes)

For each of the variables the following labels and membership functions are established:

Age:

- Universe of discourse U ={Set of all ages}
- Membership function: Sigmoid
- Tag: "Old Age"

Sigmoid function has the following equation:

$$u(x) = \frac{1}{1 + e^{-\alpha(x-y)}}$$

$$\alpha = \frac{(\ln 0.9 - \ln 0.1)}{\gamma - \beta}$$

where:
γ is the value 0.5 (as true as false).
β is the value 0.1 (almost false).
The Sigmoid function parameters are fixed by two values. First, the value at which it is considered that the statement in the predicate is true, which is set from 0.5. The second is the value for which the data is unacceptable, which is set from the value 0.1.

To define the patient has an "Old Age" was used as the value "0.5" at age $\gamma = 40$ years and the value "0.1" is $\beta = 19$ years.

Race:

- Universe of discourse *U* = {White, Mixed race, Black}
- Membership function: Function Singleton
- Tag: "White Race"

To define a patient with "White Race" is assigned a truth value to each element as follows:
White Race= {White|1, Mixed race|0.5}. Black race represents the value zero.

Hypertension:

- Universe of discourse U = {Hypertensive Detected, Risk of Hypertension, Group of no risk}
- Membership function: Function Singleton
- Tag: "Significant Hypertension"

To define a patient with "Significant Hypertension" is assigned a truth value to each element as follows:

Significant Hypertension = {Hypertensive detected|1, Risk of Hypertension|0.5}. No Risk Group represents the value zero

BMI:

- Universe of discourse U = {Set of all possible values of BMI }
- Membership function: Function Singleton
- Tag: "High BMI"

To define a patient with "High BMI" was used as the value "0.5" at BMI $\gamma = 25$ kg/m^2 and the value "0.1" is β= 17 kg/m^2.

Antecedents:

- Universe of discourse U = {Antecedents Detected, Mild Antecedents, No Risk Group}
- Membership function: Function Singleton
- Tag: "Significant Antecedents"

To define a patient with "Significant Antecedents" is assigned a truth value to each element as follows:

Significant Antecedents= {Antecedents Detected|1, Mild Antecedents|0.5}. No Risk Group represents the value zero.

Sex:

- Universe of discourse U = {Possible Sexes}
- Membership function: Function Singleton

In the case of sex as there are only two possible values we assign value 1 to males and 0 females.

Diabetes:

- Universe of discourse U = {Detected Diabetic, Risk Group, Alteration of Fasting Glucose, No Risk Group}
- Membership function: Function Singleton.
- Tag: "Degree of Diabetes"

To define the "Degree of Diabetes" is assigned a truth value to each element as follows:

Degree of Diabetes= {Detected Diabetic|1, Risk Group|0.8, Alteration of Fasting Glucose|0.5}. No Risk Group represents the value zero.

After processing all the data, we proceeded to perform the experiments using in this case the Stochastic Hill Climber as Metaheuristic Algorithm. To obtain each predicate is performed 800 iterations of Climber Hills.

The following are some of the general findings predicates obtained. All have truth values above 0.80. These results can be noted that frequently appear together, the variables advanced age and certain antecedents. It also shows a strong relationship between mass high, certain antecedents, advanced age, hypertension true and diabetes classification.

General Predicates:

$$((Race = white) \wedge ((Antecedents = true) \wedge (Age = advanced))) \rightarrow ((Classification = diabetes)\ or\ (Mass = high))$$

True value: 0.9256

$$((Race = white) \vee ((((Antecedents = true) \wedge (Age = advanced)) \wedge (Hypertension = true))) \rightarrow (((Classifications = diabetes) \vee (Mass = high))) \rightarrow (Sex = male))))$$

True value: 0.9205

$$((Antecedents = true)) \rightarrow ((((Hypertension = true) \wedge (Mass = high)) \wedge (Age = advanced)) \vee (Race = white)))$$

True value: 0.8603

$$((Antecedents = true) \rightarrow (Age = advanced))$$

True value: 0.8200

$$((Classification = diabetes)) \rightarrow ((Antecedents = true)) \rightarrow (Hypertension = true)))$$

True value: 0.8131

$$((Mass = high) \vee ((Hypertension = true) \vee (Sex = male)))$$

True value: 0.8124

The following are some rules of association obtained. All predicates have truth values above 0.80. One feature that stands out in these results is the presence in some predicates of the male sex and white race. Another feature to note is that in almost all predicates it appears influence that advanced age. They also frequently appear together the advanced age, antecedents true, diabetic classification, and hypertension true.

Association Rules:

$$(((Classification = diabetes) \wedge (Antecedents = true))) \rightarrow (Sex = male))$$

True value: 0.9532

$$((((Race = white) \wedge (Classification = diabetes)) \wedge (Age = advanced))) \rightarrow (Mass = high))$$

True value: 0.9189

$$(((Antecedents = true) \wedge (Age = advanced))) \rightarrow (((Race = white) \wedge (Sex = male)) \vee ((Mass = high) \vee (Hypertension = true))))$$

True value: 0.9166

$$((Antecedents = true)) \rightarrow ((Race = white) \vee (Age = advanced)))$$

True value: 0.9125

$$(Age = advanced) \rightarrow (Hypertension = true)$$

True value: 0.9036

$$(((Age = advanced) \wedge (Classification = diabetes))) \rightarrow (((Hypertension = true) \wedge (Mass = high)) \wedge (Race = white)))$$

True value: 0.8320

Are shown below a few rules of classification obtained in experiments. From these results it is important to note the influence of high mass, advanced age, hypertension and certain antecedents in having diabetes. Moreover, it also shows the presence of the male sex and race white in the patients suffering from this disease.

Classification Rules:

$$(((((Sex = male) \wedge (Race = white)) \wedge (Hypertension = true)) \wedge (Age = advanced))) \rightarrow (Classification = diabetes))$$

True value: 0.9741

$$(((Antecedents = true) \wedge ((Mass = high) \wedge (Hypertension = true)))) \rightarrow (Classification = diabetes))$$

True value: 0.9014

$$((((Age = advanced) \wedge (Antecedents = true)) \wedge (((Hypertension = true) \wedge (Sex = male)) \wedge (Race = white)))) \rightarrow (Classification = diabetes))$$

True value: 0.8907

$$((((Mass = high) \wedge ((Antecedents = true) \wedge (Hypertension = true))) \wedge (Age = advanced))) \rightarrow (Classification = diabetes))$$

True value: 0.8713

$$(((Antecedents = true) \wedge (Age = advanced))) \rightarrow (Classification = diabetes))$$

True value: 0.8347

$$(((Race = white) \wedge ((Sex = male) \wedge (Age = advanced)))) \rightarrow (Classification = diabetes))$$

True value: 0.7173

The predicates obtained by supervised learning are shown below. The truth values achieved are also lower than in the previous cases. The main feature shown in these results is the influence of being old, have antecedents, and suffer from hypertension, with the onset of diabetes. Also observed certain influences of the white race in the problem.

Supervised learning:

$$((Age = advanced) \rightarrow (Classification = diabetes))$$

True value: 0.8941

$$(Race = white) \wedge (Age = advanced)) \rightarrow (Classification = diabetes))$$

True value: 0.8632

$$(((Antecedents = true) \wedge (Age = advanced)) \rightarrow (Classification = diabetes))$$

True value: 0.8574

$$(((Hypertension = true) \wedge (Age = advanced)) \rightarrow (Classification = diabetes))$$

True value: 0.7041

$$(((Antecedents = true) \wedge (Age = advanced)) \rightarrow (Classification = diabetes))$$

True value: 0.7014

$$(((Sex = male) \wedge (Age = advanced)) \rightarrow (Classification = diabetes))$$

True value: 0.6912

Observations:

In these experiments the majority of predicates reach of the truth value above 0.80. In predicates obtained the main relationships found are:

- Relationship between advanced age, presence of hypertension and obesity.
- Relationship between suffering from antecedents and have advanced age.

- Relationship between suffering from antecedents, hypertension, obesity and have diabetes.
- Relationship between the male sex and the presence of diabetes.
- Relationship between the white race and the presence of diabetes.

Besides taking as measure the high values of truth, we conducted a study of the real characteristics of diabetes for comparison with the results obtained. According to investigations, obesity increases the risk of diabetes and the risk of developing hypertension. Diabetes and hypertension commonly coexist; the appearance of both is common in elderly [5, 9]. Comparing the predicates obtained and the real characteristics of diabetes observed many similarities. Therefore obtained predicates truthfully describe the relationships associated with diabetes.

With respect to sex and race, irrespective of which any person may have diabetes. Therefore, the influence of these two features in the analyzed data can be considered novel discovery.

6 Conclusion

The experimental result indicates that this approach has facilities for obtaining logical predicates that reflect reality. This proposal does not replace existing methods for the discovery of knowledge, but provides a general and flexible approach that enables a new way to extract knowledge.

In the near future the plan is to optimize the tool to obtain different structures of knowledge and combine the use of different metaheuristics algorithms. In addition it's intended to investigate in greater volumes of data and make comparisons with results from other knowledge discovery tools.

References

1. Konar, A.: Artificial Intelligence and Soft Computing: Behavioral and Cognitive Modeling of the Human Brain. CRC Press LLC (2000)
2. Rosete, A.S.: Una solución flexible y eficiente para el trazado de grafos basada en el Es-calador de Colinas Estocástico. PhD thesis, ISPJAE (2000)
3. Dubois, D., Prade, H.: Fuzzy sets and systems: theory and applications. Academic Press, New York (1980)
4. Messerli, F., Bell, D., Bakris, G.: El carvedilol no modifica el peso ni el Índice de masa corporal de los pacientes con diabetes tipo 2 e hipertensión. American Journal of Medicine 120(7), 3–62 (2007)
5. Zimmermann, H.J.: Fuzzy Set Theory and its applications. Kluwer Academic Publishers (1996)
6. Koza, J.R.: Genetic Programming II: Automatic Discovery of Reusable Programs. The MIT Press (1994)

7. Espín, R.A., Fernández, E.G.: La lógica difusa compensatoria: Una plataforma para el razonamiento y la representación del conocimiento en un ambiente de decisión muticriterio. In: Plaza, Valdés (eds.) Multicriterio para la Toma de Decisiones: Métodos y Aplicaciones, pp. 338–349 (2009)
8. Espín, R.A., Mazcorro, G.T., Fernández, E.G.: Consideraciones sobre el carácter normativo de la lógica difusa compensatoria. In: Evaluación y Potenciación de Infraestructuras de Datos Espaciales para el desarrollo sostenible en América Latina y el Caribe, Idict edn., pp. 28–40 (2007)
9. Randall, S.: A Guide to Artificial Intelligence with Visual Prolog. OutskirtsPress (2010)
10. Zegarra, T., Guillermo, G., Caceres, C., Lenibet, M.: Características sociodemográficas y clínicas de los pacientes diabéticos tipo 2 con infecciones adquiridas en la comunidad admitidos en los servicios de medicina del hospital nacional cayetanoheredia. Scielo 11(3), 3–62 (2000)

Categorization of Unlabelled Customer-Related Data Using Methods from Compensatory Fuzzy Logic

Sven Kölpin and Daniel Stamer

Abstract. This paper introduces an approach to enable categorization of unlabelled customer-related data by using methods from Compensatory Fuzzy Logic. At first, a very general and adaptable model, which is used to analyze customer-related data, is demonstrated. This model is called the WWWA(A)-model. It describes the basic and general attributes of all customer-related data. Furthermore, this paper introduces a general workflow that shows how the categorization of unlabelled customer-related data with Compensatory Fuzzy Logic and the WWWA(A)-model should be done. The workflow is very different from other data-mining concepts due to the focus on human knowledge. In order to prove the usability of the WWWA(A)-model and the phase model, data is being analyzed. It is shown how the introduced model can be utilized to derive sets of customer classes from unlabelled customer-related data.

1 Introduction

Gaining a higher standard of knowledge about customers is very important for companies as mostly already stored data is not sufficient [4].

There are numerous reasons why information cannot be gathered directly, though such knowledge is of vital importance for the strategic decisions making process in companies. To extract this knowledge, which is indirectly embedded in stored data on customers, actual expert knowledge is needed. This focuses on the evaluation of available data on customers.

In most cases it is impossible for companies to gather the necessary knowledge directly by querying their customers. On the one hand customer's rights often

Sven Kölpin · Daniel Stamer
Carl von Ossietzky University of Oldenburg, Germany
e-mail:{Sven.Koelpin,Daniel.Stamer}@uni-oldenburg.de

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_12,

prevent a company to improve knowledge in this way because the personal information on individuals, especially in a non-anonymous context, is protected [1].

On the other hand most customers are just not willing to reveal personal information. Main reasons are concerns about personal privacy issues or expected high time expenditure for answering questionnaires or surveys. However, sourcing direct information is important for companies.

In order to generate additional knowledge from customer-related data, methods from multi-valued logic can be used that enable the opportunity to categorize groups of customers. These categories express the additionally gained knowledge by evaluating a customers' membership in a certain class. The problem with different methodologies from the field of multi-valued logic is that most of these models are very difficult to apply practically.

In addition to the expert knowledge from the application domain, information is also needed on how to use the different methods from the logic sciences. Furthermore, it is still necessary to manually execute all calculations, which are needed by a certain method from the applied field of multi-valued logic.

Different methods of multi-valued logic have been implemented in various application domains [8]. A certain problem is that these implementations are highly specialized on the very purpose they have been constructed for.

This is why the central aim of this work is to generate a very general but more useful model that enables the categorization of customer-related data using multi-valued logic. Such a model allows setting the main focus on expert knowledge. Hence, an additional mathematician or logician might not be needed in the process. It also strives to eliminate the need to manually calculate the applied logic model.

To provide a practical example, in which the proposed model can be applied, the reader is encouraged to think of a supermarket chain located somewhere in central Europe. Such a market generates a huge amount of billing data, however, it often lacks in having an automated way of preparing this data or to extract hidden knowledge from these data sets. The billing data of a supermarket contains a lot of information that can be seen as customer-related data that the proposed model can be applied to this domain. For example, it can be used to analyze the data to answer strategic questions as “Who are the customers that come to our market?”

2 Utilized Methods

The main idea of the introduced concept is to generate new knowledge from existing datasets by using the basic concepts of Fuzzy Clustering and Compensatory Fuzzy Logic. These two concepts are described in this section.

2.1 Fuzzy Clustering

A cluster is “a set of entities which are alike, and entities from different clusters are not alike"’ [3]. This means that the idea of (automated) clustering is to group certain data-objects with same properties into one or more clusters by using clustering algorithms. In the approach of this paper, this classification is done automatically.

The use of Fuzzy methods in the context of clustering has the advantage of the creation of intermediate classifications [8]. Each data-object can belong to each cluster with a certain degree of membership. The problem with other clustering methods, e.g. k-means clustering, is that they create binary or "crisp" classifications. This means that an object, or in this case a customer, can only be assigned to one cluster at a time. In the context, this could possibly lead to wrong assumptions, because a customer can have properties that allow him to be in several clusters. In Fuzzy Clustering each object has a certain membership coefficient represented by $u_{i,j} \in [0,1]$ where i is the i^{th} cluster and j the j^{th} object satisfying the two constrains represented in formula (1) and (2).

$$\sum_{i=1}^{K} u_{i,j} = 1, \forall j \tag{1}$$

$$0 < \sum_{j=1}^{N} u_{i,j} < N, \forall i \tag{2}$$

Formula 1 assures the same overall weight of each data-object and formula 2 makes sure that there are no empty clusters [9]. These constraints are the basis for all further calculations of Fuzzy Clusters in this paper.

2.2 Fuzzy Logic

In the conventional Boolean logic theory, the truth value of a predicate p only maps to exactly one of the values of the set {0, 1}, which means it can either be true or false.

In contrast to this, Fuzzy Logic is a form of multi-valued logic which uses truth ranges from the interval [0, 1] to express a certain degree of truth of a specific predicate. Typically, the truth values are defined by the scale of a specific predicate. Typically, the truth values are defined by the scale shown in table 1 [2].

Table 1 Typical scale for Fuzzy truth values

Fuzzy truth value	Semantic meaning
0	false
0,1	nearly false
0,2	very false
0,3	somewhat false
0,4	more false than true
0,5	as true as false
0,6	more true than false
0,7	somewhat true
0,8	very true
0,9	nearly true
1	true

The crucial advantage of using Fuzzy Logic combined with the shown scale is that the knowledge of experts can be easily transformed into Fuzzy values. These values can then be used to calculate an approximate truth value of a predicate. This shows that, because of its power to elaborate linguistic models, Fuzzy Logic is a good methodology to model knowledge and to solve real world problems [5]. Fuzzy Logic can be seen as a communication basis between knowledge experts and decision-makers who are in the case of customer classification the managers of companies.

Fuzzy logic uses membership functions to calculate the truth values of specific predicates. For an element x of X, the membership degree of x to a Fuzzy-set is defined by the membership function $u_A(x)$ where $u_A(x) \in [0,1]$. A membership function is a very useful way to combine the knowledge of an expert with numeric Fuzzy values. Figure 1 shows an example for a membership function.

After the truth values for the predicates have been found, rules for decision-making need to be defined. This is mostly done by using conjunctions (3) disjunctions (4) and negations (5). $u(p)$ represents the truth value of p.

$$u(p \wedge q) = \min(u(p), u(q)) \tag{3}$$

$$u(p \vee q) = \max(u(p), u(q)) \tag{4}$$

$$u(\neg p) = 1 - u(p) \tag{5}$$

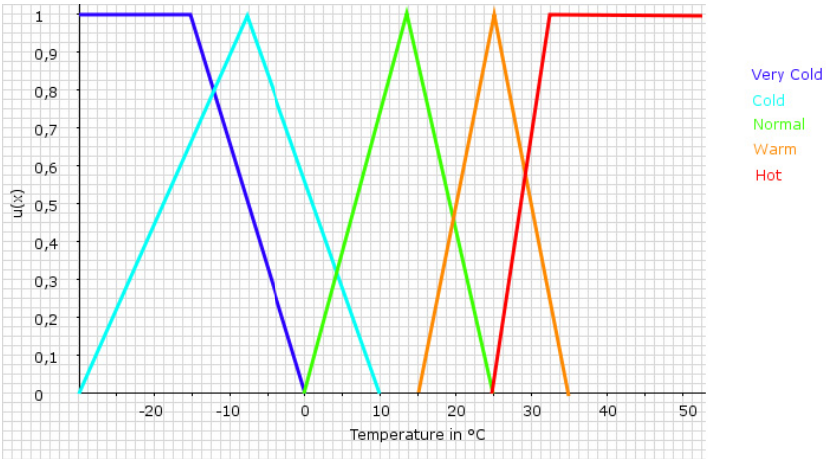


Fig. 1 Example membership function

By using the specific values for the membership degree of a predicate and the operations to connect them, rules can be applied to calculate the truth values for certain decisions. For the example above, the following rules could be created:

```
IF hot USE nothing
IF warm AND NOT hot USE summer jacket
IF (very cold OR cold) USE winter jacket
```

2.3 Compensatory Fuzzy Logic

Compensatory Fuzzy Logic (CFL) is a new approach for handling the decision-making process. It is based on the ideas of the classic Fuzzy Logic methodology but another definition for the logical operators is suggested [2]. These new definitions support the idea that "the increase or decrease of a conjunction or disjunction truth value, resulting from the truth value change of one of their components, can be traded off by the corresponding decrease or increase of the other" [5]. CFL achieves a complete generalization of Bivalued Logic [2].

Because of the performance of CFL, this paper will use the logical operators suggested by this model. In the following, the CFL conjunction (6) and CFL disjunction (7) operators are described.

$$c(x_1, x_2, \dots, x_n) = (x_1 \times x_2 \times \dots \times x_n)^{\frac{1}{n}} \tag{6}$$

$$d(x_1, x_2, \dots, x_n) = 1 - [(1 - x_1) \times (1 - x_2) \times \dots \times (1 - x_n)]^{\frac{1}{n}} \tag{7}$$

3 Properties of Customer-Related Data

In order to develop a model to classify data entities of a customer, a certain awareness of what a customer actually is, has to be identified. The following definition of a customer will clarify the understanding of customers in this context.

"A customer is a person who buys goods or services from a shop or business" [6].

For it is obvious, customers are regarded in a very general understanding. In addition to this definition, the ambiguity from the term "customer-related data" needs to be removed. In this work the focus is on all data, which is, as the terms states, related to the customer in its very role as a customer. Since the earlier definition explains that only as soon as a person purchases a good or service he becomes a customer. Hence, to a customer and his purchases will be refered, when talking about customer-related data.

With this concrete understanding of these terms, the properties of customer-related data can be further investigated.

3.1 Storage

Companies usually store customer-related data in the form of structured relational database patterns [4]. Customer entities have a relationship with billing data of

their purchases in stores, where they have purchased different items for instance. These entities are usually designed very differently and it is to determine, which attributes are common in these designs.

The definition of a customer states that a person is a customer when this person purchases items of any kind (goods or services) from a business. The matter of the business is definitely clear, since it will always be the business that tries to categorize its customers. Since persons only become customers, if they purchase something, these two different entities (person and purchase) always have a relationship with each other.

This concludes that the simplest form of input needs to link data between a customer and at least one item the customer has purchased. The customer may be anonymous, but the actual purchase can still be identified.

3.1.1 Generalized Form

In a generalized form these data sets basically appear in the form that is portrayed in figure 2. This form assumes that a purchase is always associated with four fixed attributes.

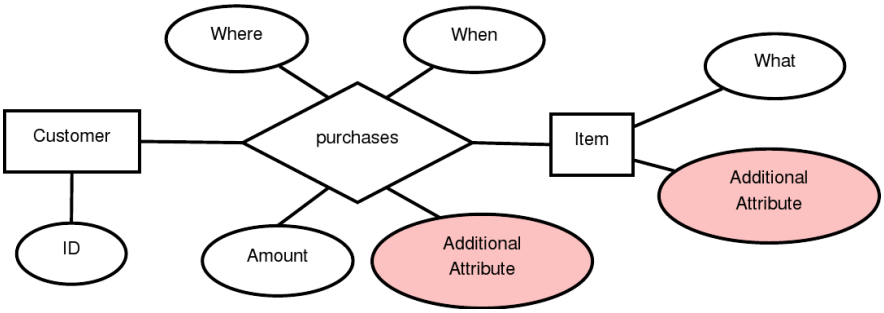


Fig. 2 Generalized form of entities and their relationships in input data

The four basic attributes are defined in the WHAT, WHEN, WHERE AMOUNT (WWWA(A))-model which is described in the following:

- **WHAT**: The first attribute "what" focuses on the type of a good or service that actually has been purchased. Good data may be a unique identifier or a unique, unambiguous description in a textual form.
- **WHEN**: The attribute "when" describes the point of time when the purchase has been made. The data type should be a time stamp.
- **WHERE**: The attribute "where" refers to the place where a purchase has been conducted.
- **AMOUNT**: The attribute "amount" describes the quantity of items that have been purchased by a customer. This could be represented as an integer (a person bought 3 pizzas) or as a double (a person bought 1.5 liters of water).

Due to the fact that this model of customer-related data is very abstract and generalized, another attribute has been added that it can be adapted to all use cas-

es. This attribute is called the "Additional Attribute". In the example of the customer-related data of a mobility service provider, the additional attribute could contain information about the distance that a customer has travelled.

Furthermore, the model in figure 2 assumes that a customer can always be addressed by an ID. If such a pseudo identity of a customer cannot be determined, the ID of the actual purchase can always be given.

3.2 Domain Expert Knowledge

There is a lot of very distinct expert knowledge on the evaluation of the data sets given in companies. This knowledge is mostly available from people, who do not have a technical background and do not come from computer science or other fields of practical or theoretical logic.

Before any kind of calculation is being done, target categories need to be defined by the experts such as the managers of a company. These target categories will then be populated with the individual memberships of each customer object. It is important to understand that technically all customer objects appear in each class, but their actual membership of $u(x)$ within a certain class might be 0.0 or another very small value. This demonstrates that their actual membership is very improbable, however, not totally impossible. In the following, a sample is shown of how these categories might look like within the example of the super market.

- *Children* - When mostly sweets and soft drinks are purchased on the cash desk, it is very likely that the customer in front of it is underage.
- *Elderly people* – This group of customers mostly tend to buy goods in very small portions, thus they are able to store them for a long time.
- *Family* – In this case the buyer may purchase huge amounts of food and beverages to cover the needs of a family.
- *Student* - This class might buy goods, which have the aim to solve a purpose, such deep frozen pizza and occasionally alcoholic beverages on Fridays or Saturdays. We are aware of the fact, that this is actually a pretty strong stereotype, but it might still be of interest.
- *Vegetarian* - It might be very interesting to see how many vegetarians actually shop in a particular market, since this number of customers is steadily increasing in recent years.

Indeed, the model needs to be provided with the membership functions, which allow the classification of customers. These function come in the form, which is portrayed in figure 1.

At last, a number of rules are necessary, which will be used to sort the customers into their classes. The general appearance of these classes is shown earlier and a more concrete example, which suits the example of the supermarket should look as following.

```
Student WHEN timeOfPurchase IS very late AND
numberOfBoughtProducts IS few
```

4 General Approach

A general phase model for the classification of customer-related data is shown in the figure 3. The proposed workflow consists out of five steps which are explained in the following.

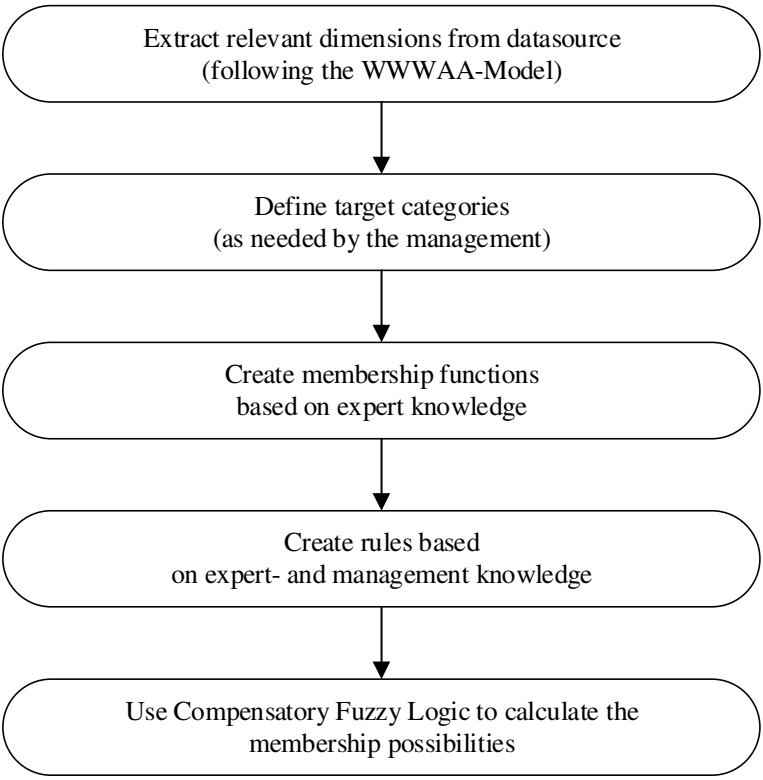


Fig. 3 Phase model for the classification of customer-related data

- **Extract relevant dimensions:** In this step, the relevant dimensions of the customer-related data need to be extracted from a company's data source. Most of the times, data will be stored in relational databases. However, there is no restriction to what form the data source needs to have in this model. It can be a (relational) database, a flat file or even data objects from any object-oriented programming language. The important thing is that at least the WHAT, WHEN WHERE and the AMOUNT (WWWA) will be extracted that a Fuzzy categorization is possible.
- **Define target categories**: The next step is the definition of the target categories. This stage of the model is very different from other data-mining strategies due to the fact that the categories or clusters are defined by a person before data has been analyzed by an algorithm. An example for defining target categories

can be taken from the use case of a supermarket. To improve their marketing strategies, the management could be interested in finding out, who the customers of their markets are. In this case, possible target classes for the clustering are for instance “students”, “employees”, “teenager” and “housewives”.

- **Create membership functions**: The definition of membership functions requires expert- and domain knowledge. This is why in this phase a lot of communication with such experts needs to be done by the management. First, the required amount of knowledge needs to be gained, for example by using statistical methods or surveys. On the basis of this knowledge, the membership functions that allow calculating a Fuzzy membership-degree of a data object can be defined. For the use case of the supermarket, statistical data can be used to define the required membership functions. An example membership function is shown in figure 4.
- **Create Rules**: After the membership-functions on at least all dimensions of the WWWA-model have been defined, the rules that describe, when a customer-object belongs to a specific category, need to be created. These rules should combine the different dimensions that a reliable categorization is possible. As in the step before, the management and domain-experts need to be involved in this phase. Conjunctions and Disjunctions from Compensatory Fuzzy Logic will be used at this point.
- **Calculate membership possibilities of the datasets**: The last phase is the automated calculation of the Fuzzy Clusters. By using the defined rules, membership functions and target categories, an implementation of this approach should be able to calculate automatically the clusters and to visualize the results in an appropriate way. It should also be possible to use the results of the clustering as a new basis for creating new knowledge.

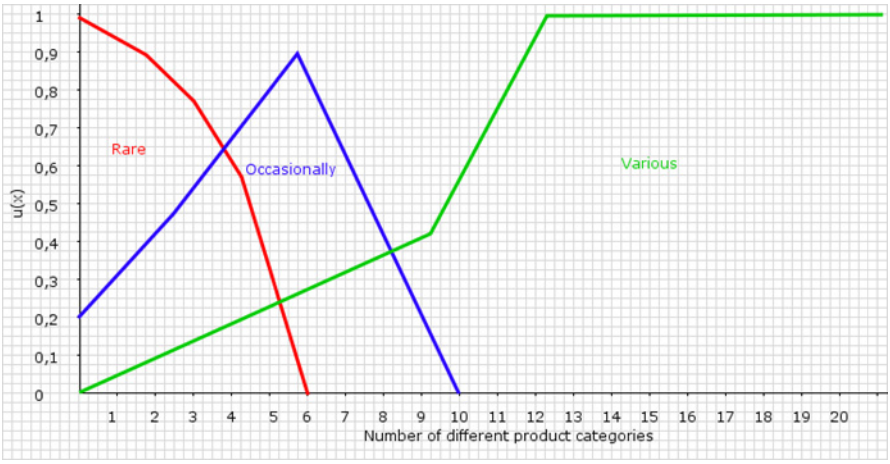


Fig. 4 Use case: Membership functions for a supermarket

5 Practical Use Case

The introduced WWWA(A)-model and the proposed workflow has been implemented in an application named “CFL customer categorization tool”. Random billing data from a supermarket is restructured into the XML-Format that it represents the WWWA(A) model. The following figure 5 shows a small extract from the WWWA(A) XML-File.

```
<datamodel>
    <entry>
        <customer>Student_1</customer>
        <purchase>
            <when>
                <minute>41</minute>
                <hour>21</hour>
                <day>13</day>
                <month>09</month>
                <year>2005</year>
            </when>
            <where>
                <name>Combi Ammerlaender Herrstrasse</name>
                <street>Combi Ammerlaender Herrstrasse</street>
            </where>
            <items>
                <item>
                    <what>Pizza</what>
                    <amount>
                        <number>3</number>
                    </amount>
                </item>
                <item>
                    <what>Beer</what>
                    <amount>
                        <number>6</number>
                    </amount>
                </item>
            </purchase>
```

Fig. 5 XML-Representation of the WWWA(A)-model

The application is able to use this XML-structure as a data model and the proposed workflow-model for the categorization of unlabelled customer-related data.

We tested the proposed workflow model with 10.000 training data sets, using only three target categories (Teenager, Housewives and Students). We managed to make the right categorization in about 71% of the time.

The quality of the categorization of the training data sets mainly depends on the rules that connect the different membership functions. After our first attempt of a categorization, we formulated more complex rules and managed to get better results.

The second approach showed an improved percentage of correct categorization. This time we were able to match the right category in about 77% of the test cases. However it also appeared to us, that the percentage of successful categorization also depends on the number of target categories. A relatively small set of target categories allows a relatively precise categorization. A huge variety of target categories causes an opposite result.

It is therefore conceivable to use a proportionally well designed composition of well-defined rules and a relatively small set of target categories.

6 Conclusion

A general data model for the categorization of customer-related data was introduced. This model, called the WWWA(A)-model, was used in a concrete use case to show its usability.

The introduced workflow model, which enables the categorization of customer-related data following the WWWA(A)-model, was described and tested. The practical use case showed that the quality of the results mainly depend on the rules that an expert creates. In the future, it could be possible to use genetic algorithms to improve the rules that have been created by a human expert.

References

1. Agre, P.E., Rotenberg, M.: Technology and Privacy: The New Landscape. The MIT Press (1998)
2. Espin, R., Fernandez, J., Marx-Gomez, J., Lecich, M.: Compensatory Fuzzy Logic: A Fuzzy normative model for decision making. Investigación Operativa. Universidad de la Habana (2), 188–197 (2006)
3. Everitt, B.: Cluster Analysis. Social Science Research Council, London (1980)
4. Grothe, M., Gentsch, P.: Business Intelligence - Aus Information Wettbewerbsvorteile gewinnen. Addison-Wesley (2010)
5. Javier, G., Espin, R., Laura, V.: A framework for tissue discrimination in Magnetic Resonance brain images based on predicates analysis and Compensatory Fuzzy Logic, pp. 1–16. TSI Press (2008)
6. Pearsall, J., Hanks, P.: Oxford Dictionary of English. Oxford University Press (2010)
7. Rosete, A., Espin, R., Marx-Gomez, J.: A general approach for knowledge discovery based on Compensatory Fuzzy Logic and metaheuristic search (2006)
8. Xu, R., Wunsch, D.: Clustering. IEEE Press (2009)
9. Zadeh, L.: Fuzzy Sets. Information and Control, 338–353 (1965)

Knowledge Discovery by Fuzzy Predicates

Taymi Ceruto Cordovés, Alejandro Rosete Suárez,
and Rafael Alejandro Espín Andrade

Abstract. With the rapid growth of databases in many modern enterprises, data mining has become an increasingly important approach for data analysis. The operations research community has contributed to this field significantly, especially through the formulation and solution of numerous data mining problems as optimization problems. This paper provides a survey of a hybrid version of metaheuristics and data mining with the purpose of obtaining fuzzy predicates. We believe that the patterns obtained by this technique represent a combination free of logical operators that other algorithms can't obtain. Finally, in this paper we apply the method in a good dataset and show the conclusions.

1 Introduction

The traditional method of turning data into knowledge relapses on manual analysis and interpretation. This form of manual probing of a data set is slow, expensive, and highly subjective. In fact, as data volumes grow dramatically, this type of manual data analysis is becoming completely impractical in many domains. Databases are increasing in size in two ways: the number of records in the database and the number of attributes. This job is certainly not one for humans; hence, analysis work needs to be automated, at least partially.

The knowledge discovery in databases (KDD) has evolved from the intersection of research fields such as machine learning, pattern re-cognition, databases, statistics, artificial intelligence (AI), knowledge acquisition for expert systems, data visualization, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in the context of large data sets [19, 20].

KDD refers to the overall process of discovering useful knowledge from data, and Data Mining (DM) refers to a particular step in this process. The additional

Taymi Ceruto Cordovés · Alejandro Rosete Suárez · Rafael A. Espín Andrade
"Jose Antonio Echeverria" Higher Technical Institute,Cuba
e-mail: {tceruto,rosete}@ceis.cujae.edu.cu,
rafaelespin@yahoo.com

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_13,

steps in the KDD process, such as data pre-paration, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data [13].

2 Data Mining

DM is the extraction of implicit, previously unknown, and potentia-lly useful information from data. It is the application of specific algorithms for extracting patterns from data. Strong patterns, if found, will likely generalize to make accurate predictions on future data [13].

The discovered patterns should be valid on new data with some degree of certainly. We also want patterns to be novel (at least to the system and preferably to the user) and potentially useful, that is, lead to some benefit to the user or task. Finally, the patterns should be understandable, if not immediately then after some postprocessing.

The knowledge discovery goals are defined by the intended use of the system. We divide the discovery goal into **prediction**, where the system finds patterns for predicting the future behavior of some entities, and **description**, where the system finds patterns for presentation to a user in a human-understandable form [19].

Below are the main types of mining [3, 19, 39, 41, 42]:

1. **Association Rule Mining** (Apriori, Genetic Algorithms, CN2 Rules): rule sets are the most ancient knowledge representations, and probably the easiest to understand. It involves the discovery of rules used to describe the conditions where items occur together are associated. Usually the condition of a rule is a predicate in certain logic, and the action is an associated class, meaning that we predict action for an input instance that makes true condition. It often used for market basket or transactional data analysis.
2. **Classification and Prediction** (CART, CHAID, ID3, C4.5, C5.0, J4.8): involves identifying data characteristics that can be used to generate a model for prediction of similar occurrences in future data. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descen-ding from that node corresponds to one of the possible range of values for this attribute.
3. **Cluster Analysis** (K-Means, Neural Networks, Expectation-Maximization): attempts to look for groups (clusters) of data items that have a strong similarity to other objects in the group, but are the most dissimilar to objects in other groups.

In a general way it can be said that these algorithms use a limited knowledge representation; because the structure of the knowledge to be obtained has few variation possibilities. For example: the rules are only conditional relation among conditions and conclusions. This restricts the possibility to obtain a new

knowledge, where the main logical relation is the conjunction or disjunction or equivalence or any free combination of logical operators.

Each knowledge representation language restricts the space of possible solutions because of the limitations of its definition. It means that any learning algorithm and knowledge representation can be only the best algorithm in a certain subset of problems.

For this reason the main objective of this work is to define a new way for knowledge discovery in databases, which allows obtaining fuzzy predicates as more flexible way of representing knowledge, using different metaheuristic algorithms.

3 Background

In order to obtain predicates, three main approaches are relevant from the literature:

Inductive Logic Programming (ILP) [35] has been defined as the intersection of inductive learning and logic programming. It is a discipline which investigates the inductive construction of first-order clausal theories from examples and background knowledge. ILP inherits its goal: to develop techniques to induce hypotheses from observations (examples) and to synthesize new knowledge from experience.

ILP has limitations as a classical machine learning techniques. It needs a set of observations (positive and negative examples), background knowledge, hypothesis language and covers relation.

Genetic Programming (GP) [24, 25, 30] is based on Genetic Algorithms. The main idea is to obtain a mathematic expression that relates some variables with a given target variable. The mathematic expression is often expressed as a tree, where the internal nodes are operators (such as addition, multiplication, etc), and the terminal nodes are variables or constants. Each tree is evaluated in each example according to the error between the result of its application and the target variable.

GP has as limitations the use exclusive of genetic algorithms, the learning is supervised and the trees that are obtained can vary of size (it implies that it is necessary to implement limits in the growth). Our proposal is different in two senses: learning can be unsupervised and other algorithms (not genetic algorithms) may be used.

Discovering Fuzzy Association Rules [1, 7, 8, 9, 10, 18, 21, 22, 26, 27, 28] The linguistic representation makes those rules discovered to be much natural or human experts to understand. The definition of linguistic terms is based on set theory and hence we call the rules having these terms fuzzy association rules. In fact, the use of fuzzy techniques has been considered as one of the key components of DM systems because of the affinity with the human knowledge representation. This approach is based on Genetic Algorithms or Ant Colony Optimization and the structure of the predicate is predefined (rules).

As our proposal is to obtain any sort of fuzzy predicates, none of these notable approaches are directly usable. The predicates to be obtained are not advocated to relate to a specified variable, nor to have a specific structure, but to obtain fuzzy predicates with great truth values in the given examples. To the best of our knowledge, there has not been another attempt to obtain fuzzy predicates before.

4 Knowledge Discovery by Fuzzy Predicates

This paper is based on two components of Softcomputing. Basically, Soft Computing is not a homogeneous body of concepts and techniques. It's a collection of methodologies that aim to exploit the tolerance for imprecision and uncertainty to achieve tractability, robustness, and low solution cost [6, 36, 39, 40, 45].

The viewpoint that we will consider here is other different to the one outlined by Zadeh in 1994. For us its principal constituents are **fuzzy sets**, probabilistic reasoning, neural networks and **metaheuristics** [40]. In large measure, these components are complementary, not competitive. It is becoming increasingly clear that in many cases it is advantageous to combine them. The hybridization in the Soft Computing context favors and enriches the appearance of original procedures which can help resolve new problems, like this. Softcomputing is likely to play an increasingly important role in many application areas.

Our data-mining method can be viewed as three primary algorithmic components: (1) model representation, (2) model evaluation, and (3) search.

4.1 Model Representation

Model representation is the language used to describe discoverable patterns. If the representation is too limited, then no amount of examples can produce an accurate model for the data. It is important that a data analyst fully comprehend the representational assumptions that might be inherent in a particular method.

For representation, we choose a compact description of relations based in Normal Forms, because in classic logic any predicate may be represented in these forms. It worth clarifying that the equivalences that are obtained from this alternative are more true than false in Multi-valued logic, but not absolutely true. In particular, we use the Conjunctive and Disjunctive Normal Forms with connectives and modifiers for intensifying or moderating the value of the variable [5, 15, 16, 23].

We also decided to use to represent the genotype a mixture of Michigan approach with Iterative Learning [14]. In the Michigan approach an individual is a predicate and the whole population is the solution to the problem. In the other hand, in Iterative Learning an individual is a predicate, like in Michigan, but the solution provided by the algorithm is the best individual of the population; although the final solution is the concatenation of the predicates obtained by running the algorithm several time.

We use a chromosome fixed-length and Integer Coding, but in reality the predicate has variable size because we add a special value that indicates the absence of that variable in the predicate. Each variable involved in the predicate can take different values according to the following scale.

Scale
0 it isn't in the predicate
1 it's in the predicate
2 it appears denied
3 it appears with the modifier very (it intensifies the value of the variable)
4 it appears with the modifier hiper (it intensifies more the value of the variable)
5 it appears with the modifier something (effect contrary of the very)

In the chromosome the variables can appear twice or more times, but in this case it include two or more clauses. For that reason the two last positions in the chromosome represent the normal form type (1-CNF, 2-DFN) and the number of clauses.

A code example with its corresponding predicate is shown next:

$(X_1 \wedge \neg X_2) \vee (X_0 \wedge (X_1)^2)$

X_0	X_1	X_2	X_0	X_1	X_2	NFT	NC
0	1	2	1	3	0	2	2

4.2 Model Evaluation

Model-evaluation criteria are quantitative statements (or fit functions) of how well a particular pattern. This is a key factor for the extraction of knowledge because the interest obtained depends directly on them. Furthermore, quality measures provide the expert with the importance and interest of the predicates obtained.

Fuzzy representation is a way to more completely describe the uncertainty associated with concepts. For each solution the unique aspect that was considered is the truth value, which depends on the number of clauses, variables and rows of the data set. The truth value is calculated using Fuzzy Logic (FL). Ever since Zadeh [44] suggested the use of the 'min-max' operators, but many researchers added and continue to add various new interpretations to the representation of combined concepts introducing new computational formulas for 'AND', 'OR' and 'IF... THEN', etc.

The main feature of the FL is that it doesn't give a unique definition of the classic operations as the union or the intersection. We studied about the appropriate fuzzy operators for decision-making. For example, Zadeh operators (min-max) are insensitive. In this case, the change of one variable does not change the value of the result. The probability operators aren't idempotent; the conjunction of two variables, with the same values, does not result in the same number, in fact the results with a lower value [31, 32].

We also studied other operators such as Hamacher, Einstein, Lukasiewicz, and Drastic and determined that they have the same problem, because they are associative [33, 34]. On the contrary, the compensatory fuzzy logic is sensitive and idempotent [11, 12]. This aspect is very important for the correct interpretation of the results and for this reason we studied another type of operators. We researched the Harmonic mean, the Geometric Mean and the Arithmetic Mean, and their dual. All these have a value between the maximum and minimum.

4.3 Search

Search strategy is very important and before defining it, the first thing that it is necessary to keep in mind it is the dimension of the space of the search. Search method consists of two components: (1) parameter search and (2) model search. In many cases the data-mining problem has been reduced to purely an optimization task: find the patterns that optimize the evaluation criteria.

Metaheuristics represent an important class of techniques to solve, approximately, hard combinatorial optimization problems for which the use of exact methods is impractical. Heuristic methods such as Tabu Search, Simulated Annealing, Hill Climbing, Genetic Algorithm [4, 17, 29, 38] have been used each of them in isolation, or in combination.

Many successful applications have been reported for all of them. According to NFL Theorem [43] it is impossible to say which is the best of all algorithms. It depends on the encoding of the problem, the correct selection of the objective function as well as this of the operators. So, the only possibility is to make experiments with different parameters.

In particular, we use BiCIAM [2, 37]. It is Open Source library that represents a unified model of metaheuristic algorithms based in "Generate and Test" paradigm. It has been developed in our university. The most important aspect is that it separates the problem of the algorithm and defines the characteristics of the problem only once. This library allows also the comparison of algorithms.

5 Experimental Results

To evaluate the usefulness of the proposed approach some experiments have been carried on a real-world database extracted from the UCI Machine Learning Repository. It is a source of very good data mining datasets at http://www.ics.uci.edu/~mlearn/MLOther.html.

That site also includes references of other good data mining sites.

The real-world database "**Adult**" was extracted from the census bureau database in 1994 by Ronny Kohavi and Barry Becker, Data Mining and Visualization, Silicon Graphics.

This database consists of 48842 records with 14 attributes each one. But to develop the different experiments, we extracted the 4 attributes from them: age,

hours-per-week capital-gain and capital-loss. The 4 continuous attributes we mapped to ordinal attributes as follows:

- **age**: cut into levels Young (0-25), Middle-aged (26-45), Senior (46-65) and Old (66+).
- **hours-per-week**: cut into levels Part-time (0-25), Full-time (25-40), Over-time (40-60) and Too-much (60+).
- **capital-gain and capital-loss**: each cut into levels None (0), Low (0 < median of the values greater zero < max) and High (>=max).

The following values have been considered for the parameters of each approach (Hill Climbing):

30 independent runs of 500 iterations, each one
14 bits per gene for the codification,
In particular for genetic process:
20 individuals, 10 truncation
0.9 as crossover probability
0.5 as mutation probability

Several experiments have been carried to analyse the fuzzy predicates obtained by the proposed approach. Analysing the results we can highlight the following conclusions:

- The results obtained by the proposed approach, presenting a good relationship between the size of the search space and the results obtained, and getting a good truth value (quality measures > 0.8).
- Our approach which allows us to obtain diversity types of predicates in the same run. But it also returns repeated solutions, that should be eliminated in a postprocessing phase.
- The results show how the proposed approach obtained knowledge with high truth value. Furthermore, the interpretability and visualization of the predicates is maintained in a high level. For that reason we will continues work in that direction.
- On the other hand, we have to take care with the relationship between the runtime and size of the search space, because now the runtime isn't totally reasonable.

6 Conclusion

In this paper, a new form of discovering knowledge has been considered that to extract fuzzy predicates from data given. To do that, we have proposed the use of specific representation model and metaheuristics algorithms for the search.

Here, we present our conclusions:

- The learning process is not supervised.
- The form of the predicates isn't totally restricted

- We uses different fuzzy operators (although compensatory is privileged by its properties)
- There is more than one search method available.

We haven't found yet in the literature a method that has these same characteristics.

Acknowledgments. Many people have helped us, and we relish this opportunity to thank them. We have received generous encouragement and assistance from the academic staff members of the Iberian-American Network named Eureka.

References

1. Alcalá-Fdez, J., Alcalá, R., Gacto, M.J., et al.: Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms. Fuzzy Sets and Systems 160(7), 905–921 (2009)
2. Alvarez, A., Gonzalez, Y.: Biblioteca de clases para la integración de algorit-mos metaheurísticos, Alvarez&Gonzalez. GRIAL. CUJAE, Pregrado (2009)
3. Berry, M., Linoff, G.: Data Mining Techniques John Wiley & Sons (2004)
4. Blum, C., Roli, A.: Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. ACM Computing Surveys 35(3), 268–308 (2003)
5. Bruno, A.: Normal forms. Mathematics and Computers in Simulation 45 (1998)
6. Cabrera, I.P., Cordero, P., Ojeda-Aciego, M.: Fuzzy Logic, Soft Computing, and Applications. In: Cabestany, J., Sandoval, F., Prieto, A., Corchado, J.M. (eds.) IWANN 2009, Part I. LNCS, vol. 5517, pp. 236–244. Springer, Heidelberg (2009)
7. Chan, K.C.C., Au, W.-H.: Mining fuzzy association rules. In: Chan, Au (eds.) Proceedings of the Sixth International Conference on Information and Knowledge Management, pp. 209–215. ACM, Las Vegas (1997)
8. De Cock, M., Cornelis, C., Kerre, E.E.: Fuzzy Association Rules: A Two-Sided Approach FIP (International Conference on Fuzzy Information processing: Theories and Applications), 385–390. DeCock&et (2003)
9. Delgado, M., Marín, N., Sánchez, D., et al.: Fuzzy Association Rules: General Model and Applications. IEEE Transactions on Fuzzy Systems 11(2) (2003)
10. Delgado, M., Manín, N., Martín-Bautista, M., et al.: Mining Fuzzy Association Rules: An Overview. In: Nikravesh, M., Zadeh, L.A., Kacprzyk, J. (eds.) Soft Computing for Information Processing and Analysis. STUDFUZZ, vol. 164, pp. 351–373. Springer, Heidelberg (2005)
11. Espín, R., Marx, J.C., Mazcorro, G., et al.: Compensatory Logic: A Fuzzy Approach to Decision Making. In: Fourth International ICSC Symposium on Engineering of Intelligent Systems (EIS 2004), Espín&et, Island of Madeira (2004)
12. Espín, R.A., Fernández, E.: Compensatory Fuzzy Logic: A Frame for Reasoning and Modeling Preference Knowledge in Multicriteria Decision- Making. Information Sciences (2010)
13. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM 39(11) (1996)

14. Garrelli, J.M.: Pittsburgh Genetic-Based Machine Learning in the Data Mining: Representations, generalization, and run-time. Computer Science Department. Garrelli, Barcelona, Universitat Ramon Llull. Doctor: 352 (2004)
15. Gehrke, M., Walker, C., Walker, E.: Some comments on fuzzy normal forms. FUZZ IEEE 2 (2000)
16. Gehrke, M., Walker, C.L., Walker, E.A.: Normal forms and truth tables for fuzzy logics. Fuzzy Sets Syst. 138(1), 25–51 (2003)
17. Glover, F., Melián, B.: Tabu Search. Revista Iberoamericana de Inteligencia Artificial 19, 29–48 (2003)
18. Gyenesei, A.: A Fuzzy Approach for Mining Quantitative Association Rules. Gyenesei, Turku Centre for Computer Science (2000)
19. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann (2006)
20. Herrera, F., Carmona, C.J., González, P., et al.: An overview on Subgroup Discovery: Foundations and Applications. Knowledge and Information Systems (2010)
21. Hong, T.-P., Kuo, C.-S., Chi, S.-C.: Mining association rules from quantitative data. Intelligent Data Analysis 3(5), 363–376 (1999)
22. Hong, T.-P., Lee, Y.-C.: An Overview of Mining Fuzzy Association Rules. In: Bustince, H., Herrera, F., Montero, J. (eds.) Fuzzy Sets and Their Extensions: Representation, Aggregation and Models. STUDFUZZ, vol. 220, pp. 397–410. Springer, Heidelberg (2008)
23. Kandel, A., Zhang, Y.-Q., Miller, T.: Knowledge representation by conjunctive normal forms and disjunctive normal forms based on n-variable-m-dimensional fundamental clauses and phrases. Fuzzy Sets and Systems 76, 73 (1995)
24. Konar, A.: Artificial Intelligence and Soft Computing: Behavioral and Cognitive Modeling of the Human Brain. CRC Press, Boca Raton (2000)
25. Koza, J.R.: Genetic programming: a paradigm for genetically breeding populations of computer programs to solve problems. Koza, Universidad de Stanford (1990)
26. Kuok, C.M., Fu, A., Wong, M.H.: Mining fuzzy association rules in databases. SIGMOD Rec. 27(1), 41–46 (1998)
27. Mata, J., Alvarez, J.L., Riquelme, J.C.: Mining numeric association rules with genetic algorithms. In: Proc. of the Conf. Mata, pp. 264–267. ICANNGA, Praga (2001)
28. Mata, J., Alvarez, J.-L., Riquelme, J.-C.: Discovering Numeric Association Rules via Evolutionary Algorithm. In: Chen, M.-S., Yu, P.S., Liu, B. (eds.) PAKDD 2002. LNCS (LNAI), vol. 2336, pp. 40–51. Springer, Heidelberg (2002)
29. Melián, B., Moreno Pérez, J.A., Moreno Vega, J.M.: Metaheurísticas: una vi-sión global. Revista Iberoamericana de Inteligencia Artificial 2(19), 7–28 (2003)
30. Mesbah, S., Toony, A.A.: Applications of Genetic Programming in Data Mining. World Academy of Science, Engineering and Technology (17) (2006)
31. Mizumoto, M.: Fuzzy Sets and their Operations II. Information and Control 50(2) (1981a)
32. Mizumoto, M.: Fuzzy Sets and their Operations I. Information and Control 48(1) (1981b)
33. Mizumoto, M.: Pictorial representations of fuzzy connectives, part I: cases of t-norms, t-conorms and averaging operators. Fuzzy Sets and Systems 31 (1989a)
34. Mizumoto, M.: Pictorial representations of fuzzy connectives, part II: cases of compensatory operators and self-dual operators. Fuzzy Sets and Systems 32 (1989b)
35. Muggleton, S., De Raedt, L.: Inductive Logic Programming: Theory and methods. The Journal of Logic Programming 19(suppl. 1), 629–679 (1994)

36. Papadrakakis, M., Lagaros, N.D.: Soft computing methodologies for structural optimization. Applied Soft Computing 3 (2003)
37. Paredes, D., Fajardo, J.: Biblioteca de clases para la unficación de algoritmos metaheuristicos basados en un punto. GRIAL, CEIS, Paredes&Fajardo, Ciudad Habana, Instituto Superior Politécnico "José Antonio Echeverría". Pregrado (2008)
38. Talbi, E.-G.: Metaheuristics: from design to implementation, John Wiley & Sons, Inc. (2009)
39. Venugopal, K.R., Srinivasa, K.G., Patnaik, L.M.: Soft Computing for Data Mining Applications. Springer, Heidelberg (2009)
40. Verdegay, J.L., Yager, R.R., Bonissone, P.P.: On heuristics as a fundamental constituent of soft computing. Fuzzy Sets and Systems 159(7), 846–855 (2008)
41. Wang, L., Fu, X.: Data Mining with Computational Intelligence. Springer (1998)
42. Witten, I.H., Frank, E.: Data Mining Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2005)
43. Wolpert, D., Macready, W.: No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation 1(1), 67–82 (1997)
44. Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338–353 (1965)
45. Zadeh, L.A.: Soft Computing and Fuzzy Logic. IEEE Software 11(6), 48–56 (1994)

Innovative Wind Farm Control

Mischa Böhm, Oliver Norkus, and Deyan Stoyanov

Abstract The change from nuclear power to renewable energy currently provides a large part of the ongoing discussions in politics, science and business. From the renewable energies' standpoint, wind energy shall be a key technology in the future energy mix. Concerning offshore wind energy in Germany there are some projects currently carried out on construction of wind farms and further more are planned. However, large uncertainties still exist due to the gap of experience in the operations of offshore wind farms. It has not been investigated yet, with which key figures they can be controlled, nor what aspects of wind farm management could be automated. Moreover, the use of methods and techniques of BI in this domain have not been researched yet and the existing utilization is just realized in a decent way. For these reasons the University of Oldenburg started to join this field of research. To establish in this area there was a research on a ratio analysis and on a flexible, scalable and modern BI reporting, which entered prototypical the domain of wind farm control, based on a dashboard and a mobile application. This chapter gives an overview of the results of this research.

1 Indicators for Wind Farm Management

Performance measurement systems are among the most important tools for decision making. To improve the control and the reporting of wind farms it is necessary to ensure that the key performance indicators are substantial investigated and standardized. The central challenge was to develop a performance measurement system for wind farm control. This challenge is settled in the field of analysis-oriented business intelligence.

The aim of the indicator analysis was to design a structured performance measurement system. The result is an ordered set of indicators for the wind farm control.

Mischa Böhm · Oliver Norkus · Deyan Stoyanov
Carl von Ossietzky University of Oldenburg, Germany
e-mail: {mischa.boehm,deyan.a.stoyanov}@uni-oldenburg.de,
oliver.norkus@gmail.com

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_14,

The approach to design a performance measurement system for wind farm control was to proceed in a three-stage development. First, relevant literature was identified and studied ([2, 4, 9, 10, 11, 15]). In summary, the management, the relevant processes and the question, which are the relevant key figures for controlling a wind farm, are rather unstructured and less standardized, often spontaneously and not optimized (c.f. p. 842 ff. [5]). This aspect Mr. Gößmann, CTO of EnBW Regional AG[1], also accentuate in an interview entitled *Intelligent Networks are an opportunity for innovative Companies*:

> "Especially by the fact that even with the large number of requirements no ready solution exists, there are great chances [...]." (p.20, [9])

Within the literature studies the relevant knowledge has been extracted. Subsequently, on the basis of extracted knowledge from the literature a theoretical approach for the performance measurement system for wind farm control was designed. In the next step this theoretical approach underwent an evaluation and optimization by experts.

In the first phase the subject was the study of literature. Through a comprehensive review of relevant literature in the domains wind energy plants, wind power, wind turbines and wind farms the knowledge for designing a first version of the performance measurement system was extracted. At this stage relevant institutions of these domains were already contacted and gave literature recommendations.

In the next phase the indicator system was performed in a first version. This first version represents a theoretical approach because the design was carried out on the basis of the extracted knowledge. Under this concept the following steps were performed:

Definition: The performance indicators have been defined based on their syntactic and semantic shape, by use and targets. The monitoring and control subject was calculated and displayed. The result of this first step of designing the performance measurement system was the key indicators and their definition.

Selection and categorization: Each of these figures was described by a fact sheet. The set of indicators was factually and logically divided into categories and subcategories. In addition to the factual and logical grouping the subdivision was also based on mathematical relationships between the individual indicators. The result of this step was the orderly and assorted performance measurement system in a theoretical approach, and the description of any indicator by ratios titles and further attributes.

The performance measurement system for the wind farm control in the version of the theoretical approach was the entry point into the third phase: *optimization and evaluation*. To make this project of performance measurement system known and to gain further evaluation partner, it was published into a German scientific journal of system for measuring and controlling [7].

[1] EnBW Regional AG is a subsidiary of EnBW AG, Germany's third largest power company.

The evaluation partners can be segmented into three areas:

- Wind farm operators, primarily in Germany, but also from Italy
- manufacturers of wind turbines and their components
- provider of software for the operation of wind farms

With each evaluation partner the following steps were implemented. The particular result of the current evaluation phase represents the object to be optimized in the next evaluation iteration.

1. Presentation of the presentment, architecture and structure of the performance measurement system.

Discuss and refine the appearance, architecture and structure.
Presentation of the individual indicators of performance measurement system.
Discussion of the individual figures especially with regard to target values, control intervals and calculating and dealing with discrepancies.
Optimization of single figures, in addition of the indicator system to specific figures, discussion of key figures.

After the evaluation and optimization within the involvement of experts the performance measurement system was finally designed.

Regarding the architecture of a performance measurement system is to distinguish between the type of linkage of the individual indicators and the type of derivation. The type of linkage of the individual indicators differentiated between a hierarchical and a no-nonsense logic operation. By type of derivation is to distinguish between empirical-inductive and deductive-logical [8]. The individual figures for the performance measurement system for wind farm control mainly are linked in a qualitative context. Also available with a factual and logical link these figures are in conjunction with each other and affect mostly independently control the object. As an example, the oil values of the generator may be mentioned. The oil pressure, oil temperature and oil level are factually and logically connected but interact not quantitative with each other. So far there is not a hierarchical link between these technical indicators. The business figures are linked hierarchically. The individual cost components (maintenance costs, transportation costs, etc.) can be combined to measure the total cost in addition. The difference between the total cost and total revenue is the profit. The indicators so far based on a hierarchical, quantitative as well as a factual and logical link. The plan for approach the design of performance measurement system has shown that derivation should base on an empirical-inductive and deductive-logical analysis. The challenge is to integrate the knowledge of the literature studies with the practical knowledge, which is represented through the experts' opinions. The logical-deductive derivation was based on selected and recommended literature. The empirical-inductive derivation has been realized through interviews and group discussions.

The performance measurement system for wind farm control consists of a set of individual indicators. These individual indicators were categorized according to their application objective. The categorization based on the application objectives

as well as on the structure of the performance measurement system. There can be distinguishing three different application objectives. The first application objective is the maximum extraction of power from the wind by an optimal control of each wind turbines. In this area the single wind turbine represents a monolithic system. The parameters are determined through measurement systems at the wind turbines and directly relate to a specific wind turbine. The second application objective is concerning the operational management of the multi-wind turbine wind farm. The practical application of this objective is to minimize the downtime on the basis of optimal processes for maintenance, best possible cost situation and minimum use of resources like materials, tools and employees. The third application objective of the indicator system is to inform the executive management, the investors and further stakeholders like banking and insurance houses about costs, revenues and profits. So, on one hand the management and on the other hand decision-making can handle with the optimal and actual information. Following these three goals the performance measurement system for wind farm control consists of three areas. Each application objective represents a target area in the performance measurement system. The first area includes figures based on the technical aspects of a wind turbine, the second area deals with processes, materials and personnel management and the third summarizes aspects of management. From first to last area technical aspects decrease and economic aspects grow. These three areas are described below.

The first area – the technical control - of performance measurement system covers all aspects of the technical operation of a wind turbine and is called technical control. The key figures that directly influence the control technology of a wind turbine are variable. In this area in addition to this motion data there are also master data. This master data describes the wind turbine control systems in general and concern to the technical control only indirectly as basic data. Besides the master and transaction data, the first area consists of another group of indicators: the weather data. Master data are constant throughout all phases of a wind turbine and affect the control technique only indirectly. The master data can be divided into general and technical master data. General master data are for example the manufacturer and the type of construction, as the date of commissioning and the coordinates of the wind power plant. Technical master data include the hub height, the nominal power of the turbine as well as the position of the rotor blades and the number of rotor blades. Transaction data are by their nature variable. The control circuits such as the yaw drive, the security system and the yaw control rides based on the transactional data. Furthermore disturbance events are reported based on the transaction data. The specific information to the disturbance events forms the basis for maintenance management. The movement data of a wind turbine can be distinguished by the affected component or by the control loop. The control loop wind direction control is associated with the indicators yaw angle and pitch angle. The availability, speed and the oil temperature are each indicators of each component as gearbox, generator and turbine. At the power generating turbine the indicators current, voltage and frequency are measured. The degree of oscillation is

measured at the nacelle and the rotor blades. Relating to the safety system, the current position of the brake and the oil values of the brake system are important.

Besides the master and transaction data, the regulation of a wind turbine is strongly addicted to the weather. There is a closed link between the transaction data and the weather. Without the information of weather measurement cannot the control circuits cannot operate. For example, the wind direction tracking needs the wind direction and blade yaw control and the security system requires the wind speed. To recognize and recite storms and other hazardous weather situations are in addition to the wind speed, the humidity, the air temperature and the air density relevant.

The second area - the commercial control - includes all aspects of the process of the maintenance and therefore necessary materials, labor, tools and transport and associated maintenance data. Furthermore, the parameters of the wind farm network and the power grid, as well as the availability and the weather forecast data are associated with this area.

The area of commercial control founded on the area of technical control.

The ratios of each individual wind turbines reporting the faults and failures are input parameters for the field of commercial control. In the field of operations management it is about correction the faults and failures and to minimization the downtime. Therefore there are two different scenarios for undertake the maintenance: the operator can perform the maintenance themselves or a service partner is involved and the maintenance is outsourced. Regardless of the maintenance scenario processes such as maintenance, inspection, repair and improvement resources are necessary for a rugged operation. These resources relate to the material, the appliance of transport, the tool kit and the staff. As part of a damaged-oriented maintenance strategy, damage is a service event in the expression of a repair action. In the time-and state-oriented maintenance strategy the service event is an inspection or maintenance action. In any case of service event materials, tools, staff and an appliance of transport is assigned. These actions are done on the basis of maintenance and deployment plans.

Another area of management is the feed-in. Wind farms feed their generated power mostly through a substation or a central feed-in component into the grid. This requires that the network parameters with the network indicators of the wind farm match. The grid operator purports the network parameters. The feed-in component has to abide by these guidelines so that the grid does not collapse. Especially in the field of offshore wind farms the journey to a wind turbine is very expensive and weather-dependent. For the transport a ship or a helicopter can be used only under proper weather conditions. Similarly, the execution of maintenance work is weather-dependent, off- as well an on-shore. The rope access as well as the operation of cherry picker requires suitable weather conditions. Therefore the prediction of the weather data is relevant to plan the service events.

The application goal of the second area is to minimize downtime of the wind farm so as to maximize yield. To achieve the goals specific performance indicators are considered, such as the real and energetically availability. The real availability represents the observation period without the failures caused downtime.

The energetically availability is a given period minimized by the wind calm time. Basis of the calculation are the default and operating times, at which the operating hours are divided into full-load and part-load hours.

The third area of performance measurement system relates to the management of the wind farm. The indicators in this section are uncoupled from technical aspects, they relate to economic aspects.

The goal of this area is to inform the management, as well as investors, to learn about for example the current and historical cost, revenue and profits in order to decide well informed about strategic and tactical aspects. As in other companies: For wind turbines depreciation shall be made. For the land on which WKA is built a rent is to be due. Salaries and employee-related costs of e.g. maintenance personnel are subsumed under personnel costs. Costs of materials and tools are maintenance costs. The level of transport costs between onshore and offshore Wind farms. Other types of costs are the insurance, administrative and capital costs. The sum of the costs is compared with income earnings. The monetary income is the mathematical product of the feed-in and feed-in capacity in watt and euro per watt.

In these three areas the indicators for a wind farm control are pictured.

Once the wind farm control can be realized on the basis of a universal set of indicators further steps can be addressed. The classical wind farm control elements can be extended so that the wind farm control becomes innovative and intelligent under the involvement of other disciplines. The merging of the topics business intelligence and wind farm control could result in the new field of research intelligent wind farm control. In addition to the further automation of control mechanisms, a prediction of errors in annexes and a revenue forecast based on wind forecasts and error estimates would become possible. Furthermore an innovative and flexible reporting can be constructed on the base of the performance measurement system.

2 Dashboard Reporting

The installation and operation of new wind farms are more complex and more expensive than the existing onshore wind farms. For this reason it is necessary to get a hold of the costs and the revenues of the wind farms, to analyze the break-even point for further investment in offshore wind farms. Under these circumstances reports must be created to deliver the relevant information and measures to the management of the wind farm.

To get deeper in the context of the wind farm management the three different areas will be explained in detail, while each of this areas have their own specific requirements for the reporting.

The first area is the regulation. In this area processes will be initialized to technically and physically regulate a wind farm. A report in regulation of wind farms needs technical real-time data and displays very detailed technical data [5, 7].

The second one is the operational management. This area is working with sensor data to ensure the correct operations of the wind farm and displays all working

processes of the wind farm and in particular wind power plants. This area requires reports, which displays processes of the operation, the maintenance and the service. These aspects conserve the short-term planning of a wind farm [5, 7].

The last area is the management of a wind farm. The management needs information about the profitability, the revenues and the availability of wind farms to support strategic decisions in the development of offshore wind farms. These reports are based on actual and historical data, which are on a high abstract level [6, 7].

In this work the focus is on the management area. Hence the following aspects will focus on the specific requirements in the reporting for the management of an offshore wind farm. There are some very critical factors for this area. The availability of power plants in offshore wind farms is still unknown and not totally discovered in comparison to onshore wind farms. Therefore it is important for the management to be informed about this availability. Other important facts are the generated and the expected revenues of the offshore wind farm. Reports will compare those revenues with each other and will give huge information about the actual profitability of the wind farm [6].

For a fitting representation of this data a dashboard was chosen as the optimal platform. It displays the information in form of a cockpit in a good-looking way with interactive charts and graphics. With a dashboard all relevant data are in view at a glance.

For the development of the dashboard a BI-Suite was chosen. In this case SAP BO with the dashboard development tool SAP BO Xcelsius was the choice.

SAP BO Xcelsius has some advantages compared to other solutions, because it offers many different great looking graphical items to display the information and it offers a good possibility to create an interactive report based on live data. These characteristics are well suited for reporting to management and for this specific case.

Another advantage or disadvantage comes with the properties of standard software solution, which will not be explained any further in this context.

Speaking of the whole SAP BO BI Suite the tool SAP BO Universe, a semantic layer between the front-end reports and the back-end data sources, enables the reuse of an established connection to the relevant data source in different reports or in the creation of different reports in different tools [13].

In the conceptual work of this dashboard it is mandatory to define concrete functional and non-functional requirements. The most important requirements specialize the way in which the data will be displayed in a dashboard. It is obligatory, that the revenue and the expected revenue of a power plant, a cluster in the wind farm and of the whole wind farm will be displayed. These revenues and expected revenues must be distinguishable into days, months and a year. The sum of the revenues should be displayed in form of a bar chart. Additionally the average revenue of a wind farm should be calculated. This average will be related to the generated revenue of each power plant. So if divergences occur on a power plant or on a wind farm the dashboard should make this visible for the user in form of graphical representation or notification.

On the technical site of the non-functional requirements the dashboard has to be interactive and displays the data graphs in real time to fulfill its purpose. The development should base on an exchangeable and extensible data source and should be implemented with a SAP BO Universe. Regarding the user interface the dashboard needs four different views on the relevant data. The first should display the revenue/ expected revenue of a power plant, a cluster or the whole wind farm, depending on the selection of the user, in a year and the allocation per month. Just like this the second view should display the revenue/ expected revenue of a power plant, a cluster or the whole wind farm in a month and the allocation per day. The third view should display the accumulated revenues/ expected revenue per month within a year. On the last view the user can choose a month and the average revenues of the wind farm should be displayed. The deviations per power plant from the average revenues of the wind farm have to be shown as well.

At the starting point of the development it is important to consider the data source and how it could be used in the creation of the dashboard or the report. A recently built new wind farm center in the North Sea 45 km north of the German island *Borkum* delivers the required data and serves as a great research object.

In this case the operational data of the wind farm are continuously saved in a SQL database. For the development an extract of this database with all original data sets was exported. In detail a database dump was extracted as a Flat File. This data set consists of real time sensor data from two different power plant manufacturers, which are working in this concrete wind farm.

The first step in the realization of the dashboard, the Flat File was uploaded and integrated into a SAP BW system as a DataSource. The next step was taken to transform the two different data types to one consistent data type for the reporting and to ensure a good performance in the reporting later on. The DataSource was loaded and mapped into a DataStore-Object. Afterwards the transformed data was published as a MultiProvider for further extensions of this InfoProvider and to create a possibility to access this data via the SAP BO system.

The next step is the integration of this data into the SAP BO system or to enable the access on the MultiProvider from the SAP BO system. A SAP BO Universe, as mentioned earlier, is the semantic layer to integrate the data from different sources and it allows user to create reports out of business objects instead of SQL queries. In the context of this specific wind farm there are millions of operational data. This data - from half a year of operations - was integrated into SAP BW Flat File. Now the data can dynamically be loaded into the system with the help of SAP Universe. To insure the performance of the report and the aggregation of the relevant data in reports later on, the aggregation behavior in the universe has to be set accordingly. In the first approach the data access should be realized with a *Query-as-a-Webservice* to directly and dynamically load the necessary data in the dashboard with a parameterized query at run time. But this proceeding was not performant at all, even after changing the settings of the query and the universe. Instead the data should now be preloaded and aggregated to create a performant data access. Therefore SAP Web Intelligence reports were created based on the requirements of the dashboard and the data was preloaded from the

Universe into this. Briefly explained a Web Intelligence report is a web based report, which allows an analysis of relational and OLAP databases and the creation of web and ad-hoc reports [14]. After that the relevant data was loaded from the Web Intelligence reports into Live Office. In this spreadsheet all measures were calculated and subsequently the data was imported into the Xcelsius dashboard.

In the development process of the dashboard it was important to dynamically bind the data source to the graphical elements. With this the extensibility can be maintained and further features can be added in future. The underlying data can be updated, too. Regarding the appearance of the dashboard, the design based on a GUI design guide and was accompanied by recurring reviews.

As a result a dashboard, which fulfills the requirements, was developed. On this dashboard the expected and generated revenues can be displayed for a specific time interval and for a specific wind power plant, a cluster or the whole wind farm. The representation in form of charts and bars are displayed in real time and interactively. In addition the dashboard offers a good basis for further features. Especially the created Universe, the Web Intelligence reports and the Live Office document offer a good perspective for further reports and maybe for a comprehensive business BI solution for the wind farm management.

Prospective users reviewed this dashboard and evaluated it. In sum the design is attractive and the dashboard offers all-important features, which are easily to use and very intuitive thanks to the interactivity of this dashboard. For an optimal use it is preferable that technical reports are linked to the dashboard to reason the possible deviations.

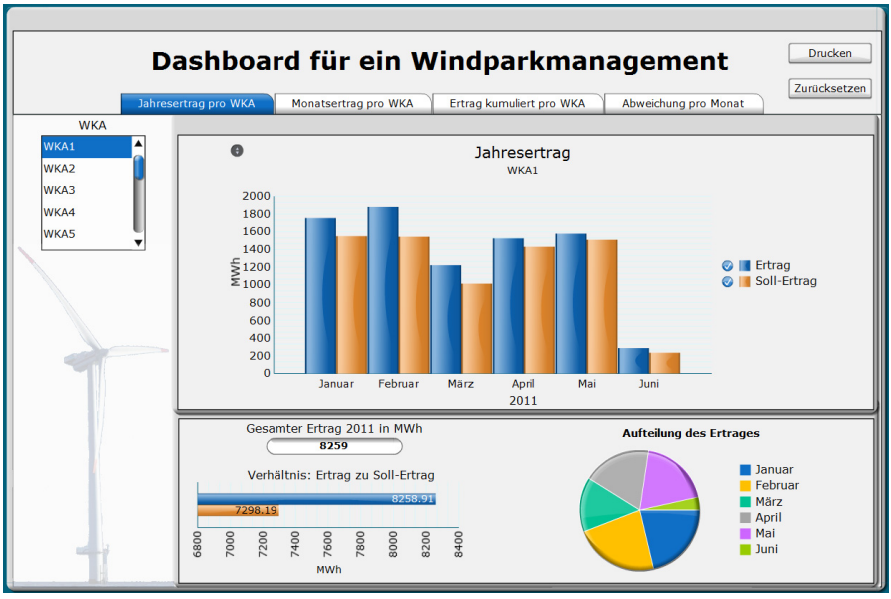


Fig. 1 Screenshot of the developed dashboard for the wind farm management

3 Mobile Reporting

Businessmen and decision makers face many challenges every day and without accurate and up to date information they couldn't succeed. Business life these days is much more dynamic and mobile. Therefore decisions must often be made on the fly and outside the office. For this reason the mobile business and mobile reporting are some of the most popular topics in the business environment. The modern wind farm management is no exception and the demand for mobile solution is steadily growing.

Mobile reporting faces a wide range of challenges; some of them have a higher priority. For the successful implementation of a mobile solution there are many requirements to be fulfilled. On one hand some of them are standard resp. very common, like: user-oriented and decision-supported visualization of data, economical use of resources and visualization of standardized and harmonized information at any time and location. On the other hand there are very specific requirements for the wind farm management, like: visualization of information about the produced annual or monthly energy, information about the wind farm and the wind turbines.

Furthermore mobile reporting, resp. mobile business intelligence is characterized by a large number of factors and can be realized with different methods. There are four approaches, which pursue mainly the same goals but differ in the technical realization and their functionalities: web page (web application), reports resp. dashboards, mobile server and native applications. The web page is the simplest form for making information from an IT system available on mobile devices. Every internet compatible device with internet browser can be used for this type of reporting. Reports and dashboards are mobile oriented web pages, which offer more comfortable business intelligence usage. Layout and information are defined and customized for specific devices or common specifications, like screen resolution. For the third type of mobile reporting a mobile server is used to automate the process of layout and information customization for the specific devices. This type of server coordinates the layout visualization and the information flow between mobile device and IT system. *Native application* is the last method for the realization of mobile reporting. It is a standalone client software for specific operation systems, such as *Android*, *iOS*, *Windows Mobile*, *Blackberry OS*, etc. and it is installed on a mobile device.

For the realization of "mobile reporting for wind farm management" case scenario a native application was developed. The fourth type of mobile business intelligence is most difficult to implement and it is applicable only on a single platform (for example iPhone with iOS). However this is the most powerful solution, because there are more or less no restrictions for the developer. A native application can support camera, GPS, store data and more on the mobile device. It can access external database systems, connect with third party software and a lot more.

This mobile solution was a supplement to the previously mentioned dashboard reporting and most of the concept was inherited. The key indicators and the most of the requirements are the same, but the solution differs in realization and

structure: dashboard reporting has its basis on standard software while mobile reporting bases on custom development. For this reason the software architecture was one of the important concept components. It has two layers, the application layer and the technical platform, whereas the whole concept bases on standard server-client architecture. The application layer focuses on the management and covers the four major reports: Annual, Monthly, Cumulated Yield and Monthly Deviation. Additionally the mobile solution makes some main information about the wind farm and the wind turbines available. Because of that the application has three main screens or also called intends: *Log-On, Control* and *Report*. The first step is to successfully log on to the application. After that the second screen appears. With three tabs the user can access information about the wind farm (*WP*), wind turbines (*WINDKRAFTANLAGE*) different reports (*Reports*).

Mobile WindFarmManagement (mWFM) was developed as a native application based on this concept for mobile reporting. The mobile operating system *Android* had been chosen as a platform for the realization, because of the market share, the rapidly growing positions and well-structured developer community. Additionally open-source software and widely distributed *JAVA* technology are used for the development of *Android* applications. Open-source software had been chosen also for the implementation of the server database system. MySQL is one of the most distributed free database system and together with the *Apache* server, which share is around 55 Percent of all internet servers, is commonly used combination of server-database system for different web-applications. In addition apache server supports PHP, which had been chosen as a technology for the interface. For the mWFM development the easy to install integrated environment XAMPP was used. This tool includes *MySQL, Apache, PHP, Tomcat* and much more. As for the developer environment the choice fell on Eclipse IDE with ADT (Android Development Tools) plug-in and Android SDK with the AVD (Android Virtual Device) Manager.

The development process can be divided in several stages. The first step is to set the foundations - to shape the main layout of the application. Android applications use XML to define the layout and java to implement the functions, but the layout can be defined also as JAVA code. In Eclipse the layout can be shaped with a user-friendly GUI, where the single elements are put together per drag and drop and XML code is generated automatically. Once the layouts are shaped the developer can proceed with the second step - the declaration of activities and element behavior in Java. Android provides many pre-defined Java-Classes and Methods, which cover a huge range of functions and activities. Additional functions can be implemented with third party libraries, for example for graphical diagrams, charts, etc. When the main activities are declared, the developer can proceed with the third step – database and information configuration. There are two main objects in this matter – database and data-flow, which are very important for a reliable and performant reporting. For the mWFM application two types of database were tested – normalized and denormalized. At first a denormalized database with only one table was used, in which the data was stored. The wind farm data is saved in five-minute tact, so over one million data-sets of information are stored annually.

The result was a database with a single table and big amount of data-sets, which reflected a very poor performance with over five seconds delays of information visualization. This concept would work very well, if it was enhanced with additional business intelligence technologies. However to keep the realization as simple as possible no further business intelligence technologies were used as middle-ware. For this reason the database was normalized and the data was stored in several tables. For example the data for the annual yield report was stored in a single table with only 72 (twelve months by six turbines) data-sets. The performance was much better and the calculated delay-times were less than half a second. Compared to the database there were less problems with the data-flow, which was developed with a PHP interface. This approach is very simple and stable on one hand and there are predefined Java-Classes on the other hand, which allow direct connection between the Android application and the PHP interface.

The application basics were implemented with these three steps. The fourth step was the realization of the reports and the optimization of the information visualization. For mobile reporting the graphical visualization is even more important, than for the classical reporting. Information must be optimally structured for the small mobile device screen. Very often overflow of displayed information can confuse the end-user and make the application useless. Keeping this in mind, charts, tables, numerical information, etc. must be customized for mobile devices and mobile reporting. Unfortunately Android offers no directly usable Java classes and methods for the graphical visualization, so a third party library was used. To test different functions and visual effects, the reports were built separately and with different structures. The annual yield report displays the information in a bar diagram. The monthly yield report displays the values also in a bar diagram, but with a zoom function, so the user can display data on a daily basis. The cumulated yield report displays the information in a line-and-dot diagram. The last report, *monthly deviation*, displays the values in a modified bar diagram, where positive values are visualized over the x-axis (0-axis) and negative values are displayed below the x-axis (0-axis). In all four reports a table displays the target-actual comparison.

The last development process stage of the mWFM application was the final test and the evaluation of the product. In the *AVD Manager* the developer can configure a virtual mobile device and directly install the application on it. This allows an instant application test and result evaluation. However in the last step, the application was exported as *.apk-Data*, which is the standard Android file format, and installed on different mobile devices (HTC, Samsung, etc.). To use the mWFM there are only two restrictions – a screen resolution with 800x480 pixels and an *Android OS* above version 2.2.

The result of few months' research and development was a fully operational Android native reporting application for wind farm management. It displays anywhere and anytime common information about a wind farm and wind turbines, for example capacity, geographical position, etc. The highlights of the mWFM application are the four reports, which visualize critical data in a user-friendly way not only as numerical data, but also as charts and diagrams. The application was very positively evaluated by branch experts and there are already demands for further development of this application with more functions and for additional platforms.

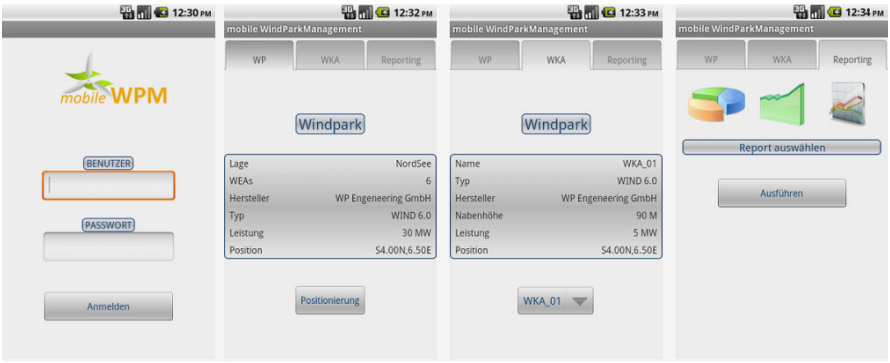


Fig. 2 Screenshots of the developed mobile application for the wind farm management

Fig. 3 Screenshots of the developed mobile application for the wind farm management

References

1. Ebert, P.: Messtech drives Automation. GIT Press, Release 9 (2011)
2. Fouhy, K.: MM MachineMarket – the Industry Magazine. Vogel Business Media, Release 12 (2011)
3. Franke, W.D.: Compendium Renewable Energy. Institute of Management-, Market and Medianinformation (2010)
4. Gasch, R., Twele, J.: Wind Power Plants: Fundamentals, Design, Construction and Operation. Press Vieweg + Teubner, 7. Auflage (2011)
5. Hau, E.: Wind Turbines: Fundamentals, Technologies, Application, Economics. Springer (2008)
6. Lyding, P., Faulstich, S., Kühn, P.: Monitoring Offshore Wind Energy Use: Status Quo - how are the offshore wind farms faring? In: EWEA 2011, Brussels, Belgium, March 14-17 (2011), `http://www.iset.uni-kassel.de/abt/FB-I/publication/2011-021_Monitoring_offshore-Paper.pdf` (last accessed: March 25, 2011)
7. Marx Gómez, J., Norkus, O.: The Secret of Offshore Wind Farms – Research overview of the Wind Farm control. In: [1], pp. 22–23 (2011)

8. Mensch, G.: Finance controlling. Oldenbourg Press, 2.Auflage (2008)
9. Munde, A.: Intelligent Networks are an opportunity for innovative enterprises, 20–21 (2011)
10. Muuß, T.: Surveillance Methods, Inspection types and Status Tests inWind Power Plants – What is necessary? GL Operating 24(7) (2007)
11. Muuß, T.: Condition Monitoring Systems for Wind Power Plants – Deployment and-Certifikation. DGZfP- Annual Meeting, Proceedings 94-CD (2005)
12. n.a., Federal ministry of environments, nature protection and reactor safety. Short info renewable energies (2010), http://www.bmu.de/erneuerbare_energien/kurzinfo/doc/3988.php
13. Rohloff, R.: Universe Best Practices (2007), http://www.sdn.sap.com/irj/boc/index?rid=/library/uuid/10dccc85-6864-2b10-23a4-f7aefc2b8deb (last accessed: May 03, 2011)
14. SAP AG, 2009: SAP BusinessObjects for SAP NetWeaver BI. In: Participant Handbook (2009)
15. Schilling, G.: Factbook regarding the economical meaning of renewable energy. In: [3], pp. 28–31 (2010)
16. Starck, M., Winkler, W.: Analysis of uncertainties in the Calculation of Revenue in Wind Farms (2002), http://www.dewi.de/dewi/fileadmin/pdf/publications/Publikations/strack_unsicherheiten_dewek2002.pdf (last access: March 25, 2011)

A Tool for Data Mining in the Efficient Portfolio Management

Vivian F. López, Noel Alonso, María N. Moreno, and Gabriel V. González

Abstract. In this work we perform a tool to data mining in the portfolios analysis. We perform an automatic data survey to draw up an optimum portfolio, and to automate the one year forecast of a portfolio's payoff and risk, showing the advantages of using formally grounded models in portfolio management and adopting a strategy that ensures, a high rate of return at a minimum risk. The use of neural networks provides an interesting alternative to the statistical classifier. We can take a decision on the purchase or sale of a given asset, using a neural network to classify the process into three decisions: buy, sell or do nothing.

1 Introduction

The decision-making process about when, how and what to invest is one of the most evocative issues in the investment world. These decisions challenge the investor's entire range of knowledge, which is always complicated, but particularly nowadays, when the exchange markets are highly volatile. Such is the case of portfolio management, which is still performed as craftwork. Selection is made according to the investor's favorite assets, or following the manager's ratings according to her experience, knowledge or intuition, but seldom based on formal grounds. This means that investors maintain inefficient portfolios that are not adjusted to the expected risk-payoff ratio.

Currently, portfolio analysis can be approached from two points of view. First, we have portfolio selection, which Harry Markowitz introduced in 1952 [10]. The second aspect is portfolio management aimed at finding the optimal structure. Today, financial market problems are often solved using artificial intelligence. Despite the great deal of effort already put into making financial time series predictions [9], support vector machines [6], neural networks [14], prediction rules [4], genetic

Vivian F. López · Noel Alonso · María N. Moreno · Gabriel V. González
Departamento Informática y Automática, University of Salamanca, Spain
e-mail: {vivian,noelalonso1,mmg}@usal.es

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_15,

algorithms [1] and multiagent system [12], the prediction of a stock market index is still difficult to attain. The main reason for the complexity of this task is the lack of autocorrelation of index value which changes even in a one-day period.

According to Markowitz, the selection is grounded in the simple observation of prices that maximize the expected payoff at a given level of risk. Although the information is growing day by day, its in-depth processing is very complicated and not within easy reach of the average investor, who is usually unable to capture and interpret the data. In this work we perform an automatic data survey to draw up an optimum portfolio, to estimate the market risk and, at a given moment, to help the decision process regarding an asset. The main objectives of the system are:

1. To automate the one year forecast of a portfolio's payoff and risk, showing the advantages of using theoretically grounded models in portfolio management and adopting a strategy that ensures a high rate of return at minimum risk.
2. To make the correct decision in the purchase or sale of a given asset, using a neural network to classify the process into three decisions: buy, sell or do nothing.
3. To use On-Line Analytical Processing (OLAP) [13] for obtained portfolios.

The automatic implementation is warranted by two fundamental reasons: The application of the models for portfolio management implies a lot of numerical calculations, impossible to perform without computing facilities and data mining, although there exist in the market some tools for this purpose, they are out of reach of most of the managers or traders of this field.

2 Portfolio Theory

Markowitz established the aim of setting up the menu of the possible payoff-risk combinations that are eligible, giving as the decision variable the weight or ratio assigned to each asset (W). Grounded in these ideas, the process of selecting the optimum portfolio can be summarized in the following steps:

1. Specification of a body of assets and investment funds to be considered for the portfolio screening.

Asset analysis by means of the estimation of their expected payoffs, variances and covariances.

Determination of the investor's efficient frontier and indifference curves.

Derivation of the optimum portfolio.

Analysis of the Risk Evaluation (VaR) [8] of the optimum portfolio.

These steps are briefly described in Section 3. In Section 4 a description of the system is made. Section 5 deals with the neural network training, the OLAP for portfolios analysis is briefly described in Section 6 and finally the conclusions.

3 Portfolio Selection

Let there be an investor with a budget to be employed in the buying of assets to maximize their expected utility. The Stock Exchange will provide him or her with a lot of investment choices, as many as shares. Nevertheless, the investor must deter-mine the share combination which, while maximizing the proposed objective, uses up the entire available budget. That is, he or she must know what assets to buy and how much to spend on each one of them.

3.1 Asset Analysis

Following Markowitz, the first step starts with observation and experience and finishes with some specific ideas on the future behaviour of the available assets, particularly concerning the density functions of the future payoffs of the shares.

3.2 Computing the Historical Payoff

Let us see how to compute the historical payoff (Rit) of an asset i in a given period of time t. Let $P_{i(t-1)}$ be the price of asset i at the end of period $t-1$, that is, at the beginning of period t. Assuming that we buy the share at this moment, it will be the purchase price. Let dit be the cash-flow obtained by the asset in period t. Finally, let Pit be considered as the price of the share at the end of period t or, in our case, its selling price. The payoff obtained in period t will be computed as in Equation 1.

$$R_{it} = \frac{P_{it} - P_{i(t-1)} + d_{it}}{P_{i(t-1)}} \tag{1}$$

3.3 Derivation of Efficient Border

Once the individual features of each asset are known, we study features that will comprise the portfolio. For this purpose, we will assume that we have *n* possible assets, each of them with its mean and variance, as representative of its payoff and risk. A portfolio is a set of assets so it will also have a payoff and variance different from those of its components. Portfolio payoff, R_c will be a function of the different random variables of payoff of the constituent assets and thus will itself be a random variable. Let us compute the risk. To this end, we will compute the portfolio payoff variance $V(R_c)$ as a function of the assets payoff variance σ_i^2, as in Eq.2:

$$V_c = \left[\sum_{i=1}^{n} R_i\right] = \sum_{i=1}^{n} {\omega_i}^2 {\sigma_i}^2 + \sum_{i=1}^{n} \omega_i\, \omega_j \sigma_{ij} = \sum\sum_{i=1} \omega_i \omega_j \sigma_{ij} \tag{2}$$

That is, the portfolio payoff variance will depend on the covariances of the assets payoffs.

3.4 Computing the Optimum Portfolio

Once the expected values and variance (risk) of payoff are known, we must decide on the optimum portfolio to choose. We will follow the process defined in the mean-variance rule: compute the efficient portfolios and select the portfolio that maximizes the utility of the decision maker. There are several ways to compute the efficient portfolio borders. Markowitz proposes, among others, the following one, maximize Rc, produced in Eq.3:

$$R_c = \sum_{i=1}^{n} \omega_i R_i \tag{3}$$

As is apparent, the problem is approached in terms of quadratic programming where a minimum for an investment risk has to be found, with a given payoff level and no possibility of debt.

3.5 Analysis of Risk Evaluation of Optimum Portfolio

The concept of Risk Evaluation (VaR) [11] comes from the need to measure with some level of confidence the percentage of loss that a portfolio will undergo in a predefined time. This is an estimation of the maximum loss that the portfolio can have. In the system implemented, the VaR is calculated for each asset by the Normal Delta Method [15], chosen because it is considered the simplest one to estimate since it requires only the market values, the final portfolio positions and the variance-covariance matrices. To calculate VaR the steps are:

a) Identify the purchase value for each asset.
b) Check that the changes in the associated values for the asset follow a normal distribution.
c) Compute the variance and covariances for each portfolio asset.
d) Compute the portfolio variance.
e) The VaR is calculated by multiplying the portfolio variance by the corresponding factor to the confidence level (1.65 in this case)

Thus, the VaR is a useful tool in that it gives investors a more precise criterion for judging the work done by portfolio managers. Furthermore, it allows us to monitor and control the risks through time by verifying that the short term portfolio does not diverge from the long term objectives.

4 System Description

Based on this experience, a new multiagent model was developed for the automatic efficient management of investment fund portfolios that takes into account the

history over a given period, adapts to dynamic market conditions and upgrades itself via the web periodically. The arquitecture of the multiagent system coordinates three kinds of agents [3]: Interface, Analysis, and Information agents [12], as shown in Figure 1.

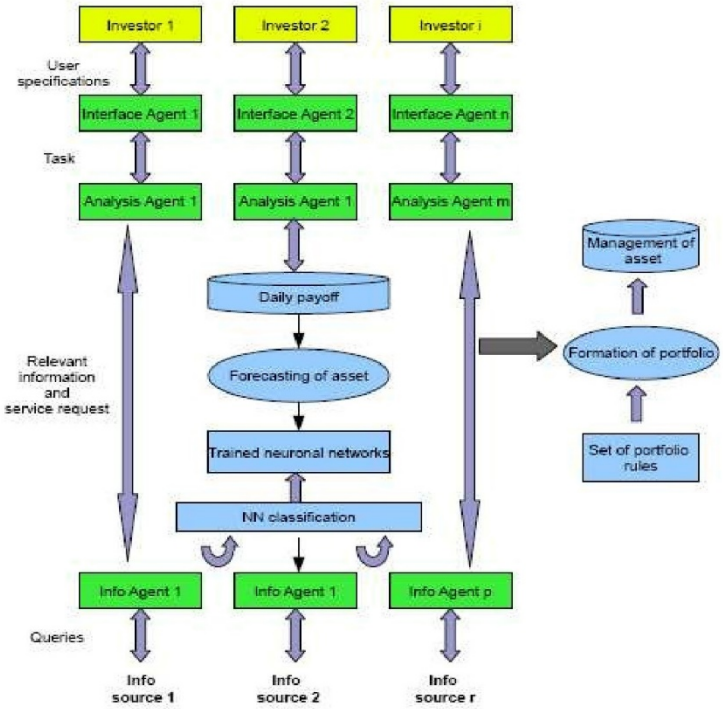


Fig. 1 The arquitecture of the multiagent system

In the system each investor is represented as an agent that can keep the investment for a certain amount of time according to individual preferences. Interface Agents interact with the user receiving user specifications and delivering results, in our model each agent distributes their wealth and divide it among the agents to learn its investment strategy from the sample data. Analysis Agent makes autonomous in-vestment decisions in each investment period. Accordingly, each agent is optimized for the best possible forecasting of asset analysis in order to place profitable and risk adjusted decisions. The portfolio manager agent is targeted to find the optimum portfolio (optimal risk-profit on the efficient frontier) according to Markowitz theory [10]. Based on this knowledge, the Information Agents, decide what information is needed and initiate collaborative searches with other

agents. In addition the Analysis Agent suggests the investor what assets to buy and the amount to be invested in each one to obtain a bigger payoff and a lower risk. Besides that it must indicate what is most suitable for the asset according to the daily evolution of prices and payoffs of each asset: keep, sell or buy.

The model offers a methodology for the composition of efficient portfolios, becoming a more appropriate tool for investment decision making. The tool is a simple but efficient IDE for investment decision making. The tool al lows users to make as-set analysis, visualize, interpret any asset and can offer a scale of portfolios through an easy-to-use graphical user interface. In summary, the main features are, as shown in Figure 2:

2. Assets: data base asset management to be considered for the portfolio screening.

Optimize: to draw up an optimum portfolio.

Portfolios: to efficient management of investment fund portfolios, and see the portfolio evolution.

Alerts: about VaR of the optimum portfolio.

Reports: to show OLAP of portfolios.

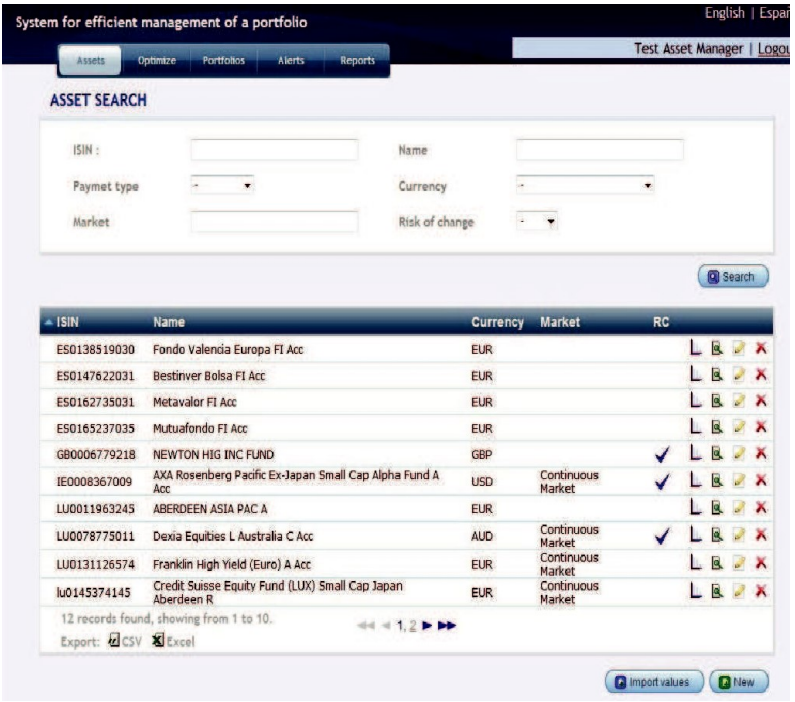


Fig. 2 System for efficient portfolio management

In next section, the aforementioned functions are implemented:

4.1 Specification of the Asset Set

The system can work with any portfolio on the stock-exchange market, so there are as many possibilities as assets. To perform the prediction computations, for each as-set in the portfolio, a data base is defined with the following fields: ISIN code (fund registration), name of the asset, estimated time in portfolio (inversely dependent on risk), date of portfolio participation, value of participation, payoff, equivalent yearly rate (APR), observed volatility, market distribution and number of days to take into account.

The data sets for performing the forecasting study of profitability and risk in a portfolio of values uses the 14 funds of different managers of the above mentioned values that were taken from the Fibanc Mediulanum Banking Group Platform All Funds [5]. The number of days to bear in mind is determined by the least amount of all the observations of each one of the 14 founds. For the particular study we used the set of assets appearing in Figure 2.

4.2 Obtaining the Ideal Portfolio

With the previous data the historical payoff is computed for each asset for a period of 321 days, and the following values are obtained: daily payoff (with respect to previous day), daily volatility (standard deviation), average daily payoff, daily profit or loss and VaR for each asset.

With the results obtained in the historical payoff phase, minimum variance point (MVP) is determined inside the boundary of production possibilities. To do this, the assets are initially given random weights and two restrictions are implicitly imposed:

1. The client has to spend 100% of the available money.
2. No negative weights are allowed.

The user has the possibility to add his or her own restrictions, as shown in Figure 1. For the entire portfolio the MVP is computed. This gives us the minimum standard deviation of the portfolio (minimum risk). Later on, the portfolio average daily payoff (MRP) is calculated. This is the Sharpe Ratio simplified by considering that the risk free interest rate is 0% (this is the case with the Goverment Bonds). Finally the MRP/MVP ratio is computed (maximum slope of the straight line) to maximize the payoff/risk ratio or, equivalently, maximum payoff at a minimum risk.

In summary, we try to find the right weights for each one of the portfolio components so that the agent can choose the best distribution. Once the payoff and risk are calculated, we select the efficient portfolio and compute the VaR. Let us suppose that the investor has decided to invest 30000 Euros. The first thing to do is to randomly distribute this amount among 14 investment assets in order to calculate the profitability and the volatility of this distribution in the portfolio. Later the

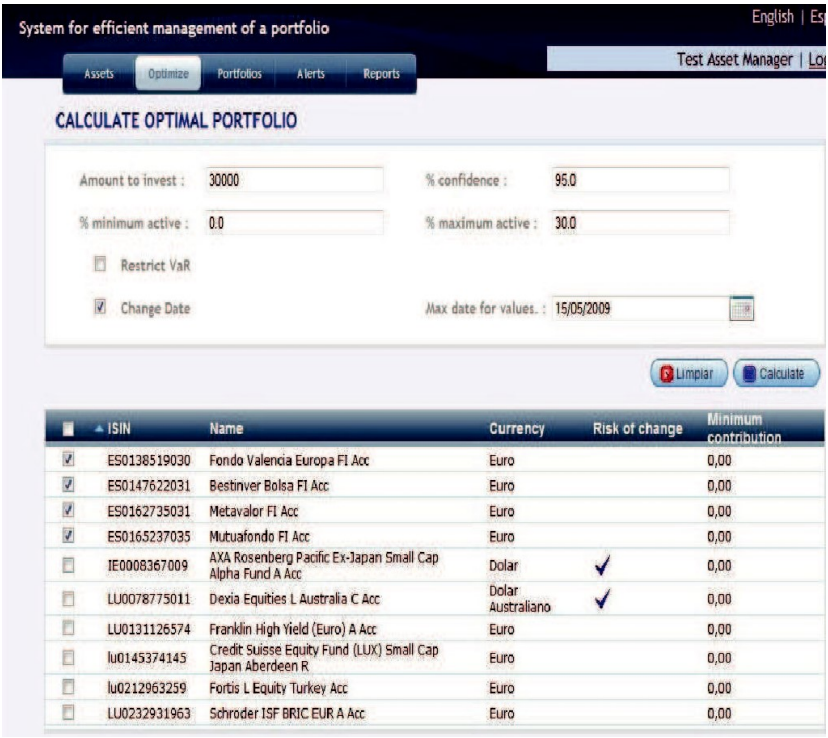


Fig. 3 The user has the possibility to add his own restrictions for efficient portfolio

computations previously mentioned are performed according to the fixed restrictions. The weights which maximize the portfolio of each of the considered assets are found. To obtain the results we click directly on the calculate button shown in Figure 3.

With this information we can obtain the ideal portfolio, as shown in Figure 4. Taking into account the history of observations in a fixed period for 14 assets and the previous calculations, Figure 4 shows the final amount to be allocated to each asset and the time that it must remain in the portfolio to obtain a bigger payoff and lower risk. As can be seen, the amount of Euros is very different from the one initially assigned. In Figure 4 it appears beside the amount that it is necessary to invest in each asset and the percent to keep it, the APR profitability in one year that it is possible to obtain and the volatility in 431 days. The assets whose final amount is 0 Euros, are those not recommended to buy. The system also returns the estimated profitability in 431 working days, ensuing from 8% and from 11% in one year which means 3419.9129 Euros. The VaR analysis is shown too in the Figure 4.

Fig. 4 Final amount and time necessary for efficient portfolio

4.3 Portfolios

In this module each user can view his portfolios and the profit ability of these developments. The users or asset manager can also configure portfolios to receive alerts on asset values. In portfolio detail, if by pressing the button view graph a pie chart of the distribution of the assets within the portfolio, will be shown, as shown in Figure 5.

4.4 Alerts

You can also configure your portfolios to receive alerts. Alerts are notifications that are sent by email, or you can see from the application. The alerts are generated when the profitability of the portfolio exceeds certain limits. From this screen already con-figured alerts and a form to add new settings can be seen. To add a new configuration the minimum rate and/or the maximum level of profitability that we want to reach in portfolio, must be defined.

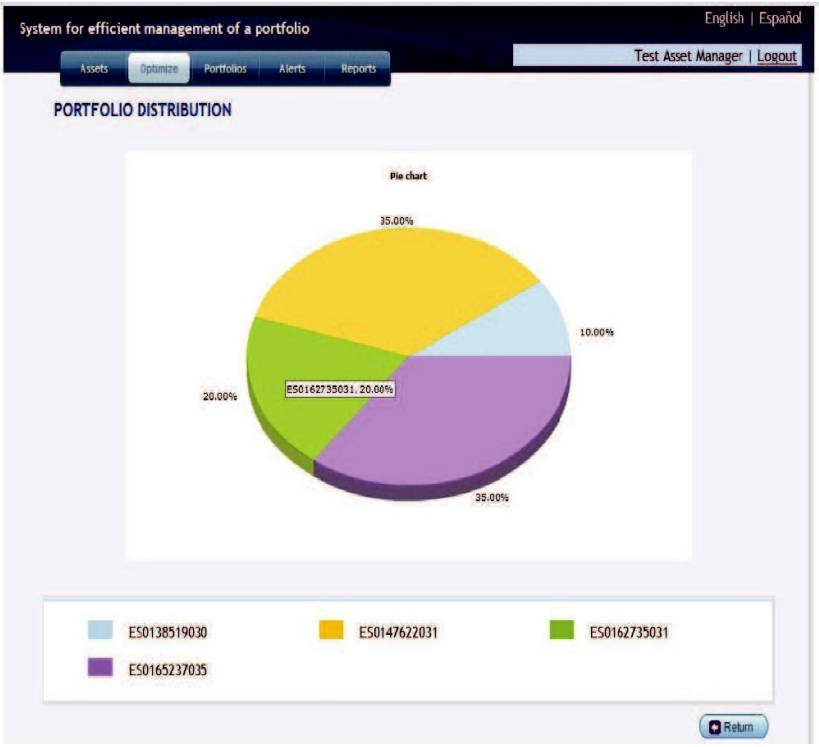


Fig. 5 The pie chart of the distribution of the assets

5 Neural Network Classification

In [2] it is pointed out that the market knows everything. In consequence, we must always study it as a source of maximum available information and thus take decisions of buying or selling the stock. It is not necessary to know all this information: we must simply study the evolutions of prices that are formed. The evolutions will indicate to some degree the likely direction that the prices are going to take in the future, since the market remembers the formations of prices that have taken place throughout history and, they will probably occur again with identical consequences on most occasions.

As soon as all the combinations of the list of assets are obtained, in order to guarantee maximum profitability and the minimal risk, it would be desirable to be able to classify the state of the price in a certain period, bearing in mind its behavior in a previous period and to be able to know if it goes down, up or keeps constant within the fixed period. It might help the investor to take a decision to buy, sell or do nothing.

With this aim, for every asset we train a perceptron neural network with a single hidden layer [7]. For our case the significant input parameters are the value of daily participation, payoff and daily payoff. With them the net is trained to learn the behavior of the prices in a one-year period, classifying them into three classes according to their daily profitability: Class 0 (do nothing), Class 1 (sell), Class 2 (buy).

In the training phase we use 70% of the available information and the remaining 30% is used for the validation. The net has three input neurons, corresponding to the significant input attributes and three output neurons (c lasses). The number of neurons in the hidden layer is a parameter to play with to achieve a tradeoff between efficiency and speed of training. In our case, with three neurons an acceptable result is reached.

Once the architecture of the net is defined, we train it using the Weka tool [16]. To do this it is necessary to fix some parameters that takes part in the training process. These parameters always depend on the problem to be solved and after performing some simulations the learning rate is fixed at 0.3 the momentum at 0.1 and the number of training cycles is 30. After training, we perform an estimation of the results provided by the network through the test patterns, and we verify that the number of examples correctly classified depends on the fund in question, ranging between 96% and 100%, as we show in the results of the experiments, the error in the estimation of the classes being 0.03.

We observed that the network correctly classified the validation pattern. Once the net has been trained with the prices and final earnings, it can be consulted with any other input value in future periods, and they will be classified to help in the decision making on an asset.

For neural network classification, we performed two fundamental experiments, consisting of training a neural network for every fund and another one with the information of all the funds in the same period used in the analysis of the portfolio. The worst results on the number of examples classified correctly were obtained by the net that included all the funds for the analyzed period, which could only correctly classify 95.24 % of the cases presented. With one different net for each fund the results range from 96.90 % corresponding to the Fidelity Fund up to 100 % of the majority, as can be seen in Table 1.

Table 1 Examples correctly classified for each asset

Name Assets	Precision	Name Assets	Precision
Franklin H.Y. ”A”	100.00	Dexia eq l aust ”c”	98.25
Dws Invest Bric Plus	100.00	Ubam Us Equi Value A	100.00
Aberdeen Asia Pac ”A”	97.90	Sch Eur Dyn Grwth A	100.00
Fortis l Eq Turk ”C”	100.00	Newton Hig Inc	100.00
Cre Suis.Cap Jp ”H”	98.30	Ing(l)inv Eur h.d ”x”	100.00
Challenge Country Mix (S)	98.60	Challenge Financial Fund	100.00
Challenge Germany Equity	97.30	Fidelity Eur S.C. ”E”	96.90

The results obtained by means of neural networks were contrasted with those derived from a statistical method. Several approaches were considered based on statistical time series processing and curve adjustments. Curve fitting of data of each one of the investment funds was used. The objective was to find the curve that best fits the data, and use it as a model.

To each one of the investment fund datasets in the case study the curve is function of the independent variable (*x*) value participation (*valor p*) and from it the dependent variable is (*y*) daily profit or loss (result). The rest of the variables was not taken in account to include them in the model because they are significantly dependent to each other.

A trial version of the TableCurve2D tool by SYSTAT Software Inc. [17] was used to carry out the curve fitting for each investment fund. The Figure 6 shows the graphs of the obtained models, as well as their corresponding mathematical expressions, for the best fund (*Ing(l)invEurh.dx*). We can see that classification errors are 38, worse than neuronal networks classification. Results were poor so our conclusion was to use nonparametric approaches, like neural networks, which can learn and adapt to new conditions. The classification errors with neural networks were much better in all the cases.

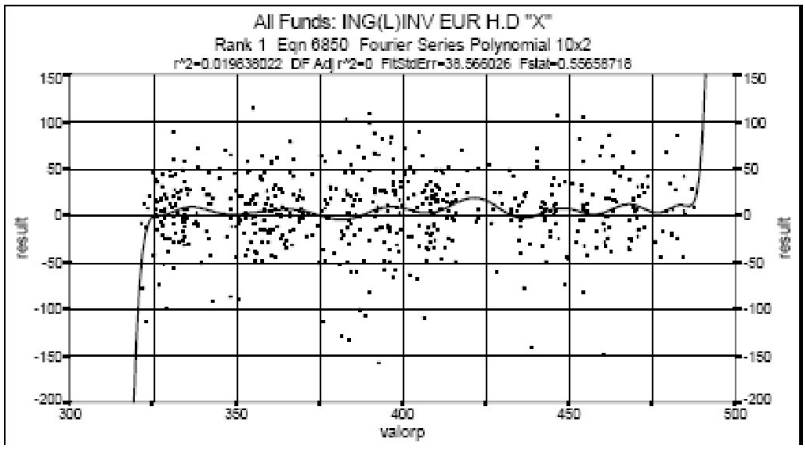
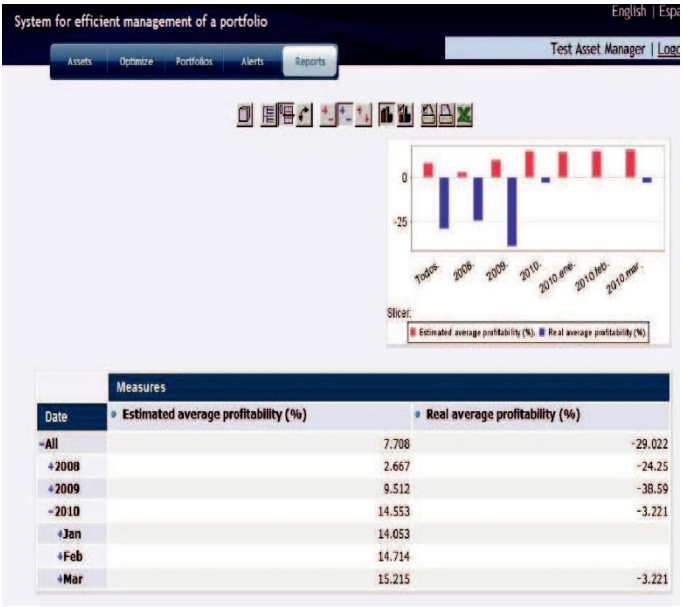


Fig. 6 Graphs of the obtained models for fund (Ing(*l*)invEurh.d "*x*")

6 OLAP for Portfolios Analysis Obtained

The OLAP for portfolios analysis is a technique often used in the Business Intelligence (BI). The goal is to allow a multidimensional analysis of high volume databases to a special analysis from different points of view (in this case about assets and portfolios). The OLAP is a data mining tool that allows to create new views and customize portfolio to make reports that suits the needs of investor, as shown in Figure 7.



	Measures	
Date	Estimated average profitability (%)	Real average profitability (%)
All	7.708	-29.022
2008	2.667	-24.25
2009	9.512	-38.59
2010	14.553	-3.221
Jan	14.053	
Feb	14.714	
Mar	15.215	-3.221

Fig. 7 OLAP for portfolios analysis

7 Conclusions

A system was created on a formal theorical basis which automates the forecast of the profitability and risk of a portfolio of assets over a one year period, by adopting a strategy that guarantees high profitability and minimal risk for the investor, without restriction in the number and types of assets.

This model offers a methodology for the composition of efficient portfolios, becoming a basic tool for investment decision making. The financial adviser, according to the type of investor (risk adverse, average risk or risk lover), can offer a scale of portfolios with a certain yield, in view of risk level.

The system is able to suggest the asset the investor should buy and the time that it must remain in the portfolio to be profitable. As a consequence, this management is more efficient and achieves better results. Moreover the computer system makes the numerous calculations for the application of the models governing the management mentioned above, as well as the periodic upgrading of the information bases. This system can adapt itself to new trends, since it keeps training with new information, so it can therefore adapt to dynamical market conditions taking into account the good results of previous periods.

The use of neural networks provides an interesting alternative to the statistical classifier. With the results described in previous tables it is clearly shown that with the neural networks classifiers a high level of accuracy can be achieved.

References

1. Allen, F., Karjalaine, R.: Using Genetic Algorithms to Find Technical Trading Rules. Journal of Financial Economics, 245–271 (1999)
2. Codina, J.: Manual de Analisis ´ Tecnico´ (5ta. Edicion), Inversor Ediciones, S. L. Madrid (2007)
3. Decker, K., Sycara, K., Zeng, D.: Designing a multi-agent portfolio management system. In: Proceedings of the AAAI Workshop on Internet Information Systems (1996)
4. Dempster, M.: Computational Learning Techniques for Intraday fx Trading Using Popular Technical Indicators. IEEE Transaction on Neural Networks, 744–754 (2001)
5. FibancMediulanum Banking Group, http://www.fibanc.es
6. Gestel, T., Suykens, J.: Financial Times Series Prediction Using Least Squares Support Vector Machines Within the Evidence Framework. IEEE Transactions on Neural Networks, 809–820 (2001)
7. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall (1999)
8. Jorion, P.: Value at Risk: The New Benchmark for Controlling Market Risk. McGraw-Hill (2000)
9. Kodogiannis, V., Lolis, A.: Forecasting Financial Times Series Using Neural Network and Fuzzy System-Based Techniques Neural Computing & Applications, 90–102 (2002)
10. Markowitz, H.: Portfolio Selection: Journal of Finance 7 (1952)
11. Minnich, M.: A Primer on VaR Perspectives on Interest Rate Risk Management for Money Managers Traders (1998)
12. López, V.F., Alonso, N., Alonso, L., Moreno, M.N.: Multiagent System for Efficient Portfolio Management. In: Demazeau, Y., Dignum, F., Corchado, J.M., Bajo, J., Corchuelo, R., Corchado, E., Fernández-Riverola, F., Julián, V.J., Pawlewski, P., Campbell, A. (eds.) Trends in PAAMS. AISC, vol. 71, pp. 53–60. Springer, Heidelberg (2010)
13. OLAP, http://olap.com
14. Tino, P., Schittenkopf, C.: Financial Volatility Trading Using Recurrent Neural Networks. IEEE Transactions on Neural Networks, 865–874 (2001)
15. Vilariño, A.: Tubulencias Financieras y Riesgos de Mercado. Prentice Hall, Madrid (2001)
16. WEKA, http://www.cs.waikato.ac.nz/ml/weka/index.html
17. SYSTAT Software Inc., http://www.systat.com/products/TableCurve2D/

Educational Data Mining: User Categorization in Virtual Learning Environments

Angel Cobo Ortega, Rocío Rocha Blanco, and Yurlenis Álvarez Diaz

Abstract. This chapter provides a brief overview of applying educational data mining (EDM) to identify the behaviour of students in virtual teaching environments and presents the results of one particular b-learning initiative. The purpose of the research is to explore the usage patterns of asynchronous communication tools by students. The initiative was developed over one semester on an operations research course that was taught using a b-learning methodology at a School of Economic Sciences. In particular, active, non-active, collaborative and passive users were identified using a soft clustering approach. A fuzzy clustering algorithm is used to identify groups of students based on their social interaction in forums and the temporal evolution of this classification during the semester is presented.

1 Introduction

The use of new information and communication technologies (ICTs) offers a new way of producing, distributing and receiving university education and complements traditional teaching and learning methods [14]. In recent years there has been considerable interest in incorporating virtual environments into teaching strategies and methodologies. These methodologies are being revised, incorporating blended learning as a combination of face-to-face and virtual teaching.

Angel Cobo Ortega
Department of Applied Mathematics and Computer Science, University of Cantabria, Spain
e-mail: acobo@unican.es

Rocío Rocha Blanco
Department of Business Administration, University of Cantabria, Spain
e-mail: rochar@unican.es

Yurlenis Álvarez Diaz
Universidad de Holguín "Oscar Lucero Moya", Cuba
e-mail: yalvarez1982@gmail.com

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_16,

Blended instruction, also known as b-learning, is an approach that combines the benefits of online and classroom instructions to improve distance learning environments where learners can be easily disoriented due to a lack of communication or direct guidance [15]. In these new teaching models, technology is a vehicle that delivers instruction, facilitating the creation of interactive learning environments, intelligent tutoring systems and allowing for integration into the educational processes of a large number of e-learning resources. There are a growing number of courses taught using Computer-Supported Collaborative Learning (CSCL) tools, such as Moodle, WebCT and Blackboard. These tools generate large repositories of data that can be explored to understand how students learn. Data include interesting and valuable information, such as the number of times, frequency and physical and temporal distribution of accesses, pages visited, activities carried out… In a b-learning model this information can be combined with indicators obtained directly by the teacher regarding classroom activities. Educational Data Mining (EDM) focuses on data mining techniques for using these data to address important educational questions. These techniques can be useful for predicting student performance, identifying behaviour patterns, defining individualised learning strategies and recommendation systems to improve students learning.

In particular, online asynchronous discussion forums in CSCL environments play an important role in the collaborative learning of students. Thousands of messages could be generated in a few months in these forums, with long discussion threads and many interactions between students. Therefore, CSCL tools should provide a mean to help instructors evaluate student participation and analyse the structure of these interactions [4, 18].

In this chapter we present the results of a study on the social interaction of students during a one-semester b-learning initiative. Several variables were used to dynamically identify groups of students responding to common patterns. The remainder of the chapter is organised as follows. In the following section, various systems for categorising users in social environments are summarised, along with a review of the classification strategies used in e-learning processes. The next section includes an explanation of the role of data mining techniques in the knowledge discovery processes in educational contexts. After that, the focus turns primarily on soft computing algorithms that can be used in clustering problems and their application in EDM. The last two sections present details of the b-learning initiative, describing the data compilation process, pre-processing and the analysis performed. This is followed by a description of the results and the conclusions drawn.

2 User Categorisation in Social Communication Environments

Learning on the Internet is highly compatible with social constructivism, emphasising the manner in which students actively construct their knowledge on the basis of what they already know through social interaction in learning contexts,

rather than passively receiving knowledge in a pre-packaged form [24]. Several studies have been conducted in the scientific community, not only regarding the benefits of social learning environments, but also in order to identify what the patterns and types of user behaviour are in such environments. Previous studies have demonstrated that monitoring and interpreting online learning activities can ultimately lead to enriched user and learner models. Research by Hong [9], defined some measurable parameters for an online class and established that performance can be defined from the documents, total discussions and percentage of open documents in the following ways: active/passive activity, active/passive day and active/passive student. Other authors such as [7] characterise users as: participative, deep, useful, non-collaborative, having initiative, skilled or communicative, focusing on the collaboration activity level and the student performance indicators.

[21] carried out a clustering approach and applied data mining to the data provided by a CSCL backed database and built analytical models which summarised interaction patterns to obtain profiles of student behaviour. As a result, six clusters of students were identified: high initiative and high participation in forums; medium/low initiative; low initiative; high initiative and low participation in forums; average in all areas; and low initiative, average participation in forums, extreme (low/high) participation in chat rooms.

More recently, the study conducted by [22] focuses on Web 2.0 and its potential in higher education and identified two different taxonomies of users. The first, according to the type of interaction on the social network (social, popular, communicator, sponge, in love, and ghost) and the second, in terms of computer skills demonstrated. Another study conducted in the field of the social Web by Forrester Research, Inc characterises social computing behaviour on a scale, with six levels of participation: creators, critics, collectors, assemblers, spectators, inactive, according to the monthly activity performed by the users. The categorisation of social media carried out by [23] of InSites Consulting follows this line of research, but based on the log-in frequency and frequency of social media activity. As a result of this investigation, four social media user types were identified: addicts (high log-in and activity frequency), voyeurs (high log-in, but low activity frequency), special occasions (low log-in, but high activity frequency) and passive users (low log-in and activity frequency).

3 Educational Data Mining

Simply stated, Data Mining (DM) refers to extracting or "mining" knowledge from large amounts of data [8]. This area has attracted a great deal of attention in the information industry and in society as a whole in recent years and data mining techniques have been applied in a wide variety of areas. In particular, the increasing use of information and communication technology also allows the interactions between students, with their instructors and with educational resources to be recorded, so that mining techniques can be used to gain a deeper understanding of the learning process and make proposals for improvements.

There are many applications or tasks in educational environments that have been resolved through data mining. Educational Data Mining (EDM) is a field that applies statistical methods, automatic learning and DM algorithms over different types of educational data [19] and is very useful for understanding students better and the settings in which they learn [2].

According to [19] EDM has emerged as a research area in recent years involving researchers all over the world from different research related areas, which are as follows:

- Offline education tries to transmit knowledge and skills based on face-to-face contact and also studies how humans learn from a psychological point of view. Psychometrics and statistical techniques have been applied to data, such as student behaviour/performance, curriculum, etc., which was gathered in classroom environments.
- E-learning and learning management systems (LMS). E-learning provides online instruction and LMS also provides communication, collaboration, administration and reporting tools. Web mining (WM) techniques have been applied to student data stored by these systems in log files and databases.
- Intelligent tutoring systems (ITS) and adaptive educational hypermedia systems (AEHS) are an alternative to the put-it-on-the-web approach, trying to adapt teaching to the needs of each particular student. DM has been applied to data picked up by these systems, such as log files, user models, etc.

This author also presented an overview of related work in the area of EDM, describing the most relevant studies in the area and grouping them into eleven categories according to their tasks/categories and objective of the studies as follows: analysis and visualisation of data, providing feedback for supporting instructors, recommendations for students, predicting student performance, student modelling, detecting undesirable student behaviour, grouping students, social network analysis, developing concept maps, constructing of educational material, planning and scheduling. A different viewpoint is presented by [2] who suggests four key areas of application for EDM: improving student models, improving domain models, studying the educational support provided by learning software, and scientific research into student learning; and five approaches/methods: prediction, clustering, relationship mining, distillation of data for human judgment and discovery from models. Regarding the categorisation work in EDM, [19] divide it into statistics and visualisation, and web mining, that can be further split into clustering, classification and outlier identification, rules for association and sequential patterns and text mining.

Over the last few years, a trend towards the combined use of data mining techniques and automatic learning for analysing activity data can be observed [19, 20]. Systems based on individually treating user activities tend to be aimed either at predicting student performance or identifying different types of students, as well as their characteristic interaction behaviour and how this behaviour relates to learning [11]. These techniques are also used to deduce important characteristics of collaborative students [1] and to discover the behaviour and models in various student profiles about how they navigate and work in their virtual courses [6] or how they solved problems using a software tutor based on the student data logs [10].

4 Soft Computing and Fuzzy Clustering

Cluster analysis or clustering is one of the major problems in data mining and a common technique for statistical data analysis used in many fields [12]. This term covers a wide variety of techniques for delineating groups or clusters in data sets. Clusters should capture the natural structure of the data. Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible [25]. The potential of clustering algorithms is to reveal the underlying structures in data and can be exploited in a wide variety of applications, including classification, image processing and pattern recognition, modelling and identification. In particular, educational data mining techniques seek to identify categories or behavioural patterns in students.

Many clustering algorithms have been introduced in literature [16]. A widespread classification system subdivides these techniques into two main groups: hard (crisp) or soft (fuzzy) clustering. In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. However, in fuzzy clustering, data elements can belong to more than one cluster and each element is associated with a set of membership levels which indicate the strength of association between that data element and a particular cluster. Due to the fuzzy nature of many practical problems, a number of fuzzy clustering methods have been developed following the general fuzzy set theory strategies outlined by [26]. Fuzzy set theory deals with the representation of classes whose boundaries are not well defined. The key idea is to associate a membership function that takes values in the interval [0,1], where 0 corresponds to no membership in the class and 1 corresponds to full membership. Thus, membership is a notion which is intrinsically gradual instead of being abrupt as in conventional Boolean logic.

The concept of fuzzy partition is essential for cluster analysis and identification techniques that are based on fuzzy clustering. Given a data set $Z = x_1, x_2, \ldots, x_n$, the objective of clustering is to partition the data into c clusters. Using fuzzy set theory, a fuzzy partition of Z is defined by a matrix $U = \left(u_{ij}\right)_{cxn}$ where the ith row contains values of the membership function, satisfying the following conditions:

$$u_{ij} \in [0,1], 1 \leq i \leq n, 1 \leq j \leq c$$

$$\sum_{j=1}^{c} u_{ij} = 1, 1 \leq i \leq n \ \ and \ 0 < \sum_{i=1}^{n} u_{ij} < n, 1 \leq j \leq c$$

One of the best known methods of fuzzy clustering is the Fuzzy c-Means method (FCM), initially proposed by Dunn [5] and made more widespread by Bezdek [3] and other authors; Kruse et al. [13] presented an overview on fuzzy clustering.

FCM provides a method that shows how to group data from a multidimensional space into a specific number of different clusters; it is based on the following optimisation problem:

$$\begin{cases} \min \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \\ 0 \le u_{ij} \le 1, \qquad \forall i, j \\ \sum_{j=1}^{c} u_{ij} = 1, \qquad \forall i \end{cases}$$

where $(c_1, c_2, \ldots, c_n)$ are the centroids of the clusters, which can be defined by a given matrix or randomly chosen, and u_{ij} is the degree of membership of x_i to the cluster j. Finally, the parameter m is a real number greater than 1 which acts as a weighting factor called fuzzifier. Normally the Euclidean distance is used, but any other rule ||*|| which expresses the dissimilarity between any measured data and the centre can be used. [16] presents a review of different distance functions used in clustering. One of the drawbacks of FCM is the requirement for the number of clusters, c, to be specified before the algorithm is applied. In literature, methods for selecting the number of clusters for the algorithm can be found [17].

Fuzzy partitioning in FCM is carried out through an iterative minimisation of the objective function under the above fuzzy constraints, involving successive calculations (updates) of the prototypes (centroids) and the partition matrix by applying the technique of Lagrange multipliers. The values of several FCM parameters need to be set up in advance: the number of clusters (c), the distance function ||·||, the fuzzification factor (m), and the termination criterion (maximum number of iterations).

5 Case Study: Knowledge Extraction in a b-Learning Initiative Using Fuzzy Clustering

In order to explore the possibilities of making use of soft clustering algorithms to identify student behaviour in a b-learning process, the following sections present the results of applying data mining techniques using data extracted from a virtual class environment. The case analysed corresponds to a one-semester introductory course on optimisation and operational research into the degrees of economy and business administration. All of the participants were first-year students and the course used a b-learning strategy that combined face-to-face and on-line activities. The on-line activities were supported by CSCL Moodle. In particular, on-line asynchronous forums in this virtual course make students actively engage in sharing information and perspectives by interacting with other students, with social interaction playing an important role in the collaborative learning of the students.

The process of data mining in educational settings can be split into the following phases [20]: data collection, data pre-processing, application of data mining techniques and the interpretation, evaluation and deployment of the results. According to this proposal, the following sections summarise the actions performed at each phase.

5.1 Data Collection

In the virtual optimisation course, different social interaction spaces (forums) were activated during the semester. The students performed three different actions in these spaces: read a message in the forum, reply to a message in the forum and start a new thread. At the end of the semester, a total of 2008 messages were generated in these forums, containing 185 discussion threads with a large number of interactions between students. The average values of messages read, replies and new threads per student were 180.76; 5.33 and 0.54 respectively, with a high dispersion or variability between students. The standard deviations from the previous magnitudes were 311.02; 13.98; 1.18.

Using Moodle administration tools, the activity indicators of the 342 students were obtained. For each student, data was collected on participation in forums in different time periods during the semester. In order to analyse the activity over the semester, 9 two-week periods were considered. In each period p, a three-dimensional activity vector $A_{sp}=(a_{sp1},a_{sp2},a_{sp3})$ was obtained for each student s, where a_{sp1} is the number of messages read per student s in period p, a_{sp2} the number of replies given by the student, and a_{sp3} is the number of new threads initiated by the student in the period. Figure 1 shows the average values and standard deviations over the 9 two-week periods in the semester. There were two periods where activity was clearly lower. The first of these, period 4, corresponds to the Easter holiday period and the second one, period 9, corresponds to the weeks after the final assessment of the course. The period of highest activity is the one just before the final assessments.

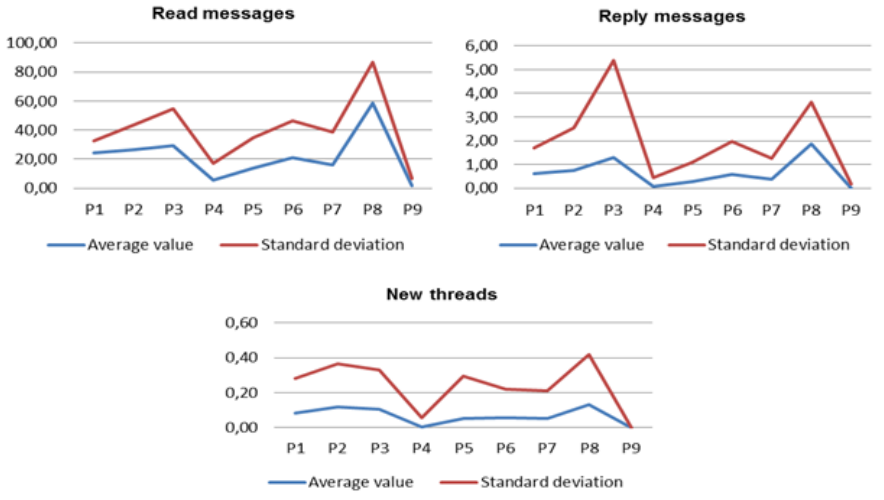


Fig. 1 Average values and standard deviations of the numbers of messages read, replies and new threads per student

5.2 Data Preprocessing

Data are often normalised before clustering, in order to remove discrepancies from the scale. In our case, a unity-based transformation was used, so all data values will take on a value of 0 to 1.

$$\alpha_{spi} = \frac{a_{spi} - \min\{a_{spi} | 1 \le s \le 342\}}{\max\{a_{spi} | 1 \le s \le 342\} - \min\{a_{spi} | 1 \le s \le 342\}}$$

where $s=1,2,\ldots,342$ represents a student; $p=1,2,\ldots,9$ a period, and $i=1,2,3$ an activity variable.

5.3 Application of Data Mining Techniques

After normalisation, the data are ready to apply data mining techniques. A great variety of capable data mining software packages are available and, in this work, R software was used. R is a free open source software environment for statistical computing and graphics; it is also a functional language and environment to explore statistical data sets. R implements a wide variety of clustering algorithms; the Fuzzy C-Means (FCM) algorithm, implemented in package 'e1071', was selected. One of the drawbacks of FCM is the requirement for the number of clusters (c) to be given before the algorithm is applied. We decided to use c=4 to identify four types of students based on their activity in forums:

- Inactive: students who did not use CSCL communication tools and did not participate at all in online communities.
- Passive: users who displayed minimal activity and did not share, comment or produce any new content themselves. These students only read posts in forums and did not share their experiences or consult academic issues. They were using the communication tools to keep their finger on the pulse of forums.
- Active: students that actively contributed to forums. They were constantly involved in creating and participating in groups, replying to messages from their classmates and trying to help them.
- Collaborative: they were much more interested in networking and connecting to others. These students were willing to create content, new open debates and participate in threads created by other students.

In this work, the temporal evolution of these student profiles or groups was analysed. In order to perform this analysis, we performed 9 runs of the fuzzy c-means (FCM) clustering algorithm, where each run corresponded to one two-week period and using $A_{sp}=(\alpha_{sp1}, \alpha_{sp2}, \alpha_{sp3})$ for $s=1,\ldots,342$ as data vector.

The R *cmeans* command has several parameters which need to be configured: the data matrix, where columns correspond to variables and rows to observations; the number of clusters or initial values for centroids; the maximum number of iterations; the distance measure to use; and the degree of fuzzification. In the case of the last three parameters, a maximum number of 500 iterations, the Euclidean

distance and a degree of fuzzification of $m=2$ were used. The cluster centroids can be seen as prototypical objects (prototypes) of the cluster; these prototypes are usually not known beforehand and are obtained by the clustering algorithm at the same time as the partitioning of the data. However in our case, we have initially provided 4 student prototypes: inactive, passive, active and collaborative (see Table 1).

Table 1 Initial cluster centroids

Student	Prototype
Inactive	(0,0,0)
Passive	(1,0,0)
Active	(1,1,0)
Collaborative	(1,1,1)

Inactive students have virtually no activity in the forums, so the vector (0,0,0) can be considered as a prototype vector for this group of students. Passive students may have a high activity with regard to reading in forums, but no creative activity, which is represented by the vector (1,0,0). For active students, it was considered appropriate to initially allocate the value 1 to the variables corresponding to reading and replying, but not to the variable associated to the generation of new debates. Finally, the prototype of a collaborative student is obtained by giving the maximum value for the three variables. These initial prototypes were updated by the FCM algorithm as shown in Table 2. This table shows the centroids of the cluster after running the FCM for each two-week period.

Table 2 Centroids of the cluster (prototypes) after the running the FCM algorithm

p	Inactive student's prototype	Passive student's prototype	Active student's prototype	Collaborative student's prototype
1	(0.027, 0.004, 0.000)	(0.195, 0.053, 0.002)	(0.542, 0.571, 0.451)	(0.125, 0.098, 0.506)
2	(0.025, 0.003, 0.000)	(0.107, 0.089, 0.499)	(0.222, 0.242, 0.017)	(0.126, 0.074, 0.984)
3	(0.018, 0.002, 0.000)	(0.242, 0.033, 0.008)	(0.681, 0.266, 0.109)	(0.130, 0.047, 0.529)
4	(0.015, 0.005, 0.000)	(0.790, 0.148, 0.000)	(0.293, 0.731, 0.000)	(0.151, 0.167, 0.971)
5	(0.013, 0.001, 0.001)	(0.109, 0.117, 0.031)	(0.885, 0.877, 0.022)	(0.330, 0.310, 0.521)
6	(0.012, 0.002, 0.000)	(0.135, 0.079, 0.002)	(0.963, 0.775, 0.022)	(0.155, 0.188, 0.999)
7	(0.025, 0.012, 0.000)	(0.087, 0.040, 0.989)	(0.172, 0.396, 0.971)	(0.889, 0.507, 0.981)
8	(0.023, 0.011, 0.001)	(0.179, 0.160, 0.009)	(0.177, 0.192, 0.324)	(0.247, 0.261, 0.685)
9	(0.004, 0.000, 0.000)	(0.983, 0.000, 0.000)	(0.205, 0.998, 0.000)	(0.109, 0.000, 0.000)

As can be observed in Table 2, the prototypes of the 4 types of students change significantly in relation to the initial ideal values, also being observed differences in the different time periods

Although the FCM algorithm calculates the degrees of membership of each group, it also returns a hard clustering array, assigning each student to the group with the closest centroid (prototype). According to this criterion, the majority of students are included in the group of inactive students (see Figure 2). A total of 167 students (49%) are assigned to the inactive group of students in all periods analysed fortnightly. Ignoring the inactive students in each period, Figure 3 shows the evaluation of the other types of students during the 9 two-week periods of the semester. In this figure two periods with higher inactivity can also be seen (periods 4 and 9); these periods correspond to the Easter school holidays and the period immediately after the final exams. Period 9 stands out due to the greater degree of involvement of collaborative students; this may be related to the fact that, having completed their academic activities, students used the forum to exchange views on the final development of the course.

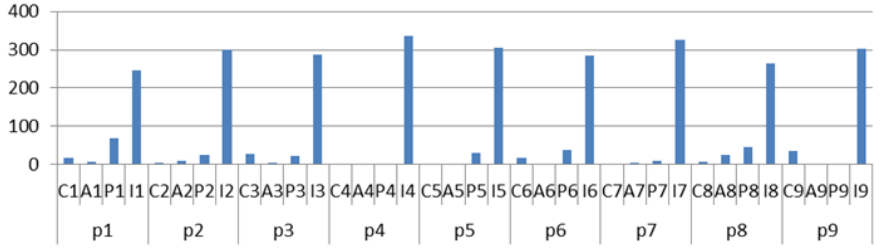


Fig. 2 Closest hard clustering, with the distribution of students in four groups (collaborative, active, passive and inactive) over the semester. Most of them are included in group I (inactive) and P (passive)

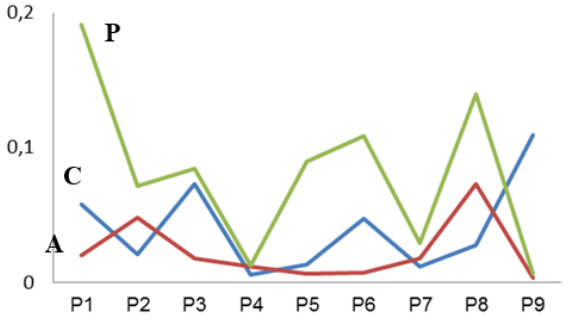


Fig. 3 Evolution of passive (P), active (A) and collaborative (C) students during the semester

The advantage of the FCM fuzzy approach is that each student has membership values associated to each group considered. Sometimes it is difficult to mark the boundary between active and collaborative students or between a passive student and a student with a very low activity. That is reason why the use of the fuzzy clustering approach is particularly suitable for this type of study. Figures 4 and 5

show examples of active and collaborative students, respectively. In both cases it can be observed how the student profile may vary slightly from one period to another, but overall there seems to be a strong link with a type of student. Another observed difference is the fact that the active student maintains a certain level of activity even in periods of low global activity (periods 4 and 9).

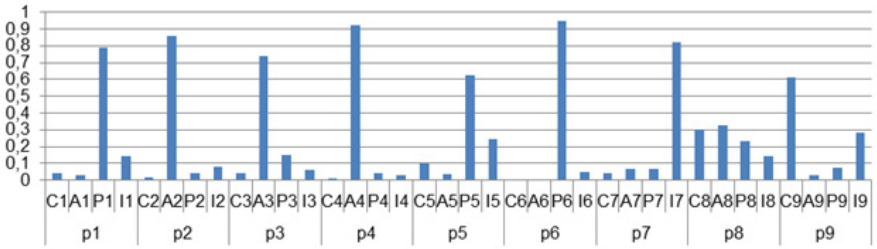


Fig. 4 Example of an active student, with membership levels in the 9 two-week periods

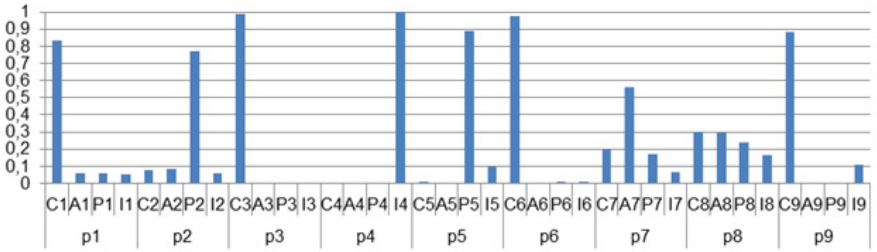


Fig. 5 Example of a collaborative student, with membership levels in the 9 two-week periods

5.4 Evaluation of the Results

As a summary of results of the practical experience developed, it could be stated that the predominant profile of student is inactive, although there is a large group of students who regularly use the virtual learning platform to enhance their learning process. It is also noted that the profile may vary from one period to another depending on various factors (activity in the face-to-face classroom, proximity of assessment periods, holiday periods,...). It seems that most students were unable or unwilling to take advantage of the potential of communication forums in the online course. It should be noted that the data correspond to a b-learning experience, in which students also had hours of classroom contact with the teacher and their classmates, so the use of the forums in the virtual learning platform was not so essential. That is the reason why most students did not make regular use of these communication tools. Ignoring the students who were found to be inactive, it seems as though students adopted more passive positions at the beginning of the course, limiting themselves to consulting, reading and observing activity in the forums. As the course progressed, these positions tended to move towards more active and collaborative attitudes. In the period just before the final evaluations

(period 8), the degree of cooperation between students increased considerably. In that period, students needed more help and saw the forums as a useful tool for this.

In order to analyse the relationship between social interaction and academic performance, the final success rates for the course were analysed. This analysis confirmed the perception that the behaviour patterns of students in CSCL showed a marked influence on their academic performance, but does not necessarily means that those students with a tendency to take a more passive position in these environments will get worse results.

In conclusion, this paper has tried to show the potential of data mining techniques to extract knowledge in teaching-learning environments. In addition, soft computing has shown itself to be a very suitable tool for identifying behaviour patterns, where the difference between some patterns and others is not so clear.

References

1. Anaya, A.R., Boticario, F.G.: Towards Improvements on Domain-independent Measurements for Collaborative Assessment. In: Proceedings of the 4th International Conference on Educational Data Mining, pp. 317–318 (2011)
2. Baker, R.S.: Data mining for education. In: McGaw, B., Baker, E., Peterson, P. (eds.) International Encyclopedia of Education, 3rd edn., vol. 7, pp. 112–118. Elsevier, Oxford (2010)
3. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York (1981)
4. Carmagnola, F., Osborne, F., Torre, I.: User data distributed on the social web: how to identify users on different social systems and collecting data about them. In: Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems. ACM, New York (2010)
5. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3, 32–57 (1973)
6. Garcia, D., Zorrilla, M.: E-learning Web Miner: A Data Mining Application to Help Instructors Involved in Virtual Courses. In: Proceedings of the 4th International Conference on Educational Data Mining, pp. 323–324 (2011)
7. Gaudioso, E., et al.: Cómo gestionar la colaboración en el marco lógico colaborativo en un entorno de aprendizaje adaptativo basado en web. Revista Iberoamericana de Inteligencia Artificial 8(24), 121–129 (2004)
8. Han, J., Kamber, M.: Data Mining.Concepts and Techniques, 2nd edn. Morgan Kaufmann Publishers (2006)
9. Hong, W.: Spinning your course into a web classroom – Advantages and challenges. In: International Conference on Engineering Education, Oslo, Norway (2001)
10. Johnson, M., Eagle, M., Joseph, L., Barnes, T.: The EDM Vis Tool. In: Proceedings of the 4th International Conference on Educational Data Mining, pp. 349–350 (2011)
11. Kardan, S., Conati, C.: A Framework for Capturing Distinguishing User Interaction Behaviours in Novel Interfaces. In: Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), pp. 159–168 (2011)
12. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc., Hoboken (2008)

13. Kruse, R., Hoppner, F., Klawonn, F., Runkler, T.: Fuzzy Cluster Analysis. John Wiley & Sons (1999)
14. López, M.V., Pérez, M.C., Rodríguez, L.: Blended learning in higher education: Students' perceptions and their relation to outcomes. Computers & Education 56(3), 818–826 (2011)
15. Oh, E., Park, S.: How are universities involved in blended instruction? Educational Technology & Society 12(3), 327–342 (2009)
16. Pedrycz, W.: Knowledge-Based Clustering. John Wiley & Sons, Inc. (2005)
17. Pham, D.T., Dimov, S.S., Nguyen, C.D.: Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 219(1), 103–119 (2005)
18. Rabbany, R., Takaffoli, M., Zaïane, O.R.: Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques. In: Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, pp. 21–30 (2011)
19. Romero, C., Ventura, S.: Educational Data Mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 40(6), 601–618 (2010)
20. Romero, C., Ventura, S., García, E.: Data mining in course management systems: Moodle case study and tutorial. Computers and Education 51(1), 368–384 (2007)
21. Talavera, L., Gaudioso, E.: Mining Student Data to Characterize Similar Behavior Groups in Unstructured Collaboration Spaces. In: Workshop on AI in CSCL, pp. 17–23 (2004)
22. Valenzuela, J., Valerio, G.: Redes sociales en línea:'primeros pasos hacia el e-learning 2.0? Virtualeduca (2010)
23. Van Belleghem, S.: Social Media around the world. InSigths consulting (2010), http://www.slideshare.net/InSitesConsulting/social-media-around-the-world-3547521 (accessed March 9, 2012)
24. Wang, M.J.: Online collaboration and offline interaction between students using asynchronous tools in blended learning. Australasian Journal of Educational Technology 26(6), 830–846 (2010)
25. Witten, I., Frank, E.: Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers (2005)
26. Zadeh, L.: Fuzzy Sets. Information and Control 8, 338–352 (1965)

Part III

Knowledge Management and Decision Making

Company Sustainability Reporting: Decision Making Model Utilising Compensatory Fuzzy Logic

Desislava Milenova Dechkova and Roani Miranda

Abstract. Sustainability Report (SR) can be used form organizations to achieve better understanding, measurement and evaluation of risks and opportunities, having financial and nonfinancial nature (e.g. biodiversity, corporate governance, corporate social responsibility, natural resources depletion, etc.). Successful assessment of organization's ability to answer the question "*To issue a Sustainability Report or not?*" should be based on systemic approach. However, lack of such methodology is among the most important obstacles when satisfying stakeholders' informational needs. This chapter structures Decision Making (DM) activities answering the above question through three indices: Readiness, Convenience to issue SR and Usefulness of SR. Algorithmised Global Reporting Initiative G3.1 performance indicators mapped with UN Global Compact sustainability principles create database for that, which is further conceptualized through here composed logical predicates.The presented model assists DM in organizations, through implementation of Compensatory Fuzzy Logic (CFL) as knowledge engineering tool. CFL was chosen for its interpretability through language and its good capacity to reflect quantitative and qualitative features of used predicates. The proposed model allows taking into account strategic preferences of organization, its stakeholders and external knowledge aiding SR introduction.

1 Introduction

Sustainability management means facing successfully on a long-term environmental, economical and social challenges of an enterprise. In today's world, the issue

Desislava Milenova Dechkova
Carl von Ossietzky University of Oldenburg, Germany
e-mail: desislavamd@gmail.com
Roani Miranda
"Jose Antonio Echeverria" Higher Technical Institute, Cuba
e-mail: rmiranda@ind.cujae.edu.cu

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_17,

of sustainability in all its aspects: cultural, social, environmental and economical, is inevitable and continuously gains on importance.

Sustainability is a wide and very complex issue. As such it has a number of reasons to be in company's spotlight. Although it is not a new topic [15, 26], sustainability is presently enforced as a business case going beyond the initial ecological point of view and embraces social and economic issues in steadily increasing number of organizations [3, 4].

"Sustainability nears a tipping point" is the conclusion from a research report of the MIT and the Boston Consulting Group presented in the beginning of 2012. Offering lessons to managers who are either already trying to develop a sustainability agenda or wondering whether they should, it shows that companies who elaborate their strategies in coherence with the principles of sustainable development (SD) are gaining on competitiveness, improve their image and in general, secure their future market place [18]. Legitimate consequence – it is no longer possible to neglect the sustainability topic, but it becomes a compulsory part for a successful future-proved planning.

Measuring performance leads to better management. Consequently, in the corporate reporting practice sustainability reporting (SR) gains increasing attention. In this regard, while defining a company as sustainability leader, the newest survey of Sustainability and GlobalScan points out *transparency/ communication* as the second highest ranked characteristic. Leading criteria is *commitment to sustainability values*, while the third place is taken by *sustainable products/ services and integration of sustainability in the core business model* [29].

This is backed up on the shareholders' level through increase of proposals to companies, concerning corporate social responsibility issues with 23% in the period 2000 to 2010 [7]. Furthermore, consideration of corporate citizenship is reflected in the companies' DM from both - investment and purchasing management and supplies chain management professionals with 40% and 44% - respectively as important, and 42% for both groups as very important. Business leaders worldwide see sustainability as central to their business: 93% of CEOs, and 98% of those in consumer goods in particular, believe that sustainability issues will be important to the future success of their business [7].

Finally, if resulting from the companies' actions as: workflows optimization; investment in research and development; education of employees in understanding and applying sustainability principles; and implementation of gained knowledge on stakeholders' demands, laying down of SD in the business strategy leads to positive effects in the society and has vital impact on companies' credibility and success [29]. Yet, for many enterprises, the decision for communicating their goals and achievements in a sustainability report is impeded by many obstacles.

In the general case, guidelines and prerequisites, which have to be taken into account from companies, are often seen as too complex and time consuming issue, therefore, if possible, preferably avoided. Vagueness of the DM processes, together with the constant striving for better profit, while enrolling principle of sustainability, puts enormous pressure on the top management.

However, regulatory demands, recognized opportunities brought by a first mover positioning to mention some, are strong stimulus when mastering modern threads and risks. New conditions necessitate new tools and strategies. Changes in the organizational environment require responsive behavior. They can create positive impact on business and be engine for new services and product portfolios, achieved through research and development in e.g. Green Technology; aware supplier selection; fostering stakeholder communication as way to recognize and utilize new trends, etc.

2 Elements of the Decision Making Problem and Its Solution

2.1 Outlining the Problem

Sustainability has great number of facets and this makes it difficult to get fast overview on it. For beginners, especially small and medium enterprises, it can become an overwhelming task. The steps needed to extract useful information from guidelines, trends and best practices are highly time and resources consuming. Despite awareness of the chief management for the need to tackle sustainability in the company's strategy on a long term, real actions are often put on hold, or insufficiently performed. To speed up this process, a standardized, yet tailored to particular company's conditions implementation pattern is needed.

Useful and universally applicable method could be elaboration of Sustainability Report. Its crucial merits go far beyond positive public relations; while crystallizing out which bricks built the company SR can be promoted successfully: employee motivation, definition of plans for environmental actions, laying down sustainability in the core strategy are some of the crucial benefits on internal level, to mention some. Providing transparency and well founded communication base with the stakeholders, are some of the natural outcomes of reaching out wider publicity.

Sustainability Report reflects great number of concerns, whose sound fulfilment is prerequisite for profitable business with a long-lasting success. It has the assets to be valuable tool for better company DM [32].

In order to ease a pro-reporting decision, should be assessed which readiness a company has, to issue a useful SR. At the same time it is recommendable for the preparation stage to be able to point out strengths and weaknesses on important topics concerned, and in the best case all this should be presented in an easily comprehensible way. In this regard, indices summarizing the gathered information can be suitable solution.

Decision making process demands multidimensionality in the strategic management requires suitable methods, which allow setting of preferences. Simultaneously, complexity of strategic thinking and involved knowledge, available in the company, should be profoundly tackled. This can't be covered using models, which describe the reality with classic approaches of decision theory e.g. basing on normative thinking.

Proper approach for solving this problem is the usage of knowledge engineering. It allows 'capturing' and transformation of knowledge from literature and experts in a formal model [11]. Such Knowledge Engineering (KE) method should enable a DM model, which reflects knowledge on strategic preferences involved, specific knowledge about the problem and include standards for sustainability report.

The elaboration of a mathematical model translating qualitative in quantitative statements to valuate KPIs – typically described in Sustainability Report and defining sustainable development, can support organizations throughout the decision-making process towards introduction of Sustainability Reporting. Such approach was not found in the literature till the moment [25].

2.2 Decision-Making Models Review

The formal methods enabling multidimensional appreciation of criteria and distinguishing strategic organizational preferences are complex and frequently involve non-linear trade-offs between attributes and preferential dependence [13]. Furthermore, strategic thinking requires: i) Dealing with uncertainty; ii) Involvement of tacit and explicit knowledge gained form best practices, literature and on company internal level. Finally, strategic thinking has to express concepts, dependences and associations with other properties and variables influencing the DM attributes. The stated requirements cannot be satisfied using models that only norm or describe reality using classic approaches of decision theory, e.g. models based on normative thinking [6].

Knowledge Engineering KE was defined as the main approach to solving the problem of this chapter. It was performed in particular through knowledge gained from the literature till now. Knowledge Engineering is to be seen as extension of expert system's ideas, which would allow 'capturing' and transforming of knowledge from the literature and experts' opinions into a formal model [8]. Consequently, the selected DM method should reflect the complex preferences of the decision maker, the strategic and specific knowledge about the problem, including standards and frameworks for sustainability reporting. Review of popular DM methods should aid in finding the best solution to this task.

Properties of prominent DM models and CFL are summarized and their behavior is evaluated according several essential criteria: which are the mathematic objects involved; ability for language modelling; allowance for compensation; complex problems' treating aptness and veto option when attribute has a 'very bad[1]' value. The results are summarized in Table 1. It makes explicit, that CFL has ability to give solutions to complex DM problems with any logical structure, based on experts' opinion (human source) Sensitivity combined with the possibility to interpret according a standard scale allows CFL to deal better with knowledge expressed in the subtle ways from the human language. Hence, predicates with the operators of CFL were selected to be the mathematical tool for this study.

[1] A 'very bad' attribute will lead to veto, thus the result will not be improved through compensation from the rest of the attributes can enable.

Table 1 Properties of Decision Making Methods

Criteria / Method	Mathematic objects involved	Usage of language	Compensation allowed	Suitable to treat complex problems	Veto option if attribute value is 'very bad'
Additive Valued function and additive utility function	Functions	No	Additive (always)	Low	No
Multiplicative valued function and multiplicative utility function	Functions	No	Multiplicative (only if the attribute value is not "very bad")	High	Yes
Analytical Hierarchy Process AHP	Eigenvalues and Eigenvectors theory	No	Always	Low	No
Descriptive Methods: Electre, Promethe, etc.	Preference relations	No	Yes, only if attribute value is not "very bad"	Low	Yes
Rough Sets	Getting rules from data	No	No	High, if the modelling is from data	No
Delphi	Used together with Statistic model of consensus	Non applicable	Non applicable	Non applicable	Non applicable
Fuzzy Logic & Fuzzy Sets	Membership Functions and Rules from data or from human sources	Limited, only rules	Through defuzzification method	High	No
Compensatory Fuzzy Logic	Membership functions & getting rules from data or human sources	Yes, by any logical structure	Yes, but only if the attribute value is not 'very bad'	High	Yes

2.3 Sustainability Reporting Frameworks

To provide better applicability of the DM model, it was selected to utilize two of the most prominent SR schemes. These are Global Reporting Initiative (GRI) and United Nations Global Compact Communication on Progress (GC). In 2009 GRI in its G3 version was used for publishing of 1400 reports worldwide from mainly large companies. GC on the other hand was adopted as a framework in more than 6000 companies – many of them Small and Medium Enterprises (SME) [33]. The two frameworks will be as next briefly described.

GRI has three standard disclosure parts: profile, management approach and performance indicators (PI). For achieving most actual results GRI G3.1 launched in 2011 is used as basis. GRI G3.1 can have three levels of application, as shown in Figure 1. In level A GRI G3.1 performance indicator part it includes 84 PI, divided into core (55) and additional (29) ones, and distributed in six dimensions. Furthermore 10 sector specific supplements are created. They describe through in detail going PIs the following branches: Airport Operators, Construction and Real Estate, Event Organizers, Electric Utilities, Financial Services, Food Processing, Media, Mining and Metals, NGOs, Oil and Gas. If applicable to the company profile, sector supplements PIs should be regarded in the SR.

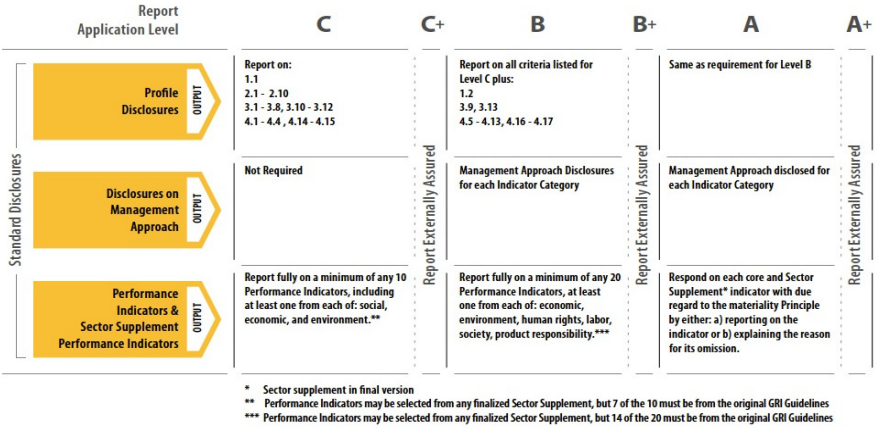


Fig. 1 GRI 3.1 Application Levels [17]

GC has solely 10 principles grouped in four topics – human rights, labor, environment and corruption [31]. Their fulfilling may vary from brief addressing to comprehensive sustainability report. Furthermore, there is a linkage between GRI and GC, since the latter can be produced using GRI PIs [16]. A sustainability report, as far as it fulfils the more detailed GRI requirements, can be listed than under both frameworks.

3 Decision Making Model

The goal of this chapter is to ease DM process towards adoption of Sustainability reporting by companies[2], which are beginners in the field of SR. It shall:

- universally applicable;
- reflect preferences of internal and external stakeholders;
- deliver tangible result on ability to produce SR;
- highlight strengths and weaknesses upon required KPIs.

[2] Throughout this chapter we use "company" „enterprise", „organisation", and other company describing terms interchangeably.

To achieve the set goal, established merge of the two prior described SR guidelines will be utilized. Further on, 10 steps procedure will be proposed. Finally tree compound indexes for Readiness, Usefulness and Convenience of SR will be derived.

Fuzzy logic interpretability when using language is well known. CFL is a new approach, which belongs to mathematical fuzzy logic and improves that property (Table 1). Henceforth, CFL is the selected knowledge engineering tool for solution of the problem discussed here.

To enable 'straight forward' implementation of the results in internationally acknowledged SR, we propose to apply level C of GRI requirements in the proposed model.

GRI G3.1. reporting scheme demands compulsory fulfilling of the first and third standard disclosures (see Figure 1 and in detail Figure 2):

- Profile Disclosures - I;
- Performance Indicators and Sector Supplements Performance Indicators disclosures - III.

In order to extract profounder insights useful for good DM towards sustainability, the proposed model implements mapping of the requirements for fulfilling GRI G3.1 level C criteria, with the ten principles of GC (as prescribed in [16]).

GRI G3.1 level C requires form a company to describe 42 requirements in Standard Disclosure I (Profile Disclosure) as 'reported', or if not, state 'reason for omission'. This first block of questions consists of two groups:

- 12 Basic Requirements – Strategy, Analysis and Organizational Profile, are seen as essential and therefore is assumed that company will always be able to answer them. For the purpose of the study, they will not be explicitly questioned, and taken a priori as possible to be reported on.
- 30 General Requirements (GR) – Report Parameters and Governance, Commitment, Engagement have to be answered only by internal stakeholders of the company (experts, consultants, etc.).

General requirements must be than quantified through giving a true value 0 (no) or 1 (yes) for each requirement, whereas a requirement can have either a '0' or an '1' as an answer, henceforth, the options (0, 0) and (1, 1) are not acceptable.

Further on, Standard Disclosure III is reflected. It was already defined, that it consists from all 55 obligatory KPIs required in GRI G3.1, and 9 additional KPIs, which round up the 'in accordance' criteria with GC. The total number of KPIs in the dimensions Economic, Environmental and Social is then 64.

Using the concepts found in the literature, these KPIs are modelled and algorithmised with four questions. The latter have to be answered, in best case, form both Internal (company) and External (society, investors, suppliers, etc.) stakeholders.

Some examples for stakeholders' perception of important issues connected with SD a company are: 1) give account for company's sustainability performance and activities; 2) improve internal processes to enhance company's performance; 3) engage with stakeholders about sustainability performance; 4) demonstrate company's management on sustainability performance, etc. [14].

Summarizing, for the correct evaluation of every KPI (i) four essential questions were set:

- *How true it is that KPI is important for the company – $I_C(i)$?* KPI is important if: Gives account for company's sustainability performance and activities, and/or Improves internal processes to enhance company's performance, and/or Engages with stakeholders about sustainability performance (e.g. anchoring of renewable energy sources in the company energy-supply-mix is a priority).
- *How true it is that KPI is important for company's external stakeholders – $I_S(i)$?* KPI is important if: Gives account for company's sustainability performance and activities, and/or Improves internal processes to enhance company's performance, and/or demonstrates company's management on sustainability performance, and/or Engages with stakeholders about sustainability performance (e.g. the company supports activities to enable stakeholders (e.g. employees) to embrace and integrate Car-Sharing and E-Mobility, powered by renewable energy sources).
- *How true is that company has the possibility to account for Performance KPI – $PA(i)$?* If company has competences and/or structures and/or workflows, supported by IT, allowing obtaining and processing of data, which describes KPI (e.g. there is accurate data available in order to trace costs on energy sourcing and consumption, etc.).
- *How true is that company's achievement is good enough in regard to KPI – $GA(i)$?* If company is able to comply with KPI, showing good benchmarking and/or improvement on it is laid down in the company's strategy (e.g. energy consumption has been reduced, due to newly built photovoltaic installation, and/or a photovoltaic installation built in the next report period should reduce the energy consumption from non-renewable primary sources).

KPIs are quantified through giving true value between 0 and 1 to the four questions. The scale used for estimating of the true value from internal experts and external stakeholders was simplified and the step was increased to 0.25 (instead the "normal" 0.1). This was decided for two reasons: 1) enabling viability of such interview in the application part of this study, bearing in mind that the number of KPIs is relatively large; and 2) as KPIs' number can be even increased, if the company decides to add on a next step own or further GRI sector supplement KPIs. Nevertheless, it is strongly recommendable to use the 0.1, since it will deliver much more sensitive results.

The five steps of the answer scale are understood as follows:

- 0 - *absolutely false* - No doubts about it, I don't feel any resistance to consider it false.
- 0.25- *more false than true* - I am not sure, but if I have to say true or false I prefer to say false.
- 0.5- *as true as false* - I am not sure, yet I am feeling resistance to say either true or false.
- 0.75- *more true than false* - I am not sure, but if I have to say true or false, I prefer to say true.
- 1- *absolutely true* - No doubts about it, I don't feel any resistance to consider it true.

Every interviewed person (e.g. expert, stakeholder, etc.) has been considered in the model as a single entity, in order to guarantee that each preference is evaluated individually. Through used independence assumption this prevents successfully dynamics present within the group settings. Regarding the type of organization where the model should be applied, a varying number of stakeholders should be addressed. Relevant stakeholders have been identified consistent with organization's strategy and related literature [2, 9, 10 p.574, 12 p.21, 20, 30].

To accomplish the general objective of this work, customized table based on required information for GRI G3.1 level C was created. The calculations needed comprehensive and able to represent complete predicates software. Fuzzy Tree Studio [23] satisfies these criteria profoundly, offering in addition graphical editor for visualizing of decision trees (refer to Fig. 4.). FTS Is able to calculates value of predicates through several FL computing systems. In order to achieve robust results [24 p.6], the proposed model shall make use of Geometric Mean Based Compensatory Logic (GMBCL), which serves CFL especially good.

4 Mathematical Description of the Predicates

Gathering of initial data from the stakeholders according the selected SR schema – GRI G3.1 level C is shown through the first two steps of Figure 2.

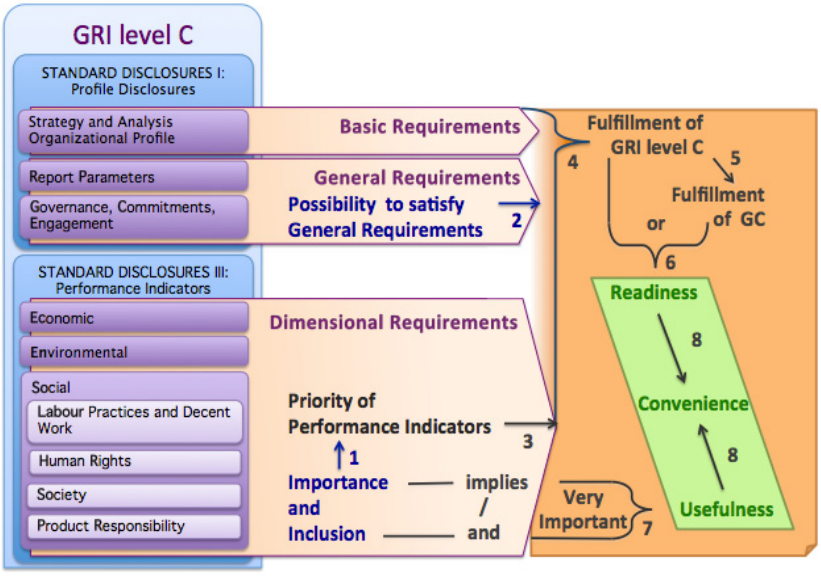


Fig. 2 Construction of raw data supply. Steps implementing the decision making model

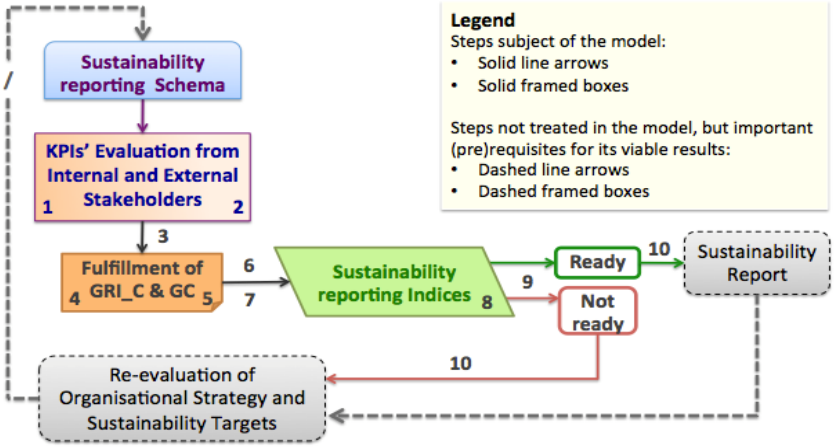


Fig. 3 Schema of the decision making model

Company experts and external stakeholders have to participate solely in the generation of the initial data needed to fulfil steps 1 and 2. Steps 3 - 8 are described through the predicates constituting the proposed model and can be performed through selected software. Schematic workflow of the proposed DM model is presented in Figure 3.

In step 9, a decision 'Ready'/'Not Ready' can be taken, resting upon results of the three indices (Readiness, Usefulness, Convenience). Step 10 describes decision taken from the testing organization with two possible outcomes:

- To produce SR, if the indices result with one, or
- To start improvement process, if the latter should be zero.

Step 9 and 10 will not be defined through predicates, since they are specific product of the decision of the Chief Management in the company applying the model.

4.1 Importance of Performance Indicator for the Organization

Importance of KPI i is observed if KPI is important for the organisation itself, or for its stakeholders. These conditions guarantee objectivity and reflection of the opinion of the organization and its environment. Here data is collected directly through an interview and then summarized in a general importance.

$$I(i) = I_C(i) \vee I_S(i) \tag{1}$$

Where:
i - GRI Key Performance Indicator
$I(i)$ - Importance of KPI i
$I_C(i)$ - Importance of KPI i for the company
$I_S(i)$ - Importance of KPI i for the stakeholders

4.2 Inclusion of Performance Indicator in the SR

KPI Inclusion in the SR can succeed, if it is possible to account the indicator and there are good achievements according it. This raw data is provided from the company in through the above-specified questions towards each KPI. In order to disclose according$i$, inclusion must be $InSR(i) > 0.75$. This condition is the one taken into account while calculating the compound predicates for fulfilling of GRI level C and fulfilling requirements of GC.

$$InSR(i) = PA(i) \wedge GA(i) \tag{2}$$

Where:
$InSR(i)$ – Inclusion of KPI i in the SR
$PA(i)$ – Possibility to Account KPI i
$GA(i)$ – Good Achievement of the company in regard to KPI i

The results of Eq. 1 and Eq. 2 are achieved through calculation of the importance evaluation given by the **c**ompany and its external **s**takeholders. Through definition of Importance and Inclusion of a KPI, step 1 of the proposed process is fulfilled.

4.3 Priority of a Performance Indicator for the Organization

Priority of KPI is described through importance of an indicator and its inclusion. In order to disclose according i, it is proposed that KPI should have priority with value $P(i) > 0.75$. Yet, even with lower priority, a KPI can be taken into account, if it is in accordance with the strategic vision of the company or is very important for it or the stakeholders. This predicate fulfils step 3 of the process.

$$P(i) = I(i) \wedge InSR(i) \tag{3}$$

Where:
$P(i)$ – Priority of KPI i for the company

4.4 Fulfilment of GRI G3.1 Level C

GRI G3.1 level C is fulfilled when the GR (standard disclosure I – profile disclosures) and Dimensional Requirements (DR) (standard disclosure III – KPIs) are satisfied.

General Requirements are prioritized by only company internal stakeholders. The given data from the company only and describe its ability to comply or to point a reason to omit particular requirement.

Ability to report according DR is taken from the Inclusion values of every KPI. Dimensional requirements have to achieve a sum of ten – $GRI\_DR\_sum(x)$, whereas at least one KPI for each dimension – economic, environmental and social can be disclosed.

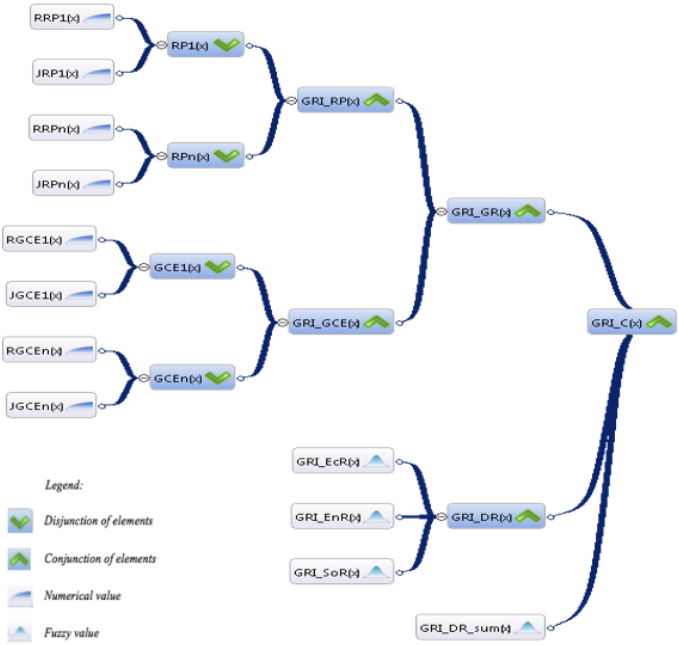


Fig. 4 Satisfaction of GRI level C from organization x. Generic version

FTS bases on graphical building of predicates and thus provides through comprehensive visualization a very good understanding of connections among the predicates in the decision tree. In this sense, Figure 4 provides an overview on fulfilment of GRI G3.1 level C.

In order to simplify the image, the Profile Disclosure j is represented through its 1 and n values in the predicate components: $RRPj(x)$, $JRPj(x)$, $RGCEj(x)$, $JGCEj(x)$.

$$GRI\_C(x) = GRI\_GR(x) \wedge GRI\_DR(x) \wedge GRI\_DR\_sum(x) \quad (4)$$

Where:

$GRI\_C(x)$ – Fulfilment of GRI C level from company x

$GRI\_GR(x)$ – GRI General Requirements for company x

$GRI\_DR(x)$ – GRI Dimension Requirements for company x

$GRI\_DR\_sum(x)$ – Sum of applicable GRI Dimension Requirements for company x

As previously explained, the first group of the Standard Disclosure I (Basic Requirements) should be easily answered. Therefore it is expected that organization's strategic management will fulfil them without explicitly questioning the latter.

$GRI\_GR(x)$ are met, if the report parameters plus government, commitment and engagement parts are accomplished (Eq. 4a) and (Eq. 4d). Fulfilling of *GR* is defined only by the internal stakeholders.

$$GRI\_GR(x) = GRI\_RP(x) \wedge GRI\_GCE(x) \tag{4a}$$

Where:
$GRI\_RP(x)$ – GRI General Requirement Report Parameters part satisfaction for company *x*.
$GRI\_GCE(x)$ – GRI General Requirements – Governance, Commitment and Engagement part satisfaction for company *x*.
This assessment mirrors step 2 of the proposed procedure.
For the report parameters the predicate is described as follows:

$$\begin{aligned} GRI\_RP(x) &= \forall_j\big(RPi(x)\big) = \forall_j\big(RRPj(x) \vee JRPj(x)\big) \\ &= \hat{j}\big(RRPj(x) \vee JRPj(x)\big) \end{aligned} \tag{4b}$$

Where:
j – GRI Profile Disclosure
$RPj(x)$ – Possibility to satisfy profile disclosure j of the Report Parameters part for company x
$RRPj(x)$ – Possibility to Report on profile disclosure j of the Report Parameters part for company x
$JRPj(x)$ – Possibility to Justify not reporting of profile disclosure j of the Report Parameters part for company x
$\hat{j}$ – Conjunction over all KPIs
Governance, commitment and engagement are expressed as:

$$\begin{aligned} GRI\_GCE(x) &= \forall_j GCEj(x) = \forall_j\big(RGCEj(x) \vee JGCEj(x)\big) \\ &= \hat{j}\big(RGCEj(x) \vee JGCEj(x)\big) \end{aligned} \tag{4c}$$

Where:
$GCEj(x)$ – Possibility to satisfy profile disclosure j in the Governance, Commitment and Engagement part for company x
$RGCEj(x)$ – Possibility to report on profile disclosure j in Governance, Commitment and Engagement part for company x
$JGCEj(x)$ – Possibility to justify omission of the profile disclosure$j$ in the Governance, Commitment and Engagement part for company x
Dimensional requirements are stated as:

$$GRI\_DR(x) = GRI\_EcR(x) \wedge GRI\_EnR(x) \wedge GRI\_SoR(x) \tag{4d}$$

Where:
$GRI\_EcR(x)$ – GRI Requirements on Economic KPIs for company x
$GRI\_EnR(x)$ – GRI Requirements on Environmental KPIs for company x
$GRI\_SoR(x)$ – GRI Requirements on Social KPIs for company x

The predicates $GRI\_DR\_sum(x)$, $GRI\_DR(x)$, $GRI\_EcR(x)$, $GRI\_EnR(x)$, $GRI\_SoR(x)$ are modeled by sigmoid membership functions, in accordance with the requirements of GRI G3.1 level C, including KPIs with true value of the predicate $InSR(i) \geq 0.75$.

Herewith, step 4 of the process is accomplished.

4.5 Fulfilling of GC

If there are KPIs corresponding to each GC principle, with good enough company achievements and possibilities to account them, organization can fulfil GC.

Since the principles of GC are explained through the GRI KPIs, the condition for true value of the predicate $InSR(i)$ greater than 0.75 will be here further maintained. Taking the proposition of the GRI framework, that at least one KPI per dimension must be fulfilled in order to comply with GRI G3.1 level C as lemma, it is proposed, that at least one KPI per GC principle has to be accomplished, in order to disclose according GC principle. Thus, fulfilling of GC will complete step 5 of the proposed process, expressed with the following predicate:

$$GC(x) = \forall_k\left(\exists_{i\epsilon P_k}(GA(i) \wedge PA(i))\right) = \bigwedge_k\left(\bigvee_{i\epsilon P_k}(GA(i) \wedge PA(i))\right) \quad (5)$$

Where:
$GC(x)$ – Satisfaction of GC from company x
P_k – GC principle k
k – Order number of the GC principle
$\bigwedge_k$ – Conjunction over all Principles
$\bigvee_{i\epsilon P_k}$ – Disjunction over all KPIs, which describe Principle k

4.6 Readiness for Introducing of Sustainability Report

Readiness to introduce SR is present, if an organization is ready to achieve GRI G3.1 level C or satisfies the GC principles. $RSR(x)$ bases on inclusion of enough number of KPIs in order to have complete GRI level C or GC report. The value of the index is in its ability to assign level of readiness to comply according one of the two frameworks or in the best case according both of them. Furthermore through the cross-reference table in 16], a focused improvement of issues related to the lagging KPIs can be achieved.

$$RSR(x) = GC(x) \vee GRI\_C(x) \quad (6)$$

Where:
$RSR(x)$ – Readiness of company x for making a Sustainability Report
Readiness fulfils step 6 of the procedure.

4.7 Usefulness of Sustainability Report

Sustainability report is useful, if all KPIs included in the SR are important, or there are some of them very important (Eq. 7). Very important KPI are calculated using the standard way of modelling of modifier 'very' [9] (Eq. 7a). These conditions prevent 'greenwashing' and avoid producing of a report, where the KPIs are simply 'checked' for the sake of the volume and level of the report. Another argument is that in this manner, crucial issues for stakeholder (especially external ones) can be successfully addressed and reported. Further merit from managerial point of view for the proposed index, is given possibility for targeted aligning of actual performance with the required one.

$$U(x) = \forall i\big(InSR(i) \rightarrow I(i)\big) \vee \exists i\big(InSR(i) \wedge VI(i)\big) \tag{7}$$

Where:
$U(x)$ – Usefulness of Sustainability Report for company x

$$VI(i) = I^2(i) \tag{7a}$$

Where:
$VI(i)$ – Very important KPI i
This predicate describes step 7 of the process.

4.8 Convenience for Elaborating Sustainability Report

Convenience to elaborate a SR exists, if the company is ready to introduce such report and it is useful for it. Convenience is a natural outcome of the first two indices. It unifies the virtues of Readiness and Usefulness and gives 'on a glance' answer of the question how favourable SR can be for certain organization.

$$C(x) = RSR(x) \wedge U(x) \tag{8}$$

Where:
$C(x)$ – Convenience for Sustainability Report elaboration

The predicate $C(x)$ accomplishes step 8 of the process and is the final stage of the mathematical description.

The proposed evaluation scale suggests to interpret Convenience in the interval $0.5 \leq C(x) < 0.75$, as 'it is more true than false that the SR is convenient'; respectively, if $C(x) \leq 0.75$, the recommendation to the company to issue a SR, will have a 'very true' value.

5 Utilization of the Results

The proposed method gives structured approach to a complex DM problem. It involves company experts (internal stakeholders) and addresses active involvement of the external stakeholders, in accordance with the requirements of GRI.

The model derives significant topics for a future SR and helps shapes its '*scope*'. Through definition of *Importance* of the KPIs, the results of the model make explicit, which is the needed '*content'* in the SR, for both the company and its stakeholders. The parameters *good achievement* and *possibility to account,* define the '*quality*' of the reported data, setting the results in a tangible perspective of their potential utilization. Hence, parameters can be described through KPI (if data available), or receive a narrative explanation (if e.g. KPI improvement is agreed as priority in the company vision).

This approach can be very effective to highlight topics for incentives in the organization, in order to advance towards diverse activities aiding SD: energy efficiency; waste reduction; interaction with the local community; and emphasize on a non-discriminative working conditions, etc.

The boundary of the SR can be defined according the range, set by the invited for participation in the interview stakeholders and the company itself. However, there is a strong recommendation to involve the company's value chain on a broad perspective.

Because of the large number KPIs reflected, this approach can be further useful, if the company should decide to apply level B or A (with more required KPIs) of the GRI framework. Moreover, GRI G4's development shows that the three levels A, B and C will be replaced with reporting 'in accordance' to the framework and with profound definition of 'materiality'. Consequently, companies implementing the proposed model shall have the option to test themselves upon all core KPIs. For that important reason, the limitation 'only 10 KPIs' was deliberately omitted and the wide spectrum of KPIs was offered.

Furthermore through the large number proceeded KPIs, the company receives opportunity for comprehensive measurement of business practices related to the KPIs. This enables not only optimization of performance, but also recognition of related procedures, processes, or rules, which empower good behavior of the KPIs, thus beneficial for the company. Sensitivity of the results will be meaningfully higher, if the scale for evaluation of the KPIs has step of 0.1 in the interval [0,1]. Accordingly, a long-term strategy for SD can be empowered by new clustering of activities or informed investing and concentration on important issues for the stakeholders and abandoning of not relevant ones.

If KPIs envisaged for inclusion can be described with real-time data, and a database empowers the calculations of the model, then up-to-date trends in the indices' values can be observed. All three indices are created with the premise to facilitate informed DM for the chief management of organization, on its way to discuss and assess feasibility to adopt SR. They can be easily embedded in e.g. sustainability management dashboard and provide instant and easy to apprehend information on company progress towards SR.

After fulfilling of step 8 of the process, decision towards SR can be taken – step 9 (refer to Figure 3). For beginners it is recommendable to concentrate on four to eight KPIs in the different dimensions, whereas sector supplements should be also reviewed [30]. In this sense, a loop in the DM process, resulting in depth approach on previously envisaged vital for the organization's strategy KPIs is recommended – step 10 in Figure 3.

The proposed model can be part of organizational 'Backcasting-toolkit', defining where company wants to be, in the sense of SD. The model can quantitatively determine existing implementation of sustainable practices and highlight issues that have to be better addressed in the future. It can be used threefold, depending on organization's preparation and advances in the field of sustainability.

- For companies *not acquainted with SR* practices and structure it is a good base to see interrelated issues of their business, with "learning along the way" added value.
- For organizations that *have experience in e.g. environmental or corporate social reports* it can be test for an "upgrade" means sustainability report.
- For enterprises *already releasing sustainability report*, the method can give more insights on issues not taken into account till the moment, but important to company and stakeholders. Additionally, achievable and accountable performance indicators can be further discovered.

Feasibility of the proposed model will be tested in the following section where it shall be applied in real company case.

6 Company Case Study

Three companies[3], with possibly different characteristics according Legal form and beginners or not experienced at all in regard of SR, were interviewed:

Table 2 Characteristics of participating companies

Company	Country	Ownership/ Legal Form/ Range of operation/ Size	Contact Person
A	Germany	Private/ GmbH/ National/ SME	Executive assistant of the company management
B	Cuba	State/ Empresa Estatal Socialista/ National/ SME	Operative Director,Technical Director, HR Manager
C	Cuba	Private/ Company Group / Multi national/ Large (national branch)	Sales, Logistics, Accounting and Controlling Mangers

Except for company C, whose mother concern is issuing sustainability report since 2004, all the questioned companies had no published SR before answering the questionnaire. The companies received 30 GR+64 KPI positions questionnaire, which was filled with some till exhaustive explanation aid from the authors.

Detailed results for company A are presented in Table 3. Obtained information from the interviews with companies and stakeholders was processed through the

[3] All respondents preferred anonymity and confidentiality for their answers. They are presented through profile specifying characteristics in Table 2.

set of predicates discussed in section 6. The correspondent statements were modelled with FTS. KPIs, feasible to be included in a SR are shown for the three companies in Table 4. Summarized results of the SR-Indices are exhibited in Table 5.

Companies' results are interpreted through five different states a KPI can have (see Table 4 and Table 5). The authors, with regard to the performed KE, defined KPIs' states, giving brief recommendations on possible following up procedures. To allow easier implementation e.g. in sustainability dashboard tool, five colors have been used.

Green: KPI must be included, if InSR ≥0.75 and P≥0.75

Inclusion of KPIs from this group is a must. They contribute to higher publicity of important practices, already supporting SD in the company. It is recommendable to maintain the performance on these parameters and set incentives to support continuous good achievements.

Yellow: it is recommendable to include KPI, if I≥0.75 and PA≤0.75 or GA≤0.75

Some KPIs should be included because of their importance, despite that in the moment of estimation, there is no possibility to account them or the company has no good achievements on them. This shall guarantee transparency of the report and reflect properly the informational needs of the stakeholders. Consequently, if these KPIs should be disclosed, the SR should include commitment where the conditions to comply on them with a data set in future shall be laid down.

The Chief Management of the company should work towards: achieving competences to manage these indicators, since they reflect both internal and external stakeholders' anticipations and expectations for a well-done sustainability performance; or improved data-collection activities and comply with robust information according here outlined KPIs.

Blue: KPI can be included, if Ic≥0.75 or Is≥0.75 and GA≤0.75

Theses KPIs are important, but cannot be achieved on the demanded from the stakeholder's level, are particularly dependent on improvement in achievements of the company towards them. Further filtering criteria is that blue KPIs appear critical for only one group of the interviewed parties. They show profounder distinction of the informational needs addressee (internal or external stakeholders). It is recommendable to include blue KPIs in the SR, in order to guarantee transparency. Furthermore, action plans and incentives for better achievements of these KPIs in the company agenda, thus enable inclusion with positive sign in the SR should be elaborated.

Purple: KPI can be included, if Ic≥0.75 or Is≥0.75 and PA≤0.75

Same as the blue group, with alterity that purple KPIs demand improved data-collection, in order to comply on them with robust data sets. The same recommendations as above can be given.

Red: KPI can be included, if InSR≥0.75 and P≤0.75

KPIs are not prominent for the corporate strategy. It is advisable to revise them and if the assumptions from the first check are confirmed on a broader perspective (e.g. complete value chain), they can be dropped in order to free assets for more important improvements.

Table 3 represents Values of the KPIs given by all internal and external stakeholders (*Ic, Is, GA, PA*) and calculated results (*I, VI, InSR, P*). Regarding the GRI G3.1 guideline, obligatory KPIs are presented in white and the additional ones, in grey lines. The pattern background of the indicators relates to the KPI's state.

Table 3 Detailed results for company A. Values of KPI *(i)* given by Stakeholders (*Ic(i)*, *Is(i)*, *GA(i)*, *PA(i)*) and calculated predicates (*I(i)*, *VI(i)*, *InSR(i)*, *P(i)*)

KPI (i)	*Ic*	*Is*	*GA*	*PA*	*I*	*VI*	*InSR*	*P*
EC1	0,75	0,75	1	1	0,75	0,5625	1	0,9086
EC2	1	1	0,5	0,5	1	1	0,5	0,63
EC3	0,75	0,5	0,5	0,5	0,646446609	0,417893219	0,5	0,5447
EC4	0	0	0	0	0	0	0	0
EC5	0,25	0,25	0,5	0,25	0,25	0,0625	0,3536	0,315
EC6	1	1	1	1	1	1	1	1
EC7	0,25	0	0	0	0,133974596	0,017949192	0	0
EC8	1	1	1	1	1	1	1	1
EN1	1	0,75	1	1	1	1	1	1
EN2	1	1	1	1	1	1	1	1
EN3	1	0,75	0	0	1	1	0	0
EN4	1	1	0,75	0,75	1	1	0,75	0,8255
EN5	1	0,5	1	1	1	1	1	1
EN6	0,75	0,75	0,75	0,75	0,75	0,5625	0,75	0,75
EN7	0,5	0,5	0,5	0,75	0,5	0,25	0,6124	0,5724
EN8	0,25	0,25	0,5	1	0,25	0,0625	0,7071	0,5
EN11	1	1	1	1	1	1	1	1
EN12	1	1	1	1	1	1	1	1
EN16	0,75	0,75	1	1	0,75	0,5625	1	0,9086
EN17	0,5	0,5	0,5	0,5	0,5	0,25	0,5	0,5
EN18	0,75	0,75	0,5	0,5	0,75	0,5625	0,5	0,5724
EN19	0,25	0,25	0,5	0,5	0,25	0,0625	0,5	0,3969
EN20	0	0	0,25	0,25	0	0	0,25	0
EN21	0,25	0	0,5	0,5	0,133974596	0,017949192	0,5	0,3223
EN22	0,25	0	1	1	0,133974596	0,017949192	1	0,5117
EN23	0	0	0	0	0	0	0	0
EN26	0,75	0,5	1	1	0,646446609	0,417893219	1	0,8647
EN27	1	0,75	1	1	1	1	1	1
EN28	0	0	1	1	0	0	1	0
EN30	1	0,5	1	1	1	1	1	1

Table 3 (*continued*)

LA1	0,5	0,25	1	1	0,387627564	0,150255129	1	0,7291
LA2	0	0	1	1	0	0	1	0
LA15	0,25	0	1	1	0,133974596	0,017949192	1	0,5117
LA4	0	0	1	1	0	0	1	0
LA5	0	0	1	1	0	0	1	0
LA7	0,25	0	1	1	0,133974596	0,017949192	1	0,5117
LA8	0,25	0	1	1	0,133974596	0,017949192	1	0,5117
LA10	0,5	0	1	1	0,292893219	0,085786438	1	0,6641
LA13	0,25	0	1	1	0,133974596	0,017949192	1	0,5117
LA14	0,25	0,5	1	1	0,387627564	0,150255129	1	0,7291
HR1	0,5	0,5	1	1	0,5	0,25	1	0,7937
HR2	1	1	1	1	1	1	1	1
HR3	0,25	0,5	1	1	0,387627564	0,150255129	1	0,7291
HR4	0	0	1	1	0	0	1	0
HR5	0,75	0,75	1	1	0,75	0,5625	1	0,9086
HR6	1	1	1	1	1	1	1	1
HR7	0,75	0,75	1	1	0,75	0,5625	1	0,9086
HR8	0	0	1	1	0	0	1	0
HR9	0	0,5	1	1	0,292893219	0,085786438	1	0,6641
HR10	0,25	0,25	1	1	0,25	0,0625	1	0,63
HR11	0	0	1	1	0	0	1	0
SO1	0,5	0,5	0,75	0,75	0,5	0,25	0,75	0,6552
SO9	0,75	0,75	0,5	0,5	0,75	0,5625	0,5	0,5724
SO10	0,5	0,5	0,5	0,5	0,5	0,25	0,5	0,5
SO2	0	0	0	0	0	0	0	0
SO3	0	0	0	0	0	0	0	0
SO4	0	0	0	0	0	0	0	0
SO5	0,25	0	0	0,25	0,133974596	0,017949192	0	0
SO6	0	0	0	0	0	0	0	0
SO8	0,25	0	0	0,25	0,133974596	0,017949192	0	0
PR1	0,75	0,75	0,75	0,75	0,75	0,5625	0,75	0,75
PR3	1	1	1	1	1	1	1	1
PR6	0,5	0,25	0,25	0,5	0,387627564	0,150255129	0,3536	0,3646
PR9	0	0	0	0	0	0	0	0

Legend of the used patterns, with summary of the matching KPIs to each state for the participating companies is presented in Table 4.

Table 4 Companies' results. KPIs, fulfilling proposed states

	KPIs fulfilling the state		
KPI State	Company A	Company B	Company C
Green	EC1, EC6, EC8, EN1-2 EN4-6, EN26-27, EN30, HR5-7	EC3, EC7, EN2, EN6, EN12, EN16, EN22-23, EN26-27, EN30, LA2, LA5, LA8, LA10, HR1, HR4-6, HR9, SO2-5, SO8-10, PR9	EC1-2, EN1, EN4-7, EN26-30, LA1-2, LA7-8, LA10, LA13, LA15, HR1, HR4-6, HR9, SO2-5, SO8-10, PR9
Yellow	EC2, EN3, EN18, SO9	EC1-2, EN17-21, LA4, LA15, PR1, PR3	-
Blue	-	EN5	-
Purple	-	EC8, EN3-4	-
Red	SO1	EC4, EC6, HR2, HR7-8, HR10	EC4-5, EC7, LA14, HR1-11, SO5-6

Readiness to implement **GRI level C** is **1**. The **Readiness** to comply according **GC** is **0**, since all KPIs describing 10-th Principle of GC have inclusion 0. It can be assumed, that should the company like to comply towards GC, improvements connected with the KPIs: SO2 – 6, describing Principle 10 [16], have to be made. This might be also feasible in the case of SO5, where the company gives a slight importance of 0.25.

The three Indices defining ability to issue a Sustainability Report have the values: **Convenience- 1; Usefulness- 1;** and **Readiness- 1**.

Consequently, it is very convenient for the company to introduce SR, since company C is ready for GRI level C and the SR would be very useful. The analysis of results gives full recommendations for advancing towards SR and maintaining of good levels of the concerned KPIs.

Table 5 Company results. Fulfilment of GRI level C and GC and SR Indices

Fulfilment / Company	GRI_C	GC	Readiness	Usefulness	Convenience
A	1	0	1	1	1
B	1	0.8776	0.9368	1	0.9679
C	0,9865	1	1	1	1

7 Conclusions and Outlook

This chapter presented method to support informed DM, based on concrete metrics, towards introducing Sustainability Reporting in organizations.

The novelty of the approach lays in the adoption of CFL in the mathematical description of the problem and creation of three compound indices. Each predicate component was deductively explained, using argumentative analysis and knowledge

engineering, modelling from linguistic expression of the knowledge involved and its transformation in pseudo code, using CFL. Implementing CFL is important contribution to the knowledge base, since it is easily applicable for its main characteristic – modelling through language, and furthermore, allows conceptualization and measurement, which cannot be obtained through non-compensatory fuzzy logic. It demonstrated translation of qualitative opinions of the stakeholders (internal and external) into quantitative values; and used them for ranking of priorities, facilitating instantaneously the organization with information on its Status Quo according ability to issue a SR.

The company case study underlined the usefulness of this model for the following decision making problems:

- Which KPIs should be included in the SR, taking into account the company strategy and the stakeholders interests?
- Which other KPIs could be included, contributing to the completeness of the SR and the image of the company?
- Is the company ready to report, according the two most important standards for SR?
- Is the SR based on the derived set of KPIs useful for the organization?
- Is it convenient for the company to introduction SR?

The proposed method enables organizations to assess own status quo according list of profile disclosures developed from international expert panels and seen as essential for sustainable way of doing business, service-providing, etc. It gives base for making informed choices, even if organization should appreciate it is still not ready for issuing a SR. Through asking simple questions and merging own comprehension with that of the stakeholders, awareness, hidden potentials can be captured and developed. It can also prove profound sustainability connected activities in specific fields, which were not appreciated till the moment as positive communication worth it assets. Finally, communication on covered GRI G3.1 disclosures and/ or GC principles is highly recommendable. Gained publicity makes the organizations not only more responsible corporate citizen, but pushes through generally further actions while adopting sustainability.

Possible directions for future research:

- Application in large number of companies from different business fields: to obtain new experiences according its applicability; contextualization and complementation of various business sectors can be beneficial and therefore recommendable.
- Ontology of global indexes: universal tool to estimate level of readiness of a company to incorporate SR. Based on statistically significant applications' number, set of specific "universal" questions can be developed. It will reflect: Size (corporations, small and medium enterprises) and Field of work (producing, services, logistics, etc.). Through the achieved answers, readiness for developing sustainable business on strategy level will be assessed and recommendation for short, middle or long term SR implementation will be give. Results can be also useful for Backcasting activities.

- KPIs can be grouped in their clusters as provided form GRI and offer another type of generalization of Readiness, Usefulness and Convenience. Such approach can be used to serve e.g. KING iii, Sustainability Dashboard, etc. [2, 24, 34]. The results can be then applied directly to the chosen reporting schema e.g. in STORM.
- The GRI G4 (fourth generation of GRI to come in May 2013) can be reflected through redefined Inclusion of the KPIs and describe larger scope of the report throughout the complete value chain. Inclusion will be then constituted additionally to the possibility to account KPI '*or*' the good achievements according it; from an '*and*' relation between possibility to account KPI '*or*' Very Important status of KPI.
- GRI G4 Inclusion predicate:

$$InSR\_G4(i) = (PA(i) \wedge GA(i)\,) \vee (PA(i)\ \wedge VI(i)) \tag{9}$$

Acknowledgments. The companies interviewed in the case study had critical role for this chapter. They enabled us to test and evaluate the created model in real life conditions. The authors thank for the dedicated time and friendly support.

References

1. Bouchet, A., Pastore, J.I., Espin Andrade, R., et al.: Arithmetic Mean Based Compensatory Fuzzy Logic. International Journal of Computational Intelligence and Applications 10, 231–243 (2011), doi:10.1142/S1469026811003070
2. Deloitte, King III Maturity Dashboard Supporting your drive for good corporate governance (2010)
3. Deloitte Development, Drivers of long-term business value and strategy (2012)
4. Deri, C., Connolly, P., Maw, L., Arndt, T.: Corporate Responsibility & Sustainability Communications (2008)
5. Ebert, U., Welsch, H.: Meaningful environmental indices: a social choice approach. Journal of Environmental Economics and Management 47, 270–283 (2004)
6. Einhorn, H.J., Hogarth, R.M.: Behavioral Decision Theory: Processes of Judgment and Choice. Annual Review of Psychology 32, 53–88 (1981)
7. Environmental Leader, Environmental and Energy Data Book Released for Q4 · Environmental Management & Energy News · Environmental Leader (2011)
8. Espin Andrade, R.A., Chao Bataller, A., Marx Gómez, J., Racet Valdés, A.: Fuzzy Semantic Transdisciplinary Knowledge Discovery Approach for Business Intelligence. In: Espin Andrade, R.A., Marx Gómez, J., Racet Valdés, A. (eds.) Towards a Transdisciplinary Technology for Business Intelligence, 1st edn., pp. 13–34. Shaker Verlag GmbH, Germany (2011)
9. Espín Andrade, R.A., Fernándes, E., Gonzaléz, E.: Un Sistema Lógico para el Razonamiento y la Toma de Decisiones: La Lógica Difusa Compensatoria Basada en la Media Geométrica. Revista Investigación Operacional, 230–245 (2011)
10. Espín Andrade, R.A., Gonzalez Caballero, E., Fernandez Gonzalez, E.: Compensatory Inference Systems. Soft Computing for Business Intelligence (2012)

11. Espin Andrade, R.A., Vanti, A.A., Marx Gómez, J., Racet Valdés, A.: SWOT-OA Fuzzy Analysis for Strategic Plan Evaluation and Decision Making Support. In: Espin Andrade, R.A., Marx Goméz, J., Racet Valés, A. (eds.) Towards a Trans-disciplinary Technology for Business Inteligence, pp. 89–111. Shaker Verlag, Aachen (2011)
12. Freeman, R.E.: What is Stakeholder Theory? - R. Edward Freeman (2009), http://www.youtube.com/watch?v=bIRUaLcvPe8 (accessed May 30, 2012)
13. French, S.: Decision Theory: An Introduction to the Mathematics of Rationality. Halsted Press, New York (1986)
14. Futera, KPMG, Sustain Ability, Reporting Change: Readers & Reporters Survey (2010)
15. Graedel, T.E., Allenby, B.R.: AT&T. Industrial Ecology 416 (1995)
16. GRI & UN Global Compact, Making the connection. The GRI Guidelines and the UNGC Communication on Progress. UN Global Compact and GRI (2007)
17. GRI, GRI Application Levels, 1–5 (2011)
18. Haanaes, K., Reeves, M., Strengvelken, I., et al.: Sustainability Nears a Tipping Point. Findings from the 2011 Sustainability & Innovation Global Executive Study and Research Project MIT Sloan 19 (2012)
19. Havas Media, Meaningful Brands global factsheet (2011), http://www.havasmedia.com/our-thinking/meaningfulbrands/meaningful-brands-global-factsheet/ (accessed July 5, 2012)
20. Holmberg, J., Robèrt, K.-H.: Backcasting from non-overlapping sustainability principles — a framework for strategic planning. International Journal of Sustainable Development and World Ecology 7, 291–308 (2000)
21. Kaptein, M., Schwartz, M.S.: The Effectiveness of Business Codes: A Critical Examination of Existing Studies and the Development of an Integrated Research Model. Journal of Business Ethics 77, 111–127 (2007), doi:10.1007/s10551-006-9305-0
22. NEEF, Business & Environment, GreenBiz Group (2011) Toward Engagement 2.0: Creating a More Sustainable Company Through Employee Engagement, 1–46 (2011)
23. Passoni, L.I., Meschino, G.J., Gesualdo, S., Monjeau, A.: Fuzzy Tree Studio: Una Herramienta para el Diseño del Tablero de Mando para la Gestión de Áreas Protegidas. In: III Taller Internacional de Descubrimiento de Conocimiento, Gestión del Conocimiento y Toma de Decisiones, Santander, pp. 10–11 (2011)
24. Project Group STORM, STORM Projektabschlussbericht. 1–684 (2010)
25. Seuring, S., Müller, M.: From a literature review to a conceptual framework for sustainable supply chain management. Journal of Cleaner Production 16, 1699–1710 (2008), doi:10.1016/j.jclepro.2008.04.020
26. Sietz, M., Seuring, S.: Ökobilanzierung in der betrieblichen Praxis (Gebundene Ausgabe), 190 (1997)
27. Singh, R.K., Murty, H.R., Gupta, S.K., Dikshit, A.K.: Development of composite sustainability performance index for steel industry. Ecological Indicators 7, 565–588 (2007), doi:10.1016/j.ecolind.2006.06.004
28. Singh, R.K., Murty, H.R., Gupta, S.K., Dikshit, A.K.: An overview of sustainability assessment methodologies. Ecological Indicators 15, 281–299 (2012), doi:10.1016/j.ecolind.2011.01.007

29. Globescan, S., Hamilton, R., Erikson, J., Coulter, C.: Sustain Ability Globescan, KR (2012) Learning from leaders, http://www.2degreesnetwork.com/groups/managing-sustainability/resources/learning-leaders-2012-sustainability-leaders-survey-webinar-recording/ Accessed 19 Apr (2012
30. SustainAbility, FBDS, UNEP, Road to credability: A study of sustainability reports in Brazil, 1–44 (2010)
31. UN GlobalCompact, 10 Principles of GlobalCompact Communication on Progress (2000), http://www.unglobalcompact.org/AboutTheGC/TheTenPrinciples/index.html
32. Utopies Stratégie & développement durable, Sustainability reporting at crossroads Reporting Trends Survey, Paris (2012)
33. Van Wensen, K., Broer, W., Klein, J., Knopf, J.: The state of play in sustainability reporting in the EU. Programme for Employment and Social Solidarity - PROGRESS, 1–173 (2011) (2007-2013)
34. Willard, B.: Sustainability Dashboard, http://www.sustainabilityadvantage.com/sustainability_dashboard/dashboard-3-2.php (accessed May 1, 2012)

Type-2 Fuzzy Logic in Decision Support Systems

Diego S. Comas, Juan I. Pastore, Agustina Bouchet, Virginia L. Ballarin and Gustavo J. Meschino

Abstract. Decision Support Systems have been widely used in expert knowledge modeling. One of the known implementation approaches is through definition of Fuzzy Sets and Fuzzy Predicates, whose evaluation determines the system's output. Despite Type-1 Fuzzy Sets have been widely used in this type of implementation, there are uncertainty sources that cannot be adequately modeled when using expert knowledge minimizing their effect on system's output, especially when it comes from several experts opinions. Type-2 Fuzzy Sets deal with fuzzy membership degrees, which can represent adequately the typical uncertainties of these systems. In this chapter, we generalize the operators of Fuzzy Logic in order to evaluate Fuzzy Predicates with Type-2 Fuzzy Sets and we define measures to assess the degree of truth of these predicates to define the theoretical background of the Decision Support Systems using this methodology. We present an example application of decision-making and a brief discussion of the results.

1 Introduction

Decision Support Systems (DSS's) have shown to be helpful to model the expert knowledge. They have a set of input variables, whose are evaluated and processed to achieve results leading to be supportive in a process of decision making [2, 16].

Diego S. Comas · Juan I. Pastore · Agustina Bouchet ·
Virginia L. Ballarin · Gustavo J. Meschino
Grupo de Procesamiento Digital de Imágenes, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Argentina
e-mail:{diegoscomas,juan.pastore,agustina.bouchet,vballari,gustavo.meschino}@gmail.com

Diego S. Comas · Juan I. Pastore · Agustina Bouchet
Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Argentina

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_18,

DSS's have been used in a wide range of applications, such as risk assessment, churn prediction as a warning of clients who are likely to discontinue the use of a service, employee performance evaluation, analysis of company's competitivity [2, 6], medical diagnosis support based on image processing [11, 13, 29], image segmentation and voice recognition [30].

An interesting approach to implement a DSS consists on expressing knowledge as a set of Fuzzy Predicates (FP's). FP's make relations between input variables using logical connectives. The output of the DSS is achieved by evaluating these relations [1, 3, 16]. The degree of truth of a FP is intended to be into the real interval [0,1] and it is obtained by Multi-valued Logics (ML's) [10, 27] applied to the membership values given by Fuzzy Sets (FS's). In this paradigm, the quality and usability of the system depends on the adequacy of the proposed model to the available knowledge. Once designed, a DSS is not able to discover any knowledge but they are a logic implementation of it [16].

There are several uncertainties sources that should be considered in DSS's designing to minimize their unexpected effects [20, 26]. In first place, there exist uncertainties in the data used as inputs, due to the fact that data are themselves uncertain or there are some perturbations in the acquirement system. In second place, not less important, while designing the predicates set, experts' uncertainties in the way they use words to describe the knowledge should be considered: the same word could lead to different meanings for different experts [19, 20, 22].

Type-1 Fuzzy Sets (T1FS's) determine a unique membership value to each set element, being it into the [0,1] interval [27, 28]. This allows defining gradually the degree of truth of one predicate using a variable, and this is useful to model the knowledge. However, having only a value of membership (or degree of truth) often it is not enough to consider some special uncertainties in a DSS [14, 18, 19, 21]:

- Meanings of the words used in the predicates may be uncertain;
- Considering experts' opinions, similar concepts may be considered differently by different experts. These differences cannot be accurately modeled by a T1FS;
- There may be input perturbations producing uncertainties that are not able to be modeled by a degree of truth only;
- Data used to optimize the DSS could be contaminated by noise, effect that is not able to be minimized by only a degree of truth.

Type-2 Fuzzy Sets (T2FS's) are an extension of T1FS's. The membership value (or degree of truth) is determined by a T1FS [18, 20, 28]. These sets allow better uncertainties models in DSS's based on experts' knowledge, minimizing their effect [14, 19, 20]. Though different opinions may produce different membership functions, using T2FS's a unique model can be designed anyway [14, 21].

In this chapter, we present a T2FS's theoretical analysis, as a T1FS's extension. We define and generalize the fuzzy conjunction, disjunction and complement operations between T2FS's membership values, using the *"Zadeh's extension principle"* [28]. We analyze the particular case of the Interval valued T2FS's

(IT2FS's). We also define the theoretical background of T2FS-based DSS and we define measures needed in comparing degrees of truth resulting after evaluation of T2FS-based FP's. Finally, we present a practical example and we analyze briefly the results.

2 Type-1 Fuzzy Sets

In this section, T1FS's are defined, including the operations to compute fuzzy relations, thus introducing concepts and notation that will be used forward in this chapter.

Definition #1: A T1FS A, over a universe X_A, is characterized by a *Membership Function* (MF) $\mu_A : X_A \rightarrow [0,1]$ and it can be represented as [10, 27]:

$$A = \int_{x \in X_A} \mu_A(x)/x, \tag{1}$$

where $\int$ represents the joint of the elements into the set A [28]. The MF μ_A fully defines FS A.

Definition #2: A *Classic Set* (CS) is a particular case of a T1FS whose MF μ_A is 1 for each element into the set [27, 28].

Dealing with Fuzzy Logic (FL), *conjunction*, *disjunction*, *complement* and *implication* operators are used to evaluate relations between Fuzzy Variables (FV's) into a compound FP. They are applied to the degree of truth of a FV (simple predicates) in order to compute the degree of truth of the compound FP's [16]. They are defined as follows [9]:

Definition #3: A function $D: [0,1]^2 \rightarrow [0,1]$ is a *fuzzy disjunction* if it is increasing on both arguments and fulfills the binary truth table of the disjunction:

$$\begin{aligned} D(1,1) = D(0,1) = D(1,0) = 1 \\ D(0,0) = 0 \end{aligned} \tag{2}$$

Definition #4: A function $C: [0,1]^2 \rightarrow [0,1]$ is a *fuzzy conjunction* if it is increasing on both arguments and fulfills the binary truth table of the conjunction:

$$\begin{aligned} C(0,0) = C(0,1) = C(1,0) = 0 \\ C(1,1) = 1 \end{aligned} \tag{3}$$

There exist non-commutative functions satisfying definitions #3 and #4. However, conjunctions and disjunctions should be symmetric to be used in the definition and evaluation of FP's. That is why *triangular norms* (*t-norms* and *t-conorms* or *s-norms*) are widely preferred, since they satisfy previous definitions and also additional restrictions, such as being symmetric [9]. *T-norms* generalize conjunction and *s-norms* generalize disjunction [5]. Table 1 and Table 2 summarize some known operators who meet the conditions to be *t-norms* and *s-norms*.

Table 1 Some known t-norms

T-Norm	Expression
Standard	$t(a,b) = \min(a,b)$
Algebraic product	$t(a,b) = ab$
Bounded product	$t(a,b) = \max(0, a+b-1)$
Drastic product	$t(a,b) = \begin{cases} a & si\ b = 1 \\ b & si\ a = 1 \\ 0 & otherwise \end{cases}$
Dubois and Prade	$t(a,b) = \frac{ab}{\max(a,b,\gamma)},\ 0 < \gamma < 1$
Hamacher's product	$t(a,b) = \frac{ab}{\gamma+(1-\gamma)(a+b-ab)},\ \gamma \geq 0$
Einstein's product	$t(a,b) = \frac{ab}{2-a+b-ab}$
Yager's product	$t(a,b) = 1 - \min\left(1, ((1-a)^\gamma + (1-b)^\gamma)^{\frac{1}{\gamma}}\right), \gamma > 0$

Table 2 Some known s-norms

S-Norm	Expression
Standard	$s(a,b) = \max(a,b)$
Algebraic sum	$s(a,b) = a + b - ab$
Bounded sum	$s(a,b) = \min(1, a+b)$
Drastic sum	$s(a,b) = \begin{cases} a & si\ b = 0 \\ b & si\ a = 0 \\ 1 & otherwise \end{cases}$
Dubois and Prade	$s(a,b) = 1 - \frac{(1-a)(1-b)}{\max(1-a, 1-b, \gamma)}, \quad 0 < \gamma < 1$
Hamacher's sum	$s(a,b) = 1 - \frac{a+b+(\gamma-2)ab}{1+(\gamma-1)ab},\ \gamma \geq 0$
Einstein's sum	$s(a,b) = \frac{a+b}{1+ab}$
Yager's sum	$s(a,b) = \min\left(1, (a^\gamma + b^\gamma)^{\frac{1}{\gamma}}\right), \gamma > 0$

Definition #5: A function $c: [0,1] \rightarrow [0,1]$ is a *fuzzy complement* if it satisfies, for any $a, b \in [0,1]$ [5]:

a. Boundary condition: $c(0) = 1$ and $c(1) = 0$;
b. Symmetry: If $a \leq b$ then $c(a) \geq c(b)$;
c. Involutional property: $c\big(c(a)\big) = a$.

Table 3 summarizes some known operators who meet the conditions to be fuzzy complements.

Definition #6: A function $I: [0,1]^2 \rightarrow [0,1]$ is a *fuzzy implication* if it is decreasing on its first argument, increasing on its second one and fulfills [9]:

$$\begin{gathered} I(0,0) = I(0,1) = I(1,1) = 1 \\ I(1,0) = 0 \end{gathered} \tag{4}$$

Fuzzy implication can be written based on combinations between fuzzy conjunctions, disjunctions and complements.

Table 3 Some known functions able to be fuzzy complements

Fuzzy complement	Expression
Standard	$c(a) = 1 - a$
Sugeno	$c(a) = \frac{1-a}{1+\lambda a},\ \lambda > -1$
Yager	$c(a) = (1 - a^{\omega})^{\frac{1}{\omega}},\ \omega > 0$

Often it is expected that the degree of truth obtained for a compound predicate after application of a fuzzy operator is sensitive to the degree of truth of all the simple predicates it involves, keeping its linguistic meaning. Compensatory Fuzzy Logic (CFL) is a ML model that resigns some FL conjunction and disjunction features in order to give operators being sensitive and idempotent, compensating the degrees of truth [1, 4, 6].

Two models satisfying CFL have been defined: Geometric Mean Based Compensatory Fuzzy Logic (GMBCFL) [1-4] and Arithmetic Mean Based Compensatory Fuzzy Logic (AMBCFL) [6]. Their operators are defined in Tables 4 and 5.

Table 4 Geometric Mean Based Compensatory Fuzzy Logic (GMBCFL) operators

Operator	Expression
Conjunction	$c(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_n)^{\frac{1}{n}}$
Disjunction	$D(x_1, x_2, \ldots, x_n) = 1 - [(1 - x_1)(1 - x_2) \ldots (1 - x_n)]^{\frac{1}{n}}$
Complement	$N(x_i) = 1 - x_i$

Previous definitions will be recalled on the following sections in order to define T2FS's, extend fuzzy operators to work with these sets and define a structure and operations on T2FS-based DSS's.

Table 5 Arithmetic Mean Based Compensatory Fuzzy Logic (AMBCFL) operators

Operator	Expression
Conjunction	$c(x_1, x_2, \ldots, x_n) = \left[\min(x_1, x_2, \ldots, x_n)\frac{1}{n}\sum_{i=1}^{n} x_i\right]^{\frac{1}{2}}$
Disjunction	$D(x_1, x_2, \ldots, x_n) = 1 - \left[\min(1 - x_1, 1 - x_2, \ldots, 1 - x_n)\frac{1}{n}\sum_{i=1}^{n}(1 - x_i)\right]^{\frac{1}{2}}$
Complement	$N(x_i) = 1 - x_i$

3 Type-2 Fuzzy Sets

In this section, we define T2FS's, we give their main features and we make a formal extension of the *conjunction*, *disjunction* and *complement* operators for this type of FS's.

Definition #7: Let be A a T1FS whose MF is $\mu_A(x)$: $X_A \rightarrow [0,1]$. A Footprint of Uncertainly (FOU) is:

$$FOU_{\mu A} = \bigcup_{x \in X_A} \left\{ \left[\underbrace{\mu_A(x) - \varepsilon_x}_{\varphi_x^-}, \underbrace{\mu_A(x) + \delta_x}_{\varphi_x^+} \right] : 0 \leq \mu_A(x) - \varepsilon_x \leq \mu_A(x) + \delta_x \leq 1 \right\} \tag{5}$$

$$FOU_{\mu A} = \bigcup_{x \in X_A} J_x = \{[\varphi_x^-, \varphi_x^+]\}$$

where $FOU_{\mu A}$ is defined as a uncertainty bounded-region over the MF associated to the T1FS A. It defines, for each $x \in X_A$, a set of membership values by means of a *primary membership function* J_x [7, 18].

The *Upper Membership Function* (UMF) is defined as:

$$\bar{\mu}(x) = \varphi_x^+ \quad \forall x \in X_A \tag{6}$$

The *Lower Membership Function* (LMF) is defined as:

$$\underline{\mu}(x) = \varphi_x^- \quad \forall x \in X_A \tag{7}$$

Definition #8: A T2FS, denoted as $\tilde{A}$, is defined as [7, 18, 20, 21]:

$$\tilde{A} = \int_{x \in X_{\tilde{A}}} \int_{u \grave{o} J_x \subseteq [0,1]} \mu_{\tilde{A}}(x,u) \Big/ (x,u) = \int_{x \in X_{\tilde{A}}} \left[\int_{u \grave{o} J_x \subseteq [0,1]} \mu_{\tilde{A}}(x,u)/u \right] \Big/ x, \tag{8}$$

where x is the primary variable with domain $X_{\tilde{A}}$; $u \in U$ is the secondary variable with domain J_x in each $x \in X_{\tilde{A}}$ and $\mu_{\tilde{A}}: X_{\tilde{A}} \times U \rightarrow [0,1]$ is the secondary membership value for $\tilde{A}$. A T2FS can model a FOU over the T1FS-MF and so assign a secondary membership value for each element. Figure 1 shows the FOU for a T2FS and his UMF and LMF.

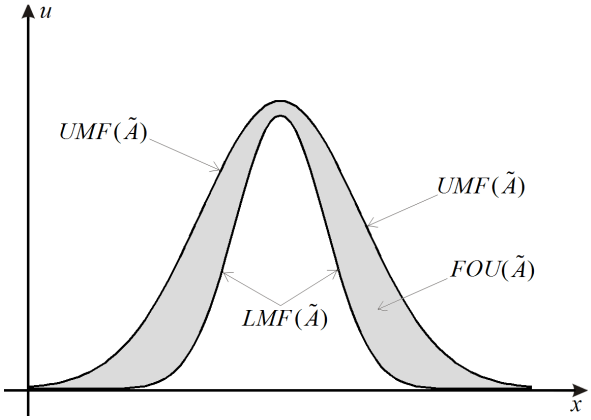


Fig. 1 FOU for a T2FS

Definition #9: An *Interval Type-2 Fuzzy Set* (IT2FS) is a particular case of a T2FS, where the membership degree is represented by a CS; namely $\mu_{\tilde{A}}(x,u) \equiv 1, \forall x \in X_{\tilde{A}}, u \in U$ [18]. IT2FS's are totally defined by their FOU and they allow modeling of uncertainties assigning an Interval of Membership Values (IMV) for each element [7, 17].

Definition #10: A *Vertical Slice* (VS) or *secondary membership function* of a T2FS $\tilde{A}$ for a fixed $x \in X_{\tilde{A}}$, denoted as $\tilde{A}_x$, is the T1FS formed by the primary membership grades and the secondary membership grades for the element x in $\tilde{A}$ [7, 18]:

$$\tilde{A}_x = \int_{u \partial J_x \subseteq [0,1]} \mu_{\tilde{A}}(x,u)/u \tag{9}$$

Figure 2 shows an example of T2FS-MF $\mu_{\tilde{A}}$, for discrete domains $X_{\tilde{A}}$ and U. For $x = 1$ the VS associated to $\tilde{A}$, noted $\tilde{A}_1$, is:

$$\tilde{A}_1 = 0.5/0 + 0.35/0.2 + 0.35/0.4 + 0.2/0.6 + 0.5/0.8 \tag{10}$$

A VS defines a set of membership values for each element of a T2FS, joining primary membership grades to secondary membership grades. Therefore, conjunctions, disjunctions, complements and implications that have been used to define

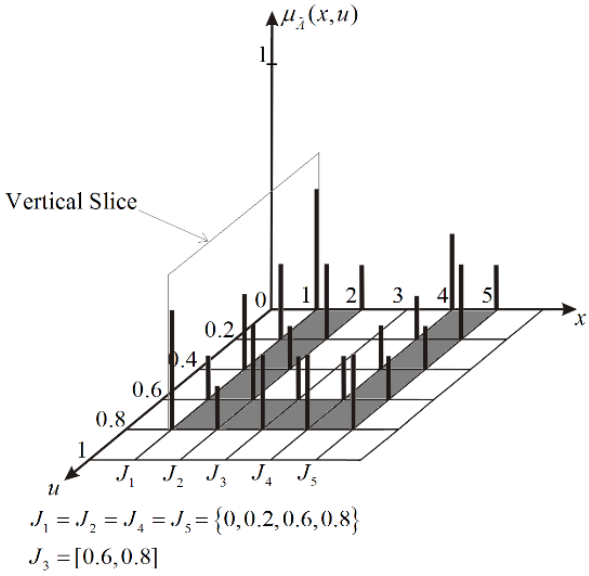


Fig. 2 Type-2 membership function for discrete domains $X_{\tilde{A}}$ and U

membership values in $[0,1]$ must be extended to be able to deal with VS. After this, degrees of truth of FP defined with T2FS – FV's could be computed [15, 23, 28].

Hereinafter in this section, *Zadeh's Extension Principle* [21, 28] will be applied, in order to extend the previous operation to VS's.

Let $\tilde{A}_x$, and $\tilde{B}_y$, be two VS's with MF's given by $\mu_{\tilde{A}_x}: U \to [0,1]$, and $\mu_{\tilde{B}_y}: V \to [0,1]$ and let $f: U \times V \to W$ be a function defined over the primary membership grades, i.e. the elements of the VS's. According to the *Zadeh's Extension Principle*, we can define a new T1FS A over W:

$$A = \int_{w \in W} C(\mu_{\tilde{A},x}(u), \mu_{\tilde{B},y}(v))/f(u,v), \tag{11}$$

where C is a fuzzy conjunction, $(u,v) \in U \times V$ and $w = f(u,v)$. Thereby, the operations defined for degree of truth into $[0,1]$ can be extended to evaluate FP's using T2FS's. Applying the equation (11) and using the operations previously defined, conjunction, disjunction and complement of VS's can be defined as:

Definition #11: Conjunction between VS's $\tilde{A}_x$ and $\tilde{B}_y$, denoted as $\tilde{A}_x \wedge \tilde{B}_y$, is defined by:

$$\tilde{A}_x \wedge \tilde{B}_y = \int_{C_2(u,v) \in W} C_1(\mu_{\tilde{A}_x}(u), \mu_{\tilde{B}_y}(v))/C_2(u,v), \tag{12}$$

where $C_2: U \times V \to W$ is a fuzzy conjunction between the primary membership grades, with $U, V, W \subseteq [0,1]$ and $C_1: [0,1]^2 \to [0,1]$ is a fuzzy conjunction between the secondary membership grades.

Definition #12: Disjunction between VS's $\tilde{A}_x$ and $\tilde{B}_y$, denoted as $\tilde{A}_x \vee \tilde{B}_y$, is defined by:

$$\tilde{A}_x \vee \tilde{B}_y = \int_{D(u,v)\in W} C(\mu_{\tilde{A}_x}(u), \mu_{\tilde{B}_y}(v))/D(u,v), \tag{13}$$

where $D: U \times V \to W$ is a fuzzy disjunction between the primary membership grades, with $U, V, W \subseteq [0,1]$ and $C: [0,1]^2 \to [0,1]$ is a fuzzy conjunction between the secondary membership grades.

Definition #13: Complement of a VS $\tilde{A}_x$, denoted as $\neg\tilde{A}_x$, is defined by:

$$\neg\tilde{A}_x = \int_{c(u)\in W} \mu_{\tilde{A}_x}(u)/c(u), \tag{14}$$

where $c: U \to W$ is the primary grade fuzzy complement, with $U, W \subseteq [0,1]$.

When T2FS's are interval-valued (IT2FS's), then VS's are defined by IMV's and therefore the operations between them are simplified: it is required to operate only between the bounds of the intervals [17, 18, 20, 21, 28].

4 T2FS-Based Decision Support Systems

In this section, we define and generalize FP's to be able for operation with FV's considering T2FS's. We also define how to deal with DSS's using them.

Definition #14: Let $\left\{\tilde{A}^{\lambda}_{x_\lambda}\right\}_{\lambda=1,\dots,n}$, with $x_\lambda \in X_{\tilde{A}^\lambda}$, be a family of VS's and let $\left\{\mu_{\tilde{A}^{\lambda}_{x_\lambda}}\right\}_{\lambda=1,\dots,n}$ be the family of MF's associated, with $\mu_{\tilde{A}^{\lambda}_{x_\lambda}}: U_{\tilde{A}^{\lambda}_{x_\lambda}} \to [0,1]$; a FP over X, denoted as $P: \left\{\tilde{A}^{\lambda}_{x_\lambda}\right\}_{\lambda=1,\dots,n}$, where X defines the set of elements $\{x_\lambda\}_{\lambda=1,\dots,n}, x_\lambda \in X_{\tilde{A}^\lambda}$ and B is a VS, can be generalized as:

$$P(X) = \sum_{i\in I} \sum_{\{\lambda_1,\dots,\lambda_i\}\in M_i} \varphi_{\{\lambda_1,\dots,\lambda_i\}}\left(\tilde{A}^{\lambda_1}_{x_{\lambda_1}},\dots,\tilde{A}^{\lambda_i}_{x_{\lambda_i}}\right), \tag{15}$$

where $I \subset \{1,\dots,n\}$; $M_i \subset \Gamma_{n,i}$, being $\Gamma_{n,i}$ the set of all different subsets, of size i, of set $\{1,\dots,n\}$ where order does not matter; $\varphi_{\{\lambda_1,\dots,\lambda_i\}}$ represents any logical operation over each combination of VS's $\{\lambda_1,\dots,\lambda_i\}$ and $\sum$ represents any logical operations combination between each previous VS's combinations.

We consider two FP-based DSS's [2, 16]:

a. Multiple options represented as predicates to make a decision. Output is determined by evaluating a set of FP's, taking the one whose degree of truth is a maximum;
b. A unique FP to be evaluated using a data set (cases). Output is a ranking of degrees of truth for each case.

As seen in Definition #14, the degree of truth of a FP, when it is defined over FV's modeled by T2FS's, has a VS as result, which is a T1FS. Consequently, we need a function defined over each VS to have a measure (a real number) to allow determining the output of the DSS.

IT2FS-MF given in Definition #9 assign an IMV to each element of the set; i.e. each VS in an interval $E = [a, b]$ with $a, b \in [0,1]$. This type of FS is enough to model the majority of uncertainties often presented in DSS's. They have been widely used in different Type-2 Fuzzy Systems applications [11, 13, 17, 19, 20, 26, 29].

In this work, we propose defining a *Measure over Interval of Membership values* (MIM) as follow:

Definition #15: Let $E = [a, b]$, with $a, b \in [0,1]$, be a IMV associated to a VS of an IT2FS; $\lambda(E) = g(a, b)$, with $g: [0,1]^2 \to [0,1]$, is a MIM if and only if it meets the following conditions $\forall a, b, c, d \in [0,1]$:

a. $g(a, b) \geq 0$;
b. $g(a, b) = 0$ if and only if $a = b = 0$;
c. $g(a, b) = 1$ if and only if $a = b = 1$;
d. Symmetry: $g(a, b) = g(b, a)$;
e. If $|b - a| = |d - c| \ \wedge \ a + b \leq c + d \ \Rightarrow \ g(a, b) \leq g(c, d)$.

Functions showed in Table 6, whose proof is trivial, are some of the functions that satisfy the conditions defined to be MIM.

Table 6 Functions satisfying the MIM definition

Function Number	Expression
1	$g(a, b) = \frac{a+b}{2} \sup(\{a, b\})$
2	$g(a, b) = \begin{cases} \alpha\lvert b - a\rvert + (1 - \alpha)\sup(\{a, b\}) & \text{if } (a, b) \neq (1,1), \alpha \in [0,1] \\ 1 & \text{otherwise} \end{cases}$
3	$g(a, b) = \begin{cases} \frac{a+b}{2}\left(1 - \sup(\{a, b\})\right) & \text{if } (a, b) \neq (1,1) \\ 1 & \text{otherwise} \end{cases}$
4	$g(a, b) = \begin{cases} 1 - 2\frac{\lvert b-a\rvert}{a+b} & \text{if } 2\lvert b - a\rvert > a + b \\ 0 & \text{otherwise} \end{cases}$
5	$g(a, b) = \begin{cases} \alpha(1 - \lvert b - a\rvert) + (1 - \alpha)\sup(\{a, b\}) & \text{if } (a, b) \neq (0,0), \alpha \in [0,1] \\ 0 & \text{otherwise} \end{cases}$

Under these considerations, we propose defining the structure of a DSS based on FP's defined over FV's associated to IT2FS's as follows:

Definition #16: A DSS based on FP's defined over FV's associated to IT2FS's can be defined for: a set of IT2FS's $\left\{\tilde{A}^{\lambda_i}\right\}_{\lambda_i=1,\ldots,n}$, a set of FP's $\{P_i\}_{i=1,\ldots,m}$, with $P_i: \left\{\tilde{A}^{\lambda_i}_{x_{\lambda_i}}\right\}_{\lambda_i=1,\ldots,n} \to B_i$, $\tilde{A}^{\lambda_i}_{x_{\lambda_i}} = \left[a_{\lambda_i}, b_{\lambda_i}\right]$ and $B_i = [a_i, b_i]$ with

$a_{\lambda_i}, b_{\lambda_i}, a_i, b_i \in [0,1]$, and a MIM $\lambda(B_i) = \mathrm{g}(a_i, b_i)$. The DSS output is given by (according to the DSS types):

a. The option corresponding to the FP $P_k(X)$ whose measure $\lambda(B_k)$ is a maximum of $\{\lambda(B_i)\}_{i=1,\ldots,m}$ (when multiple options are considered);
b. The ranking of measures $\{\lambda(B_k)\}_{k=1,\ldots,M}$ of the FP (when a unique FP is considered for a set of different cases $\{(X_k)\}_{k=1,\ldots,M}$).

Definition #16 may be extended to DSS's based on FV's associated to T2FS's by defining a measure on T1FS's.

5 Application Example

In this section we present the implementation of a DSS based on IT2FS's to define a ranking of the twenty nine Urban Areas (UA) of Argentina given its *Social Welfare Level* using the databases (DB's) of *Instituto Nacional de Estadísticas y Censos* (INDEC) and *Dirección Nacional de Política Criminal – Ministerio de Justicia y Derechos Humanos* (DNPC/MJDDHH) [8, 12]. IT2FS's and FP's were defined using expert's knowledge about available information for each of the variables in the DB's. The FP that defines the output of DSS to a UA X is:

$$P(X) = P_1(X) \wedge P_2(X), \tag{16}$$

where $P(X)$ is the FP "The area X has high social welfare level", $P_1(X)$ is "The area X has good social conditions" and $P_2(X)$ is "The area X has good housing conditions". FP's $P_1(X)$ and $P_2(X)$ are defined from others simple FP's directly related to the DB's variables as follows:

$$\begin{aligned} P_1(X) &= \left(P_{11}(X)^{0.5} \wedge \left(P_{12}(X) \wedge P_{13}(X)\right)^{0.5}\right) \wedge \left(P_{14}(X) \wedge P_{15}(X)\right) \wedge P_{16}(X) \\ P_2(X) &= P_{21}(X) \wedge \left(P_{22}(X) \wedge P_{23}(X) \wedge P_{24}(X) \wedge P_{25}(X) \wedge P_{26}(X)\right) \end{aligned} \tag{17}$$

where $P_{12}(X) \wedge P_{13}(X)$ is "The area X has high public safety", $P_{14}(X) \wedge P_{15}(X)$ "The area X has a high education level" and $P_{22}(X) \wedge P_{23}(X) \wedge P_{24}(X) \wedge P_{25}(X) \wedge P_{26}(X)$ defines "The area X has a proper home". Table 7 defines the meaning of each predicates used in (16). For $P_{11}(X)$ and $P_{12}(X) \wedge P_{13}(X)$ an hedge was applied in order to define $P_1(X)$, using the square root of the VS, resulting in new FP's that can be interpreted as: "The area X has a slightly higher level of health" and "The area X has a slightly higher public safety" respectively.

IMV was evaluated and scored for each UA included in the DB, determining the truth degree of FP *"The area has high social welfare level"*. In order to establish a ranking among the set of UA's, we used the MIM number 5 given in Table 6. The MIM can be analyzed as follows:

- If $\alpha = 1$, $\lambda(E)$ measures the uncertainty of the VS, given by the range of IMV, then the greater the uncertainty, the lower the value of $\lambda(E)$;

- If $\alpha = 0$, $\lambda(E)$ allows evaluating the degree of truth of the VS, then the higher the degrees of truth (closer to 1), the higher values of $\lambda(E)$;
- Values of α between 0 and 1 are allowed to take into account both the uncertainty of the IMV as their degrees of truth.

For this example, $\alpha = 0.5$ was set. Conjunctions between VS's were performed using the AMBCFL conjunction. Table 8 shows a list of the most representative UA's in the analyzed DB, which represent areas with different production structure [24]; the ranking number for each UA (depending on their *social welfare level*); the IMV of $P(X)$ and his MIM.

Table 7 Meaning of simple FP's defined from expert knowledge

Fuzzy Predicate	Meaning
$P_{11}(X)$	The area X has a high level of health
$P_{12}(X)$	The area X has a low amount of voluntary manslaughter
$P_{13}(X)$	The area X has a low amount of wrongful death
$P_{14}(X)$	The area X has a high level of education
$P_{15}(X)$	The area X has a high educational level
$P_{16}(X)$	The area X has a decent level
$P_{21}(X)$	The area X has a relatively high income level
$P_{22}(X)$	The area X has a low level of overcrowding
$P_{23}(X)$	The area X has a good quality of roof-floor
$P_{24}(X)$	The area X has good access to water service
$P_{25}(X)$	The area X has a good quality of bathing
$P_{26}(X)$	The area X has a tenure of housing right

The ranking obtained for the UA was compared to the *Índice de Desarrollo Humano* (IDH) computed for the argentine provinces [25], which analyzes the education levels, health and income in order to get the level of human development in each district. According to this comparison, we can establish the following:

- *Ushuaia - Rio Grande* (located in 1st place) belongs to a province (Tierra del Fuego) that historically was placed 2nd in the IDH ranking;
- *Gran La Plata* (located in 7th place) is the capital of the province of Buenos Aires, which has jurisdiction located in the top 10 provinces;
- *Gran Catamarca* (located in 16th place), which is in the province of Catamarca and is located between positions 11th and 16th in the IDH ranking;
- *Gran Resistencia* (located in 29th place) belongs to the province of Chaco, which historically is among the last 3 places of the IDH ranking.

Therefore, the ranking of UA's obtained in the developed DSS is consistent with the order established by the IDH.

Table 8 Obtained ranking for most representative UA's

Ranking Number	Urban Area	$P(X) = E = [a, b]$	$\lambda(E)$
1	Ushuaia - Río Grande	[0.6663 , 0.8311]	0.8332
7	Gran La Plata	[0.5135 , 0.7175]	0.7567
16	Gran Catamarca	[0.3889 , 0.5637]	0.6944
29	Gran Resistencia	[0.2938 , 0.4489]	0.6469

6 Conclusions

In this chapter we analyzed T2FS's as an extension of T1FS's, which allows modeling uncertainties, mainly in the DSS's construction, both in system input data as on the expert's knowledge used. We covered various types of uncertainties that may be found in the DSS. The set of operators of the FL and CFL, which help assess relationships established by FP's, was extended and generalized for use T2FS's. We fully defined the theoretical structure of the DSS's based on T2FS's, providing some measures to analyze the degree of truth of the IMV's, which allowed determining the output of a DSS when IT2FS's are used. We presented an application example of a DSS using IT2FS's. The paradigm tackled in this chapter provide a comprehensive framework for the use of T2FS's in DSS's, that enables exploiting the enormous advantages of T2FS's in modeling of expert's knowledge.

Acknowledgments. Authors wish to thank Eugenio Actis, Bachelor of Science in Economics, for their helpful assistance on the example presented in this chapter, both in the development of the fuzzy system and the results analysis.

References

1. Andrade, R.E., Gómez, J.M., Téllez, G.M., González, E.F.: Compensatory logic: A fuzzy approach to decision making. In: Proceedings of 4th International Symposium on Engineering of Intelligent Systems (EIS 2004), Madeira, Portugal (2004)
2. Andrade, R.E., González, E.F.: La Lógica Difusa Compensatoria: Una Plataforma para el Razonamiento y la Representación del Conocimiento en un Ambiente de Decisión Multicriterio, Análisis Multicriterio para la Toma de Decisiones: Métodos y Aplicaciones (2009)
3. Andrade, R.E., Téllez, G.M., González, E.F., Marx-Gómez, J., Lecich, M.I.: Compensatory logic: A fuzzy normative model for decision making. Investigación Operativa 27, 188–197 (2006)
4. Andrade, R.E., Téllez, G.M.: Consideraciones sobre el carácter normativo de la lógica difusa compensatoria. Infraestructuras de Datos Espaciales en Iberoamérica y el Caribe, Cuba (2007)
5. Batyrshin, I.Z., Kaynak, O., Rudas, I.: Generalizad conjuntion and disjunction operations for fuzzy control. In: EUFIT 1998, Aachen, Germany (1998)

6. Bouchet, A., Pastore, J., Andrade, R.E., Brun, M., Ballarin, V.: Arithmetic Mean Based Compensatory Fuzzy Logic. International Journal of Computational Intelligence and Applications 10, 231–243 (2011)
7. Comas, D.S., Meschino, G.J., Pastore, J.I., Ballarin, V.L.: A survey of medical images and signal processing problems successfully solved with the application of Type-2 Fuzzy Logic. Journal of Physics: Conference Series (2011)
8. DNPC/MJDDHH: Dirección Nacional de Política Criminal. Ministerio de Justicia y Derechos Humanos, DNPC/MJDDHH (2011), http://www.jus.gob.ar/ (accessed March 26, 2012)
9. Drewniak, J., Król, A.: A survey of weak connectives and the preservation of their properties by aggregations. Fuzzy Sets and Systems, 202–215 (2010)
10. Dubois, H., Prade, D.: Fuzzy Sets and Systems: Theory and Applications. Academic Press Inc., New York (1980)
11. Garibaldi, J.M., Ozen, T.: Uncertain Fuzzy Reasoning: A Case Study in Modelling Expert Decision Making. IEEE Transactions on Fuzzy Systems 15, 16–30 (2007)
12. INDEC: Base de Microdatos - Encuesta Permanente de Hogares. EPH-INDEC (2011), http://www.indec.mecon.ar/ (accessed March 26, 2012)
13. John, R.I., Innocent, P.R., Barnes, M.R.: Type 2 fuzzy sets and neuro-fuzzy clustering of radiographic tibia images. In: Proceedings of the Sixth IEEE International Conference on Computational Intelligence, Anchorage, AK, USA (1998)
14. Karnik, N.N., Mendel, J.M., Liang, Q.: Type-2 Fuzzy Logic Systems. IEEE Transaction on Fuzzy Systems 7, 643–658 (1999)
15. Karnik, N.N., Mendel, J.M.: Operations on Type-2 Fuzzy Sets. Fuzzy Sets and Systems 122, 327–348 (2001)
16. Li, H.X., Yen, V.C.: Fuzzy Sets and Fuzzy Decision-Making. N.W. Boca Raton (1995)
17. Liang, Q., Mendel, J.M.: Interval Type-2 Fuzzy Logic Systems: Theory and Design. IEEE Transaction on Fuzzy Systems 8, 535–550 (2000)
18. Mendel, J., John, R.I.B.: Type-2 Fuzzy Sets Made Simple. IEEE Transactions on Fuzzy Systems 10, 117–127 (2002)
19. Mendel, J.: Fuzzy sets for words: a new beginning. In: 12th IEEE International Conference on Fuzzy Systems, Saint Louis, MO (2003)
20. Mendel, J.M.: Type-2 fuzzy sets and systems: an overview. IEEE Computational Intelligence Magazine 2, 20–29 (2007)
21. Mendel, J.M.: Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions. Prentice-Hall, Upper-Saddle River (2001)
22. Mendel, J.M.: Uncertainty, fuzzy logic, and signal processing. Signal Processing, 913–933 (2000)
23. Mizumoto, M., Tanaka, K.: Some Properties of Fuzzy Sets of Type 2. Information and Control 31, 312–340 (1976)
24. PNUD: Aportes para el Desarrollo Humano 2002. PNUD, Buenos Aires (2002)
25. PNUD: Aportes para el Desarrollo Humano en Argentina 2009. PNUD, Buenos Aires (2009)
26. Wagner, C., Hagras, H.: Uncertainty and Type-2 Fuzzy Sets and Systems. In: UK Workshop on Computational Intelligence (UKCI), Colchester, UK (2010)
27. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
28. Zadeh, L.A.: The Concept of a Linguistic Variable and its Application to Approximate Reasoning-I. Information Sciences, 199–249 (1975)
29. Zarandi, M.H., Zarinbal, M., Izadi, M.: Systematic image processing for diagnosing brain tumors: A Type-II fuzzy expert system approach. Applied Soft Computing 11, 285–294 (2011)
30. Zeng, J., Liu, Z.-Q.: Type-2 fuzzy hidden Markov models and their application to speech recognition. IEEE Transaction on Fuzzy Systems 14, 454–467 (2006)

Fuzzy Predictive and Reactive Scheduling

Jürgen Sauer and Tay Jin Chua

Abstract. Fuzzy techniques are widely used to model and use imprecise and incomplete knowledge. This paper shows how Fuzzy techniques can be used to solve scheduling problems in which not only imprecise information is present but also preferences of users have to be regarded. The first part deals with the predictive scheduling in chemical industry where preferences of human schedulers have to be regarded. The second part presents an approach for a reactive scheduling system in which Fuzzy techniques are used to select between alternatives.

1 Introduction

Scheduling is the task of finding an assignment of activities to resources regarding several constraints like due dates, time windows, capacities and availabilities of resources and so on. Predictive scheduling creates schedules in advance assuming a stable environment. This is seldom the case in reality, therefore reactive scheduling copes with the repair of schedules that have become invalid due to events occurring in the scheduling environment. Predictive scheduling problems are often solved using optimizing approaches, in which important aspects are neglected, e.g. the preferences of users or the priorities of clients. This paper investigates the possibilities of using Fuzzy techniques to model such pieces of knowledge and to use them in finding the best solutions for the scheduling problem at hand. It is also important to separate the scheduling problems in predictive and reactive ones because the goals differ typically.

In a first part we will focus on predictive scheduling and use a typical case from the chemical industry to show where and how preferences are useful to find the best suited schedule. The case bases on the observation that in chemical industry

Jürgen Sauer
Carl von Ossietzky University of Oldenburg, Germany.
e-mail: sauer@wi-ol.de
Tay Jin Chua
Singapore Institute of Manufacturing Technology, Singapore.
e-mail: tjchua@simtech.a-star.edu.sg

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_19,

but also in other production areas like precision industry [1] we often find several alternative ways of producing one item. This is because the equipment is doubled or because there are multi-purpose machines that can be used for different production steps within the production process. Therefore a set of alternative production descriptions may be used to produce the same product, these are called "production variants". On the other hand single steps within these production processes may be performed on different machines, which are called "alternative machines". This leads to a combinatorial amount of alternative ways of producing one item. But most often not all of these possible processes are seen equally. Therefore preferences exist for some of them, mostly given by experience. The standard scheduling approaches used to schedule production do not regard these preferences. Thus we need a technique to represent and use the (often hidden) knowledge about preferences in production processes.

Additionally, the reactive scheduling tasks have to be regarded as well. In reactive scheduling it is necessary to react as fast as possible to the events that have occurred at the shop floor. But not only speed is important, other goals are [3]:

- Keep as much as possible of the existing schedule in order to avoid "shop nervousness"
- Find schedules of good quality.

The main task of reactive scheduling is to rearrange the schedule to the new needs. In the process of rearrangement operations or orders have to be taken from the schedule and scheduled again at another place. Here it is crucial to find the "important operations", which should not be moved in order to keep a good schedule or to find less important ones that can be moved. On the other hand it is basic to find the "most important" events, i.e. those that should be tackled first. Again it is necessary to find a technique to represent and use the knowledge about what an important order or event is.

In both cases Fuzzy based techniques can be used to represent and process the knowledge that can be used to solve the problem. Several approaches have already been presented most of them use some kind of imprecise data in the scheduling process ([15], see also section 3). A basic feature of Fuzzy based techniques is the use of linguistic variables, which are used to express natural language concepts like "good, better, best" to describe aspects of objects and to use them by Fuzzy rules to find solutions for the problems to be solved.

This paper presents two approaches using Fuzzy techniques to solve scheduling problems. The first shows how to represent and use preferences in a predictive scheduling case and the second one shows how to use Fuzzy techniques to find important or less important orders in a reactive scheduling case.

The rest of the paper is organized as follows. First predictive and reactive scheduling as well as Fuzzy techniques and their usage are described in more detail (sections 2 and 3). Then in sections 4 and 5 the Fuzzy scheduling approach for using preferences in predictive scheduling is presented together with some preliminary results. Sections 6 and 7 describe the Fuzzy reactive scheduling approach and some evaluation results.

2 Predictive and Reactive Scheduling

Scheduling problems can be found in different application domains, this ranges from production and all other processes within a supply chain like transportation or storage [19] to service oriented processes e.g. in hospitals or schools [14]. If we look at production scheduling, the task of scheduling is the temporal assignment of resources to activities (production steps) taking into account several constraints. Constraints can be divided in Hard Constraints and Soft Constraints. Hard Constraints have to be regarded in order to get a valid schedule. Soft Constraints may be relaxed still leading to a valid schedule but typically with less quality. Additionally, schedules can be compared and evaluated using specific evaluation functions. And, finally, events from the environment have to be regarded because they often affect the existing schedule. Thus scheduling problems can be described using a 7-tuple (R, P, O, HC, SC, G, E) [19] with

`R is the set of resources`
every resource has only a limited capacity.

`P is the set of products`
With P all the possible production processes for P are described. Therefore production variants and alternative machines/ resources to be used within the production steps are given. A variant defines one possible production process for product P and the process consists of several steps (operations) to be performed in given sequence. Every operation can use one of the alternative resources and needs a given duration.

`O is the set of orders`
For every order the amount needed and the time window in which it shall be fulfilled is given. Also a priority value may be added.

`HC is the set of Hard Constraints {XE "Hard Constraint"}`
Hard Constraints have to be regarded in order to have a valid schedule. Most of them are technical constraints, e.g.

- Resources to be used
- Sequences of operations to be obeyed

`SC is the set of Soft Constraints {XE "Soft Constraint"}`
Soft Constraints should be obeyed in order to get a "good" schedule. They may be relaxed.

`G is the set of goal functions`
Goal functions are used to evaluate schedules. Most of them are due date based and do not regard aspects like preferences or priorities. Often a set of functions may be appropriate. The selection of goal functions is therefore a difficult task.

`E is the set of Events`
Events are caused by changes in the scheduling environment, e.g. the breakdown of a machine. As a consequence the schedule becomes invalid or inconsistent. Here is the starting point for reactive scheduling.

Predictive scheduling assumes a static environment and leads to schedules that contain a temporal assignment of resources to activities which is typically shown in a Gantt chart. The schedule is evaluated using several functions and often an optimal solution regarding one of these functions is wanted. In real world scenarios the environment is not static, quite the contrary, many events occur which may have a positive or negative effect on the schedule. The task of adapting the schedule to the permanently changing environment is called reactive scheduling. Main goals are here the maintenance of stability, i.e. most of the schedule should remain as before but the new schedule shall also be as good as before.

The problem space of such a scheduling problem can be represented with an AND/OR tree [18]. The tree consists of two types of nodes connected by directed edges. The solution of a problem represented with the tree is composed by the rules:

- With an AND node all direct successors are in the solution
- With an OR node only one of the direct successors is in the solution.

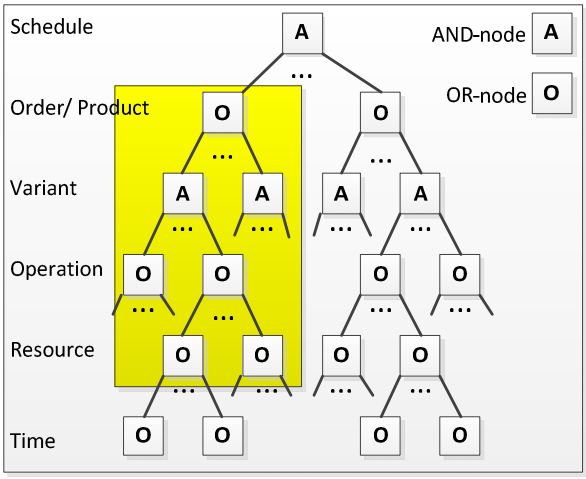


Fig. 1 AND-OR tree representing a scheduling problem

The AND/OR tree not only shows the complexity of the problem area, it can also be used to show the production description of a product with its production variants, the operations within the variants and the alternative resources for every operation (this is the marked area in figure 1). Every variant defines one possible production process for the product. Table 1 presents some sample data for a production description of product “5” which can be produced in two ways (variants 0 or 1) each with seven steps (operations) and several alternative resources.

Table 1 Database Table with Production Description of Product 5x

Product	Variant	Operation	Resources	Duration
5	0	1	[10 12 15 4]	1
5	0	2	[18 19 21 8 9]	1
5	0	3	[15 12 10]	1
5	0	4	[16 5 6]	1
5	0	5	[13 2 3]	1
5	0	6	[17 7 16]	1
5	0	7	[14 2 3]	1
5	1	1	[39 56 36]	1
5	1	2	[41 42 53 54]	1
5	1	3	[37 47 36]	1
5	1	4	[48 38 28]	1
5	1	5	[35 44 45 46]	1
5	1	6	[49 48 27]	1
5	1	7	[44 45 46]	1

Some of the constraints are already represented within the AND/OR tree, e.g. the root node (AND) defines, that all orders have to be scheduled. The problem size of a problem represented by an AND/OR tree can be estimated using the scheme presented in [17]. With the data from table 1 this leads to about $2*3^7$ (2 variants with 7 operations, each with 3 alternative resources) production alternatives for product "5". But typically not all of them are equal in the sense of delivering the same production quality. So the best production processes have to be found and marked for preferential use. These preferences for using specific variants or resources are derived from experiences like

- This is the process we have used all the time
- The workers are used to this process
- The connections between the resources are best
- Pre- and post-effort is minimal.

Table 2 gives an excerpt of a schedule that could have been created using the data from table 1 and an order for producing product "5" eighteen times between

Table 2 Part of a Schedule

Product	Variant	Operation	Resource	Start	End
5	1	1	39	312	330
5	1	2	54	330	348
5	1	3	36	348	366
5	1	4	38	366	384
5	1	5	35	384	402
5	1	6	49	402	420
5	1	7	46	420	438

310 and 440. The schedule shows the temporal assignment (Start, End) of resources (Resource) to the activities given by the production description (Product, Variant, Operation). The schedule is valid if all Hard Constraints are fulfilled, it is called consistent if all constraints are regarded.

3 Fuzzy Techniques

Since the beginning of the 1990s Fuzzy techniques are used and investigated for solving scheduling problems. Fuzzy logic and its application is well suited for modeling and using dynamic and imprecise knowledge especially in combination with rule based approaches and heuristics [4, 9, 10, 12]. Imprecise statements like "A is better than B but not so good as C" are quite usually, even in the scheduling activities of human schedulers. Expressions like "small, medium, large" are called "linguistic variables" and are represented by Fuzzy sets describing a degree of similarity between the object and a prototype [4]. The rule based combination of such expressions can be defined by Fuzzy rules. The Fuzzy set defines a gradual membership of an element in a set using a membership function between 0,1 which can be interpreted as the probability with which a given value matches the statement. Fuzzy rules are used to combine the imprecise expressions in order to find solutions. Thus in our case Fuzzy systems are specific rule based systems.

Figure 2 shows the membership functions for a set called "machinecount" with linguistic variables "less, few, many". With this function it is expressed how many alternative resources there are available for a specific operation. These expressions are then used together with others to find out which are the most or less important alternatives to be used in the schedule, e.g. in a rule like "IF machinecount is many THEN importance is low".

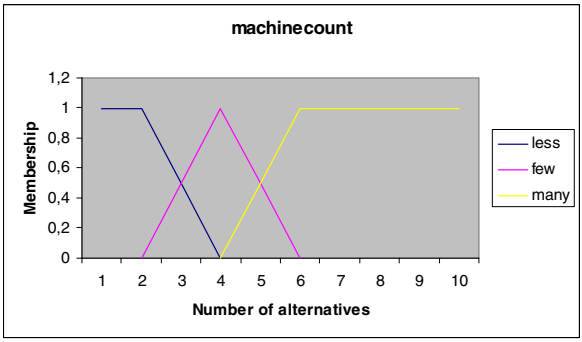


Fig. 2 Membership Function "machinecount"

Figure 3 shows how Fuzzy control works. Three main steps have to be performed:

1. Input values are mapped to the Fuzzy sets (Fuzzification).
2. Appropriate Fuzzy rules (using the input parameters) are selected and applied. This is performed by the interpreter which uses rules and Fuzzy Sets from the Knowledge Base.

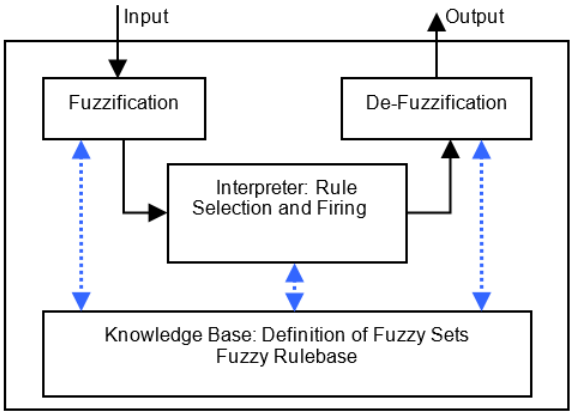


Fig. 3 Principle Architecture of a Fuzzy-Controller

3. The resulting Fuzzy values are transformed to crisp values which are the output of the controller.

Several approaches have been presented in which Fuzzy techniques are used to solve scheduling problems. The approaches typically focus on specific aspects of the scheduling problem at hand, e.g. [2, 5, 7, 8, 13, 15, 21]

- Imprecise definition of due dates
- Imprecise definition of durations
- Fuzzy evaluation criteria
- Precedence relationships
- Cumulative data for global scheduling.

In the approach presented here we will concentrate on preferences (of users) and on reactive scheduling.

4 Fuzzy Predictive Scheduling Regarding Preferences

As argued in section 2 all the possible production alternatives may be used but only few of them are really desirable and therefore preferred by the personnel. The knowledge about these combinations is often from experience and therefore not described precisely. To create a scheduling approach that regards those preferences it is necessary to make the knowledge explicit and provide a structure to use it. Fuzzy logic provides this and will therefore be used to design and implement an approach for predictive scheduling regarding users preferences. Fuzzy control is embedded in a basic heuristic which also shows how these two approaches can be combined and extended.

When creating a Fuzzy system the main challenges are:

- Find the aspects of the problem to be expressed in linguistic terms. It is crucial to find the variables that can help to detect what is important, critical, to be preferred etc.

- Find the rules that combine the input expressions to those expressions that lead to the "best" solution, e.g. that detect the best alternative to be used.
- Integrate the fuzzy control in a strategy to come to a complete system.

Additionally, to evaluate the approach it is necessary to select evaluation functions that will really rate the desired goals.

Algorithm 1: general order-based scheduling algorithm

```
WHILE orders to schedule DO
  select order (FIFO);
  select variant (FIFO);
  WHILE operations to schedule DO
    select operation (FIFO);
    select resource (FIFO);
    select interval (First Fit);
    schedule operation;
  END WHILE
END WHILE
```

The general idea of the approach is that it combines an order based scheduling strategy with fuzzy logic to select the preferred orders, variants and resources [16]. The basic strategy is adopted from [19] and is shown in Algorithm 1. The algorithm aims to find a good solution for a scheduling problem represented by an AND/OR tree as shown before. The performance and quality of the algorithm depends on the selection of the "right" objects. In Algorithm 1 some standard selection rules are in brackets that can be used as a reference (FIFO: first in first out, using first of a given list; First Fit: use the first value that fits).

Algorithm 2: Order based scheduling algorithm using Fuzzy controllers

```
    Calculate the sequence of orders to be scheduled
    (Fuzzy-Control 1)
WHILE orders to schedule DO
  select order (use first of sequence);
  WHILE variants to schedule DO
    select variant (Fuzzy-Control 2);
    WHILE operations to schedule DO
      select operation (FIFO);
      select resource (Fuzzy-Control 3);
      select interval (First Fit);
      schedule operation
    END WHILE
  END WHILE
END WHILE
```

To include the Fuzzy control it is necessary to decide how and where preferences are useful. If we look at the problem structure the selection of orders, variants and resources (even intervals) are options. The strategy we want to realize is:

- if possible use preferred variants
- if possible use preferred resources
- Additionally, to find a schedule that meets the due dates: try to schedule difficult orders first, easy ones later. Here difficult means that we have only little alternatives and a small time window for scheduling.

This leads to Algorithm 2 in which three Fuzzy controls for the selection of preferred variants, resources and for selecting difficult orders are integrated. This can be implemented e.g. with different sets of Fuzzy sets and rules used by one controller. Next step is to find the variables that can express preferences and difficulty of schedules and the rules to be used to find the desired results, i.e. difficult orders, preferred variants and preferred resources.

For Fuzzy-Control 1 variables are used that express the number of alternative variants and resources and the remaining time buffer for the placement of the operations (slack):

- `flextime{low, middle, high}`:
  rates the time buffer
- `variantcount {less, few, many}`:
  rates the number of production variants
- `machinecount {less, few, many}`:
  rates the number of alternative resources (see figure 2)
- `difficulty {low,lowmiddle,middle,middlehigh,high}`:
  is the result value and rates the difficulty of the order

Rules that are used to find the most difficult order (which then will be scheduled first) are:

```
IF variantecount IS less AND machinecount is less AND
flextime is low THEN difficulty IS high;
IF variantecount IS less AND machinecount is less AND
flextime is middle THEN difficulty IS middlehigh;
IF variantecount IS less AND machinecount is few AND
flextime is low THEN difficulty IS high;
```

For the Fuzzy-Controls 2 and 3 it is important to find a rating of the preferences. Figure 4 shows a simple membership function for “preference”. It uses penalties that are given for not using the preferred alternatives. The production

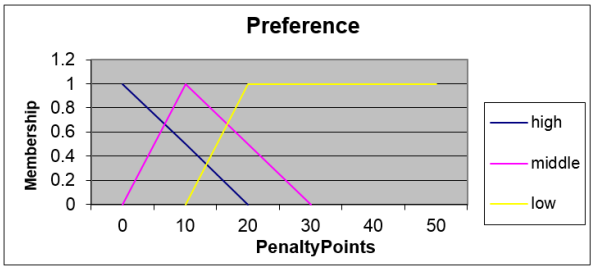


Fig. 4 Membership Function for Preferences

description is therefore extended with the penalty values. This means: the most preferred variant has a penalty value of 0, the second preferred has 10, the third 20, etc. These penalties are then fuzzified in the "preference" Fuzzy set. The same is done with the alternative resources. The penalties can also be used in an evaluation function which will be shown later on.

In the second Fuzzy control the Fuzzy sets "preference" and "importance" are used to rate and find the best production variant.

- `preference {low, middlehigh, high}`:
  rates the given preference of a variant
- `importance {low,lowmiddle,middle,middlehigh,high}`:
  is the result value and rates the importance of the variant, the most important will be selected

A sample rule used here is:

```
IF preference IS high THEN importance IS high;
```

The third Fuzzy control is quite similar to the second one. It uses penalty values as inputs but also information about the use and demand of the resources in order to find the most preferred resource that is also available and not so demanded. This helps to find schedules that also will meet the due dates. Fuzzy sets are

- `isBusy: {low, high}`
  rates the occupation of the resource
- `occurrencerate: {low, middle, high}`
  rates the demand of the resource
- `preference: {low, middle, high}`
  rates the preference using the penalty points
- `importance: {low,lowmiddle,middle,middlehigh,high}`
  is the resulting rating for the resource to be chosen.

Examples of rules used are:

```
IF isBusy IS low AND preference is high THEN importance
IS high;
IF isBusy IS high AND preference is low THEN importance
IS middle;
```

5 Evaluating Preferences

The next question is how to evaluate schedules with a focus on preferences. The classical evaluation functions are mainly time focused and use values like tardiness or lateness. These are not useful if we look at preferences. Therefore we have to define a new function basing on the penalties for not using the preferred variants or resources. We simply add the penalties of the scheduled variants and resources leading to an overall penalty for the orders resp. the whole schedule. The lower the penalty points the better is the schedule:

In the production description the variants and alternative resources are written in the order of their preference (VariantNr, MachineNr), so we can use this to calculate the variant and resource penalty:

```
VariantPenalty = VariantNr * 10,
MachinePenalty = MachineNr * 10
```

The penalty of the schedule is the sum of all order penalties. The order penalty is the sum of its variant and resource (machine) penalties.

$$PenaltySum = \sum_{i=1}^{n} \left(VariantPenalty_i + \sum_{j=1}^{m} MachinePenalty_{ij} \right)$$

n = number of orders
m = number of operations per order

After implementing the approach in Java using the Fuzzy-Package JFuzzyLogic (http://sourceforge.net/projects/jfuzzylogic) [21] it was evaluated using some sample data from chemical industry. To compare the results some other schedules were calculated using heuristics as shown in algorithm 1 and meta-heuristics (SA: simulated annealing, TA: threshold acceptance, GD: grand deluge algorithm, TS: Tabu search) implemented in another project [22].

Table 3 Results of the Tests

10 Orders	Lateness	Penalties	**54 Orders**	Lateness	Penalties
FIFO	327	14370	FIFO	8065	126170
SA	290	12840	SA	6961	111190
TA	290	13810	TA	7035	111430
GD	290	14110	GD	7387	113860
TS	290	12150	TS	6811	118790
FuzzyLT	298	14490	FuzzyLT	6918	114370
FuzzyPrio	306	**13100**	FuzzyPrio	12201	**95860**

Table 3 shows the results of the tests using different sets of orders (from 10 to 54) out of 54 possible products (more results are in [16]). The functions used for the comparison “Sum of lateness“ and “penalties“. Additionally, the Fuzzy based system used two sets of rules. One is focusing on providing schedules with good results regarding “lateness” (FuzzyLT) and one is focusing on “preferences” (FuzzyPrio), as shown above. For the meta-heuristics the best results out of 15 runs are shown. Runtime was measured only for the biggest set of orders. The heuristic algorithm needs about 1 second, the Fuzzy system about 10 seconds and the meta-heuristics about 400 seconds.

The evaluation shows that the Fuzzy based approach (FuzzyPrio) outperforms the other ones in the area it is designed for but not in the other cases. If we focus on time based goals like lateness without looking at preferences, then the other algorithms are better. But we can see that with another rule set (FuzzyLT) it is possible to get also good results for the time based goal function.

6 Fuzzy Reactive Scheduling

When looking at the dynamics of the scheduling environment, it becomes clear that an approach is needed to cope with the changing environment in order to adapt the schedules to these changes. This means that we have to develop a system that can react to the events of the scheduling environment. When reacting to the events the three main goals of reactive scheduling have to be regarded: reaction time, stability and quality of the schedule. Such an approach was presented in [20] and will be used as a basis for a Fuzzy based system for reactive scheduling.

The first important step in reactive scheduling is to find the events that are important for scheduling or rescheduling decisions. Examples of such events are listed in Table 4 (see also [20]). The events may cause invalid or inconsistent schedules, e.g. because resources are no longer available or additional orders have been placed.

Next step is to find how to react to the events. In [20] a general algorithm for event handling has been presented:

Table 4 Possible Events

Event	Description	Event	Description
E1	new order	E9	change of variant
E2	order canceled	E10	change of resource intensity
E3	start time of order changed	E11	new resource
E4	due date of order changed	E12	resource deleted
E5	change of amount	E13	shift changed
E6	change of priority	E14	outsourcing
E7	breakdown, maintenance	E15	order splitted
E8	change of product data		

Algorithm 3: General reactive scheduling algorithm

```
WHILE events queued
    select event; propagate effects of event
    IF constraints violated
    THEN  select constraint violation
          select appropriate reaction
    END IF
    delete actual event
END WHILE
```

The following section shows how a Fuzzy based approach can be used to implement the tasks of event selection and handling (line 2 – 7). The general idea is to group the events regarding the necessary actions. This is supported by the following observations: Some events need immediate reaction and therefore have to be treated directly. Some do not need any rescheduling activity but may have positive consequences for reactions. Some events cause follow up events. Some have the consequence of inserting new orders into the schedule thereby replacing other ones. For all it is important to find an order in which they should be handled.

Table 5 Groups of Events

Group	Description of effect	Events
K1	Event leads to space in schedule	(E11), (E14)
K2	Event leads to removal of operation/order, free space in schedule, and creates new events	(E2), (E7), (E8), (E12)
K3	Event leads to changes in data without direct consequences	(E6)
K4	Event leads to removal of operation/order or only to changes in data without a direct effect	(E3), (E4), (E5), (E9), (E10), (E13)
K5	Event is a “new” order and leads to a number of reactions	(E1)

Table 5 gives an overview on the five categories we use to group the events and the events that belong to these groups. The event groups shall be treated in the sequence they are listed, because K1 leads to space in the schedule which may be positive for other actions, K2 leads to space in the schedule but creates subsequent events, K3 has no direct effect, K4 leads to space in the schedule but may be treated after K2, in K5 all the new or withdrawn orders that have to be inserted newly or again.

This leads to the general strategy of the Fuzzy based approach:

Algorithm 4: Fuzzy reactive scheduling algorithm

```
WHILE events queued
    IF event in K1-K4
    THEN propagate effects of event
    END IF
```

```
      IF event in K5
      THEN  select most important order
            select optimal place for rescheduling
      END IF
  END WHILE
```

The main area where to use the Fuzzy approach is the selection of orders to be newly scheduled or to be rescheduled and the orders that may be replaced during the rescheduling process. Thus we have to model what are the aspects of the important orders and what are the aspects of orders that may be replaced during rescheduling, i.e. we need Fuzzy sets to rate the importance of orders and Fuzzy sets to rate if an order is replaceable, both together with the appropriate Fuzzy rules.

The following Fuzzy sets can be used for both tasks:

- `Delay (very high, high, no)`:
  rates the actual delay of the order
- `Priority (high, normal, low)`:
  rates the priority of the order
- `Urgency (high, normal, low)`:
  rates the urgency of the order (due date)
- `Buffer (high, normal, low)`:
  rates the planning buffer that remains
- `Variants (high, normal, low)`:
  rates the number of variants
- `Complexity (low, normal, high)`:
  rates number of operations
- `AlternativeResources (high, normal, low)`:
  rates number of alternative resources that are available
- `AlreadyReplaced (no, low, normal, often)`:
  rates if the order has been replaced already

and as resulting Fuzzy sets:

- `Importance (very high, high, normal, low, very low)`:
  rates the importance of the order to be scheduled
- `Replaceable (very good, good, normal, bad, worse)`:
  rates how easy it is to replace an order

The rule set that is used to find the most important order to be scheduled contains the following rules:

```
IF Delay is very high AND Priority is high THEN Impor-
tance is very high
IF Delay is high AND Priority is high THEN Importance
is very high
IF Delay is no AND Priority is low AND Urgency is high
THEN Importance is normal
```

```
IF Delay is no AND Priority is high AND Urgency is low
THEN Importance is low
```

The rule set to be used to find a replaceable order tries to identify orders that can be replaced easily, because they were not replaced before, have enough buffer, have enough alternative to be scheduled, etc. Examples are:

```
IF Buffer is high AND Variants is high AND Complexity
is low AND AlternativeResources is high AND
AlreadyReplaced is no THEN Replaceable is very good

IF Buffer is high AND Variants is low AND Complexity is
high AND AlternativeResources is normal AND
AlreadyReplaced is often THEN Replaceable is bad
```

7 Evaluating Reactive Scheduling

In order to evaluate a reactive scheduling strategy it is necessary to find some criteria that can rate the main goals of reactive scheduling: time, stability and quality. To compare time runtime of the algorithms can be used, to rate the quality the classical evaluation function like tardiness or lateness may be used, and to rate stability the simplest way is to count the changes in the adapted schedule.

Another task is to find a test scenario that represents as many as possible cases of the reactive scheduling problem, thus some kind of benchmark would be advantageous. In a first step we started with a simple scenario which contains some important events to be handled [11]. The test case contains a starting set of 10 orders that have to be scheduled or a starting schedule and a sequence of events together with their description and time stamps for their occurrence.

Table 6 shows some first results of the tests. The Fuzzy reactive approach is compared with a heuristic that simply creates a complete new schedule and thus cannot react to most of the events. The table shows that the Fuzzy system can react to the events and that the quality is better than that of the predictive strategy.

Table 6 Some Results from Reactive Rescheduling

Time	Event	Heuristic Lateness	Fuzzy approach Lateness
0	Starting schedule	298	298
250	Order cancelled	-	196
267	Machine maintenance	-	196
270	Machine breakdown	-	191
275	3 New orders	429	396

Some more tests are necessary in order to evaluate the complete functionality of reactive approaches, a set of benchmarks could be appropriate.

8 Conclusion

It has been shown in several publications that Fuzzy techniques can be used efficiently to cope with uncertain and vague data that are often present in scheduling problems. In this paper two approaches using fuzzy techniques for scheduling have been presented. The first is for a predictive scheduling case in which Fuzzy sets and rules are used to model users preferences for production processes and alternative resources. Linguistic variables and rules utilizing them have been presented. These preferences are then used in an algorithm for the creation of schedules. The reactive approach uses Fuzzy techniques to rate orders that can be replaced during the rescheduling process. New goal functions have been presented that can be used to evaluate the performance of the approaches. Both approaches show that Fuzzy techniques are an appropriate methodology if e.g. preferences for alternative resources or ratings for orders to be reschedules have to be found. The evaluation also shows that it is necessary to find the right set of variables and rules for a specific case and that these rule sets are only efficient for the specific case they have been created for.

Acknowledgments. Part of the work on reactive scheduling was done during a visiting researcher fellowship at SIMTech (Singapore Institute for Manufactoring Technology). Prototypical systems have been implemented by the students Jun Huang, Ling Zou, and Angelo Maron.

References

1. Chua, T.J., et al.: Integrated Production Planning, Scheduling and Shop Floor Tracking system for High-Mix Low-Volume Manufacturing - A Consortium Approach. In: Key Engineering Materials, vol. 447-448, pp. 306–310. Transtech Publications (2010)
2. Dorn, J., Kerr, R.M., Thalhammer: Reactive Scheduling – improving the robustness of schedules and restricting the effects of shop floor disturbances by fuzzy reasoning. International Journal on Human-Computer Studies 42, 687–704 (1995)
3. Dorn, J.: Evaluating reactive scheduling systems. In: Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 2004). IEEE, New York (2004)
4. Dubois, D., Prade, H.: The three semantics for fuzzy sets. Fuzzy Sets and Systems 90(2), 141–150 (1997)
5. Dubois, D., Fargier, H., Fortemps, P.: Fuzzy Scheduling: Modelling flexible constraints vs. coping with incomplete knowledge. European Journal of Operations Research 147, 231–252 (2003)
6. Huang, J.: Fuzzy Scheduling Algorithmen als WebService. Diploma Thesis, Department fürInformatik, Universität Oldenburg (2007)
7. Kerr, R.M., Walker, R.N.: A Job Shop Scheduling System Based on Fuzzy Arithmetic. In: 2nd International Conference on Expert Systems and Leading Edge in Production and Operations Management, pp. 433–450 (1989)
8. Kerr, R.M.: Research Issues and Challenges in Fuzzy Scheduling. In: Workshop on Fuzzy Scheduling Systems, University of Linz, Department of Mathematics, Austria (1993)

9. Kruse, R., Gebhardt, J., Klawonn, F.: Foundations of Fuzzy Systems. Wiley, New York (1994)
10. Kruse, R.: Fuzzy-Systeme - Positive Aspekte der Unvollkommenheit. Informatik Spektrum 19(1), 4–11 (1996)
11. Maron, A.: Fuzzy reactive Scheduling. Diploma Thesis, Department fürInformatik, Universität Oldenburg (2008)
12. Nauck, D., Klawonn, F., Kruse, R.: Foundations of Neuro-Fuzzy Systems. Wiley, Chichester (1997)
13. Petrovic, S., Geiger, M.J.: A Fuzzy Scheduling System with Dynamic Job Priorities and an Extension to Multiple Criteria. In: Meredith, R., Shanks, G., Arnott, D. (eds.) Proceedings of IFIP International Conference on Decision Support Systems, Prato, Italy, pp. 637–646 (2004)
14. Pinedo, M.: Planning and Scheduling in Manufacturing and Services. Springer (2005)
15. Sauer, J., Appelrath, H.-J., Suelmann, G.: Multi-site Scheduling with Fuzzy-Concepts. International Journal of Approximate Reasoning in Scheduling 19, 145–160 (1997)
16. Sauer, J., Huang, J., Zou, L.: Nutzung von Präferenzenbei der Planungvariantenreicher Produktion. In: Mönch, L., Pankratz, G. (eds.) Intelligente System ezur Entscheidungsfindung, Teilkonferenz der Multi Konferenz Wirtschafts Informatik 2008, SCS Publishing House, Sydney (2008)
17. Sauer, J.: Knowledge-Based Systems in Scheduling. In: Leondes, C.T. (ed.) Knowledge-Based Systems: Techniques and Applications, vol. 4, pp. S. 1293–S. 1325. Academic Press, San Diego (2000)
18. Sauer, J.: Meta-Scheduling using Dynamic Scheduling Knowledge. In: Dorn, J., Froeschl, K.A. (eds.) Scheduling of Production Processes, Ellis Horwood. Ellis Horwood Series in Artificial Intelligence, pp. S. 151–S. 162 (1993) ISBN 0-13-075136-7
19. Sauer, J.: Modeling and solving multi-site scheduling problems. In: Wezel, W.V., Jorna, R., Meystel, A. (eds.) Planning in Intelligent Systems: Aspects, Motivations and Methods, pp. 281–299. Wiley, New York (2006)
20. Sauer, J.: Vertical Data Integration for Reactive Scheduling. KünstlicheIntelligenz 24(2), 123–129 (2010)
21. Slany, W.: Scheduling as a fuzzy multiple criteria optimization problem. Fuzzy Sets and Systems 78(2), 197–222 (1996) (special issue on fuzzy multiple criteria decision making)
22. Zou, L.: WebbasierteMetaheuristikenfürlokaleAblaufplanungsprobleme. Diploma Thesis, Department fürInformatik, Universität Oldenburg (2007)

A Computational Evaluation of Two Ranking Procedures for Valued Outranking Relations

Juan Carlos Leyva López and Mario Araoz Medina

Abstract. In this paper are examined the ranking methods Net Flow Rule (NFR) and a method based on multiobjective evolutionary algorithms (MOEA), which allow exploiting a known valued outranking relation, and are studied some kind of ranking irregularities these procedures suffer and analyzed the reasons of the phenomenon. Furthermore, the two ranking methods are evaluated in terms of one ranking test, which they fail. Hypothetical examples are described to demonstrate the occurrence of ranking irregularities. Next a computational study was implemented and discussed. The results of examinations show that i) the rates of ranking irregularities were rather significant when the ranking obtained with the two ranking procedures were compared with the simulated "correct" ranking and the associated valued outranking relation studied in this research, ii) the rates of the ranking irregularities were more considerable for the NFR rather than the MOEA procedure.

1 Introduction

Multicriteria Decision Analysis (MCDA) is widely used for selecting or ranking alternatives when they are assessed with multiple criteria [Figueira et al. 2005]. In outranking methods, we can distinguish two phases: aggregation and exploitation. Let A be the set of decision alternatives or potential actions and let us consider a valued outranking relation S_A^σ defined on AxA; this means that we associate with each ordered pair $(a,b) \in AxA$ a real number $\sigma(a,b)$ $(0 \leq \sigma(a,b) \leq 1)$, reflecting the degree of strength of the arguments favoring the crisp outranking aSb. The exploitation phase transforms the global information included in S_A^σ into a global ranking of the elements of A. In this phase, the classic method PROMETHEE II

Juan Carlos Leyva López · Mario Araoz Medina
Universidad de Occidente. Blvd. Lola Beltrán y Blvd. Rolando Arjona s/n, Culiacán, Sinaloa, México.
juan.leyva@udo.mx, mario.araoz@culiacan.gob.mx

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*, Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_20,

[1], use the ranking method Net Flow Rule (NFR) to obtain a score function. It is well known that a score function could lead to a final ordering, giving a pair-wise rank reversal effect [6].

In [5] a ranking method based on multiobjective evolutionary algorithms (MOEA) for exploiting a known valued outranking relation is introduced, with the purpose of constructing a recommendation for ranking problems within the outranking approach. The problem of obtaining the final ranking is modeled with multiobjective combinatorial optimization; the solution method rests on the main idea of reducing differences between the global model of preferences and the final ranking.

When we are working with ranking methods, different methods may yield different answers when they are fed with exactly the same valued outranking relation. Thus, in the outranking approach, the issue of evaluating the relative performance of different ranking methods is naturally raised. This, in turn, raises the question of how can one evaluate the performance of different ranking methods? Some kind of testing procedures have been developed. One such procedure is to examine the stability of a ranking method's mathematical process by checking the validity of its proposed rankings.

This present paper follows this line of research to investigate the ranking performance of the MOEA procedure, as opposed to another widely used ranking method, namely the NFR. The investigation is based on an extensive simulation experiment. Thus, the aim of this paper is to examine the ranking procedure NFR and the MOEA procedure, to study the type of ranking irregularities these procedures suffer and analyze the reasons of the phenomenon. We present an empirical study that focused on how often these ranking irregularities may happen under the NFR and the MOEA procedure, all of these under a test criterion to evaluate the performance of ranking procedures by testing the validity of their ranking results.

This paper is organized as follows: The second section discusses the test criterion that is used to test the performance of the NFR and the MOEA procedures. The third section describes one example of a hypothetical valued outranking relation for which ranking irregularities occurred by using the NFR and MOEA procedures. The fourth section describes an empirical study that focused on how often these ranking irregularities happen by using the NFR and the MOEA procedures under the test criterion. The last section presents concluding comments.

2 A Test Criterion for Evaluating Ranking Procedures

In the following a test criterion is established to evaluate the performance of ranking procedures by testing the validity of their ranking results.

Test Criterion:

An effective ranking method should not change the indication of the preference order between pairs of alternatives in a valued outranking

relation without cycles, incomparabilities, and indifferences between alternatives (i. e., the ranking of the alternatives obtained when the ranking method exploits a valued outranking relation should be consistent with the aggregation model of preferences represented by the valued outranking relation).

Suppose that E is an effective ranking procedure and has ranked the set of alternatives A in the complete rankingR_A. Then, according to the test criterion, the indication of the preference order between $(a, b) \in AxA$ in the valued outranking relation should not change in the ranking generated by the ranking procedure. Therefore, the following condition is fulfilled:

$$aP^{\lambda,\beta} \quad \text{if only if} \quad aRb$$

Where $aP^{\lambda,\beta}b$ is an asymmetric preference relation favoring a, we assume the existence of a threshold $\beta > 0$ such that if $aS_A^{\lambda}b$ and also $\sigma(b, a) \leq (\lambda - \beta)$, and $R_A$defines a weak order R on A.

The test criterion is consistent with the Principle of Correspondence [2].

Principle of correspondence. Every particular model should be included as a limit condition of another more general. The weak preference relation of the normative approach could be considered as a particular case of valued outranking relations. If λ approaches 1 and β is small, and if also S_A^{λ} is transitive and complete, making a decision using valued outranking relations should lead to the same result provided by classical decision theory.

If S_A^{λ} is a weak order on A, there is a value function $\boldsymbol{u}$ agreeing with the same preferences expressed by S_A^{λ} (cf. [4]).

3 Illustration of Ranking Irregularities with the Net Flow Rule and MOEA Procedures

In order to illustrate ranking irregularities with NFR and MOEA procedures, one test problem is shown. In the MOEA computational study, 1 trial/problem of the MOEA heuristic was generated for solving the following test problem. The test problem was finished when 10,000 populations had been generated. The population size was set to 40. The crossover probability was chosen 0.85 and the mutation probability was 0.30.

3.1 Test Problem

The valued outranking relation given by the credibility matrix (6x6) between actions $A_1, A_2, A_3, A_4, A_5, A_6$ is shown in Table 1. It was processed with the NFR and the MOEA procedures. The results obtained when we use the net flow score for ranking the alternatives are shown in Table 2.

These values suggest the final ranking: $A_3 \geq A_4 \geq A_6 \geq A_1 \geq A_5 \geq A_2$, where $A_i \geq A_j$ means A_i is preferred to A_j. It is difficult to support.

Table 1 Credibility matrix between actions $A_1, \ldots, A_6$

	A_1	A_2	A_3	A_4	A_5	A_6
A_1	1.00	0.63	0.10	0.54	0.88	0.77
A_2	0.53	1.00	0.11	0.83	0.37	0.08
A_3	0.89	0.87	1.00	0.65	0.84	0.65
A_4	0.64	1.00	0.77	1.00	0.80	0.90
A_5	0.16	1.00	0.41	0.35	1.00	0.26
A_6	0.87	0.76	0.75	1.00	0.79	1.00

Table 2 Results of the net flow score for ranking alternatives

Action	Positive outranking flow	Negative outranking flow	Net flow
A_1	2.92	3.09	-0.17
A_2	1.92	4.26	-2.34
A_3	3.90	2.14	+1.76
A_4	4.11	3.37	+0.74
A_5	2.18	3.68	-1.50
A_6	4.17	2.66	+1.51

There are two strong arguments for choosing A_6 as the best action. First, consider the subset A^{UND} composed by the unfuzzy nondominated alternatives (cf. [7]):

$$A^{UND} = \{A_i \in A: \sigma(A_i, A_j) \geq \sigma(A_j, A_i)\ \forall A_j \in A\}$$

In this case A^{UND} is not empty; $A_6 \in A^{UND}$ and the other alternatives do not belong to it. Moreover, $\sigma(A_6, A_j) \geq \sigma(A_j, A_6) + 0.10\ \forall A_j$. Second, if we consider the reduced set $A' = A - \{A_6\}$, then $A'^{UND} = \{A_4\}$. This is clearly the best alternative in this subset. But note also that in the pairwise comparison between A_6 and A_3 and between A_4 and A_3, if we take λ=0.75 and β =0.1 as consensus and threshold levels respectively, we can accept reasonably that "A_6 outranks A_3 but A_3 does not outrank A_6", and "A_4 outranks A_3 but A_3 does not outrank A_4", giving a presumed preference in favor of A_6 and A_4. In this ranking, generated with λ= 0.75, we see a clear violation of what is suggested by the global model of preferences. ObviouslyA_1, A_5, A_2, although irrelevant alternatives, play an important role in the relative ranking of A_6 and A_3.

Similar situation is observed when the ranking is generated by the MOEA procedure. Table 1 was processed with the MOEA procedure. The Restricted Pareto set $P_{known}^{restricted}$ returned by the MOEA at termination is presented in Table 3.

Table 3 Restricted Pareto set found in the solution space. The numbers in the cells represent the sub indices of the alternatives

6	4	3	4	3	3	4	4	3	4	6	4	6	6	4	4	4	6	6	4	6	4	4	3	3	4	4	6
3	6	6	6	4	4	6	3	6	3	3	6	4	4	3	6	6	4	4	6	4	6	6	6	6	6	6	4
4	3	6	3	1	1	3	6	4	6	4	3	3	3	6	3	3	3	3	3	3	3	3	4	6	3	3	3
1	1	1	1	6	6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Basing on the procedure for obtaining a recommendation from the MOEA procedure [5], the proposed ranking of the alternatives is given as follows $A_4 \geq A_6 \geq A_3 \geq A_1 \geq A_5 \geq A_2$, with $\lambda = 0.641$. The kind of analysis given above is valid also to the MOEA procedure. Note that in the pairwise comparison between A_6 and A_4, if we take λ =1.00 and β=0.1 as consensus and threshold levels respectively, we can accept reasonably that "A_6 outranks A_4 but A_4 does not outrank A_6", giving a presumed preference in favor of A_6 .

3.2 Analysis of the Ranking Irregularities with the NFR and MOEA Procedures

The main reason for the above ranking irregularities lies in the independence of irrelevant alternatives property. That is, the NFR does not fulfill this property. The above ranking irregularities examples reveal that there is not an a priori ranking of the alternatives when they are ranked by the NFR, because the ranking of an individual alternative derived by this method depends on the performance of all the other alternatives currently under consideration. This causes the ranking of the alternatives to depend on each other. Thus, it is likely that the best alternative may be different and the ranking of the alternatives may be distorted to some extent.

In test problem 1, A_3 and A_6 are rank reversed because A_1, A_5 and A_2 results irrelevant alternatives. This kind of irregular situation is undesirable for a practical decision-making problem though it is reasonable in terms of the logic of the net flow rule.

4 An Empirical Study

This section describes an empirical study that focused on establishing how often these ranking irregularities may happen for the NFR and the MOEA procedures. Some computer programs were written in Visual.Net languages in order to generate simulated valued outranking relations. The valued outranking relations were generated based on a previously simulated ranking of a set of alternatives with the indication that the ranking were consistent with the valued outranking relation (the pseudo code is presented in appendix A). The performance of the NFR and the MOEA procedures was tested under the test criterion described in Section 2.

4.1 Experimental Design

4.1.1 The Factors

The comparison of the MOEA procedure to the NFR method is performed through an extensive simulation. The simulation approach provides a framework to conduct the comparison under several data conditions and derive useful conclusions on the relative performance of the considered methods given the features and properties of the data. The term *performance* refers solely to the ranking accuracy of the methods.

The experiment presented in this paper is only concerned with the investigation of the ranking accuracy of ranking methods on experimental data. In particular, the conducted experimental study investigates the performance of the methods on the basis of 2 factors. Table 4 presents the Factors and the levels considered for each one in the simulation experiment.

Table 4 Factors investigated in the experimental design

Factors	Levels
F1: Ranking procedures	1.- NFR; 2.- MOEA procedure
F2: Size of the multicriteria ranking problems	1.- 6; 2.- 8; 3.- 10; 4.- 12; 5.- 18

4.1.2 Data Generation Procedure

An important aspect of the experimental comparison is the generation of the data having the required properties defined by the factors described in the previous subsection. In this study we proposed a methodology for the generation of the data. The general outline of this methodology is presented in Appendix A. The outcome of this methodology is a matrix and a vector consisting of a valued outranking relation and the associated ranking of alternatives which is consistent with the valued outranking relation in terms of the test criterion of section 2.

This experiment was repeated 5,000 times for each value of factor F2 (5 levels). Overall, 25,000 reference sets (Valued Outranking Relation, Ranking) were considered. Each reference set was used to calculate a ranking through the methods specified by factor F1 (cf. Table 4). Each generated ranking was compared to the corresponding ranking of the reference set, to test its performance.

4.2 Obtained Results

We had five cases of different numbers of alternatives. For each such case we run 5,000 random replications and each problem was solved by using the two ranking methods. The sample size of 5,000 was large enough to ensure statistically significant results. Some of the computational results are presented as Table 5.

For each test problem, the two rankings derived by using the NFR and the MOEA approach were analyzed in three different ways. The first way was to see whether the two rankings agreed with the “correct ranking” in the indication of the best two and best three alternative. The percentage of times the two approaches (NFR and MOEA) yielded a different indication of the best two and three alternatives is denoted as “Rate 1”, “Rate 2”, “Rate 3” and “Rate 4” respectively. For instance, when the NFR method was used and the number of alternatives was equal to 6, "Rate 1" was equal to 0.51. This means that 51% of the test problems with 6 alternatives resulted in a different indication of the best two alternatives when the ranking derived with NFR was compared to the “correct” ranking.

The second way (denoted in Table 5 as “Rate 5” and “Rate 6”) was to record the number of times the two rankings generated by NFR and MOEA were different from the “correct” ranking, without considering the magnitude of the individual differences. That is, when two rankings are evaluated, this rate would be equal to 0 if the two rankings are identical or equal to 1, otherwise (i.e., it is binary valued). Similarly as above, for test problems with 8 alternatives this rate is equal to 88% when the NFR method is used.

Table 5 Ranking Irregularity Rates of NFR and MOEA procedures to differ-ent alternatives under a Test Criterion

Alternatives	Rate 1 NFR 2 best alternatives	Rate 2 MOEA 2 best alternatives	Rate 3 NFR 3 best alternatives	Rate 4 MOEA 3 best alternatives	Rate 5 NFR	Rate 6 MOEA
6	0.51	0.21	0.67	0.30	0.82	0.54
8	0.58	0.24	0.73	0.34	0.88	0.67
10	0.62	0.28	0.78	0.41	0.89	0.70
12	0.66	0.35	0.81	0.50	0.90	0.73
18	0.72	0.64	0.87	0.80	0.99	0.90

The third way is a little more complex. It considers a weighted measure for expressing ranking discrepancies. Generally, we are considering rankings of a set of alternatives in some order, normally from the most important to the least important. For a 5-tuple of alternatives denoted by A_1, A_2, A_3, A_4, A_5, one possible ranking would be (note that for simplicity we use only the indexes): (1, 2, 3, 4, 5), that is, alternative A_1 is the most preferred alternative, alternative A_2 is the next most preferred one, etc.

Suppose that $(i_1, i_2, \ldots, i_m)$ and $(j_1, j_2, \ldots, j_m)$ are two possible rankings of m alternatives. That is, a ranking is a permutation of integers between 1 and m. Then, one of the problems examined in this subsection is how one can express the conflict or difference between these two rankings. Assume for the purpose of analysis that (1, 2, 3, 4, 5) is the "reference" ranking. This reference ranking represent the “correct” ranking Ro in the simulated ranking of the set of alternatives with the indication that ranking were consistent with the simulated valued outranking rela-

tion, i.e., from this reference ranking we construct a valued outranking relation using the process pointed in Appendix A.

Suppose that the simulated valued outranking relation have been exploited with the NFR and the MOEA procedures resulting in the complete rankings R1 and R2, respectively. If one wishes to evaluate how different the two obtained rankings are with respect to Ro, then the derived rankings could be measured with a proximity measure. In this work we used the weighted L1 norm (see Table 6).

In order to evaluate a ranking it is necessary to convert the differences in the rankings into numerical data. The method used of valuating the ranking conflict evaluates a ranking by the weighted sum of the absolute values of the differences of the assigned ranks given by a ranking method completing the ranking, where "rank" is the position given to a specific alternative within a given ranking. We call it "Rate 7 and "Rate 8". According to this measure one may wish to assign more significance to discrepancies on top rankings and less significance to discrepancies on lower rankings. Although one may use any kind of weights to achieve this goal, we used the weights (m, m-1, m-2, …, 2, 1). For instance, let us consider the "correct" ranking A= (3, 4, 5, 2, 1) and the ranking derived with NFR B=(2, 3, 5, 4, 1). Concerning ranking A, alternative 3, 4, 5, 2, 1, is in the rank 1, 2, 3, 4, 5, respectively.

The vector of ranks associated to the correct ranking is Rank A= (1, 2, 3, 4, 5). Proceeding in the same way, the vector of ranks associated to ranking B is Rank B= (4, 1, 3, 2, 5). Then, the weighted sum of the absolute values of the differences (weighted L1 norm) of the assigned ranks is given by:

d(Rank A, Rank B)= 5|1-4| + 4|2-1| + 3|3-3| + 2|4-2| + 1|5-5| = 23.

Table 6 Comparison of the NFR and MOEA procedure to different alternatives in terms of the weighted L1 norm (WL1N) under a Test Criterion

Ranking method	Mean distance WL1N	Total sum Distance WL1N	Number of times one method is better than the other	Number of times the two rankings have equal distance
Rate7:NFR (6 alternatives)	10.52	52,588	734	775
Rate8: MOEA(6 alternatives)	3.95	19,735	3491	
Rate7:NFR (8 alternatives)	22.93	114,664	693	229
Rate8:MOEA(8 alternatives)	8.84	44,180	4078	
Rate7:NFR (10 alternatives)	41.80	209,012	674	63
Rate8:MOEA(10 alternatives)	17.56	87,782	4263	
Rate7:NFR (12 alternatives)	67.90	339,503	738	36
Rate8:MOEA(12 alternatives)	35.71	178,574	4226	
Rate7: NFR(18 alternatives)	196.37	981,832	2310	27
Rate8: MOEA (18alternatives)	190.08	950,404	2663	

4.3 Analysis of the Empirical Study

The results obtained from the simulation experiment involve the ranking error rates of the methods in the reference sets. The analysis that follows is focused on the ranking performance of the methods. The error rates obtained using the reference sets provide an estimation of the general performance of the methods, measuring the ability of the methods to provide correct recommendations on the ranking of alternatives.

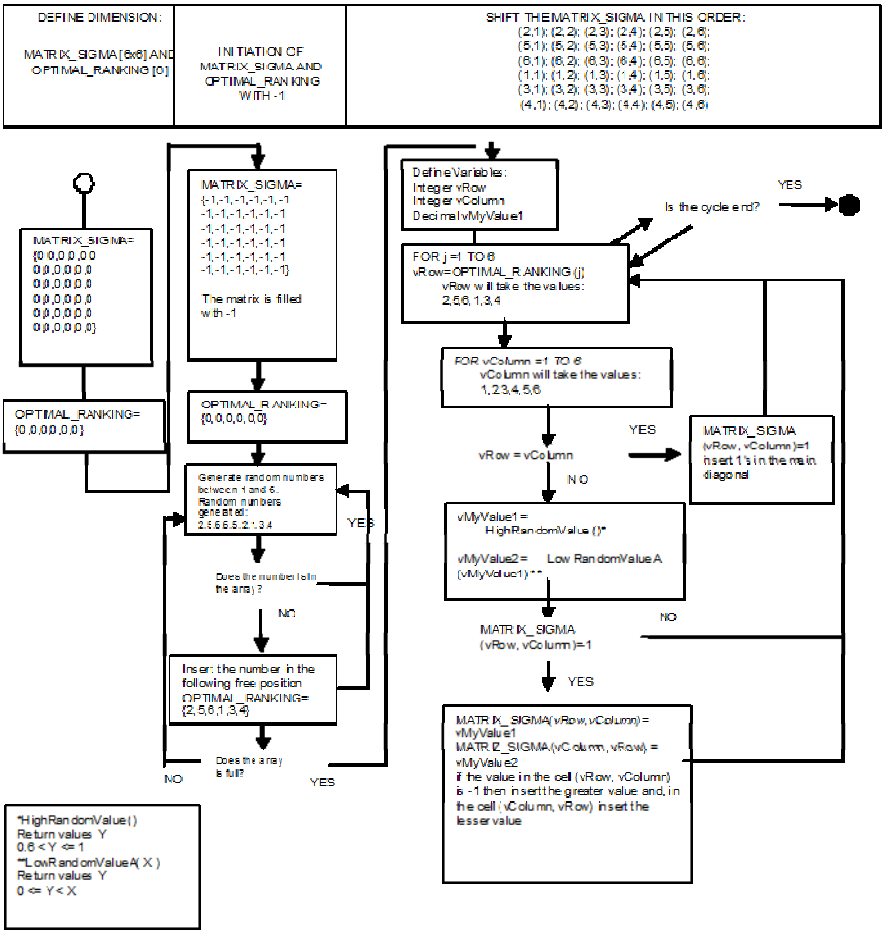
For i) the percentage of times the two approaches (NFR and MOEA) yielded a different indication of the best two and three alternatives, and ii) the number of times the two rankings derived from NFR and MOEA were different from the "correct" ranking, the MOEA procedure provides better results, i.e., the MOEA procedure provides significantly lower error rates. Also in the case of expressing ranking discrepancies, with respect to the number of times one method is better than the other, the MOEA procedure provides considerably better results than NFR (see Table 6).

The interaction which is found significant in this experiment for the explanation of the differences in the performance of the methods, involves the size of the reference set. The results presented in table 5 show that the increase of the size of the reference set (number of alternatives) reduces the performance of both methods. This is an expected result, since in this experiment larger reference sets increase complexity of the ranking problem. NFR method is more sensitive to the size of the reference set. Nevertheless, it should be noted that for each chosen reference set size, the considered MOEA procedure always perform better than the NFR method.

5 Concluding Remarks

The Net Flow Rule (NFR) and the Multiobjective Evolutionary Algorithms (MOEAs) approaches for exploiting a valued outranking relation may result in a different ranking of all alternatives when they are used on the same ranking problem. An extensive empirical analysis of this methodological problem revealed that this phenomenon might occur frequently on simulated ranking problems. The NFR and the MOEA performed in different manner in the tests. The error rates were significant and became more dramatic when the number of alternatives was increasing. Although it may be possible to know which ranking is the "correct" one, this study also proved that both ranking methods are not immune to ranking inconsistencies. The rates of the ranking irregularities were more considerable for the NFR rather than the MOEA procedure.

Appendix A. Generation Process of a Ranking and the Associated Valued Outranking Relation: Schematic Representation of a 6x6-Matrix



References

1. Brans, J.P., Vincke, P., Mareschal, B.: How to select and how to rank projects: The PROMETHEE method. European Journal of Operational Research 24, 228–238 (1986)
2. Fernandez Gonzalez, E., Cancela, N., Olmedo, R.: Deriving a final ranking from fuzzy preferences: An approach compatible with the Principle of Correspondence. Mathematical and Computer Modelling 47, 218–234 (2008)
3. Figueira, J., Greco, S., Ehrgott, M. (eds.): Multiple Criteria Decision Analysis: State of the Art Surveys. Springer (2005)

4. French, S.: Decision Theory: an introduction to the mathematics of Rationality. Halsted Press, NY (1986)
5. Leyva-Lopez, J.C., Aguilera-Contreras, M.A.: A Multiobjective Evolutionary Algorithm for Deriving Final Ranking from a Fuzzy Outranking Relation. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 235–249. Springer, Heidelberg (2005)
6. Mareschal, B., De Smet, Y., Nemery, P.: Rank reversal in the PROMETHEE II method: some new results. In: Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, Singapore, pp. 959–963 (2008) ISBN: 978-1-4244-2629-4
7. Orlovski, S.A.: Decision-Making with a Fuzzy Preference Relation. Fuzzy Sets and Systems 1, 155–167 (1978)

Selection of Evolutionary Multicriteria Strategies: Application in Designing a Regional Water Restoration Management Plan

Angel Udías, Andrés Redchuk, Javier Cano, and Lorenzo Galbiati

Abstract. Sustainability of water resources has become a challenging problem worldwide, as the pollution levels of natural water resources (particularly of rivers) have increased drastically in the last decades. Nowadays, there are many Waste Water Treatment Plant (WWTP) technologies that provide different levels of efficiency in the removal of water pollutants, leading to a great number of combinations of different measures (PoM) or strategies. The management problem, then, involves finding which of these combinations are efficient, regarding the desired objectives (cost and quality). Therefore, decisions affecting water resources require the application of multi-objective optimization techniques which will lead to a set of tradeoff solutions, none of which is better or worse than the others, but, rather, the final decision must be one particular PoM including representative features of the whole set of solutions. Besides, there is not a universally accepted

Angel Udías
European Commission, Joint Research Centre. Institute for Environment and Sustainability. Italia
e-mail:angelluis.udias@gmail.com

Andrés Redchuk
Facultad de Ciencias Empresariales. Universidad Autónoma de Chile. Santiago de Chile
e-mail:andres.redchuk@gmail.com

Javier Cano
Departamento de Estadística e Investigación Operativa. Universidad Rey Juan Carlos. Madrid. España
e-mail:javier.cano@urjc.es

Lorenzo Galbiati
Agència Catalana de l'Aigua. Barcelona. España
e-mail:lgalbiati@gencat.cat

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*, Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_21,

standard way to assess the water quality of a river. In order to consider simultaneously all these issues, we present in this work a hydroinformatics management tool, designed to help decision makers with the selection of a PoM that satisfies the WFD objectives. Our approach combines: 1) a Water Quality Model (WQM), devised to simulate the effects of each PoM used to reduce pollution pressures on the hydrologic network; 2) a Multi-Objective Evolutionary Algorithm (MOEA), used to identify efficient tradeoffs between PoMs' costs and water quality; and 3) visualization of the Pareto optimal set, in order to extract knowledge from optimal decisions in a usable form. We have applied our methodology in a real scenario, the inner Catalan watersheds with promising results.

1 Introduction

Water is a precious resource, often jeopardized by its poor quality. Watersheds are constantly subject to increasing threats such as over-exploitation of both surface and ground water, and rising levels of contamination from point and diffuse sources of pollution [8]. In this context, the development and application of new political and management strategies and methodologies, aimed at reversing the degradation in water quantity and quality, has become of vital importance.

Although the European Commission has published a number of guidance documents to ease the implementation of WFD [5-7], no specific methodology has been suggested to evaluate the practical efficiency of PoMs; nor it is mentioned how such combinations of measures should be selected in order to achieve the best cost-effective strategy. In this regard, the restoration of water quality at watershed level (considering the water bodies as management units) is related to a series of objectives that should be taken into account when defining the river basin management plan.

From a methodological point of view, water resources planning and management is a sub-field of natural resource management, in which decisions are particularly amenable to multiple criteria analysis [23]. Moreover, decisions in water management are characterized by multiple objectives and involves several stakeholders groups with different interests and goals. In this regard, decision makers are increasingly looking beyond conventional cost-benefit analysis and looking towards techniques of multi-criteria analysis that can handle a multi-objective decision environment [12].

Water Quality Models (WQM) have been widely used to assess and simulate the efficiency of PoMs in increasing the availability and quality of water. Although such models are useful for evaluating single "what-if" scenarios and testing potential management alternatives, they are unable to solve, in an automatized way, the multi-criteria (cost, water quality, water availability) optimization problems involving the selection of the best cost-effective PoM tradeoffs.

Thus, linear programming [1], non-linear programming [2] and integer programming [25] have been used as alternatives to solve the cost optimization and river quality management model for regional wastewater treatment. Some approaches also

consider the river flow as a random variable, building a probabilistic model for it [10]. However, most of the abovementioned approaches, consider only one or two water quality parameters, and, therefore, optimal decisions do not take into account the general state of the watershed regarding its contamination levels, the political strategies and the socioeconomic status of the region. Besides, the inherent nonlinearity of water quality models, the presence of integer decision variables (implementation or not of the WWTPs), and the multiple criteria that are considered simultaneously, make Multi Objective Evolutionary Algorithms (MOEA) a suitable tool to identify tradeoffs between multiple objectives. Over recent years, MOEA [3, 26] have been applied to obtain the Pareto optimal set of solutions for the multiobjective management of watershed with promising results in a single execution [15, 22].

Our proposal to deal with this type of complex problems is a new multicriteria decision support methodological tool, devised to help the decision makers with the management of water quality during WFD implementation at a catchment scale. This methodology results from the integration of various elements: (1) the Qual2k water quality model [16]; (2) a new efficient MOEA, designed to efficiently solve expensive multiobjective optimization problems [24], and a set of tools for visualization and analysis. Our model can incorporate different approaches in order to assess the overall quality of the river, promoting, in this way, new points of view between the stakeholders within the negotiation process, and improving the robustness of the final decision.

Our methodology is being currently used in practice. Specifically, it has been applied on the inner Catalan watersheds to select a robust cost-efficient PoM, in order to achieve the WFD objectives within a reasonable cost. We describe in this paper how to identify the problems on each watershed, how our tool is designed to help in the decision process, and how the optimum PoM is finally selected. We must emphasize that the results provided by our tool have been an essential contribution to the definition of the Catalan hydrological plan to achieve the WFD objectives by 2015. The PoM provided by our model has been recently approved by the Directive Board of the Catalan Water Agency (ACA, by its initials in Spanish). It will be endorsed by the Catalan Government [4], together with the River Basin Management Plan of Catalonia within the next months, being, then, implemented in practice, using, among others, the indications and conclusions obtained by our tool.

2 Multicritera Strategies Selection (MC-SS) Methodology

2.1 The Strategies

The European Directives [5-7] have, as their main motivation, the protection of the environment from the adverse effects of waste water discharges. In order to carry out these directives, the ACA has developed an urban and industrial WWTP program [18, 19], that, in a preliminary study, allowed to identify a number of suitable locations to build more than 670 WWTPs for all the Catalan internal catchments.

Nowadays, there are many reclamation technologies that provide different levels of efficiency in the removal of water pollutants [20]. For the PoM implementation analysis, ACA considered seven WWTP technology types, in terms of their nutrient removal efficiency, and the investment and operational costs, see Table 1. We consider here three different nutrients: ammonium (NH_4), nitrate (NO_3) and phosphates (PO_4). Then, in a hypothetic river with n WWTP possible locations, we would have 7^n different possible combinations of PoMs (strategies). The management solution involves finding which of these PoM combinations are efficient, according to the criteria established by the ACA for the 2010 scenario.

Table 1 Cost and nutrient removal efficiency of the WWTP technologies considered by ACA

Treatment Type	Nutrient Effic. Remov. (%)			Monthly Cost (€/m^3)	
X_T	NH_4	NO_3	PO_4	Investment	Operation
Primary	0	100	0	222 (fixed)	$-0.0001 \cdot Q_P^{0.115}$
Secondary	30	95	50	$2.758 \cdot Q_D^{-0.357}$	$4.645 \cdot Q_P^{0.337}$
Nitrif (60%)	60	10	50	$3.172 \cdot Q_D^{-0.357}$	$5.342 \cdot Q_P^{-0.337}$
Nitrdeni 70%	75	85	50	$3.447 \cdot Q_D^{-0.357}$	$5.342 \cdot Q_P^{-0.337}$
Nitde70% Pr	75	85	75	$3.447 \cdot Q_D^{-0.357}$	$5.574 \cdot Q_P^{-0.337}$
Nitde85% Pre	85	90	80	$4.137 \cdot Q_D^{-0.357}$	$5.574 \cdot Q_P^{-0.337}$
Advanced	95	85	85	$4.413 \cdot Q_D^{-0.357}$	$6.604 \cdot Q_P^{-0.337}$

Here, Q_D and Q_P are, respectively, the design and operational capacities of a WWTP in m^3/day.

2.2 Problem Formulation

Optimization problems with multiple conflicting objectives lead to a set of tradeoff solutions, each of which is no better or worse than the others. Most environmental optimization problems are of this nature. In the WFD scenario, achieving a solution usually implies determining the best tradeoffs strategies in order to satisfy the WFD's objectives within a reasonable cost.

To fix ideas, let us assume that we are dealing with an arbitrary optimization problem with M objectives, all of them to be maximized. Then, a general multiobjective problem can be formulated as follows:

$$\begin{aligned} &maximize \quad f_m(x), && m = 1, 2, \ldots, M, \\ &subject\ to: \ g_j(x) \geq 0, && j = 1, 2, \ldots, J, \\ &\qquad\quad h_k(x) = 0, && k = 1, 2, \ldots, K, \\ &\qquad\quad x_i^{(L)} \leq x_i \leq x_i^{(U)} && i = 1, 2, \ldots, n \end{aligned} \tag{1}$$

where x is the n-vector of decision variables: $x = (x_1, x_2, \ldots, x_n)^T$. In our case, x describes the waste water treatment alternatives, corresponding to each WWTP

(strategy) planned to be built in the region. The inequality and equality constraints, $g_j(x), j = 1, \ldots, J$, and $h_k(x), k = 1, \ldots, K$, together with the bounds $x_i^{(L)}$ and $x_i^{(U)}$, $i = 1, \ldots, n$, define the decision variable space *D*. We say that $f^* = (f_1^*, f_2^*, \ldots, f_M^*)$ is a Pareto optimal objective vector if there is no feasible solution x', such that $f' = (f_1', f_2', \ldots, f_M') = \left(f_1(x'), f_2(x'), \ldots, f_M(x')\right)$, satisfying $f_m^* \leq f_m'$ for each $m = 1, 2, \ldots, M$, and $f_j^* < f_j'$ for at least one index *j* in$1 \leq j \leq M$. Each decision variable $x_i, i = 1, \ldots, n$ is actually a discrete variable with 7 possible values, see Table 1. In some cases, and according to the physicochemical characteristics of the stretches, a constraint for the minimum purification treatment must be added.

$$x_i > x_{i,min} \ \forall i = 1, \ldots, n \tag{2}$$

In our specific application to the Catalan inner watersheds, we shall consider two objective functions, the first one having to do with economic factors, the second one dealing with quality aspects of the water.

2.3 The Cost Objective Function

The cost of each strategy corresponds to the sum of the investments in all the catchment WWTP, and the operation costs. The costs for each WWTP facility depend on the flow rate and the type of treatment plant, see Table 1. Then, the first objective function has the form

$$f^1 = \sum_{j=1}^{NumWWTP} \left(ICost_j + OCost_j\right) \tag{3}$$

where *j* is the WWTP index and *NumWWTP* is the total number of WWTPs. Besides, $ICost_j = f(Q_D, x_j)$ and $OCost_j = f(Q_P, x_j)$ represent the investment needed to build the *j*-th WWTP (monthly cost with a 15-year payback period), and the monthly operating costs, respectively.

2.4 The Water Quality Objective Function

The quality criteria considered are the relative concentration of NH_4, NO_3 and PO_4, according to the WFD limits. For a given river stretch, and using the WFD reference, we can evaluate the water quality according to:

$$\delta_s^k = \frac{(WFDL_s^k - AC_s^k)}{WFDL_s^k} \tag{4}$$

where $WFDL_s^k$, AC_s^k and δ_s^k represent, respectively, the WFD concentration limits, the current level of concentration, and the relative concentration of the *k*-th contaminant (k =2,3,4 stand for NH_4, NO_3 and PO_4, respectively) in the *s*-th stretch, according to the WFD's limits.

As the global river water quality depends on the quality of all the river stretches, each quality objective function (f^2 is the NH_4 river quality, f^3 is the NO_3 river quality and f^4 is the PO_4 river quality) must be computed based on the values of δ_s^k for all the river sections. There are many possible ways (metrics) to do this [9], possibly leading to significantly different results, but saying that one particular solution is better than the others is a very subjective and subtle issue. To avoid this controversy, it is possible to run our methodology using different metrics, in order to assess the objective functions of global quality of a river with respect to the three contaminants. The three metrics considered are described below:

1. *Utilitarian*

This metric considers all river sections as equivalent, and the objective is, then, to minimize the average of δ_s^k in all river sections. A usual formulation is [17]

$$\min f_u^k = \frac{1}{ns}\sum_{s=1}^{ns} \delta_s^k \quad k = 2,3,4 \tag{5}$$

where *ns* is the number of stretches.

2. *Egalitarian (Smorodinsky-Kalai)*

Another possibility is to seek an equitable strategy that tries to reduce the differences on quality in all river sections. To achieve an egalitarian solution we minimize the Smorondinsky-Kalai objective function [13]

$$\min f_e^k = \mu_k \quad k = 2,3,4 \tag{6}$$

such that

$$\delta_s^k \le \mu_k \quad k = 2,3,4; \;\; \forall s \in ns \;\; \delta_s^k \le \mu \quad \forall s \in ns$$

3. *Separate Utilities (fulfilling and unfulfilling of WFD)*

This quality function has two different approaches, depending on whether it measures the success or failure in the achievement of a good ecological status. Positive values of the metric mean that the WFD objectives are accomplished for every basin stretch. Otherwise, a negative value means that the WFD objectives are exceeded by at least one river stretch [24].

$$\min f_{su}^k = \begin{cases} \dfrac{1}{ns}\displaystyle\sum_{s=1}^{ns} \delta_s^k & \forall \delta_s^k \ge 0 \\ \dfrac{1}{nsi}\displaystyle\sum_{s=1}^{nsi} \delta_s^k & \forall \delta_s^k < 0 \end{cases} \quad k = 2,3,4 \tag{7}$$

where *nsi* is the number of stretches that do not satisfy the WFD limits

2.5 The MOEA

As we have already mentioned, evolutionary computation methods are becoming increasingly popular for the resolution of environmental problems. Especially suitable are those MOEAs for which conventional techniques are not easily adapted, including nonconvex, mixed integer, non-linear, constrained and/or noisy cost functions. In this regard, a MOEA is a heuristic search algorithm based on a population of strings (called chromosomes) that mimic the process of natural evolution. This population encodes candidate solutions to an optimization problem, called individuals, and evolves toward better solutions.

The MOEA developed in this work to optimize (select) WWTP tradeoff strategies, applies binary gray encoding [11] for each chromosome (optimization string). The length of each optimization string corresponds to a total number of genes, one for each facility. Each gene uses 3 bits to encode the 7 sewage treatment levels for each plant. After decoding the chromosome in treatment levels for each WWTP, the water quality in each stretch is forecasted by the water quality model. The associated goodness-of-fit value is assessed for each one of the cost and quality equations describe above.

The MOEA algorithm applies the usual procedures of selection (tournament), crossover (multi-point) and mutation (uniform) to generate the new population. Efficient convergence is achieved with small populations (10 chromosomes per generation) and mutation rates of 3%. For more details about the convergence of the algorithm see [24]. This MOEA algorithm also introduces elitism by maintaining an external population [3, 26]. In each generation, the new solutions belonging to the internal population are copied to the external population when they are not Pareto-dominated by any solution of this external population. If solutions for the external population are dominated by some of the new solutions, these solutions are deleted from the external population. The external elitist population is simultaneously maintained in order to preserve the best solutions found so far, and to incorporate part of the information in the main population by means of crossover. Elitism is also included in this recombination process, by selecting each of the parents through a fight (tournament) between two randomly-selected chromosomes from the external Pareto set (according to a density criterion), or from the population set (according to their ranking determined through a dominance criterion). The stopping criterion applies when no new non-dominant chromosomes appear in a significant number of generations

2.6 The Water Quality Model

Water Quality Models (WQM) aim at describing the spatial and temporal evolution of the contaminants and constituents characterizing a river flow. Many highly reliable simulation models are available today to evaluate the behaviour of physical systems, such as water bodies, with reasonable computational requirements [21]. In this work, we have used Qual2kw [16], as it represents the state of the art of the last two decades of advances in river water quality modelling and numerical computations.

A range of inputs is used in the water quality simulations, including topography, climate and predicted pressures for 2015, when the objectives of the Water Framework Directives will be effective. Specifically, the main inputs of the WQM are: the head water in all tributaries, point sources (urban, industrial, WWTP; etc), water extractions, diffuse sources of pollution, as well as physicochemical and biological parameters for waste, hydraulics (morphological elements, Manning's roughness coefficient, flow curve, flow). The inflows for the proposed WWTPs are the urban and industrials effluents; based on the information from their discharges in the last 10 years, see [24] for more details.

2.7 Application of the MC-SS

Although our methodology has been actually applied to all Catalan internal watersheds, the results presented in this work correspond to its application in the Muga basin. The Muga River has its source in the Eastern Pyrenees, at an approximate height of 1200 meters, flowing towards the Mediterranean Sea, laying its basin entirely within the region of Catalonia, Spain. The Muga River has its headwaters located in mountainous areas, whereas the middle and lower parts of the watershed are subject to Mediterranean climate, implying higher hydrological variability in these last sections. Its main channel has a total length of 64.7 km, draining a watershed of 759 km^2 (2.3% of the total area of Catalonia). It receives an annual average of 177 Hm^3 and its runoff coefficient is 0.285.

In order to apply the Qual2kw model to a river network, the river system must be divided into river elements, having roughly the same hydraulic characteristics. In each cell, the model computes the major interactions between up to 16 state variables and their values for static and dynamic conditions. In this case, the total length of the main channel of the Muga River, and its 12 tributaries is 227 km, which were divided into 54 elements of approximately 5 km length.

For this problem, the ACA considered 41 WWTP locations, each with 7 sewage treatment levels. Each gene uses 3 bits to encode these 7 possible alternatives for the decision variables. Therefore, in the Muga watershed, the number of possible WWTP locations are 41, with a chromosome length of $41 \times 3 = 123$ bits. Then, there are $7^{41} \approx 4.4 \times 10^{34}$ different possible PoM combinations (strategies). The management solution involves finding which of these PoM combinations is efficient according to the ACA estimated conditions for the 2015 scenario, and the goal is to find out which is the most efficient one, according to all the criteria.

The integrated tool (MC-SS) was executed considering simultaneously from 2 to 4 objectives (cost-ammonium-nitrates-phosphates). Runs of the algorithm were performed with different MOEA parameter configuration, using the three quality metrics described above, obtaining, in this way, different Pareto fronts for each one. In order to analyse the convergence process, we consider a MOEA stopping criterion corresponding to a maximum number of WQM evaluations, in this case 10000 evaluations. The number of points obtained for each Pareto front depends on the metric and objectives used.

3 Results

In order to make the proposed methodology useful in the decisions making process, ensuring the achievement of the objectives of the WFD, it is required to work in an efficient manner. In other words, the algorithm must converge close enough to the Pareto solutions in a reasonable number of evaluations of the objective function, making the problem amenable to being solved by low-cost computers. This is especially important in this kind of problems, where the objective function evaluation has a significant computational cost (for some large sized basins, each evaluation may take up to 15 minutes).

The success of our approach was achieved thanks to several improvements on the "standard" multi objective evolutionary techniques, which speeded up the convergence. Specifically, the main improvements in the performance of the algorithm are: (1) the steady state evolution (small population size); and (2) the elitism that allows to reach a good convergence for the Muga basin in less than 6000 evaluations of the WQM, considering simultaneously two objectives. In this regard, a significant increase in the size of the optimization problem only produced a slight increase in the number of evaluations required for our MOEA to reach convergence. On the other hand, an increase in the number of criteria (e.g., from two to four) required more than 10.000 evaluations to achieve convergence. Further improvement on the convergence speed of the MOEA (up to 50%) was achieved by choosing adequate initial strategies from the Pareto fronts obtained in previous executions (e.g., carried out with different metrics or less objectives), rather than generating them in a random way. More details about the convergence process and the configuration of the MOEA parameters can be found on [24].

The use of either water quality metric had little influence on the algorithm convergence. In this regard, we should mention that the egalitarian metric converged slightly faster, because it encompassed a lower number of efficient strategies than the two others, the reason being that only the WWTP located close to those stretches with the worst river quality had influence on the value taken by the egalitarian metric. On the contrary, changes in most of the WWTP had influence in the value taken by the other two metrics (utilitarian and separate utilities). This fact suggests us that one of the main drawbacks of the egalitarian metric is that, by using it, it is difficult to know the general status of the river, because it only informs us about the state of the worst quality stretch.

Once the Pareto frontier is delineated, it must be analyzed. However, special techniques should be used when there are more than two criteria. To accomplish that, we have used Interactive Decision Maps (IDM), see [14], to simultaneously study tradeoffs for up to 7 criteria. The number of efficient strategies provided by the MOEAhen 4 criteria are simultaneously taken into account is quite high, easily exceeding several hundreds. However, by using the IDM, this difficult shape analysis and comparison of simultaneous tradeoffs becomes quite simple. Specifically, the stakeholders performed a preliminary strategy selection, with the IDM visualization tools, and then translated it into the 2D representation. In the 2D diagram, see Figure 1, the *Y* axis represents the cost of the strategies, whereas the *X* axis

represents the water quality for each indicator according to (5), (6) or (7). When using separate utilities or egalitarian metrics, the value $X = 0$ corresponds exactly to meeting the WFD objective. The points falling on the left side of the graphs are strategies that do not satisfy WFD goals, and the points on the right side of the graphs do meet them. A positive value indicates good quality in the defined objective. However, applying the utilitarian metric has the inconvenient that it is difficult to know, from the examination of the Pareto frontier, if one specific strategy meets the limits of the WFD, because the value of the stretches of poor quality may be compensated for the value of the stretches of good quality and vice versa.

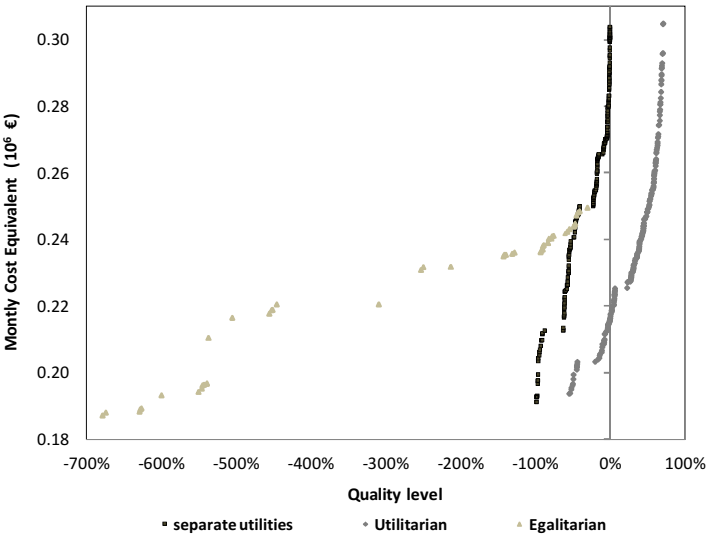


Fig. 1 Pareto fronts based on an optimization using only two objectives (cost and ammonia) for the tree quality metrics (separate utilities, utilitarian, and egalitarian)

The Pareto set is the basic knowledge resource from which the stakeholder will base the decision process, so special care should be taken in order to represent it in an intelligible, yet rigorous, manner. Exploration of the Pareto frontier helps the decision makers to understand the criteria tradeoffs, and to identify, in a direct way, a preferred criterion point.

Additional information can be obtained from the slope of these criteria quality curves (the Pareto front curves). They indicate the sensitivity of the water quality to the water treatment actions, i.e., they provide the cost increase required to achieve a unitary increase on the water quality for each strategy. Figure 1 shows the three Pareto fronts obtained for the same problem with each of the metrics discussed in this paper, considering only the cost and ammonium objectives. As we can observe, the use of one or another metric to calculate the overall river quality has a great influence on the Pareto front shape.

If we analyze Figure 1 in more detail, we see that, for the egalitarian metric, when we increase the budget in most intensive sewage PoMs by 40%, this reduces the WFD ammonium breach in the worst stretch of the river from -700% to -30%. Regarding the separate utilities Pareto metric, a similar increase on the depuration budget of around 50% lead to a drastic improvement on the average quality of those stretches not fulfilling the WFD limits by more than 100%. After such investment, the average river quality was very close to the WFD acceptance limits, and, probably, many of the river sections that, separately, did not satisfy the WFD, now they do. We can also observe that, even for the most intensive sewage PoM, it is impossible to achieve the WFD's objective satisfactorily for the ammonium criteria in this catchment. So, in this case, it would be more reasonable to select a strategy with an associated budget close to 270,000 €/month, because spending more money does not lead to a significant improvement on the water quality results. Finally, if we examine the curve corresponding to the utilitarian metric, we see that positive values are obtained for investments higher than 220,000 €/month. Nevertheless, we must keep in mind that the utilitarian metric only indicates whether the average quality of the river is good or not, but, given a positive overall value, it does not ensure a fulfilment of the WFD in all the stretches.

We have just discussed, the benefits and drawbacks of each metric used, with respect to the visual analysis of the Pareto front. But it is important to note that we must also take into account that the MOEA finds different strategies to be Pareto optimal depending on the metric considered. The utilitarian solution tends to save costs on those WWTP related to river stretches in which depurating is very expensive and *vice versa*, and, then, it weights both contributions. On the contrary, the egalitarian metric will tend to invest almost all the budget on those WWTP highly related to the most contaminated stretches, leaving the rest of the river unaffected. The separate utilities metric partially solves this problem, thanks to the fact that, if there are several stretches violating the WFD's objectives, this metric takes all of them into account (and not only the worst one). Otherwise, when all the stretches fulfill the requirements, this metric is equivalent to the utilitarian one.

In this regard, we must conclude that there is not a perfect metric to help us in the decision making process. Rather, each one can be consider better or worse than the others depending on the (subjective) point of view or the interests of each stakeholder. The main advantage of providing decision makers with different results obtained using various metrics is to reduce, as much as possible, the inherent subjectivity of the decision process. This is achieved by providing the stakeholder with efficient solutions, attained using different metrics to assess the overall quality of river waters and regarding the concentration of each pollutant considered.

By performing a deeper analysis of the decision variables (WWTP) corresponding to all the Pareto optimal solutions obtained for each metric, we can reduce further the subjectivity of the decision process. Specifically, for the basin analyzed in this work, we observe that for 14 of the 41 WWTPs implemented, the treatment level is the same for all the strategies and for any of the fronts obtained. This allows us to fix these 14 values prior to make the final decision, facilitating, in this way, the stakeholders' decision process.

Summarizing, when a final decision is to be found, each stakeholder participates in a decision process that begins by pointing out which regions of the Pareto frontier he or she has specific preferences on. Then, the decision process is followed by the negotiation phase, in which all the stakeholders reach an agreement on some strategies or regions of common interest. Before making the final decision, each of these strategies or regions must be examined in detail.

In this regard, for one selected strategy and pollutant indicator, the use of geographical information systems (GIS) to display, or summarize the information that is automatically generated by our tool might be also of great help.

For a single criterion, it is easier (and more interesting from a stakeholder's point of view) to simultaneously compare results between different strategies, for all months and stretches. In our case, from all the solutions of the Pareto front, we have preselected three strategies (PoM). The first one corresponds to low-intensive and cheap treatments, the second one is related to very intensive treatments (actually, the most expensive ones), and the third one is an intermediate solution between the first two ones. In Figure 2 we have analyzed the monthly results for the three strategies corresponding to the ammonium level at each stretch through a box plot. We can observe how the quality improves as time varies in all stretches but one, fulfilling, in this way, the WFD requirements.

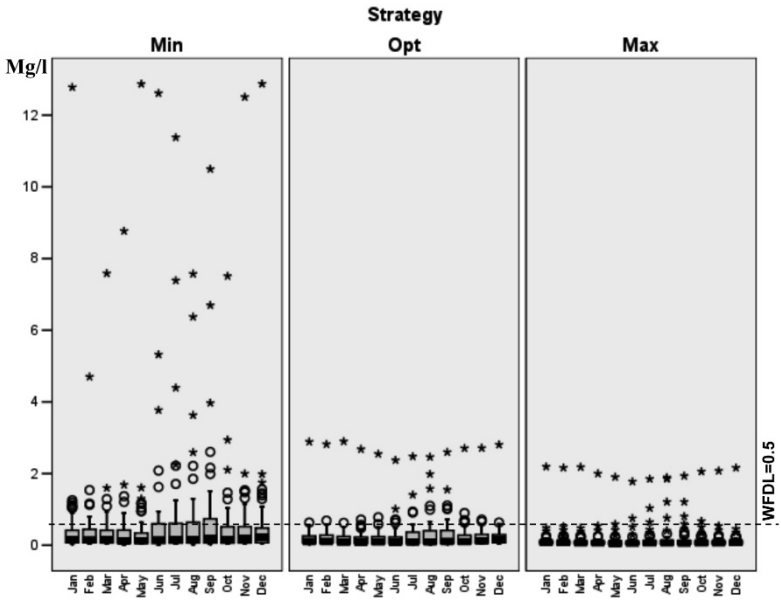


Fig. 2 Box plot for the levels of ammonium in the stretches, depending on the month and the applied purification treatment (Min, Opt, Max) (Ter basin)

4 Summary and Conclusions

A new integrative evolutionary Multi Criteria Strategies Selection (MC-SS) methodology is proposed to help in the selection of the most efficient PoMs for water resources conflicting objectives. It has been applied in the context of the implementation of the WFD in Catalonia. Based on this methodology, a hydroinformatic tool has been developed to assist in the management of water quality at a catchment scale.

The tool is an effective combination of a WQM, which estimates monthly runoff and pollutant loads in the catchments, and the MC-SS algorithm, whose main component is a multicriteria genetic algorithm especially designed and configured to find the Pareto optimal set of PoM (strategies). It is able to incorporate conflicting elements into the analysis, such as environmental objectives and economical issues. Thanks to several improvements on "standard" techniques, which have speeded up the convergence of the MOEA, the approach enables the delineation of non-dominated Pareto optimal solutions in a number of WQM executions that are small enough to be performed on a standard PC, in a timescale that meets the requirements of the Catalan Water Agency (ACA).

We have carried out a case study, taking waste water systems into account, resulting in seven different cleaning technology alternatives, which were also modelled in terms of cost and treatment for each pollutant. Therefore, and in addition to the cost criteria (operating and investment cost), three quality criteria were considered simultaneously: ammonium, nitrate and phosphate. The inherent nonlinearity of the WQM, the integer character of the decision variables (WWTP) and the four criteria simultaneously considered, make MOEA methods more efficient than conventional optimization methods in identifying tradeoffs among multiple objectives.

The selection process of PoMs through which accomplishing the WFD objectives, is a participative process. Then, our methodology has an added value, as it gets suitably integrated within the negotiation and decision processes that the stakeholders must carry out. On the other hand, the stakeholders themselves can suggest new different metrics to assess the global quality of the river water, obtaining new Pareto fronts upon running of the MC-SS. This fact facilitates the stakeholders with a greater degree of intervention on the participation process. Nevertheless, we must keep in mind that there is no perfect metric to help us in the decision making process on the whole basin, although the availability of various fronts obtained from different metrics can be of great help in the decision-making process.

The developed methodology has been shown to be an important tool to: (1) evaluate the effectiveness of the actions that are being currently undertaken to improve water quality; and (2) to provide decision makers with the capacity to explore the multi-objective nature of problems, and to discover tradeoffs amongst objectives avoiding subjectivities as much as possible. We have found this feature to be very helpful, especially during the negotiation process prior to the achievement of the final decision. The main factors intended to guarantee the success on

the implementation of the system have been: (1) users' involvement; (2) development of several evolutionary prototypes; and (3) design of a specific user-friendly interface adopted for multicriteria applications and a variety of implemented models and decision support tools.

This tool has been a key factor in the design of part of the PoMs which shall be implemented to achieve the WFD objectives by 2015 in Catalonia. For the Catalan catchments, the model and tools developed have successfully identified the problems in each watershed, for all the WFD criteria considered in this study. Indeed, application of the model has required a reasonably small number of Qual2k executions, keeping the computational time requirements within reasonable limits.

Acknowledgments. This work has been supported by ACA, by grants from MICINN (eColabora-RIESGOS), the RIESGOS-CM program S2009/ESP-1685 and MTM2011-28983-C03-01 of. Additionally, the authors are grateful to Auditorías e Ingenierías, S.A. (Auding) that has been in charge of developing the database Qual2k implemented in the model. Part of this research was done while the third author was visiting KTH, supported with grants from URJC's postdoctoral programmes.

References

1. Afkhami, M., Shariat, M., Jaafarzadeh, N., Ghadiri, H., Nabizadeh, R.: Regional water quality management for the Karun Dez River basin, Iran. Water and Environment J. 21(3), 192 (2007)
2. Cai, X., McCinney, D.C., Lasdon, L.S.: Solving nonlinear water management models using a combined genetic algorithm and linear programming approach. Advances in Water Resources 24(6), 667–676 (2001)
3. Deb, K.: Multi-objective Optimization Using Evolutionary Algorithms. John Wiley, Hoboke (2001)
4. Diari Oficial de la Generalitat de Catalunya. Decret d'aprovació del Pla de gestió del districte de conca fluvial de Catalunya. Núm. 5764, 86782–86814 (November 26, 2010)
5. European Commission 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy (2000)
6. European Commission 2001, Strategic document: Common Strategy on the Implementation of the Water Framework Directive (2001)
7. European Commission 2002, Economics and the environment. The implementation challenge of the Water Framework Directive. Policy Summary to the Guidance Document (2002)
8. European Commission 2007. Water Framework Implementation Reports. Towards Sus-tainable Water Management in the European Union (2007)
9. Fernández, N., Solano, F.: Indices de Calidad de Agua e Indices de Contaminación, p. 310. Universidad de Pamplona, Colombia (2005)
10. Fujiwara, O., Gnanendran, S.K., Ohgaki, S.: Chance constrained model for river water quality management. Jour. Environ. Eng., ASCE (1987)
11. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Pub. Co. (1989)

12. Hajkowicz, Collins, Hajkowicz, S.A., Collins, K.: A review of multiple criteria analysis for water resource planning and management. Water Resources Management 21, 1553–1566 (2007)
13. Kalai, E.: Proportional Solutions to Bargaining Situations: Inter20 personal Utility Comparisons. Econometrica: Journal of the Econometric 21 Society 45(7), 1623–1630 (1997)
14. Lotov, A.V., Bushenkov, V.A., Kamenev, G.K.: Interactive Decision Maps, Approximation and Visualization of Pareto Frontier, Applied Optimization, vol. 89. Springer (2004)
15. Muleta, M.K., Nicklow, J.W.: Decision support for watershed management using evolutionary algorithms. J. Water Resour. Plng. and Mgmt., ASCE 131(1), 35–44 (2005)
16. Pelletier, G., Chapra, S.: Qual2kw User Manual (Version 5.1): A modelling framework for simulating river and stream water quality, Olympia, WA. Washington State Department of Ecology (2004)
17. Ponsati, C., Watson, J.: Multiple-Issue Bargaining and Axiomatic Solutions. International Journal of Game Theory 26, 501–524 (1997)
18. Programa de sanejament d'aigües residuals urbanes (PSARU 2005) Diari Oficial de la Generalitat de Catalunya (DOGC). Núm. 4679 (July 19, 2006)
19. Programa de sanejament d'aigües residuals industrials (PSARI-2003) Diari Oficial de la Generalitat de Catalunya (DOGC) Núm. 3986 (October 13, 2003)
20. Qasim, S.R.: Wastewater Treatment Plants: Planning, Design, and Operation. CRC Press (1999)
21. Rauch, W., Henze, M., Koncsos, L., Reichert, P., Shanahan, P., Somlyódy, L., Vanrolleghem, P.: River water quality modelling: I. State of the art. Wat. Sci. Tech. 38(11), 237–244 (1998)
22. Ritzel, B.J., Eheart, J.W., Ranjithan, S.: Using genetic algorithms to solve a multiple objective groundwater pollution containment problem. Water Resources Res. 30(5), 1589–1603 (1994)
23. Romero, C., Rehman, T.: Natural resources management and the use of multiple-criteria decision making techniques: a review. Eur. Rev. Agric. Econ. 14(1), 6–89 (1987)
24. Udías, A., Galbiati, L., Elorza, F.J., Efremov, R., Pons, J., Borras, G.: Framework for Multi-Criteria Decision Management in Watershed Restoration. Journal of Hydroinformatics (2011), doi:10.2166/hydro
25. Yang, W., Yang, Z., Qin, Y.: An optimization approach for sustainable release of e-flows for lake restoration and preservation: Model development and a case study of Baiyangdian Lake, China. Ecological Modelling 222(14), 2448–2455 (2011)
26. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C.M., Grunert da Fonseca, V.: Performance assessment of multi-objective optimizers: An analysis and review. IEEE Transactions on Evolutionary Computation 7(2), 529–533 (2003)

Fuzzy Data-Mining Hybrid Methods for Recommender Systems

María N. Moreno, Joel P. Lucas, and Vivian F. López

Abstract. CRM (Customer Relationship Management) is one important area of Business Intelligence (BI) where information is strategically used for maximizing the value of each customer in a company. Recommender systems constitute a suitable context to apply CRM strategies. This kind of systems are becoming indispensable in the e-commerce environment since they represent a way of increasing customer satisfaction and taking positions in the competitive market of the electronic business activities. They are used in many application domains to predict consumer preferences and assist web users in the search of products or services. There are a wide variety of methods for making recommendations; however, in spite of the advances in the methodologies, recommender systems still present some important drawbacks that prevent from satisfying entirely their users. This chapter presents one of the most promising approaches consisting of combining data mining and fuzzy logic.

1 Introduction

The general target of Business Intelligence (BI) is to extend the operational use of the information to the strategic use in order to improve the business. In spite of both are valuable, and without the operational use of information a business could not exist, strategic uses can provide an added value. The degree of this value depends on the information consumer as well as on its strategic uses. In that sense, although, the applications of BI in the achievement of strategic objectives are numerous, one of the areas where BI can report more benefits is CRM (customer relationship management). In this context Business intelligence is rapidly becoming a useful tool to achieve competitive advantage and to support better business decision-making, especially in the e-commerce domain. CRM is an overused term,

María N. Moreno · Joel P. Lucas · Vivian F. López
University of Salamanca, Spain
e-mail: {mmg,vivian}@usal.es, joelpl@gmail.com

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_22,

which seems to enclose the formula to turn all contacts into customers and all customers into great customers. Analytic capabilities are the basis of CRM that allow a company to acquire a better knowledge about their customers and to increase the value of them. In addition, customer's experience can be enhanced by the results of these analytics [35].

The specific aspects of CRM suitable to be used in the e-commerce field are the following:

- Customer profiling. The aim is to treat every user in a different way depending on his individual characteristics or profiles. These profiles determine the different forms of acting, for example, in the accomplishment of the marketing campaigns directed to consumers with similar tastes.
- Personalization. It is the process of adapting the presentation of the information to the user on the basis of his profile. The web systems provided with personalization mechanisms make use of the users' profiles to select and to show to the user the content that adjusts to his profile.

Collaborative filtering. In the e-commerce systems users can receive suggestion about products based on the preferences of other users with similar tastes. The systems endowed with this kind of personalization mechanism are named recommender systems.

This chapter is focused on the application of the two last aspects in the context of recommender systems since their presence in current web systems, especially in the e-commerce domain, is becoming indispensable. Recommender systems make personalized suggestions and provide information of items available in the system. Nowadays, there is a vast amount of methods, including data mining techniques that can be employed for personalization in recommender systems. However, such methods are still quite vulnerable to some limitations and shortcomings related to recommender environment.

Collaborative filtering methods are the most used in recommender systems. They make use of information related to evaluations (or ratings) provided by users. This can cause the sparsity problem when evaluations from users are insufficient. On the other hand, traditional collaborative filtering approaches based on nearest neighbor algorithms show serious performance and scalability problems. In the last years many recommendation techniques have been proposed aiming at improving the quality of the recommendations as well as dealing with other typical drawbacks of recommender systems. Data mining techniques have been successfully applied in recommender systems to predict user preferences. They do not present performance problems since predictive models are already built when the user logs in the system and they are less sensitive to sparsity problems. However, the results vary with the selected algorithm given that they present different behavior depending on the characteristics of the dataset. Therefore, it is necessary to apply suitable algorithms in order to obtain precise recommendations.

Classification based on association (associative classification) yields better results than other data mining techniques with data from recommender systems [37]. Fuzzy logic has also been successfully applied in this field in combination with

associative classification [36]. The problem of managing numerical by data mining algorithms can be addressed by means fuzzy logic. In spite of the value of this approach has been proved in many application areas, it has not been widely explored in recommender systems. Fuzzy logic can help minimizing, or even solving, typical drawbacks of such systems. Dubois et al. [22] affirm that fuzzy logic provides high-value properties to recover items stored in a database and, as a consequence, to provide recommendations for users. The reason is the capability of fuzzy sets to manage concepts such as similarity, preference and uncertainty in a unified way, while also performing approximate reasoning. Due to such advantages, especially for uncertainty, fuzzy logic can help to minimize the sparsity problem, which is the main drawback in current recommender systems.

The aim of this chapter is to present the state of the art of hybrid recommender methodologies combining data mining and fuzzy logic. The chapter also includes the description of a hybrid methodology developed by the authors of the chapter and validated in a recommender system in the tourism field.

2 Recommender Systems' Methods

Web recommender systems are used in many application domains to predict consumer preferences and assist web users in the search of products or services. Methods used to do that have different levels of complexity, since those that recommend products based on associations between them in previous transactions to those that make recommendations based on user evaluations about products and similarity between user preferences. They can be classified into two main categories [29]: collaborative filtering and content-based approaches. The first class of techniques was based initially on nearest neighbor algorithms. These algorithms predict product preferences for a user based on the opinions of other users. The opinions can be obtained explicitly from the users as a rating score or by using some implicit measures from purchase records as timing logs [42]. In the content based approach text documents are recommended by comparing between their contents and user profiles. The weights for the words extracted from the document are added to the weights for the corresponding words in the user profile, if the user is interested in the page [29]. The main shortcoming of this approach in the e-commerce application domain is the lack of mechanisms to manage web objects such as motion pictures, images, music, etc. Besides, it is very difficult to handle the big number of attributes obtained from the product contents.

Currently there are two approaches for collaborative filtering, memory-based (user-based) and model-based (item-based) algorithms. Memory-based algorithms, also known as nearest-neighbor methods, were the earliest used [41]. They treat all user items by means of statistical techniques in order to find users with similar preferences (neighbors). The prediction of preferences (recommendation) for the active user is based on the neighborhood features. A weighted average of the product ratings of the nearest neighbors is taken for this purpose. The advantage of these algorithms is the quick incorporation of the most recent information, but they have the inconvenience that the search for neighbors in large databases is

slow [44]. Model-based collaborative filtering algorithms use data mining techniques in order to develop a model of user ratings, which is used to predict user preferences.

Collaborative filtering, specially the memory-based approach, has some limitations such as sparsity and scalability that will be commented later. Sparsity is due to the high number of ratings required for prediction and scalability is related to performance problems in the search for neighbors in memory-based algorithms. The lesser time required for making recommendations is an important advantage of model-based methods. This is due to the fact that the model is built off-line before the active user goes into the system, but it is applied on-line to recommend products to the active user. Therefore, time spent in building the model has no effects in the user response time since little process is required when recommendations are requested by the users, contrary to the memory based methods that compute correlation coefficients when user is on-line. Nevertheless, model based methods present the drawback that recent information is not added immediately to the model but a new induction is needed in order to update the model.

The quality of the recommendations for the users has an important effect on the clients' retention. Users refuse poor recommender systems which can cause two types of error: false negatives, which are products that are not recommended, though the customer would like them, and false positives, which are products that are recommended, though the customer does not like them [16]. The most serious errors are false positives, because these errors will cause negative reactions in the customers and thus they won't probably visit the site again. The use of data mining algorithms to find customers characteristics that increase the probability of buying recommended products can help to avoid these problems.

There are a great variety of data mining algorithms that can be applied in model based CF. Neural networks were the former of this kind of methods [8], which changed the nearest neighbor approach of CF methods by a classification approach. The same technique has been applied in several works such as [26] where it is combined with Case Based Reasoning, or [17] where it is used to predict consumer preferences taking into account his navigation behavior through navigation patterns extracted by means of an unsupervised web mining method. Bayesian networks constitute another technique widely used in the induction of recommendation models [20] in a single way [11] or jointly with other methods [14]. The main shortcoming of these methods is the high computational cost of building the net, especially when the amount of data is great. Although this is not a critical drawback since classification models are built off-line, when these models need to be often updated it can become a serious inconvenience. It occurs in current recommender systems due to continuous changes in the database of products and users [28].

Although Support Vector Machines (SVM) are linear classifiers originally designed for binary classification, they can be used in recommender systems [49]. The procedure consists on representing every user as a vector composed by ratings about products and building a hyperplane that separates the geometric space where the vectors are situated in classes representing groups of users of similar

preferences. Different strategies to reconstruct a multi-class classifier from binary SVM classifiers are studied [48]. In some works, SVM is used as a complementary technique to other methods. For instance, in [21] it is used to induce ranking functions from the preference judgments of each user as a previous step to the application of a clustering algorithm that builds groups of people with closely related tastes.

Clustering is also often used as a model based technique, in spite of being an unsupervised method, since the induced groups of people with similar preferences constitute a way of classification. Thus, the predictions for the active user are based in the opinions of the members of the group he belongs to. By means of fuzzy logic a user can be assigned to more than one cluster with different belonging degree and receive recommendation from more than one group. In any case, the personalization achieved is lesser than the one provided by other methods, for that reason, clustering is usually used in the preliminary exploration of the data as a previous step to the application of machine learning algorithms [43].

The works referenced in this section are just a little sample of the numerous data mining proposals to be used in collaborative filtering based recommender systems. However, the current trendy, especially in sparse contexts where ratings are insufficient, is to exploit hybrid methodologies combining content based and collaborative filtering approaches in order to take advantage of the strengths of each of them [6, 12].

In recent works, semantic information is added to the available data in order to formalize and classify product and user features, being able in this way to generate more reliable content based models that can be combined with other approaches in order to improve recommendations [9, 27].

3 Recommender Systems' Drawbacks

As commented before, collaborative filtering, especially the memory-based approach, has some limitations that have and important impact on the quality of the recommendations. First, rating schemes can only be applied to homogeneous domain information. Besides, sparsity and scalability are two important problems influencing on the quality of the recommendations [16]. Sparsity is due to the number of ratings needed for prediction is greater than the number of the ratings obtained because usually CF requires user explicit expression of personal preferences for products. The second limitation is related to performance problems in the search for neighbors in memory-based algorithms. These problems are caused by the necessity of processing large amount of information. The computer time grows linearly with both the number of customers and the number of products in the site. Model-based methods do not present this drawback since the model is built off-line before the active user goes into the system, but it is applied on-line to recommend products to the active user. Therefore, as commented before, time spent in building the model has no effects in the user response time, contrary to the memory based methods that compute correlation coefficients when the user is

on-line. Nevertheless, model based methods present the drawback that recent information is not added immediately to the model but a new induction is needed in order to update the model.

Despite the drawbacks described before may be minimized by means of data mining methods, there are other shortcomings that may occur even with these methods. The first-rater (or early-rater) problem arises when it is not possible to offer recommendations about an item that was just incorporated in the system and, therefore, has few (or even none) evaluations from users. In fact, the early rater problem is directly linked to sparsity since when a system has a high number of items, probably most of these items have never received any evaluation. Conceptually, the early-rater problem can be viewed as a special instance of the sparsity problem [25]. Sarwar et al. [42] affirm that current recommender systems depend on the altruism of a set of users who are willing to rate many items without receiving many recommendations. Economists have speculated that even if rating required no effort at all, many users would choose to delay considering items to wait for their neighbors to provide them with recommendations [5].

Analogously, such drawback also occurs with a new user joining the system, since there is no information about his preferences, it would be impossible to determine his behavior in order to provide him recommendations. Actually, this variant of the first-rater problem is also referred as the "cold-start problem" [24] in the literature. As extreme case of the first-rater problem, when a new recommender system starts, every user suffers from the first-rater problem for every item.

The grey-sheep problem [18] is another drawback associated with collaborative filtering methods. This problem refers to the users who have opinions that do not consistently agree or disagree with any group of users. As a consequence, such users do not receive recommendations. However, such problem does not occur in content-based methods, because such methods do not consider opinions acquired from other system users in order to make recommendations.

The commented problems have been treated in some works in the literature. A way of deal with sparsity and scalability problems consists on reducing the dimensionality of the database used for CF by means of a technique called Singular Value Decomposition (SVD) [47]. Barragáns-Martínez et al. [6] have adapted the proposal of Vozalis and Margaritis for a hybrid system combining content-based and CF approaches in the TV programs' recommendation domain. SVD allows increase the efficiency in the calculation of the similarities for the neighborhood formation used for generating recommendations. However, in spite of reducing the scalability problem, nearest neighbors methods are by themselves very time consuming and they require carrying out the similarity computation on-line (in recommender time), therefore they never can achieve the efficiency provided by model based CF methods that are induced off-line.

Cold-start problem has also been addressed in recent works. Most of them are focused on finding out new similarity metrics for the memory based CF approach since traditional measures such as Pearson's correlation and cosine provide poor recommendations when the available number of ratings is scarce, situation that becomes critical in the cases of the cold-start and first-rater problems. In [4] a

heuristic similarity measure based on the minute meanings of co-ratings is proposed in order to improve recommendation performance. Another similarity measure can be found in [10], which is a linear combination of simple similarity measures obtained by using optimization techniques based on neural networks.

A different approach is given in [30], where a hybrid recommendation procedure is proposed. It makes use of Cross-Level Association RulEs (CLARE) to integrate content information about domain items into collaborative filters. In that way cold-star problem can be solved by means of inducing user preferences from associations between a given item's attributes and other domain items when no recommendations for that item can be generated using CF.

Hybrid content based and CF approaches have also been applied to deal with the first rater problem. As a representative framework we can cite RSA (Fusion of Rough-Set and Average-category-rating) that integrates multiple contents and collaborative information to predict user preferences based on the fusion of Rough-Set and Average-category-rating [45].

4 Fuzzy Methods

Fuzzy approaches have been successfully used in many application areas; however, it has not been widely explored in recommender systems. In [7] fuzzy association rules were used in the tourism recommendation field. Fuzzy logic provides soft transitions between sets very suitable for tourism applications, where, a user is able, for example, to prefer a restaurant which is within a certain physical distance, but without having a fixed maximum distance. Apart from the advantages of this technique in particular domain, fuzzy logic can help minimizing, or even solving, typical drawbacks of such systems. Dubois et al. [22] affirm that fuzzy logic provides high-value properties to recover items stored in a database and, as a consequence, to provide recommendations for users. The reason is the capability of fuzzy sets to manage concepts such as similarity, preference and uncertainty in a unified way, while also performing approximate reasoning. Due to such advantages, especially for uncertainty, fuzzy logic can help to minimize the sparsity problem, which is the main drawback in current recommender systems.

Depending on the context and the type of method considered, fuzzy logic can be used both in content-based and collaborative filtering approaches. A general use of fuzzy logic with both types of methods is proposed in [22], where a case-based decision support system is implemented as the basis to contemplate situations where users do not have absolute preferences, or where preferences are expressed relatively to the context in order to be stored.

There are more specific works, such as [50], where fuzzy set methods are used to describe information in a content-based recommender system, or [15] where fuzzy concepts are used in order to recommend to users products not usually consumed. This work also addresses other situations in which there may not be enough information about the customer's past purchases and the customer may

give his specific requirements in each single purchase. In that context, fuzzy set operations are used in order to define relationships between user requirements and product features. On the other hand, fuzzy logic methods are applied in some works for developing recommendation approaches based on collaborative filtering. In [38] the notion of user session is defined as a compact temporary sequence of web accesses made by the user. These sessions are categorized using fuzzy partitions and, subsequently, recommendations are made in accordance to the categorized sessions. In [14] a comprehensive approach is established, which combines fuzzy set theory with Bayesian Networks in order to represent the ambiguity or vagueness in the description of opinions provided by users.

More recent approaches make use of fuzzy logic in combination with other techniques. For instance, in [23] semantic technologies and fuzzy logic are combined in a recommender system for investment portfolios. Recommendations are based on both psychological aspects of the investor and traditional financial parameters of the investments.

5 Recommendation Methodology

This section introduces a recommendation methodology [35, 36] that joins clustering, classification based on association and fuzzy sets. That combination aims at composing a hybrid method taking advantage of the strengths of both collaborative filtering and content-based approaches. In this way, the quality and effectiveness of the recommendations can be improved. A specific algorithm, CBA-Fuzzy, has been developed for mining the fuzzy association rules that constitute the associative classification model used for recommendation. The proposal takes into account the main drawbacks of recommender system and offers the appropriate mechanisms to minimize their effects. The methodology has two differentiated parts: the induction of the recommendation models, described in the section 3.1, and the recommendation process, presented in the section 3.2.

The first part process consist of two stages, generation of users' groups and rule set induction, which are carried out off-line, before the entry of the user in the system. The second part is responsible for classifying, at recommendation time, the active user in order to provide him personalized recommendations. In the next subsections we describe these components.

5.1 Building the Recommendation Models

The recommendation framework enclose aspects from collaborative filtering and content based methods since the recommendations to a specific user are made by comparing his preferences with the ones of other users but also taking into account features of users and products. These aspects are enclosed in the recommendation models which are built in two stages. The process is represented in figure 1 by means of two activity diagrams.

The first stage of the methodology consists on building groups of users with similar preferences and characteristics. In subsequent stages the active user is classified in one or more of these groups in order to make him recommendations according to his profile.

A clustering algorithm is applied to build the groups of users by using attributes containing demographic information about users such as age, postal code and level of education, and also attributes concerning items to be recommended, which users have rated or purchased. Additionally, users' past interactions with the system by means of implicit actions (such as time spent seeing an item and number of mouse clicks) may be taken into account. In that sense, this process may be considered as a collaborative filtering approach to provide recommendations. The information about user preferences comes from the transactions they have carried out in the system. The examples provided as input to the clustering algorithm (for instance k-means) are formed by these transactions and the corresponding attributes from users and items.

After applying the clustering algorithm, a set $G = \{g_1, g_2, g_3, \dots, g_N\}$ of users' groups is obtained, where N is a predefined number of groups which may be set according to the number of users and items available in a particular recommender system. The set G is provided as input to the next step, which is responsible for assigning an ordered list of items (or products) $P = \{p_1, p_2, p_3, \dots, p_m\}$ to each group g_i, where $i \in \{1,2,3, \dots, N\}$. The top items in each list will be the ones who better represent each group; therefore, an ordination criterion must be established. We consider the distance of each item to the centroid of the cluster, but other alternatives can also be taken account. For instance, the top items may be the ones who received better evaluation from the users of the group, or the most frequent ones (taking into account the number of purchases or given ratings) or any other criterion defined by an expert in the domain area involving the system. The ordered list of items assigned to the user groups will be supplied as input to the recommendation process, which constitutes the second part of the methodology.

The second stage in the construction of the recommendation models is the induction of the associative classification rules by means of the CBA-Fuzzy algorithm. Such algorithm was developed specifically for this methodology and constitutes its kernel given that it is responsible for generating the rules that compose the classification model employed for making recommendations. The main aspect of the algorithm is the combination of associative classification and fuzzy sets, which can provide important benefits. On the one hand, more reliable recommendations can be obtained since associative classification has a better behavior than other methods in sparse data contexts such as those from recommender systems where the number of rated products is insufficient to build the models [37]. On the other hand, the fuzzy rules allow the classification of the user in more than one group, dealing in this way with other important drawbacks of recommender systems, such as the gray sheep problem. Such rules will be responsible for classifying every new user at recommender time.

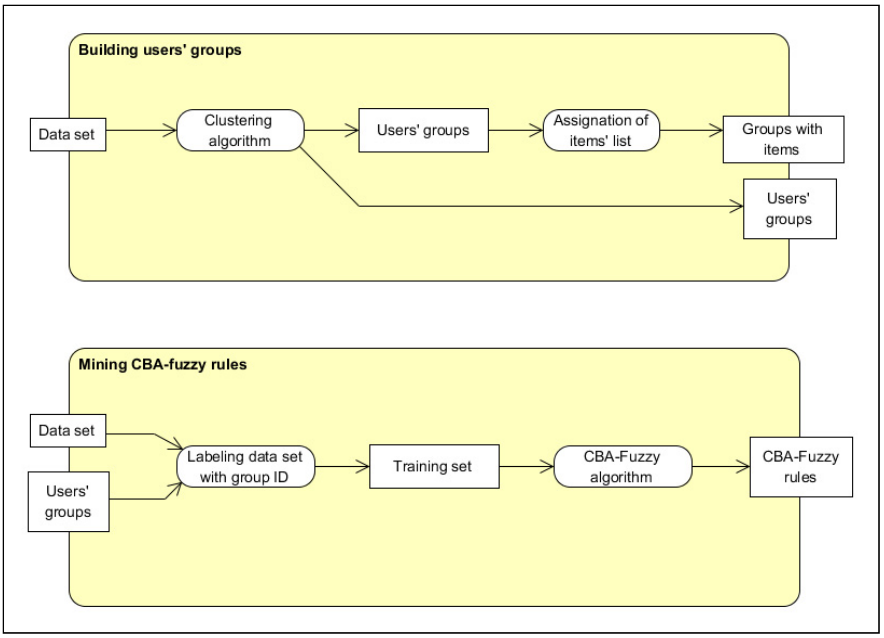


Fig. 1 Building the recommendation models

As showed in the second part of the figure 1, the rule generation process has two input sets: the groups of users provided as output by the clustering algorithm and the same input data set used for building the groups. The first activity for generating the list of classification rules is to combine the two inputs. At this point, the label attribute for classification is added to the training set. To do so, the values of the examples composing users' groups are compared to the ones of the training set in order to fulfill the label attribute. Therefore, each sample of the training set will have an identification corresponding to a group of users. In this way a new training set is created, which will be the input for the CBA-Fuzzy algorithm. The output provided by the algorithm will be a set of classification rules $R(g_i) = \{r_1, r_2, r_3, \ldots, r_p\}\ \forall g_i \in G$. Thus, the classification model will be composed of a set of class association rules available for each group of users.

Each classification rule encompasses a support and a confidence value. The confidence value expresses the degree of reliability of each rule. Therefore, before running the CBA-Fuzzy algorithm, a minimum threshold value for both measures (support and confidence) must be set up. It is recommended to set a high value for confidence and a low value for the support, especially in a scenario involving data from recommender systems, which are usually sparse and frequent itemsets might be less likely to occur.

5.2 CBA-Fuzzy Algorithm

The CBA-Fuzzy algorithm is an extension of the approach of the CBA algorithm proposed by Liu et al.[32], which is an associative classification method.

Association analysis was first introduced by Agrawal and col. [2, 3], who proposed the Apriori algorithm [1]. Afterward, the prolific research about association rules carried out has been mainly focused on simplifying the rule set and improving the algorithm performance. Associative classification was later introduced. A proposal of this category is the CBA (Classification Based on Association) algorithm [32] that consists of two parts, a rule generator based on Apriori for finding association rules and a classifier builder based on the discovered rules. CMAR (Classification Based on Multiple Class-Association Rules) [31] is another two-step method, however CMAR uses a variant of FP-growth instead of Apriori. Another group of methods, named integrated methods, build the classifier in a single step. CPAR (Classification Based on Predictive Association Rules) [51] is the most representative algorithm in this category. It is based on the algorithm FOIL (First Order Inductive Learner) proposed in [39]. FOIL learns rules to distinguish positive examples starting from negative examples.

As CBA, CBA-Fuzzy associative classification algorithm consists of two components: a rule generator (called CBA-RG) and a classifier builder (called CBA-CB). The rules' generator takes as basis the well known Apriori algorithm [3], hence the rules are generated from the so-called "frequent itemsets" that satisfy a minimum support threshold. Given that the mined rules are used for classification they must be "class association rules".

On the other hand, the classifier builder component is responsible for producing a classifier out of the whole set of rules, which involves pruning and evaluating all possible rules. Pruning is also done in each subsequent pass of the rule generator. It uses the pessimistic error rate based pruning method proposed by Quinlan [40] for the C4.5 algorithm.

The integration of fuzzy logic features in the CBA algorithm consists on changing the data input format in order to deal with fuzzy values and the calculation of the support and confidence measures. The original LUCS-KDD CBA algorithm limits the input data to have only discrete numbers on attribute values and, in addition, they have to be ordered sequentially. Hence, the algorithm requires a great pre-processing effort of input data, contrary to the CBA-Fuzzy algorithm that accepts any type of attribute value, even continuous or categorical attributes. To avoid the pre-processing step, the algorithm includes dicretization and fuzzyfication processes for continuous attributes. The discretization process for numerical attributes can be done automatically by CBA-Fuzzy either using the equal-width approach, where samples are divided into a set $V = \{v_1, v_2, v_3, \ldots, v_n\}$ of N intervals of the same length, or using the equal-depth approach, where the attribute range is divided into intervals containing approximately the same number of samples (same frequency).

The general workflow of the CBA-Fuzzy algorithm is shown in algorithm 1, where "*D*" is the dataset used as input for the algorithm (training set) and D*f* the dataset after the fuzzyfication process.

The line 1 of the algorithm represents the formation of the dataset "D" from an input data file. The following lines correspond to the discretization process. The second input parameter of line 2 represents the type of discretization the analyst wants to perform. Hence, the analyst can set up the number of intervals and the type of discretization he finds more suitable. In lines 4 and 6, the appropriate membership function used to perform the fuzzyfication process is applied according to the type of discretization selected by means of the parameter "type". In order to calculate the membership values of a sample of a discretized dataset using the equal-width approach, a triangular membership function (three parameters) is used. For the datasets discretized using the equal-depth approach, a trapezoidal membership function (four parameters) is used, because in this case some intervals are wider than others and, therefore, they encompass a region with a constant value defining an exclusive membership. During the fuzzification process, one or two membership values are assigned to each sample of the dataset, because each sample may belong to one or two intervals at the same time. The assignment of the membership value(s) depends on the proximity of the sample value to the interval range. Line 8 embodies the application of CBA-Fuzzy rule generator (CBAFuzzy-RG) to the fuzzified data and line 9 the building of the classifier by means of the classifier builder (CBA-CB).

Algorithm 1 CBA-Fuzzy's workflow

Algorithm 1 CBA-Fuzzy's workflow

1. D = processInput($inputFile$);
2. V = discretize(D, $type$, N);
3. **if** ($type$ = "equal-width")
4. **then** D_f = applyFuzzyTriang(D, V);
5. **else if** ($type$ = "equal-depth")
6. **then** D_f = applyFuzzyTrap(D, V);
7. **end if**
8. CR_s = CBAFuzzy-RG(D_f);
9. CRM = CBA-CB(CR_s);

The CBA-Fuzzy rule generation process differs from the "crisp version of CBA" in the calculation of the support and confidence measures. Instead of calculating the support of an item by counting the number of transactions in which it appear (summing 1 each time the interval it belongs to appears), the CBA-Fuzzy considers partial memberships by summing continuous values between 0 and 1 each time an interval owning (totally or partially) the item appears.

5.3 Recommendation Process

The models built previously are used for recommending items to the active user when he is on-line. Firstly, the model of class association rules is required to classify the active user and predict in this way the group or groups he belongs to. Since preferences may change as time goes by, the most recent interaction data of the active user is taken to do the classification. Therefore, data gathered from the active user's last interaction with the system, represented by a transaction "y", is checked against the rules' set. This transaction has the same attributes of those used to build groups of users, thus, the provided recommendations will be well-suited to his current preferences. Moreover, as we are using past information of the active user in order to provide him recommendations, in this context, the proposed methodology may be considered as a content-based approach. Figure 2 shows an activity diagram of this stage that is carried out at runtime, when the user is interacting with the system.

The last transaction of the active user and the model of class association rules obtained in the previous stage, are supplied as input to the recommender process. The process starts by selecting the set $R_c = \{r_1, r_2, r_3, \ldots, r_N\}$ of N rules satisfying the condition that all antecedent terms' values are matched by the user's last transaction attribute values. In the case of continuous attributes, a partial membership to the interval is considered as a match. If there are no rules (or very few ones) respecting this condition we take into account the downward closure property of association rules' support, which guarantees that for a frequent itemset, all its subsets are also frequent. In this way, we decrease the size of the itemset and successively verify if there are rules matching the condition stated before, in order to decrease the itemset's size until suitable rules are found. In fact, several authors like Toivonen et al. [46] and Liu et al. [33], argue that, usually, the more general rules are (the ones which encompass less terms), the more relevant and less ambiguous they are.

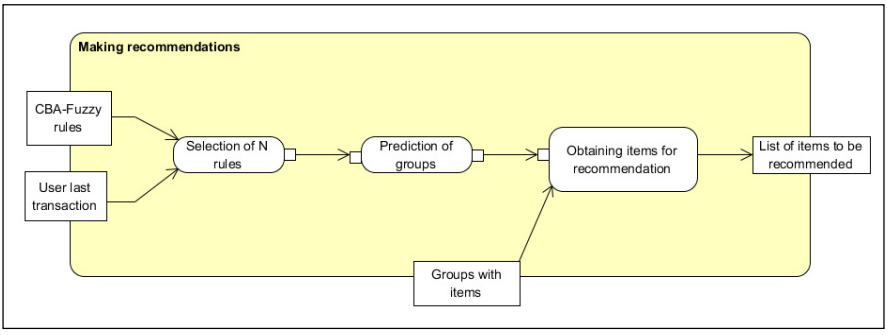


Fig. 2 Recommendation process

After obtaining R_c, the values of the label attribute (consequent term) of the rules in R_c are considered in order to obtain the possible groups (predicted groups)

to which the active user owns to. Then, his membership function to every class (group of users) found on R_c's rules consequent terms is calculated. To do so, a discriminator function "g" is defined in order to calculate the degree of truth that the active user owns to every class found in R_c. Considering that the active user is represented by a transaction "y" and "h" represents a group of users, the discriminator function can be calculated by means of the following formula:

$$g_h(y) = \sum_{1 \le k \le M, i = C_h} \prod_{j=1}^{l_k} B(j,k)[X(j,k)(y)]$$

where l_k is the number of terms (attributes) in each rule, $X(j,k)(y)$ the value taken by the attribute $X(j,k)$ in the sample "y" and $F(j,k)[X(j,k)(y)]$ its degree of membership. Hence, such function calculates the product of attributes' degrees of membership for all the rules in R_c and then sum of the results obtained in each rule.

When the discriminator values for each class (or group of users) are obtained, the system compares them in order to find the greatest. In this sense, this method is different from most fuzzy associative classification approaches since they usually predict just one class label for a given instance. Nevertheless, this procedure can consider several classes, which are the ones satisfying a minimum discriminator threshold previously established. This way, there can be "t" groups of users related to the active user. Once the active user is classified, the recommender process makes use of the sets of items assigned to the users' groups, which are generated by the processes of induction of the models and provided as input to the second part in charge of making recommendations. Given that each group is associated to a list of items and each user is associated to one or several groups, the recommendation presented to the active user is a suggestion involving the "n" best ranked items from the corresponding lists. In order to have a constant number of recommended items, "n" will be inversely proportional to "t", therefore, the more classes there are the less items are considered in each list.

In case of the active user has not done any transaction, the recommender procedure considers just the user attributes by comparing the values of such attributes with the ones of the groups. In this way, the more suitable group to the user is found and then the most accessed item in such group is verified in order to compose the transaction "y" and continue the recommender process as usual.

On the other hand, if the active user data is in the training set (he belong to a group), the classification rules are not necessary since he is already classified.

6 Conclusions

This chapter is focused on recommender systems, one of the most productive application areas of CRM (Customer Relationship Management) where strategical techniques of Business Intelligence are used for increasing customer satisfaction by means of the recommendation of products or services he is interested in. The main methodologies used in recommender systems are classified and described in the chapter jointly with their principal drawbacks.

Data mining techniques can be used to address some of these problems, however, used in a solely way, they cannot overcome some limitations presented by current recommender systems. A recent trend is to exploit hybrid methodologies combining several approaches in order to take advantage of the strengths of each of them. The chapter presents one of the most promising strategies consisting of joining data mining and fuzzy logic. In order to illustrate it, a specific hybrid methodology of such type is described in detail. The methodology has been validated in a tourism recommender system, the PSiS system [19], developed for aiding tourists to find a personalized tour plan in the city of Oporto, Portugal.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. of 20th Int. Conference on Very Large Databases, Santiago de Chile, pp. 487–489 (1994)
2. Agrawal, R., Imielinski, T., Swami, A.: Database mining. A performance perspective. IEEE Trans. Knowledge and Data Engineering 5(6), 914–925 (1993a)
3. Agrawal, R., Imielinski, T., Swami, A.: Mining associations between sets of items in large databases. In: Proc. of ACM SIGMOD Int. Conference on Management of Data, Washinton, D.C, pp. 207–216 (1993b)
4. Ahn, H.J.: A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Information Sciences 178, 37–51 (2008)
5. Avery, C., Zeckhauser, R.: Recommender systems for evaluating computer messages. Communication ACM 40(3), 88–89 (1997)
6. Barragáns-Martínez, A.B., Costa-Montenegro, E., Burguillo, J.C., Rey-López, M., Mikic-Fonte, F.A., Peleteiro, A.: A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value descomposition. Information Sciencies 180, 4290–4311 (2010)
7. Berka, T., Plößnig, M.: Designing recommender systems for tourism. In: The 11th International Conference on Information Technology in Travel and Tourism (2004)
8. Bilsus, D., Pazzani, M.J.: Learning collaborative information filters. In: 15th International Conference in Machine Learning, Bari, Italy, pp. 46–54. Morgan Kaufmann (1998)
9. Blanco-Fernández, Y., Pazos-Arias, J.J., López-Nores, M., Gil-Solla, A., Ramos-Cabrer, M., Garcia-Duque, J., Fernández-Vilas, A., Díaz-Redondo, R.: Incentivized provision of metadata, semantic reasoning and time driving filtering: Making a puzzl of personalized e-comerce. Expert Systems and Applications 37, 61–69 (2010)
10. Bobadilla, J., Ortega, F., Hernando, A., Bernal, J.: A collaborative filtering approach to mitigate the new user cold start problem. Knowledge-Based Systems 26, 225–238 (2012)
11. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pp. 43–52 (1998)
12. Burke, R.: Hybrid recommender systems: survey and experiments. User Modeling and User-Adapted Interaction 12(4), 331–370 (2002)

13. Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: A collaborative recommender system based on probabilistic inference from fuzzy observations. Fuzzy Sets Syst. 159(12), 1554–1576 (2008)
14. Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Rueda-Morales, M.A.: Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian Networks. International Journal of Approximate Reasoning 51, 785–799 (2010)
15. Cao, Y., Li, Y., Liao, X.-F.: Applying fuzzy logic to recommend consumer electronics. In: Chakraborty, G. (ed.) ICDCIT 2005. LNCS, vol. 3816, pp. 278–289. Springer, Heidelberg (2005)
16. Cho, H.C., Kim, J.K., Kim, S.H.: A Personalized Recommender System Based on Web Usage Mining and Decision Tree Induction. Expert Systems with Applications 23, 329–342 (2002)
17. Chou, P.-H., Li, P.-H., Chen, K.-K., Wu, M.-J.: Integrating web mining and neural network for personalized ecommerce automatic service. Expert Systems with Applications 37(4), 2898–2910 (2010)
18. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combining content-based and collaborative filters in an online newspaper. Technical report, Computer Science Department, Worcester, Massachusetts (1999)
19. Coelho, B., Martins, C., Almeida, A.: Adaptive tourism modeling and socialization system. In: 2009 International Conference on Computational Science and Engineering, vol. 4, pp. 645–652 (2009)
20. Condliff, M.K., Lewis, D.D., Madigan, D., Posse, C.: Bayesian mixed-effects models for recommender systems. In: Proceedings of SIGIR 1999 Workshop on Recommender Systems Algorithms and Evaluation, Berkeley, CA (1999), http://www.cs.umbc.edu/~ian/sigir99-rec
21. Diez, J., del Coz, J.J., Luaces, O., Bahamonde, A.: Clustering people according to their preference criteria. Expert Systems with Applications 34, 1274–1284 (2008)
22. Dubois, D., Hullermeier, E., Prade, H.: Fuzzy methods for case-based recommendation and decision support. Journal of Intelligent Information Systems 27(2), 95–115 (2006)
23. García-Crespo, Á., López-Cuadrado, J.L., González-Carrasco, I., Colomo-Palacios, R., Ruiz-Mezcua, B.: SINVLIO: Using Semantics and Fuzzy Logic to provide individual investment portfolio recommendations. Knowledge-Based Systems 27, 103–118 (2012)
24. Guo, H.: Soap: Live recommendations through social agents. In: Proc. of Fifth DELOS Workshop on Filtering and Collaborative Filtering, Budapest, November 10-12 (1997)
25. Huang, Z., Chen, H., Zeng, D.: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Trans. Information Systems 22(1), 116–142 (2004)
26. Ingoo, H., Kyong, J.O., Tae, H.R.: The collaborative filtering recommendation based on SOM cluster-indexing CBR. Expert Systems with Applications 25, 413–423 (2003)
27. Kim, H.N., Alkhaldi, A., El Saddik, A., Jo, G.S.: Collaborative user modeling with user-generated tags for social recommender systems. Expert Systems with Applications 38, 8488–8496 (2011)
28. Koren, Y.: Collaborative filtering with temporal dynamics. Communications of the ACM 53(4), 89–97 (2010)
29. Lee, C.H., Kim, Y.H., Rhee, P.K.: Web Personalization Expert with Combining collaborative Filtering and association Rule Mining Technique. Expert Systems with Applications 21, 131–137 (2001)

30. Leung, C.W., Chan, S.C., Chung, F.: An empirical study of a cross-level association rule mining approach to cold-start recommendations. Knowledge-Based Systems 21, 515–529 (2008)
31. Li, W., Han, J., Pei, J.: CMAR. Accurate and efficient classification based on multiple class-association rules. In: Proc. of the IEEE International Conference on Data Mining (ICDM 2001), California, pp. 369–376 (2001)
32. Liu, B., Hsu, W., Ma, Y.: Integration classification and association rule mining. In: Proc. of 4th Int. Conference on Knowledge Discovery and Data Mining, pp. 80–86 (1998)
33. Liu, B., Hsu, W., Ma, Y.: Pruning and summarizing the discovered associations. In: KDD 1999: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 125–134. ACM, New York (1999)
34. Loshin, D.: Business Intelligence. The Savvy Manager's Guide, Elsevier Inc. (2003)
35. Lucas, J.P.: Métodos de clasificación basados en asociación aplicados a sistemas de recomendación. PhD. Thesis, University of Salamanca (2010)
36. Lucas, J.P., Laurent, A., Moreno, M.N., Teisseire, M.: A fuzzy associative classification approach for recommender systems. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 20(4), 579–617 (2012a)
37. Lucas, J.P., Segrera, S., Moreno, M.N.: Making use of associative classifiers in order to alleviate typical drawbacks in recommender systems. Expert Systems with Applications 39(1), 1273–1283 (2012b)
38. Nasraoui, O., Frigui, H., Joshi, A., Krishnapuram, R.: Mining web access logs using relational competitive fuzzy clustering. In: Proceedings of the Eight International Fuzzy Systems Association World Congress (1999)
39. Quinlan, J.R., Cameron-Jones, R.M.: FOIL: A midterm report. In: Brazdil, P.B. (ed.) ECML 1993. LNCS, vol. 667, pp. 3–20. Springer, Heidelberg (1993)
40. Quinlan, R.J.: C4.5: Programs for Machine Learning. Morgan Kaufmann Series in Machine Learning. Morgan Kaufmann (1993)
41. Resnick, P., Iacovou, N., Suchack, M., Bergstrom, P., Riedl, J.: Grouplens: An open architecture for collaborative filtering of netnews. In: Proc. of ACM CSW 1994 Conference on Computer. Supported Cooperative Work, pp. 175–186 (1994)
42. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based Collaborative Filtering Recommendation Algorithm. In: Proceedings of the Tenth International World Wide Web Conference, pp. 285–295 (2001)
43. Schafer, B.J.: The application of data-mining to recommender systems. In: Wang, J. (ed.) Encyclopedia of Data Warehousing and Mining. Information Science Publishing (2005)
44. Schafer, J.B., Konstant, J.A., Riedl, J.: E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery 5, 115–153 (2001)
45. Su, J.H., Wang, B.W., Hsiao, C.Y., Tseng, V.S.: Personalized rough-set-based recommendation by integrating multiple contents and collaborative information. Information Sciences 180, 113–131 (2010)
46. Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., Mannila, H.: Pruning and grouping discovered association rules. In: MLnet Workshop on Statistics, Machine Learning, and Discovery in Databases, pp. 47–52 (1995)
47. Vozalis, M.G., Margaritis, K.G.: Applying SVD on item-based filtering. In: Proceedings of the 5th International Conference on Intelligent Systems Design and Applications (ISDA 2005), pp. 464–469 (2005)

48. Wang, A., Yuan, W., Liu, J., Yu, Z., Li, H.: A novel pattern recognition algorithm: Combining ART network with SVM to reconstruct a multi-class classifier. Computers and Mathematics with Applications 57, 1908–1914 (2009)
49. Xu, J.A., Araki, K.: A SVM-based personal recommendation system for TV programs. In: Proc. Int. Conf. on Multi-Media Modeling Conference, Beijing, China, pp. 401–404 (2006)
50. Yager, R.R.: Fuzzy logic methods in recommender systems. Fuzzy Sets Syst. 136(2), 133–149 (2003)
51. Yin, X., Han, J.: CPAR. Classification based on predictive association rules. In: SIAM International Conference on Data Mining (SDM 2003), pp. 331–335 (2003)

Fuzzy Rationality Implementation in Financial Decision Making

N.D. Nikolova and K. Tenekedjiev

Abstract. Expected utility theory is by far the best normative theory for decision making under uncertainty. It helps the decision maker find the proper balance between expected profits and risks, and has been acknowledged as a key approach to rational economic behavior of individuals. The whole measurement process in the expected utility theory is based on the solution of preferential equations, with the help of which utilities and probabilities are being elicited. However, the resulting estimates are in an interval form, which disobeys some main rationality assumptions of the theory. Therefore, fuzzy rational decision analysis in introduced as a way to unify the normative rationality with the fuzziness of real preferences. This chapter outlines a series of practical techniques dealing with the interval nature of assessed utility and probability measures, using the intrinsic optimism-pessimism attitude of the DM. Main preference-related and uncertainty-related problems are stressed.

1 Introduction

Personal financial decisions need fine balance between expected profits and risks. One of the best normative theories for decision making under uncertainty is the expected utility theory (EUT), which requires adequate evaluation of the probabilistic estimates and of the utility function. The core of EUT is utility theory, which argues that the rational decision maker (DM) should choose the alternative that best meets her preferences and risk attitude. Utility theory has been introduced and widely acknowledged as the principle of rational economic decision making. Some of its basic postulates are commonly explained in the comprehensive setup of financial and economic decisions. Monetary consequences are generally used to interpret the risk attitude of individuals, since in such a setup people tend to perform is rather similar way. It is thus only natural that the implementation area of

N.D. Nikolova · K.Tenekedjiev
N. Vaptsarov Naval Academy, Varna, Bulgaria. 73 Vassil Drumev Str.,
9026 Varna, Bulgaria
e-mail: natalianik@gmail.com, Kiril.Tenekedjiev@fulbrightmail.org

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_23,

EUT is in major investment projects, where the complexity of the rational decision analysis is well awarded by high financial return of investment in the long-term perspective [10]. Therefore it is only possible to conclude that being an approach to rational decision making, EUT is also in the core of business intelligence.

Alternatives under risk are presented as lotteries, each giving one out of a set X of consequences (prizes) with known probability. Since preferences are measured by the utility function $u(.)$, then rational choice is brought down to calculating the expected utility of each lottery (i.e. utilities of prizes weighted by their probabilities) and choosing the lottery with the highest expected utility.

The whole measurement process in the EUT is based on the solution of preferential equations. The relative preference of the DM over the objects in those equations can be described by three fuzzy sets [19]. As a result each preferential equation has an interval rather than a point estimate as a solution, which breaches some of the rationality requirements of EUT. Thus ideal rationality becomes a fiction, and only bounded rationality can be achieved by the real DM, also referred to as fuzzy rationality (the DM is called respectively a fuzzy-rational decision maker – FRDM) [19]. The theory of fuzzy rationality is a generalization of EUT, which tries to unify the normative rationality with the fuzziness of real preferences in the measurement process.

In this chapter, we outline a series of practical techniques dealing with the interval nature of assessed utility and probability measures, using the intrinsic optimism-pessimism attitude of the DM.

The main preference-related problems stressed in this chapter are: a) Analytical approximation of the one-dimensional utility function under monotonic fuzzy-rational preferences; b) Identification of the extremum interval and construction of one-dimensional utility function under two types of non-monotonic fuzzy-rational preferences; c) Testing the unity of scaling constants sum, which determines the form of the multi-dimensional utility function in the case of fuzzy rationality.

The main uncertainty-related problems stressed in this chapter are: a) Construction of ribbon distributions of discrete and continuous variables; b) Introduction of fuzzy-rational lotteries as models of alternatives with partially described uncertainty (which are generalizations of the classical-risky and the strict uncertainty lotteries); c) Approximation of fuzzy-rational lotteries by classical-risky lotteries (called Q-lotteries), which account for the optimist-pessimist attitude of the FRDM by means of strict uncertainty decision criteria (Q criteria).

2 Constructing the Utility Function

2.1 The One-dimensional Case

Any choice of action within a problem situation is expected to meet the objectives of the DM, which also predefine the structure of the consequences. In the simplest case, there is only one objective (e.g. in economic decisions that would be maximization of profit) measured by a single attribute (e.g. in economic decisions that would be the net profit), and the preferences of the DM are strictly increasing

(e.g. monetary prizes). If the attribute is continuous then the set of prizes X would be a one-dimensional (1-D) bounded interval of prizes.

If X is a 1-D piece-wise continuous set of prizes, then the smallest prize is $x_{min} = \inf(x)$, whereas the highest one would be $x_{max} = \sup(x)$. The preferences of the DM should be measured by the utility function that is increasing on the subject's preferences. The 1-D utility function $u(.)$ may be constructed in the interval $[x_{min}, x_{max}]$. To accommodate the utility of prizes in $[x_{min}, x_{max}]$ that are not in X, it is convenient to put $u(x) = 0$ for $x \in (-\infty; x_{min}) \cup (x_{max}; +\infty)$. As it is impossible to elicit the utilities of all prizes in $[x_{min}; x_{max}]$ it is reasonable to elicit only z number of nodes with coordinates $(x_{u_l}, u_l), l = 1,2, \dots, z$, where x_{u_l} and u_l are respectively the utility quantile and the utility quantile index. The utility elicitation techniques solve preferential equations by changing one parameter until compared options (prizes and/or lotteries) become indifferent [8]. Classical elicitation methods are the probability equivalence (PE) [4], certainty equivalence (CE) [1], and lottery equivalence (LE) [15]. Modern techniques are the trade-off (TO) [35], and the uncertain equivalence (UE) [30] methods.

A next step would be to construct the utility function, which might be performed using an analytical function $u = u(x, \vec{\boldsymbol{P}})$, where $\vec{\boldsymbol{P}}$ is an n-dimensional vector of unknown parameters. The method should precisely interpret the data, and should preserve the risk attitude of the DM [10, 28], modeled by the local risk aversion function $r(x) = -u''(x)/u'(x)$ [13, 23].

The literature [13, 27] proposes many analytical forms of the utility function for different prize sets and risk attitude. The work [16; 29] proposes the dependence $arctan(ax - ax_0)$, $a > 0$. It suits to a set X that contains gains and losses, because the form of its $r(.)$ corresponds to the most typical risk attitude.

The analytical approximation of $u(.)$ applies when there is a small number of elicited nodes and/or when the elicited uncertainty intervals are too wide. The analytical construction of $u(.)$ is also in position to filter the error in the subjective estimates of utilities in case the selected mathematical form can properly describe the risk attitude of the FRDM. If the optimal approximated curve passes through the uncertainty intervals of the elicited nodes of $u(.)$, then the imprecision of the elicitation would be significantly reduced. If the optimal approximated curve significantly deviates from the uncertainty intervals, then the analytical approximation must be replaced by linear interpolation.

A more complex situation arises when the DM has quasi-unimodal preferences, i.e. when there is a value x_{opt} with extreme utility (either a minimum or a maximum) within the interval of prizes [20]. There are two types of quasi-unimodal preferences – hill (with a maximum extremum) and valley (with a minimum extremum) preferences. Both occur due to two contradicting factors related to the analyzed variable. Difficulties arise when the DM has to compare values on both sides of the extreme interval – if a sufficient difference in utilities exists, the DM would be able to state preference otherwise she would be indifferent being unable to compare the options. This leads to mutual non-transitivity of preferences and even a very motivated and rational DM would express fuzzy, rather than classical

rationality. As a result, she would identify an extreme interval rather than a single value x_{opt}. Since all elicitation techniques need the reference points (the prizes with extreme utility), identifying the extreme interval is mandatory.

The models of hill and valley utility functions are based on two separate sets of assumptions that refer to the discriminating abilities of the FRDM and the characteristics of the extreme interval. Two 20-step algorithms are elaborated (one per each type of quasi-unimodal preferences), which find the extreme interval via a dialog with the FRDM [20]. Both algorithms combine the golden section search [14] and bisection [25] methods.

After the extreme interval has been identified, the next step is to construct local utility functions in the sections with monotonic preferences (the sections on each side of the extremum) using classical techniques. Finally, it is possible to construct the global utility function over the entire set of prizes by rescaling the local functions [20].

2.2 The Multi-dimensional Case

As a result of the multiple objectives that a DM has in a decision problem, consequences are usually defined as multi-dimensional vectors, whose attributes $X_1, X_2, \dots, X_d$ measure the degree to which objectives are met. If the *i*-th attribute is a random variable X_i with an arbitrary realization x_i, then prizes take the form of *d*–dimensional vectors $\vec{\boldsymbol{X}} = (x_1, x_2, \dots, x_d)$ in the set of prizes, which is a subset of the *d*-dimensional Euclidean space. The set $\{X_1, X_2, \dots, X_d\}$ may be divided into $n \in \{2, 3, \dots, d\}$ non-empty non-overlapping subsets $Y_1, Y_2, \dots, Y_n$, called fundamental vector attributes, each with an arbitrary realization $\vec{\boldsymbol{y}}_j$, which implies that $\vec{\boldsymbol{X}} = (\vec{\boldsymbol{y}}_1, \vec{\boldsymbol{y}}_2, \dots, \vec{\boldsymbol{y}}_n)$ [17]. The most preferred value of the fundamental vector attribute Y_j is $\left(\vec{\boldsymbol{y}}_j\right)_{best}$, whereas the least preferred value is $\left(\vec{\boldsymbol{y}}_j\right)_{worst}$. The work [13] proposes an algorithm to construct the multi-dimensional $u(.)$ over the multi-dimensional prizes $\vec{\boldsymbol{X}} \in \mathbb{R}^d$ as $u = u\left(\vec{\boldsymbol{X}}\right) = u(x_1, x_2, \dots, x_d)$. However, they do recommend the algorithm in the last resort, since it is practically impossible to realize when $d > 3$. A much better approach is to use an utility function of the kind

$$u\left(\vec{\boldsymbol{X}}\right) = u(\vec{\boldsymbol{y}}_1, \vec{\boldsymbol{y}}_2, \dots, \vec{\boldsymbol{y}}_n) = f\left[u_{y,1}(\vec{\boldsymbol{y}}_1), u_{y,2}(\vec{\boldsymbol{y}}_2), \dots, u_{y,n}(\vec{\boldsymbol{y}}_n)\right] \quad (1)$$

The adequacy of (1) requires that certain independence conditions of preferences over the attributes hold – preferential, utility and additive independence. Preferential independence is most common and the weakest independence condition. The strongest condition is additive independence, but also very hard to establish, and thus difficult to use in practice. Most important from a practical point of view is mutual utility independence [3], which allows to represent the *d*-dimensional utility function $u(.)$ as a polynomial of *n* utility functions $u_{y,j}(.)$ over the fundamental vector attributes [7, 9]:

$$
\begin{aligned}
u &= u(\vec{\boldsymbol{X}}) = u(\vec{\boldsymbol{y}}_1, \vec{\boldsymbol{y}}_2, \dots, \vec{\boldsymbol{y}}_n) \\
&= \sum_{j=1}^{n} k_{y,j} u_{y,j}(\vec{\boldsymbol{y}}_j) + K_y \sum_{j=1}^{n-1} \sum_{s=j+1}^{n} k_{y,j} k_{y,s} u_{y,j}(\vec{\boldsymbol{y}}_j) u_{y,s}(\vec{\boldsymbol{y}}_s) \\
&+ K_y^2 \sum_{j=1}^{n-2} \sum_{s=j+1}^{n-1} \sum_{r=s+1}^{n} k_{y,j} k_{y,s} k_{y,r} u_{y,j}(\vec{\boldsymbol{y}}_j) u_{y,s}(\vec{\boldsymbol{y}}_s)\, u_{y,r}(\vec{\boldsymbol{y}}_r) + \cdots \\
&+ K_y^{n-1} k_{y,1} k_{y,2} \dots k_{y,n} u_{y,1}(\vec{\boldsymbol{y}}_1) u_{y,2}(\vec{\boldsymbol{y}}_2) \dots u_{y,n}(\vec{\boldsymbol{y}}_n)
\end{aligned}
\tag{2}
$$

Here, $u_{y,j}(.)$ are d_j-dimensional bounded utility functions over all possible values $\vec{\boldsymbol{y}}_j$ of Y_j. The functions $u(.)$ and $u_{y,j}(.)$ are normalized so that $u_{y,j}(\vec{\boldsymbol{y}}_{j,best}) = 1$, $u_{y,j}(\vec{\boldsymbol{y}}_{j,worst}) = 0$, for $j = 1,2,\dots,n$, $u(\vec{\boldsymbol{X}}_{best}) = u(\vec{\boldsymbol{y}}_{1,best}, \vec{\boldsymbol{y}}_{2,best}, \dots, \vec{\boldsymbol{y}}_{n,best}) = 1$, $u(\vec{\boldsymbol{X}}_{worst}) = u(\vec{\boldsymbol{y}}_{1,worst}, \vec{\boldsymbol{y}}_{2,worst}, \dots, \vec{\boldsymbol{y}}_{n,worst}) = 0$. If $d_j > 1$ then each $u_{y,j}(.)$ has to be additionally decomposed [5], otherwise $u_{y,j}(.)$ has to be directly constructed. The values $k_{y,j} \in [0;1]$ are scaling constants that indicate the relative significance of each fundamental vector attribute for the preferences of the DM over the multi-dimensional prizes, and k_y is a general constant that depends on $k_{y,j}$. The $k_{y,j}$ correspond to the utility of the corner consequences $\vec{\boldsymbol{X}}_{\boldsymbol{j,corner}} = \left[(\vec{\boldsymbol{y}}_1)_{worst}, (\vec{\boldsymbol{y}}_2)_{worst}, \dots, (\vec{\boldsymbol{y}}_j)_{best}, \dots, (\vec{\boldsymbol{y}}_n)_{worst}\right]$, thus $k_{y,j} = u\left[(\vec{\boldsymbol{y}}_1)_{worst}, (\vec{\boldsymbol{y}}_2)_{worst}, \dots, (\vec{\boldsymbol{y}}_j)_{best}, \dots, (\vec{\boldsymbol{y}}_n)_{worst}\right]$, and $1 + k_y = \prod_{j=1}^{n}(k_y \times k_{y,j} + 1)$. Scaling constants are subjectively elicited [17], and take the interval form $k_{y,j} \in [\hat{k}_{y,j}^{d}; \hat{k}_{y,j}^{u}], j = 1,2,\dots,n$. The form of the multi-dimensional $u(.)$ depends on the sum of $k_{y,j}$, which calls for point estimates $\hat{k}_{y,j}$, for $j = 1,2,\dots,n$. The uniform method has been proposed to solve this problem in an analytical procedure [17], numerical approximation [33], and Monte-Carlo based simulation approximation [17]. For the non-trivial case, where $a_n = \sum_{j=1}^{n} k_{y,i}^{d} < 1$ and $b_n = \sum_{j=1}^{n} k_{y,i}^{u} > 1$, it is proposed to use a two-tail statistical test for unit value of the sum of scaling constants [12]. Then

$$
u(\vec{\boldsymbol{X}}) = \begin{cases} \sum_{j=1}^{n} k_{y,j} u_{y,j}(\vec{\boldsymbol{y}}_j), & if \sum_{j=1}^{n} k_{y,j} = 1 \\ \left(\prod_{j=1}^{n} \left[K k_{y,j} u_{y,j}(\vec{\boldsymbol{y}}_j) + 1 \right] - 1 \right) \Big/ K, & if \sum_{j=1}^{n} k_{y,j} \neq 1 \end{cases}
\tag{3}
$$

Finally, the overall multi-dimensional utility function of a FRDM may be constructed using an algorithm, proposed in [21].

3 Uncertainty and Distributions

The uncertainty in real-life problems is associated with random variables (discrete, continuous or mixed). It can be quantitatively measured in terms of probability distributions. The classical forms of probability distributions [24] apply for an ideal DM, who gives precise probability estimates. The FRDM only defines interval probability measures, which causes the introduction of ribbon distribution functions of various types [31].

A discrete random variable X only takes one of the possible values $x_1, x_2, \dots, x_r, \dots, x_t (x_1 < x_2 < \dots < x_r < \dots < x_t)$. In the case of fuzzy rationality, the 1-D ribbon discrete probability function (DPF) $f_d^R(.)$ partially quantifies the associated uncertainty and is known to lie entirely between the lower $P^d(.)$ and upper $P^u(.)$ distributional bounds: $P^d(x_r) \le f_d^R(x_r) \le P^u(x_r), r = 1,2, \dots, t$, and $0 \le P^d(x_r) \le P^u(x_r) \le 1$, $\sum_{r=1}^{t} P^d(x_r) \le 1 \le \sum_{r=1}^{t} P^u(x_r)$.

In the case of fuzzy rationality, a 1-D ribbon cumulative distribution function (CDF) $F^R(.)$ partially quantifies the uncertainty in a random variable X, and is known to entirely lie between the lower and upper border functions $F_d(.)$ and $F_u(.)$, i.e. $F_d(x) \le F^R(x) \le F_u(x)$, for $x \in (-\infty; +\infty)$. As a result of the elicitation process, a set of elicited nodes is available in two forms:

a) with an uncertainty interval for the quantile (error on the abscissa x), i.e. $\{(x_{d,l}; x_{u,l}; F_l)/l = 1,2, \dots, z\}$, where $x_{d,1} \le x_{d,2} \le \dots \le x_{d,z}$, $x_{u,1} \le x_{u,2} \le \dots \le x_{u,z}, x_{d,1} = x_{u,1}, x_{d,z} = x_{u,z}$, and $0 = F_1 \le F_2 \le \dots \le F_z = 1$. The resulting x-ribbon $F^{xR}(.)$ and its lower and upper x-bounds $F_{xd}(.)$ and $F_{xu}(.)$ are constructed via linear interpolation on the margins of the nodes:

$$F_{xd}(x) = \begin{cases} 0 & x < x_{d,1} \\ F_l & x_{d,l} = x < x_{d,l+1}, l = 1,2, \dots, z-1 \\ F_l + \left[(x - x_{d,l})(F_{l+1} - F)\right]/(x_{d,l+1} - x_{d,l}), \dots & \\ \dots x_{d,l} < x < x_{d,l+1}, l = 1,2, \dots, z-1 & \\ 1 & x_{d,z} \le x \end{cases}$$

$$F_{xu}(x) = \begin{cases} 0 & x < x_{u,1} \\ F_l & x_{u,l} = x < x_{u,l+1}, l = 1,2, \dots, z-1 \\ F_l + \left[(x - x_{u,l})(F_{l+1} - F_l)\right]/(x_{u,l+1} - x_{u,l}), \dots & \\ \dots\ x_{u,l} < x < x_{u,l+1}, l = 1,2, \dots, z-1 & \\ 1 & x_{u,z} \le x \end{cases} \quad (4)$$

$$F_{xd}(x) \le F^{xR}(x) \le F_{xu}(x)$$

b) with an uncertainty interval for the quantile index (error on the probability), i.e. $\{(x_l; F_{d,l}; F_{u,l})/l = 1,2, \dots, z\}$, where $x_1 \le x_2 \le \dots \le x_z$, $0 = F_{d,1} \le F_{d,2} \le \dots \le F_{d,z} = 1,\ 0 = F_{u,1} \le F_{u,2} \le \dots \le F_{u,z} = 1,\ F_{d,l} \le F_{u,l}$,

for $l = 2,3,\dots,z-1$. The resulting p-ribbon $F^{p^R}(.)$, and its lower and upper p-bounds $F_{pd}(.)$ and $F_{pu}(.)$ are constructed by analogy to a), via linear interpolation on the margins of the nodes.

In a multi-dimensional setup, the uncertainty, associated with a d-dimensional system of random variables $(X_1, X_2, \dots, X_d)$, is partially quantified by a d-dimensional ribbon CDF $F^R(.)$ that is known to lie entirely between two d-dimensional bounds (classical CDF) $F^d(.)$ and $F^u(.)$ (in line with the discussion in [34]). If $\vec{\boldsymbol{X}} = (x_1, x_2, \dots, x_d)$ is a d-dimensional vector with random fixed values, then for all $\vec{\boldsymbol{X}} \in \mathbb{R}^d, F^d(\vec{\boldsymbol{X}}) \le F^R(\vec{\boldsymbol{X}}) \le F^u(\vec{\boldsymbol{X}}), F^d(\vec{\boldsymbol{X}}) \le F^u(\vec{\boldsymbol{X}})$.

If consequences are described by $Y_1, Y_2, \dots, Y_n$, then a ribbon vector argument CDF (ribbon VACDF) is defined as a product of d_j-dimensional ribbon CDF of $Y_1, Y_2, \dots, Y_n$. If the event $A_j = X_1^{Y_j} \le x_1^{y_j} \cap X_2^{Y_j} \le x_2^{y_j} \cap \dots \cap X_{d_j}^{Y_j} \le x_{d_j}^{y_j}$, $j = 1,2,\dots,n-1$, then $F^R(\vec{\boldsymbol{X}}) = F_{y,1}^R(\vec{\boldsymbol{y}}_1)F_{y,2}^R(\vec{\boldsymbol{y}}_2|A_1)\dots F_{y,n}^R(\vec{\boldsymbol{y}}_n|A_1 \cap A_2 \cap \dots A_{n-1})$ for all $\vec{\boldsymbol{y}}_1 \in \mathbb{R}^{d_1}$ and for all $\vec{\boldsymbol{X}} \in \mathbb{R}^d$. If the fundamental vector attributes are probabilistically independent, then a vector independent (ribbon VICDF) is defined as $F^R(\vec{\boldsymbol{X}}) = F_{y,1}^R(\vec{\boldsymbol{y}}_1)F_{y,2}^R(\vec{\boldsymbol{y}}_2)\dots F_{y,n}^R(\vec{\boldsymbol{y}}_n)$ for all $\vec{\boldsymbol{y}}_j \in \mathbb{R}^{d_j}$.

A special case of a ribbon VACDF is the scalar argument CDF (ribbon SACDF) $F^R(.)$, where $n = d, d_j = 1,\ j = 1,2,\dots,n$, thus the multi-dimensional ribbon CDF is a product of the 1-D ribbon CDF. If $B_j = X_j \le x_j, j = 1,2,\dots,d-1$ then $F^R(\vec{\boldsymbol{X}}) = F_1^R(x_1)F_2^R(x_2|B_1)\dots F_d^R(x_d|B_1 \cap B_2 \cap \dots B_{d-1})$ for all $x_1 \in (-\infty;+\infty)$ and for all $\vec{\boldsymbol{X}} \in \mathbb{R}^d$. In case the attributes are probabilistically independent ($n = d, d_j = 1,\ j = 1,2,\dots,n$), then the ribbon VICDF transforms into the scalar independent (ribbon SICDF), such that $F^R(\vec{\boldsymbol{X}}) = F_1^R(x_1)F_2^R(x_2)\dots F_d^R(x_d)$ for all $\vec{\boldsymbol{X}} \in \mathbb{R}^d$ and for all $x_j \in (-\infty;+\infty)$.

4 Modeling and Ranking Uncertain Alternatives

In a problem under risk, the uncertainty is entirely measured by classical distributions, whereas the preferences over prizes – by a utility function. The resulting classical risky lotteries are ranked according to expected utility [10] (fig. 1).

The uncertainty in the alternatives that the FRDM faces can only be partially measured by ribbon distributions. These alternatives cannot be adequately modeled by classical risky lotteries (since those lotteries require unit sum of the probabilities, which cannot be guaranteed in the case of interval values), not to mention being ranked according to expected utility. Therefore, fuzzy rational lotteries are introduced [22], where the uncertainty is only partially measured by ribbon distributions. These lotteries are ranked in a two-stage procedure, known as Q-expected utility.

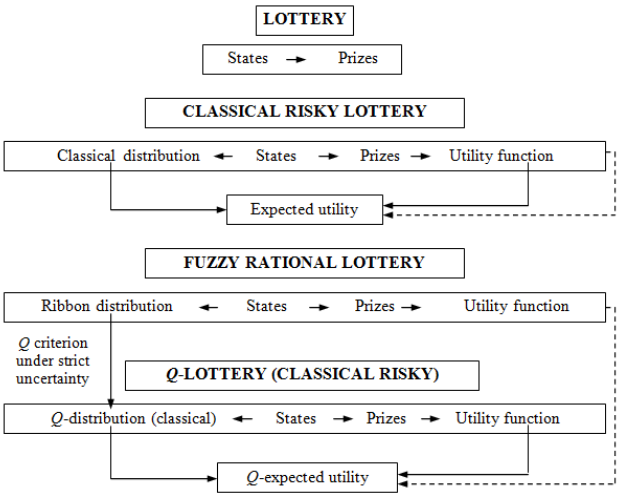


Fig. 1 Scheme of the process of ranking classical risky and fuzzy rational lotteries

In the first stage, the ribbon distribution is approximated by a classical one. According to [2] this is a task under strict uncertainty since any classical distribution that belongs to the ribbon distribution is just as likely as the other ones. The main idea is to use any of the Q criteria under strict uncertainty at that stage. A survey on the essence and adequacy of these methods is proposed in [10]. Despite their disadvantages, these methods have been well studied and are flexible in modeling alternatives taking into account the pessimism of the FRDM. The resulting approximating Q-lotteries are classical risky. In the second stage, the Q-lotteries are ranked according to Q-expected utility. Q is an arbitrary criterion under strict uncertainty. It usually stands for L, W, $\neg W$, and H_α that correspond to the Laplace, Wald, maximax, and Hurwicz$_\alpha$ criteria.

4.1 The case of Ordinary Lotteries

If the set of lotteries L and the set of prizes X are countable, then alternatives are modeled as ordinary lotteries (OL). An OL with a 1-D ribbon DPF $F_{d,i}^{R}(.)$ is *a fuzzy rational OL*. It takes the form

$$\begin{aligned} l_i^{fr} =& \ll \theta_{i,1}, P_i^d(\theta_{i,1}), 1-P_i^u(\theta_{i,1}) >, \vec{\boldsymbol{X}}_{i,1}; < \theta_{i,2}, P_i^d(\theta_{i,2}), 1-P_i^u(\theta_{i,2}) >, \\ & \vec{\boldsymbol{X}}_{i,2}; < \theta_{i,t_i}, P_i^d(\theta_{i,t_i}), 1-P_i^u(\theta_{i,t_i}) >, \vec{\boldsymbol{X}}_{i,t_i} >, \\ & 0 \le P_i^d(\theta_{i,r}) \le P_i^u(\theta_{i,r}) \le 1, r = 1,2,\dots,t_i; \\ & \sum_{r=1}^{t_i} P_i^d(\theta_{i,r}) \le 1 \le \sum_{r=1}^{t_i} P_i^u(\theta_{i,r}), i = 1,2,\dots,q. \end{aligned} \tag{5}$$

where $\theta_{i,r}$ is the event: "to get the prize $\vec{X}_{i,r}$ from the *i*-th alternative", $P(\theta_{i,r}) \in [P_i^d(\theta_{i,r}); P_i^u(\theta_{i,r})]$. The fuzzy rational OL may be ranked in two stages [32]:

1. U2sing a Q criterion under strict uncertainty, $F_{d,i}^R(.)$ is approximated by a classical DPF $f_{d,i}^Q(.)$, such that $P_i^d(\theta_{i,r}) \le f_{d,i}^Q(r) = P_i^Q(\theta_{i,r}) \le P_i^u(\theta_{i,r}), r = 1,2,\dots,t_i,\ i = 1,2,\dots,q$. In that way any fuzzy rational OL is approximated by a classical risky Q-OL: $l_i^Q = \ll \theta_{i,1}, P_i^Q(\theta_{i,1}) >, \vec{X}_{i,1}; < \theta_{i,2}, P_i^Q(\theta_{i,2}) >, \vec{X}_{i,2}; \dots; < \theta_{i,t_i}, P_i^Q(\theta_{i,t_i}) >, \vec{X}_{i,t_i} >, i = 1,2,\dots,q$.
2. The alternatives are ranked in descending order the expected utilities of the Q-OL: $E_i^Q(u|f_{d,i}^R) = \sum_{r=1}^{t_i} P_i^Q(\theta_{i,r}) u(\vec{X}_{i,r}), i = 1,2,\dots,q$.

Calculating the *Q*-expected utility of a fuzzy rational OL is brought down to estimation of $P_i^Q(\theta_{i,r})$. This has a different interpretation for a different *Q* criterion.

The Laplace probabilities $P_i^L(\theta_{i,r}), r = 1,2,\dots,t_i$, do not depend on the utility function. According to the Laplace's insufficient reasoning principle [26], if no information is available for the likelihood of a group of hypotheses, then the probability of any of the hypotheses equals to the reciprocal value of the number of hypotheses in the group. Therefore the required probabilities are weighted means of the lower and upper bounds, so that $\sum_{r=1}^{t_i} P_i^L(\theta_{i,r}) = 1$. Then

$$P_i^L(\theta_{i,r}) = [1 - \alpha_L^{(i)}] P_i^d(\theta_{i,r}) + \alpha_L^{(i)} P_i^u(\theta_{i,r}), r = 1,2,\dots,t_i$$

$$\alpha_L^{(i)} = \begin{cases} \left(1 - \sum_{r=1}^{t_i} P_i^d(\theta_{i,r})\right) \Big/ \left(\sum_{r=1}^{t_i} P_i^u(\theta_{i,r}) - \sum_{r=1}^{t_i} P_i^d(\theta_{i,r})\right) & \dots \\ & \dots \sum_{r=1}^{t_i} P_i^u(\theta_{i,r}) > \sum_{r=1}^{t_i} P_i^d(\theta_{i,r}) \\ 0.5 & \sum_{r=1}^{t_i} P_i^u(\theta_{i,r}) = \sum_{r=1}^{t_i} P_i^d(\theta_{i,r}) \end{cases} \quad (6)$$

The result is the Laplace lottery

$$l_i^L = < P^L(\theta_{i,1}), x_1; P^L(\theta_{i,2}), x_2; \dots; P^L(\theta_{i,t_i}), x_{t_i} >$$

The Wald (pessimism) criterion under strict uncertainty [6] assumes to increase the probabilities of prizes from their lowest limit (but not higher than their upper limit), initiating with the worst outcome, until the corrected probabilities sum to one. The result is the Wald lottery $l_i^W = < P_i^W(\theta_{i,1}), x_1; P_i^W(\theta_{i,2}), x_2; \dots; P_i^W(\theta_{i,t_i}), x_{t_i} >$

$$P_i^{W}(\theta_{i,r}) = \begin{cases} P_i^d(\theta_{i,r}) & \text{for } \rho(r) < \rho\left(r_W^{(i)}\right), \\ \left[1-\beta^{(i)}\right]P_i^d(\theta_{i,r}) + \beta^{(i)}P_i^u(\theta_{i,r}) & \text{for } \rho(r) = \rho\left(r_W^{(i)}\right), \\ P_i^u(\theta_{i,r}) & \text{for } \rho(r) > \rho\left(r_W^{(i)}\right), r = 1,2,\dots,t \end{cases}$$

$$\beta_{\rho(r)}^{(i)} = \begin{cases} \dfrac{1-\sum_{k=r+1}^{t_i} P_i^u(\theta_{i,\rho(k)}) - \sum_{k=1}^{r} P_i^d(\theta_{i,\rho(k)})}{P_i^u(\theta_{i,\rho(r)}) - P_i^d(\theta_{i,\rho(r)})} \text{ for } \dots \\ \qquad \dots P_i^u(\theta_{i,\rho(r)}) > P_i^d(\theta_{i,\rho(r)}) \\ 0 \text{ for } P_i^u(\theta_{i,\rho(r)}) = P_i^d(\theta_{i,\rho(r)}) \text{ and } \dots \\ \qquad \dots\left(\sum_{k=1}^{t_i} P_i^u(\theta_{i,k}) > \sum_{k=1}^{t_i} P_i^d(\theta_{i,k}) \text{ or } r < t_i\right) \\ 1 \text{ for } \sum_{k=1}^{t_i} P_i^u(\theta_{i,k}) = \sum_{k=1}^{t_i} P_i^d(\theta_{i,k}) \text{ and } r = t_i \end{cases} \quad (7)$$

$$r_W^{(i)} = arg\left\{\beta_r^{(i)} \in (0;1]\right\};\ \beta^{(i)} = \beta_{r_W^{(i)}}^{(i)}$$

The maximax criterion [11] is the opposite to the Wald criterion and the required probabilities $P_i^{\neg W}(\theta_{i,r})$ may be found using the Wald procedure by putting $u(\vec{X}_{i,r}) = -u(\vec{X}_{i,r}), r = 1,2,\dots,t_i$.

The Hurwicz approach balances the extreme pessimism and extreme optimism by an pessimistic-optimistic index $\alpha \in [0;1]$, which, measures the pessimism of the DM [36]. The $Hurwicz_\alpha$ lottery takes the form $l_i^{H\alpha} =< P^{H\alpha}(\theta_{i,1}), x_1; P^{H\alpha}(\theta_{i,2}),\ x_2; \dots; P^{H\alpha}(\theta_{i,r}), x_r >$, where $P^{H\alpha}(\theta_{i,j}) = \alpha P^{W}(\theta_{i,j}) + (1-\alpha)P^{\neg W}(\theta_{i,j})$.

4.2 The Case of Generalized Lotteries of I Type

Assume there are q alternatives that give 1-D prizes x from a piece-wise continuous 1-D set X, according to continuous or mixed probability laws. Such alternatives are modeled by 1-D generalized lotteries of I type (GL-I). A 1-D GL-I with a ribbon $F_i^R(x)$ is a 1-D fuzzy-rational GL-I: $g_i^{fr} =< F_i^R(x); x >$, for $i = 1,2,\dots,q$. These are ranked in two stages:

1. Using a Q criterion under strict uncertainty, each $F_i^R(.)$ is approximated by a 1-D classical CDF $F_i^Q(.)$ such that $F_i^d(x) \le F_i^Q(x) \le F_i^u(x), i = 1,2,\dots,q$ and for all $x \in (-\infty;+\infty)$. So, each 1-D fuzzy rational GL-I is approximated

by a 1-D classical risky GL-I, called Q-generalized (1-D Q-GL-I): $g_i^Q =< F_i^Q(x); x >$.

2. The alternatives are ranked in descending order of the expected utilities of $g_i^Q: E_i^Q(u|F_i^R) = \int_{-\infty}^{+\infty} u(x)dF_i^Q(x)$.

One-dimensional x-fuzzy rational GL-I

A special case of a 1-D fuzzy rational GL-I with a *x*-ribbon CDF is a 1-D *x*-fuzzy rational GL-I: $g_i^{xfr} =< F_i^{xR}(x); x >, i = 1,2,\dots,q$. Calculating the *Q*-expected utility of the 1-D *x*-fuzzy rational GL-I may be brought down to the following steps:

1. Using a Q criterion, $F_i^{xR}(.)$ is piece-wise linearly approximated by a 1-D classical CDF $F_i^{xQ}(.)$ with nodes

$$\left\{\left(x_l^{Q,(i)}; F_l^{(i)}\right)\middle| l = 1,2,\dots,z_i\right\},$$
$$x_1^{Q,(i)} \le x_2^{Q,(i)} \le \cdots \le x_{z_i}^{Q,(i)}, x_l^{d,(i)} \le x_l^{Q,(i)} \le x_l^{u,(i)}, l = 2,3,\dots,z_i - 1,$$
$$x_1^{Q,(i)} = x_1^{d,(i)} = x_1^{u,(i)}, x_{z_i}^{Q,(i)} = x_{z_i}^{d,(i)} = x_{z_i}^{u,(i)}$$

$$F_i^{xQ}(x) = \begin{cases} 0 & \text{for } x < x_1^{Q,(i)} \\ F_l^{(i)} & \text{for } x_l^{Q,(i)} = x < x_{l+1}^{Q,(i)}, l = 1,2,\dots,z_i - 1, \\ F_l^{(i)} + \dfrac{\left(x - x_l^{Q,(i)}\right)\left(F_{l+1}^{(i)} - F_l^{(i)}\right)}{x_{l+1}^{Q,(i)} - x_l^{Q,(i)}} & \text{for } \dots \\ & \dots\ x_l^{Q,(i)} < x < x_{l+1}^{Q,(i)}, l = 1,2,\dots,z_i - 1, \\ 1 & \text{for } x_z^{Q,(i)} \le x. \end{cases} \quad (8)$$

Thus, g_i^{xfr} is approximated by a 1-D classical risky *x*Q-generalized (1-D xQ-GL-I) $g_i^{xQ} =< F_i^{xQ}; x >$.

2. The Q-expected utility of g_i^{xfr} is calculated as the expected utility of g_i^{xQ}:

$$E_i^{xQ}(u|F_i^{xR}) = \sum_{\substack{l=1 \\ x_{l+1}^{Q,(i)} > x_l^{Q,(i)}}}^{z_i-1} \frac{F_{l+1}^{(i)} - F_l^{(i)}}{x_{l+1}^{Q,(i)} - x_l^{Q,(i)}} \int_{x_l^{Q,(i)}}^{x_{l+1}^{Q,(i)}} u(x)dx + \sum_{\substack{l=1 \\ x_{l+1}^{Q,(i)} = x_l^{Q,(i)}}}^{z_i-1} \left(F_{l+1}^{(i)} - F_l^{(i)}\right) u(x_l^{Q,(i)}) \quad (9)$$

This last task is brought down to the estimation of the inner quantiles $x_l^{Q,(i)}, l = 2,3,\dots,z_i-1$, of the classical CDF in g_i^{xQ}.

The quantiles $x_l^{L,(i)}, l = 2,3,\dots,z_i-1$, do not depend on the utility function. Following the Laplace criterion, if no information is available for the quantiles (i.e. $x_l^{d,(i)} = x_1^{d,(i)} = x_1^{u,(i)}, x_l^{u,(i)} = x_{z_i}^{d,(i)} = x_{z_i}^{u,(i)}, l = 2,3,\dots,z_i-1$), then the distribution must be uniform in the interval $\left[x_1^{d,(i)}; x_{z_i}^{d,(i)}\right]$. Let the quantile with the $F_l^{(i)}$ index of this uniform distribution be called *quantile of the complete ignorance*: $x_l^{aL,(i)} = x_1^{d,(i)} + \left(x_{z_i}^{d,(i)} - x_1^{d,(i)}\right)F_l^{(i)}, l = 2,3,\dots,z_i-1$. Let $h_l^{x,(i)}$ be the homothety of the maximal uncertainty interval under strict uncertainty of the *l*-th quantile $\left[x_1^{d,(i)}; x_{z_i}^{d,(i)}\right]$ into the actual uncertainty interval $\left[x_l^{d,(i)}; x_l^{u,(i)}\right]$. Then according to [31], $x_l^{L,(i)}$ will be the image of $x_l^{aL,(i)}$ at the $h_l^{x,(i)}$, i.e. $x_l^{L,(i)} = x_l^{d,(i)} + \left(x_l^{u,(i)} - x_l^{d,(i)}\right)\frac{x_l^{aL,(i)} - x_l^{d,(i)}}{x_{z_i}^{d,(i)} - x_1^{d,(i)}} = x_l^{d,(i)} + (x_l^{u,(i)} - x_l^{d,(i)})F_l^{(i)}$.

The Wald criterion implies to choose $x_l^{W,(i)}, l = 2,3,\dots,z_i-1$, so that to minimize the *xW*-expected utility of the lottery [18]:

$$
\begin{aligned}
E_i^{xW}(u|F_i^{xR}) &= \sum_{\substack{l=1 \\ x_{l+1}^{W,(i)} > x_l^{W,(i)}}}^{z_i-1} \frac{F_{l+1}^{(i)} - F_l^{(i)}}{x_{l+1}^{W,(i)} - x_l^{W,(i)}} \int_{x_l^{W,(i)}}^{x_{l+1}^{W,(i)}} u(x)dx \\
&\quad + \sum_{\substack{l=1 \\ x_{l+1}^{W,(i)} = x_l^{W,(i)}}}^{z_i-1} \left(F_{l+1}^{(i)} - F_l^{(i)}\right)u\left(x_l^{W,(i)}\right) \\
&= \sum_{l=1}^{z_i-1} \left(F_{l+1}^{(i)} - F_l^{(i)}\right)I_l^{xW,(i)}
\end{aligned}
\tag{10}
$$

$$
I_l^{xW,(i)} = \begin{cases} \dfrac{1}{x_{l+1}^{W,(i)} - x_l^{W,(i)}} \displaystyle\int_{x_l^{W,(i)}}^{x_{l+1}^{W,(i)}} u(x)dx & \text{for } x_{l+1}^{W,(i)} > x_l^{W,(i)}, \\ u\left(x_l^{W,(i)}\right) & \text{for } x_{l+1}^{W,(i)} = x_l^{W,(i)}, \end{cases} \qquad l = 1,2,\dots,z_i-1
$$

Since the maximax criterion is opposite to the Wald criterion, then $x_l^{\neg W,(i)}, l = 2,3,\dots,z_i-1$ are found in the same way as with the Wald transformations, using the substitution $u(x) = -u(x)$ for all $x \in (-\infty; +\infty)$.

The Hurwicz principle implies to choose $x_l^{H_\alpha,(i)}$, $l = 2,3,\dots,z_i-1$ as weighted measured of $x_l^{W,(i)}$ and $x_l^{\neg W,(i)}$ by $\alpha \in [0;1]$: $x_l^{H_\alpha,(i)} = \alpha x_l^{W,(i)} + (1-\alpha)x_l^{\neg W,(i)}$.

One-dimensional p-fuzzy rational GL-I

A special case of a 1-D fuzzy rational GL-I with a p-ribbon CDF is a 1-D p-fuzzy rational GL-I: $g_i^{pfr} < F_i^{pR}(x); x >, i = 1,2, \dots, q$. Calculating the Q-expected utility of the 1-D p-fuzzy rational GL-I may be brought down to the following steps:

1. Using a Q criterion, $F_i^{pR}(.)$ is piece-wise linearly approximated by a 1-D classical CDF $F_i^{pQ}(.)$ with nodes

$$\left\{\left(x_l^{(i)}; F_l^{Q,(i)}\right)\middle| l = 1,2, \dots, z_i\right\},$$
$$0 = F_1^{Q,(i)} \le F_2^{Q,(i)} \le \cdots \le F_{z_i}^{Q,(i)} = 1, F_l^{d,(i)} \le F_l^{Q,(i)} \le F_l^{u,(i)}, l = 2,3, \dots, z_i - 1,$$

$$F_i^{pR}(x) = \begin{cases} 0 & \text{for } x < x_1^{(i)} \\ F_l^{Q,(i)} & \text{for } x_l^{(i)} = x < x_{l+1}^{(i)}, l = 1,2, \dots, z_i - 1, \\ F_l^{Q,(i)} + \dfrac{\left(x - x_l^{(i)}\right)\left(F_{l+1}^{Q,(i)} - F_l^{Q,(i)}\right)}{x_{l+1}^{(i)} - x_l^{(i)}} & \text{for } \dots \\ & \dots\ x_l^{(i)} < x < x_{l+1}^{(i)}, l = 1,2, \dots, z_i - 1, \\ 1 & \text{for } x_{z_i}^{(i)} \le x. \end{cases} \tag{11}$$

Thus, g_i^{pfr} is approximated by a 1-D classical risky 1-D pQ-GL-I $g_i^{pQ} =< F_i^{pQ}(x); x >$.

2. The Q-expected utility of g_i^{pfr} is calculated as the expected utility of g_i^{pQ}:

$$E_i^{pQ}\left(u\middle|F_i^{pR}\right) = \sum_{\substack{l=1 \\ x_{l+1} > x_l}}^{z_i - 1} \frac{F_{l+1}^{Q,(i)} - F_l^{Q,(i)}}{x_{l+1}^{(i)} - x_l^{(i)}} \int_{x_l^{(i)}}^{x_{l+1}^{(i)}} u(x)dx + \sum_{\substack{l=1 \\ x_{l+1} = x_l^{(i)}}}^{z_i - 1} \left(F_{l+1}^{Q,(i)} - F_l^{Q,(i)}\right) u(x_l^{(i)}) \tag{12}$$

This last task is brought down to the estimation of the inner quantile indices $F_l^{Q,(i)}, l = 2,3, \dots, z_i - 1$, of the classical CDF in g_i^{pQ}.

Following the Laplace criterion, if no information is available for the quantile indices (i.e. $F_l^{d,(i)} = 0, F_l^{u,(i)} = 1, l = 2,3, \dots, z_i - 1$), then the distribution must be uniform in the interval $\left[x_1^{(i)}; x_{z_i}^{(i)}\right]$. Let the quantile index of $x_l^{(i)}$ of this uniform

distribution be a quantile index of the complete ignorance: $F_l^{aL,(i)} = \frac{x_l^{(i)} - x_1^{(i)}}{x_{z_i}^{(i)} - x_1^{(i)}}, l = 2,3,\ldots,z_i - 1$. Let $h_l^{p,(i)}$ be the homothety of the maximal uncertainty interval under strict uncertainty of the l-th quantile index [0; 1] into the actual uncertainty interval $\left[F_l^{d,(i)}; F_l^{u,(i)}\right]$. Then $F_l^{L,(i)}$ is the image of $F_l^{aL,(i)}$ at the homothety $h_l^{p,(i)}$: $F_l^{L,(i)} = F_l^{d,(i)} + \left(F_l^{u,(i)} - F_l^{d,(i)}\right)F_l^{aL,(i)} = F_l^{d,(i)} + (F_l^{u,(i)} - F_l^{d,(i)})\frac{x_l^{(i)} - x_1^{(i)}}{x_{z_i}^{(i)} - x_1^{(i)}}$.

The Wald criterion implies to choose $F_l^{W,(i)}, l = 2,3,\ldots,z_i - 1$, so that to minimize the pW-expected utility of the lottery:

$$E_i^{pW}\left(u\middle|F_i^{pR}\right) = I_{z_i-1}^{p,(i)} + \sum_{l=2}^{z_i-1} F_l^{W,(i)}\left(I_{l-1}^{p,(i)} - I_l^{p,(i)}\right),$$

$$I_l^{p,(i)} = \begin{cases} \dfrac{1}{x_{l+1}^{(i)} - x_l^{(i)}} \displaystyle\int_{x_l^{(i)}}^{x_{l+1}^{(i)}} u(x)dx & \text{for } x_{l+1}^{(i)} > x_l^{(i)} \\ u\left(x_l^{(i)}\right) & \text{for } x_{l+1}^{(i)} = x_l^{(i)} \end{cases}, l = 1,2,\ldots,z_i - 1 \quad (13)$$

The required quantile indices $F_l^{\neg W,(i)}, l = 2,3,\ldots,z_i - 1$, may be identified using the Wald procedures with a substitution $u(x) = -u(x)$ for all $x \in (-\infty; +\infty)$.

The quantile indices $F_l^{H\alpha,(i)}, l = 2,3,\ldots,z_i - 1$ are chosen as weighted measures of $F_l^{W,(i)}$ and $F_l^{\neg W,(i)}$ by $\alpha \in [0;1]$: $F_l^{H\alpha,(i)} = \alpha F_l^{W,(i)} + (1-\alpha)F_l^{\neg W,(i)}, l = 2,3,\ldots,z_i - 1$.

4.3 The Case of Multi-dimensional Fuzzy Rational GL-I

Assume there are q alternatives that give multi-dimensional (multi-D) prizes $\vec{\boldsymbol{X}}$ from a piece-wise continuous d-dimensional set X according to continuous of mixed multi-D probability laws. Such alternatives are the multi-D GL-I. A multi-D GL-I with a multi-D ribbon CDF $F_i^R(.)$ is a multi-D fuzzy rational GL-I: $g_i^{fr} =< F_i^R\left(\vec{\boldsymbol{X}}\right); \vec{\boldsymbol{X}} >, i = 1,2,\ldots,q$. It may be ranked in two stages:

1. Using a Q criterion under strict uncertainty, each $F_i^R(.)$ is approximated by a multi-D classical CDF $F_i^Q(.)$ such that for all $\vec{\boldsymbol{X}} \in \mathbb{R}^{\boldsymbol{d}}$, $F_i^d(\vec{\boldsymbol{X}}) \le F_i^Q(\vec{\boldsymbol{X}}) \le F_i^u(\vec{\boldsymbol{X}})$. In that way, any g_i^{fr} is approximated by a multi-D classical risky GL-I –Q-generalized (multi-D Q-GL-I): $g_i^Q =< F_i^Q\left(\vec{\boldsymbol{X}}\right); \vec{\boldsymbol{X}} >$.
2. The alternatives are ranked in descending order of the expected utilities of g_i^Q: $E_i^Q(u|F_i^R) = \iint_{\mathbb{R}^d} \ldots \int u(\vec{\boldsymbol{X}})\partial^d F_i^Q\left(\vec{\boldsymbol{X}}\right)$.

If the uncertainty in the prizes is described by a ribbon VICDF $F_i^R(.)$, and the preferences of the DM over $Y_1, Y_2, \dots, Y_n$ are MUVI, then g_i^{fr} may be decomposed to n fictitious d_j-dimensional fuzzy rational GL-I: $g_{y,j}^{fr,(i)} =< F_{y,j}^{R,(i)}(\vec{\boldsymbol{y}}_j); \vec{\boldsymbol{y}}_j >, j = 1,2, \dots, n$, where $F_{y,j}^{R,(i)}(.)$ is a marginal d_j-dimensional ribbon CDF of Y_j. Then base fictitious (fuzzy rational BF-GL-I) are present. They are ranked in the following steps:

1. using a Q criterion under strict uncertainty each $F_{y,j}^{R,(i)}(.)$ is approximated by a marginal dj-dimensional classical CDF $F_{y,j}^{Q,(i)}(.)$, such that for all $\vec{\boldsymbol{y}}_j \in \mathbb{R}^{d_j}$: $F_{y,j}^{d,(i)}(\vec{\boldsymbol{y}}_j) \le F_{y,j}^{Q,(i)}(\vec{\boldsymbol{y}}_j) \le F_{y,j}^{u,(i)}(\vec{\boldsymbol{y}}_j)$. So, $g_{y,j}^{fr,(i)}$ is approximated by a d_j-dimensional base fictitious Q-GL-I (d_j-dimensional BF-Q-GL-I), $g_{y,j}^{Q,(i)} =< F_{y,j}^{Q,(i)}(\vec{\boldsymbol{y}}_j); \vec{\boldsymbol{y}}_j >$;
2. the Q-expected utility of $g_{y,j}^{fr,(i)}$ is calculated as the expected utilities of $g_{y,j}^{Q,(i)}$: $E_{y,j}^{Q,(i)}\left(u \middle| F_{y,j}^{R,(i)}\right) = \iint_{\mathbb{R}^{d_j}} \dots \int u_{y,j}(\vec{\boldsymbol{y}}_j) \partial^{d_j} F_{y,j}^{Q,(i)}(\vec{\boldsymbol{y}}_j)$.
3. the Q-expected utility of the fuzzy rational GL-I with a ribbon VICDF in the case of MUVI preferences takes a multiplicative form as follows: $E_i^Q(u|F_i^R) = \frac{1}{K_y}\prod_{j=1}^{n}\left[K_y k_{y,j} E_{y,j}^{Q,(i)}\left(u \middle| F_{y,j}^{R,(i)}\right) + 1\right] - \frac{1}{K_y}$. If the preferences of the DM over $Y_1, Y_2, \dots, Y_n$ were MAVI, then the Q-expected utility would takes an additive form $E_i^Q(u|F_i^R) = \sum_{j=1}^{n} k_{y,j}\, E_{y,j}^{Q,(i)}\left(u \middle| F_{y,j}^{R,(i)}\right)$.

If the uncertainty associated with the prizes is described by a ribbon SICDF $F_i^R(.)$, and the preferences for $X_1, X_2, \dots, X_d$ are MUSI (or MASI), then g_i^{fr} can be decomposed to d fictitious 1-D fuzzy rational GL-I: $g_{x,j}^{fr,(i)} =< F_j^{R,(i)}(x_j); x_j >$, $j = 1,2, \dots, d$, where $F_j^{R,(i)}(.)$ is a marginal 1-D ribbon CDF of X_j. Then attribute fictitious (fuzzy rational AF-GL-I) are present. They are ranked as follows:

1. using a Q criterion under strict uncertainty any $F_j^{R,(i)}(.)$ is approximated by a marginal 1-D classical CDF $F_j^{Q,(i)}(.)$, such that $F_j^{d,(i)}(x_j) \le F_j^{Q,(i)}(x_j) \le F_j^{u,(i)}(x_j)$ for all $x_j \in (-\infty; +\infty)$. Then any hypothetical $g_{x,j}^{fr,(i)}$ is approximated by an attribute fictitious 1-D Q-GL-I (AF-Q-GL-I): $g_{x,j}^{Q,(i)} =< F_j^{Q,(i)}(x_j); x_j >$.
2. the Q-expected utility of $g_{x,j}^{fr,(i)}$ are calculated as the expected utilities of $g_{x,j}^{Q,(i)}$:
   $E_{x,j}^{Q,(i)}\left(u \middle| F_j^{R,(i)}\right) = \int_{-\infty}^{+\infty} u_j(x_j) dF_j^{Q,(i)}(x_j)$.
3. the Q-expected utility of the fuzzy rational GL-I with a ribbon SICDF in the case of MUSI preferences takes a multiplicative form $E_i^Q(u|F_i^R) = \frac{1}{K}\prod_{j=1}^{d}\left[K k_j E_{x,j}^{Q,(i)}\left(u \middle| F_j^{R,(i)}\right) + 1\right] - \frac{1}{K}$. If the preferences over $X_1, X_2, \dots, X_d$

were MASI, then the Q-expected utility would take an additive form $E_i^Q(u|F_i^R) = \sum_{j=1}^{d} k_j \, E_{x,j}^{Q,(i)}\left(u\middle|F_j^{R,(i)}\right)$.

5 Conclusion

This chapter introduced fuzzy rational decision analysis as a generalization of EUT, which tried to unify the normative rationality with the fuzziness of real preferences in the measurement process. It may be interpreted as the modern approach to the behavior of economic subjects, which tends to build upon the maximum quantity and quality of subjective information. It was shown that fuzzy (bounded) rationality stems from the interval nature of subjective estimates (utilities and probabilities). The interval probabilities in particular caused the introduction of ribbon distributions, and in turn – fuzzy-rational lotteries. A scheme was given in the chapter, showing that ribbon distributions should be approximated by classical ones using Q criteria under strict uncertainty in order to be able to rank fuzzy-rational alternatives. Finally, decisions were made using the Q-expected utility criterion. Interpretation of this criterion was given for the case of fuzzy-rational ordinary lotteries, for 1-D fuzzy-rational GL-I, and for fuzzy-rational multi-dimensional GL-I. The procedures to rank the other type of lotteries – fuzzy-rational generalized lotteries of II type, of III type, as well as the semi-generalized lotteries of I type (SGL-I) – are a combination of those already presented here.

As a rather new theory, fuzzy rational decision analysis has multiple problems yet to be solved. Some of these tasks are: 1) to outline the procedure to rank generalized lotteries, where the uncertainty is described by VICDF, and preferences over the attributes are either MUSI or MASI; 2) to outline the procedure to rank generalized lotteries, where the uncertainty is described by SICDF, and the preferences over the fundamental vector attributes are either MUVI or MAVI. These questions shall be a topic of further research.

Acknowledgments. This book chapter has been supported by the INPORT project (project No. DVU01-0031, Bulgarian National Science Fund of Bulgaria).

References

1. Abdellaoui, M., Barrios, C., Wakker, P.: Reconciling Introspective Utility with Revealed Preference: Experimental Arguments Based on Prospect Theory. Journal of Econometrics 138, 356–378 (2007)
2. Augustin, T.: On Decision Making under Ambiguous Prior and Sampling Information. In: de Cooman, G., Fine, T., Moral, S. (eds.) ISIPTA 2001: Proc. Second International Symposium on Imprecise Probabilities and their Applications, pp. 9–16. Cornell University, Ithaca (2001)

3. Bacchus, F., Grove, A.: Utility Independence in Qualitative Decision Theory. In: Aiello, L.C., Doyle, J., Shapiro, S. (eds.) Principles of Knowledge Representation and Reasoning, pp. 542–552. Morgan Kaufmann, CA (1996)
4. Clemen, R.: Making Hard Decisions: an Introduction to Decision Analysis, 2nd edn. Duxbury Press, Wadsworth Publishing Company (1996)
5. Engel, Y., Wellman, M.: CUI Networks: A Graphical Representation for Conditional Utility Independence. In: Twenty-First National Conference on Artificial Intelligence, pp. 1137–1142 (2006)
6. Fabrycky, W.J., Thuesen, G.J., Verma, D.: Economic Decision Analysis, 3rd edn. Prentice-Hall (1998)
7. Farquhar, P.H.: Research Directions in Multi-Attribute Utility Analysis. In: Hansen, P. (ed.) Essays and Surveys on Multi-Criteria Decision Making, pp. 63–85. Springer (1983)
8. Farquhar, P.H.: Utility Assessment Methods. Management Science 30(11), 1283–1300 (1984)
9. Fishburn, P., Roberts, F.: Mixture Axioms in Linear and Multilinear Utility Theories. Theory and Decisions 9(9), 161–171 (2004)
10. French, S.: Decision Theory: an Introduction to the Mathematics of Rationality. Ellis Horwood (1993)
11. Hackett, G., Luffrum, P.: Business Decision Analysis. An Active Learning Approach. Blackwell (1999)
12. Hanke, J.E., Reitsch, A.G.: Understanding Business Statistics. Irwin (1991)
13. Keeney, R.L., Raiffa, H.: Decisions with Multiple Objectives: Preference and Value Tradeoffs. Cambridge University Press (1993)
14. Kiefer, J.: Sequential Minimax Search for a Maximum. Proc. American Mathematical Society 4, 502–506 (1953)
15. McCord, M., De Neufville, R.: Lottery Equivalents': Reduction of the Certainty Effect Problem in Utility Assessment. Management Science 32, 56–60 (1986)
16. Nikolova, N.D.: Arctg-approximation of monotonically decreasing utility functions. Computer Science and Technology 1, 60–69 (2007)
17. Nikolova, N.D.: Uniform Method for Estimation of Interval Scaling Constants. Engineering and Automation Problems 1, 79–90 (2007B)
18. Nikolova, N.D.: Three criteria to rank x-fuzzy-rational generalized lotteries of I type. Cybernetics and Information Technologies 7(1), 3–20 (2007C)
19. Nikolova, N.D., Shulus, A., Toneva, D., Tenekedjiev, K.: Fuzzy Rationality in Quantitative Decision Analysis. Journal of Advanced Computational Intelligence and Intelligent Informatics 9(1), 65–69 (2005)
20. Nikolova, N.D., Hirota, K., Kobashikawa, C., Tenekedjiev, K.: Elicitation of Non-Monotonic Preferences of a Fuzzy Rational Decision Maker. Information Technologies and Control, Year IV 1, 36–50 (2006)
21. Nikolova, N.D., Toneva, D., Ahmed, S., Tenekedjiev, K.: Constructing multidimensional utility function under different types of preferences over the fundamental vector attributes. In: Zadeh, L., Tufis, D., Filip, F., Dzitac, I. (eds.) From Natural Language to Soft Computing: New Paradigms in Artificial Intelligence, pp. 129–149. Publishing House of the Romanian Academy (2008)
22. Nikolova, N.D., Tenekedjiev, K.: Fuzzy Rationality and Parameter Elicitation in Decision Analysis. International Journal of General Systems, Special Issue on Intelligent Systems 39(5), 539–556 (2010)

23. Pratt, J.W.: Risk Aversion in the Small and in the Large. Econometrica 32, 122–136 (1964)
24. Pratt, J.W., Raiffa, H., Schlaifer, R.: Introduction to Statistical Decision Theory. MIT Press, Cambridge (1995)
25. Press, W.H., Teukolski, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes – The Art of Scientific Computing. Cambridge University Press (1992)
26. Rapoport, A.: Decision Ttheory and Decision Behaviour – Normative and Descriptive Approaches. Kluwer Academic Publishers, USA (1989)
27. Stoyanov, S.: Optimization of Technological Objects. Sofia, Tehnika (1993) (in Bulgarian)
28. Swalm, R.: Utility Theory – Insight into Risk Taking. Harvard Business Review 44, 123–136 (1966)
29. Tenekedjiev, K.: Quantitative Decision Analysis: Utility Theory and Subjective Statistics. Doctor of Sciences Dissertation, Varna. Technical University, Varna (2004) (in Bulgarian)
30. Tenekedjiev, K., Nikolova, N.D., Pfliegl, R.: Utility elicitation with the uncertain equivalence method. Comptes Rendus De L'Academie Bulgare des Sciences 3, Book 3, 283–288 (2006A)
31. Tenekedjiev, K., Nikolova, N.D., Toneva, D.: Laplace Expected Utility Criterion for Ranking Fuzzy Rational Generalized Lotteries of I Type. Cybernetics and Information Technologies 6(3), 93–109 (2006B)
32. Tenekedjiev, K., Nikolova, N.D.: Ranking Discrete Outcome Alternatives with Partially Quantified Uncertainty. International Journal of General Systems 37(2), 249–274 (2007)
33. Tenekedjiev, K., Nikolova, N.D.: Justification and Numerical Realization of the Uniform Method for Finding Point Estimates of Interval Elicited Scaling Constants. Fuzzy Optimization and Decision Making 7(2), 119–145 (2008)
34. Utkin, L.V.: Risk analysis under partial prior information and non-monotone utility functions. International Journal of Information Technology and Decision Making 6(4), 625–647 (2007)
35. Wakker, P., Deneffe, D.: Eliciting von Neumann-Morgenstern Utilities when Probabilities Are Distorted or Unknown. Management Science 42, 1131–1150 (1996)
36. Yager, R.R.: Generalizing Variance to Allow the Inclusion of Decision Attitude in Decision Making under Uncertainty. International Journal of Approximate Reasoning 42, 137–158 (2006)

Comparing Methods of Assessing R&D Efficiency in Latin-American Countries

Catalina Alberto, Lucía I. Passoni, Claudia E. Carignano, and Mercedes Delgado

Abstract. Scientific and technological activities have a growing importance in the economic development of a region. The aim of this paper is to analyze how efficient scientific research and experimental development (R&D) are in terms of public expenditure. The effect of public expenditure on the ability to innovate, measured by the number of patents, publications and invention indices is evaluated. We propose the use of tools from the area of Operations Research as Data Envelopment Analysis and the Technique for Order Performance by Similarity to Ideal Solution, as well as methodologies from Computational Intelligence such as Self Organizing Maps to visualize the joint behavior of the different countries. The results obtained confirm an imbalance among countries in the area, imbalance that should be addressed in order to increase the global efficiency of the region.

1 Introduction

R&D can be defined as creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications. It includes basic research, applied research and experimental development

Catalina Alberto · Claudia E. Carignano
Facultad de Ciencias Económicas, Universidad Nacional de Córdoba, Argentina
e-mail: {catalina.alberto,claudiacarignano}@gmail.com

Lucía I. Passoni
Laboratorio de Bioingeniería, Facultad de Ingeniería,
Universidad Nacional de Mar del Plata, Argentina
e-mail: isabel.passoni@gmail.com

Mercedes Delgado
"Jose Antonio Echeverria" Higher Technical Institute, Cuba
e-mail: mdelgado@ind.cujae.edu.cu

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_24,

In the last years, Research and Development has become a popular term. It is undeniably true that in order to improve competitiveness and productivity countries need to assign more resources to R&D activities

Such is the case of the European Union where the role of knowledge and innovation is considered the driving force of sustainable growth. The European Research

Area's objective is to strengthen its scientific and research bases and to encourage public and private investment in R&D [3]. However, and despite the effort made to achieve these aims, there still is a significant imbalance in the scientific and technological activities across regions, which affects the sustainable development of the whole Union.

In Latin America, there is no explicit formal statement on this issue. However, it has become absolutely necessary for governments and institutions to view knowledge, and consequently R&D activities, as a way of achieving and securing the economic growth of the region. To this end, the gaps in scientific and technological activities among countries must be reduced. In other words, the more homogeneous countries are in these areas, the higher chances of success in global development for the region.

The aim of this paper is to analyze for various countries in the region, the effect of different levels of effort in public R & D on the ability to innovate, as measured by patents, publications and index of invention. To this end, we propose the use of tools form the area of Operations Research as Data Envelopment Analysis (DEA) [1] and the Technique for Order Performance by Similarity to Ideal Solution (TOPSIS), as well as the use of unsupervised neural networks (Self Organizing Maps) to visualize the joint behavior of the different countries [4].

DEA is one of many procedures used in non-parametric efficiency estimation. This method has been successfully used to estimate efficiency in R&D environments. As in the proposal of Chuang et al that describes [6] the relation of efficiency between input and output into an R&D context of the microelectronic industry in Taiwan.

Also Sharma and Thomas [11] examines the relative efficiency of the R&D process across a group of twenty two developed and developing countries from Europe, India and Asia using Data Envelopment Analysis (DEA). The R&D technical efficiency is examined in this paper using a model with patents granted to residents as an output and gross domestic expenditure on R&D and the number of researchers as inputs.

TOPSIS is widely used for solving a Multi-criteria decision-making problem. The concept of TOPSIS is choosing the best alternative according to the relative position in all of the alternatives. It means that an alternative is good if it has the shortest distance from the positive ideal solution (PIS) and the farthest from the negative ideal solution (NIS). The TOPSIS method can provide total ranking order of all decision alternatives.

Amiri *et al* [7] have proposed the integration of both methodologies: DEA and TOPSIS to develop a decision making framework to evaluate the risk of related portfolios of a particular financial market.

We also propose the use of Self Organizing Maps (SOM) [5] that is a kind of an unsupervised neural network, to transform the data that characterize the R&D behavior of the different countries to a two-dimensional grid of nodes while preserving its 'topological' structure. We compare and contrast the 'feature map' generated by the SOM with the results of the DEA and TOPSIS results.

2 R&D Activities in Latin America

In the Latin-American countries context, data of R&D expenditure show significant differences among the countries. In 2008 the expenditure on R&D as a percentage of GDP was for Brazil 1,11% while Colombia barely spent 0,14%, Panamá 0,21% and Costa Rica 0,39%. [9]

Another frequently used indicator to measure R&D activities is the number of patents granted. The figures below show the differences among countries relating number of patents to number of inhabitants (Figure 1) and to R&D expenditure in million dollars purchasing power parity (PPP) (Figure 2).The aim of this paper is to analyze the R&D activities in the countries of the region in order to detect any asymmetries.

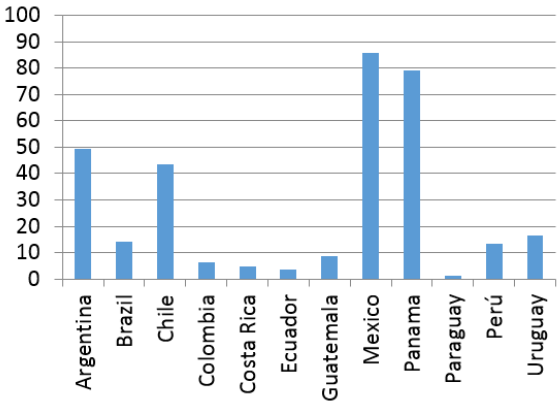


Fig. 1 Patents granted per million inhabitants (2004-2008)

Patents granted demonstrate the ability to turn R&D efforts into developments that can be exploited by industry.

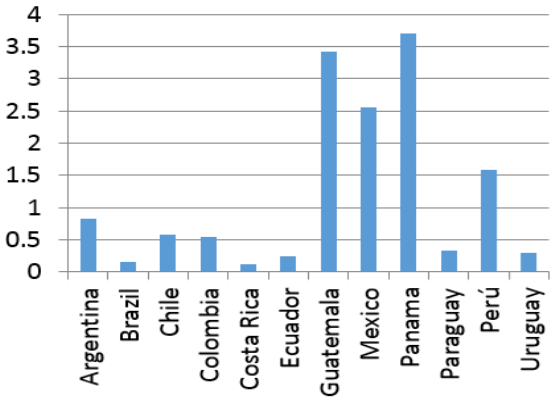


Fig. 2 Relation between patents granted and R&D expenditure (million dollars purchasing power parity (PPP))

Figure 1 and Figure 2 shows a marked asymmetry among countries in R&D budget and activity.

However, it is important to note the use of the number of patents granted as an indicator of R&D activity has some limitations. Not all innovations do necessarily result in patent grants, and their economic value may differ significantly. In some cases, a small number of patents may be very profitable, while in others, a high number of patents may be of little worth.

3 Definitions of Variables

The problem of variable selection was made from a large set of indicators for assessing the performance of the use of R&D countries. For the purpose of organizing data for use in the three proposed methods are presented as input and output variables. Data collected from average figures of period 2004 – 2008, obtained from the Ibero-American Network of Science and Technology Indicators (RICYT) database and the Report on Human Development of the United Nations Development Program (PNUD) [9].

Given the variable data availability, the analysis will be performed on the following countries: Argentina, Brazil, Chile, Colombia, Costa Rica, Ecuador, Guatemala, Mexico, Panama, Paraguay, Peru and Uruguay.

input

RDE: expenditure on R&D in millions of dollars, purchasing power parity.

outputs

PG: number of patents granted.

SCI; number of publications recorded on the multidisciplinary *Science Citation Index*[1] database.
CI: invention coefficient[2].

4 Analysis Methodologies

In this study we applied different methods to analyze the performance of countries in the use of public spending on R&D.

TOPSIS proposes an ordering based on the distance between an ideal target (and anti-ideal) and performance indicators in each country actually observed. These indicators are incorporated as criteria (inputs) to minimize and criteria to maximize (outputs).

Meanwhile DEA is a non-parametric model that calculates the performance of each country and the relative distance between an empirical production frontier (determined by the countries best positioned) and the actual performance of each country according to the observed inputs and outputs.

Self-organizing maps are used to visualize the behavior of different countries depending on their similarity (using only the outputs of the system variables) as performance indicators.

While these methods work from different assumptions, we believe that the comparison of results (obtained from the set of selected indicators) allow interesting conclusions.

4.1 Topsis

Hwang and Yoon (1995) developed TOPSIS (Technique for Order Preference by Similarity to Ideal Solution), a technique based on the idea that it is desirable for a given alternative to be as close to the ideal solution as possible and as far from the negative-ideal solution as possible [4].

An ideal solution is defined as a set of ideal values (or ratings) in relation to a given problem's attributes, even when the ideal solution is usually impossible or not feasible. Therefore, the most rational approach is to find the simultaneous minimization of distance from an ideal solution point and the maximization of distance from a negative solution point.

TOPSIS defines a similarity index (or relative proximity index) in relation to the ideal positive solution, combining the proximity to the ideal solution and the farthest distance from the ideal negative solution. The alternative that is closest to the maximum similarity to the ideal positive solution is chosen.

---

[1] There are different scientific publications bases. Some are multidisciplinary while others are discipline specific. Given that these bases are not mutually exclusive; the biggest multidisciplinary base has been considered in this paper.

[2] CI is defined as patent applications per resident per million inhabitants.

In Figure 3 the position of two alternatives a_1 and a_2 in relation to the ideal benefit attribute (a^+) and to the ideal cost or disadvantage attribute (a^-) are considered [4]. Euclidean distances to the positive ideal and the negative ideal show that, in this bi-dimensional space, a_1 is closer to a^+ and that a_2 is farthest from the anti-ideal a^-. Due to this ambiguity it is necessary to determine the similarity index of both alternatives, which will tend to maximize the relative distance from the negative ideal as compared to the sum of the distances from the positive and negative ideals respectively.

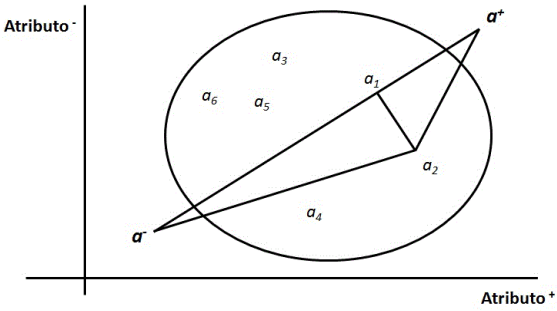


Fig. 3 Position of two alternatives A_1 and A_2, in relation to A^+ and A^-

$$A^+ = \{v_1^+, \dots\dots, v_n^+\} \tag{1}$$

$$A^- = \{v_1^-, \dots\dots, v_n^-\} \tag{2}$$

With vector vj+ showing the best values for the criteria (positive ideal), and with vj- showing the worst or least desirable values achievable for the same criteria (negative ideal), the TOPSIS method determines first the weighted normalized value for each alternative *i* with respect to criterion *j* by applying and Euclidean distance as:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}}} \tag{3}$$

The normalized coefficients r_{ij} are then weighted obtaining the following values

$$v_{ij} = w_j \times r_{ij} \tag{4}$$

Distances of each alternative i to the positive ideal S_i^+ and the negative ideal S_i^- are measured in this way:

$$S_i^+ = \sqrt{\sum_{j=1}^{n} \left(v_{ij} - v_j^+\right)^2} \qquad S_i^- = \sqrt{\sum_{j=1}^{n} \left(v_{ij} - v_j^-\right)^2} \tag{5}$$

Finally, the similarity index to the positive ideal is evaluated as the quotient:

$$C_i^* = \frac{S_i^-}{(S_i^+ + S_i^-)} \tag{6}$$

That is to say, the higher the index C_i^* is, the farther alternative *i* will be from the negative ideal in relation to the total distances from both ideals, and consequently, the more preferable its global position will be.

4.2 Data Envelopment Analysis (DEA)

DEA is a mathematical programming tool which compares the relative efficiency of units that use the same type of input to produce the same group of outputs. DEA models measure the efficiency of each unit against an empirical frontier, such as the distance quotient between a given unit and another feasible, efficient unit on the frontier. The Superefficient BCC model [1] will be used to obtain a full ranking of the universities evaluated, and then this will be used to compare results obtained using the other methodologies [1].

When trying to work out the mathematical formula of the DEA technique, it is necessary to understand the concept of efficiency known as Pareto Koopmans. This concept states that "*a unit is efficient when none of its output can be increased without increasing its inputs and none of its outputs can be decreased without decreasing its outputs*". Efficiency measurements imply a comparison of the output/input relationship.

DEA is a non parametric technique characterized by its flexibility in weight determining and by the use of multiple inputs and outputs. The method operates on the concept of "efficient frontier". Each unit (DMU) becomes more efficient when moving closer to the frontier. As the reference is on the frontier line, each unit will have different "reference units" depending on their relative position in relation to the efficient line.

4.2.1 Model of Constant Returns to Scale (CCR)

Suppose we have *n* units to evaluate (DMUs), where each DMU_j $(j = 1, \dots, n)$ produces *s outputs* and r_j $(r = 1, \dots, s)$ using *m inputs* x_{ij} $(i = 1, \dots, m)$, DEA will use the following measurement of DMU efficiency:

$$h_j = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \tag{7}$$

where v_i $(i = 1, \ldots, m)$ and u_r $(r = 1, \ldots, s)$ are the weights or *inputs* and *outputs* to calculate a weighted sum of *m inputs* and *s outputs* for the DMU_j respectively.

According to Chames, Cooper and Rhodes [2], DMU weights can be determined by the following mathematical programming problem:

$$h_0^* = \max h_0$$

If

$$\begin{gathered} h_j \leq 1, j = 1, \ldots, n \\ v_i, u_r \geq 0 \\ i = 1, \ldots, m, \quad r = 1, \ldots, s \end{gathered} \tag{8}$$

where,

$$h_o = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \tag{9}$$

represents the quotient between the weighted sum of *outputs* and the weighted sum of *inputs* for a given DMU (DMU_o), which means the resolution of as many no linear programs as DMUs. By applying this model to each unit, we will obtain *n* DEA efficiency indices, h_j^* related to each DMU, where each one will be related to optimum weights (*m* + *s*) corresponding to each *input* and each *output*.

It is obvious that the higher h_j^* is, the more efficient DMU_j performance will be. However, the highest possible value is 1- due to the restrictions imposed by the mathematical program. If $h_j^* = 1$ then DMU_j is relatively efficient.

Many authors suggest that these models can only classify units into efficient and inefficient but cannot help rank them. In order to overcome these limitations, modifications to the classic models have been developed, such as the Superefficient Model [1] and the *cross efficiency model* [10].

4.2.2 Superefficient Model

To avoid obtaining more than one DMU with DEA indices equal to one, which would Rank several units first thus making it difficult to have a strict total order, Andersen and Petersen introduced a change in the model. By excluding the $h_o \leq 1$ restriction for DMU_o, we get:

$$\begin{gathered} h_o^* = \max h_o \\ subject\ to \\ h_j \leq 1, j = 1, \ldots, n \ \ and\ j \neq 0 \\ v_i, u_r \geq 0 \end{gathered} \tag{10}$$

Then, there can be efficient DMUs with values higher than one, breaking the ties that are usual when applying DEA and which make strict ranking difficult. This model is known as superefficient.

4.3 Self Organizing Maps

Self Organizing Maps (*SOMs*) introduced by Teuvo Kohonen [5], a well known type of neural network, are a powerful tool for the visual analysis of multiple variables. SOMs are a valuable alternative in data exploration which can be used to complement traditional statistical methods SOMs are a particular type of neural network in which neurons (also called "cells" in this context) work in a two-dimensional arrangement [12, 13].

The dimension of each cell in the network is identical to that of the input vectors (patterns). Each cell is trained and associated to a weight vector called "prototype vector". In the first stages of the map, before training, prototype vectors are given either random values or values that vary linearly on the map depending on the variables of the training data set.

The BMU (*Best Matching Unit*) is the cell whose prototype vector is most similar to the input pattern (using a, generally Euclidean, distance criterion)

During training, prototype vectors are adjusted according to an iteration pattern by using the following equation:

$$W_j(n+1) \leftarrow W_j(n) + \eta(n) h_{ji}(n) \left[X(n) - W_j(n) \right] \tag{11}$$

where n is the iteration number, j is the neuron index considered in the iteration, W_j is the cell's prototype vector j, $\eta(n)$ is the learning rate, $h_{ji}(n)$ is the BMU neighborhood function BMU and $X(n)$ is the input vector (pattern) in the iteration n. The higher the number of iterations, the lower is the learning rate and the neighborhood function scope.

Once SOM training is complete, there are different ways of visualizing information. Each variable value of each cell's prototype vector can be analyzed and represented in separate maps by using different colors or shades of grey. These maps are topologically related and are usually called "feature maps". Alternatively, variable numerical values can be included in each cell instead of, or as a complement of, colors to obtain a more detailed analysis.

Map training quality can be evaluated by considering how closely prototype vectors represent training data. To do this, quantification errors must be defined:

$$E_Q = \frac{1}{N} \sum_{i=1}^{N} \| x_i - m_i \| \tag{12}$$

where x_i is a pattern (datum), m_i is the inner condition of that pattern's BMU and N is the number of patterns in the training set.

In order to assess whether data topology has been preserved, topographical errors must be defined. For all training patterns, the best matching cell in the map (BMU) and the second best matching cell in the map (2nd BMU) must be worked out. If these two cells are adjacent, then there is an error. Total error can be normalized in a 0 to 1 form, where 0 means the perfect preservation of the topology:

$$E_T = \frac{1}{N}\sum_{i=1}^{N} u(x_i), \tag{13}$$

where $u(x_i)$ is 1 if 2nd BMU is not an adjacent cell (topographical position) and 0 if it is [8].

Consequently, during SOM training a sample space projection is built and a map which preserves the grid multivariable pattern topology is generated. Information on group or cluster distance is visualized in the Unified Distance Matrix projection.

The Unified Distance Matrix is worked out from the trained map, taking into account the fact that the map is a bi-dimensional array of cells whose values have been adapted during learning. Distance matrices are typically used to show the mapping clustering. To that end, a matrix where cells are colored according to the distance from neighboring units is shown.

Similar colors are used for cells at similar distances so that groups or clusters can be recognized as areas in the map where there is a small distance among cells and there is a separation zone (area where there is a big distance among cells) from one group to the next.

To analyze each variable individual behavior within the map, component plans are visualized. These help us to find out how each variable behaves in relation to the cluster identified in the distance matrix. Variable plans provide visual information which helps correlate the simultaneous behavior of all the variables considered in the model.

5 Results

5.1 TOPSIS

TOPSIS was applied considering as data model the partial efficiency measures calculated from the ratios of output/ input for each country (PO/RDE, SCI/ RDE and CI/ RDE). To evaluate the distance between a country and the ideal solution and negative-ideal solution used the Euclidean distance measure. The ideal solution is formed by the best partial efficiencies observed country (PO/ RDE, SCI/ RDE and CI/ RDE) between the countries analyzed, while the negative-ideal solution is the composed of the values of these measures have the worst performance among all considered. The similarity index obtained could be interpreted as a measure of performance of each country in relation to theses partial efficiencies.

Similarity index obtained are shown in the Table 1. The ranking shows that Panama, Paraguay and Guatemala have a good performance in relation to R&D expenditure (RDE). Brazil and Ecuador have low performance appearing with lower similarity index.

Table 1 TOPSIS Indices

Country	Index
Panama	0,7777
Paraguay	0,5472
Guatemala	0,5404
México	0,3637
Uruguay	0,3626
Peru	0,2935
Costa Rica	0,2637
Chile	0,2535
Argentina	0,2057
Colombia	0,1493
Ecuador	0,0620
Brazil	0,0056

5.2 DEA

In order to obtain a complete ranking of the countries analyzed, the superefficient CCR model output oriented was applied. It was considered RDE as input and PG, SCI and CI variables as outputs. The objective was to obtain a ranking of the efficiency of expenditure considering multiple outputs (Table 2). Results show that Panama and Paraguay have values higher than 1 (superefficient). Guatemala, Chile, Uruguay and Argentina too have a good performance (higher than 0,70). Brazil and Ecuador the lowest obtained results (less than 0,35).

Table 2 DEA Efficiency Indices

Country	Efficiency
Panamá	2,5150
Paraguay	1,4940
Guatemala	0,9220
Chile	0,7510
Uruguay	0,7430
Argentina	0,7080
Mexico	0,6910
Colombia	0,6400
Costa Rica	0,5560
Peru	0,5380
Ecuador	0,3480
Brazil	0,3330

5.3 *Self-Organized Maps*

A Self-Organized Map was trained with the PG, SCI and CI variables, in relation to the expenditure on R&D in millions of dollars, purchasing power parity RDE. The aim was to evaluate expenditure efficiency as shown by productivity. The map design offered a well preserved topographical space (topographical error = 0.06) and had a grid of 15x 15 cells. Figure 4 a) shows the distance matrix as a 3D visualization and the impacts labeled with the names of the countries whose cases were studied.

In Figure 4 b) the chosen colors show similarities among cells. Impacts were also labeled with countries names.

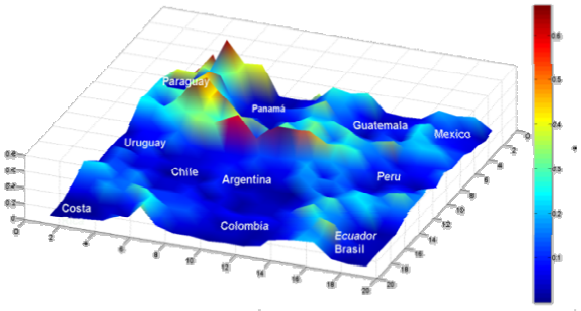


Fig. 4 a) Self-organized maps. Distance matrix and impacts

Fig. 4 b) Self-organized maps. Map coloring shows cell similarity

By analyzing Figures 4 a) and 4 b) we can see there are areas with different behaviors. In Figure 4 a) (Unified Distance Matrix), we find that Paraguay and Panama is, in Euclidean terms, are far away from the rest of the countries, since its impact region is distinctly separate by high value cells (in bright colors).

In Figure 4 b) we also see Paraguay is different from other countries (light green). In Figure 4.a) Paraguay is surrounded by high distance cells. In this figure, a clear difference between the upper left corner -Panamá and Paraguay, and the

bottom right corner, where Brazil and Ecuador are. All the other countries share similarity areas.

In Figures 5, 6, and 7 show maps of the standardized variables SCI, PO and CI related to the spending on R&D (RDE), respectively. On maps of variables remains the topographic location of each country. Thus we can assess the level of each variable in the map and interpret similarities and differences between the cases.

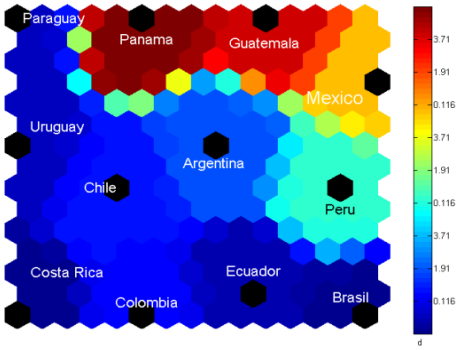


Fig. 5 Variable PG/RDE Map. Color Bar shows values

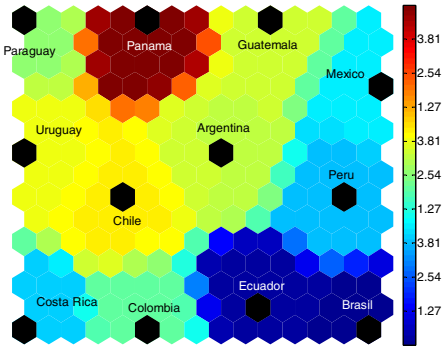


Fig. 6 Variable SCI/RDE Map. Color Bar shows values

Looking at the results of the mapping of the analyzed variables generated by the SOM is displayed a particular behavior of Panama, which has a high level of variables PG/RDE SCI/RDE, in relation to the other countries analyzed, and presented an average level of the variable CI/RDE. Colombia, Argentina, Chile and Uruguay have strongly homogeneous behavior in Fig 4, low level of patents granted, and some variability in maps and SCI variables CI. A homogeneous region is where impact Brazil and Ecuador which have similar behaviors in the three variables

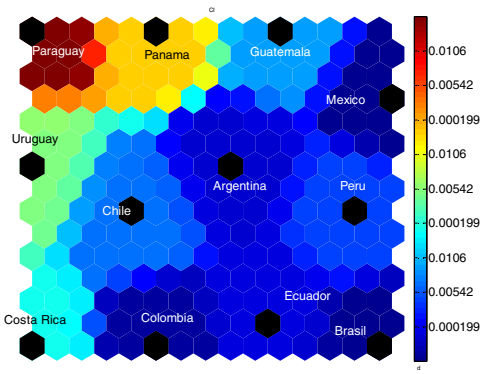


Fig. 7 Variable CI/RDE Map. Color Bar shows values

6 Discussions

By comparing the results obtained using the three different approaches, we could see the efficiency ranking did not suffer significant variations. Neither were there major differences in the cardinal comparison of results. Differences arose from the way in which the different methods work. DEA is a mathematical programming method which analyses the problem globally and in which all DMUs are compared at the same time, contrasting the input/output relation under evaluation against all other units in the problem.

The TOPSIS results, calculated from data considered partial efficiencies measures, shows to Panama, Paraguay and Guatemala as the country's best positioned in the ranking of performance. It notes that Brazil and Ecuador have the values furthest from the ideal and near negative-ideal. They observed a group of countries with good performance indices are very close together (Peru, Costa Rica, Chile and Argentina), which is consistent with that shown by the SOM map.

Note that the results of the DEA method show a behavior similar to those obtained by TOPSIS, best positioned countries were again Paraguay, Guatemala and Chile meanwhile Brazil and Ecuador remain in the last places in the ranking.

The performed analysis denoted that a) the DEA method is relatively sensitive to the input and output specifications; b) the TOPSIS model which relates Product and Resource has proved to offer rankings which are the closest to DEA's, and whose global ranking agrees with this method in the scoring of the most efficient and least efficient units. These facts support the choice of this method as control of the estimates given by DEA.

Self-Organized Maps results, designed with the product variables in relation to investment, graphically show country distribution according to their RDE efficiency. The possibility of visualizing results and variable behavior simultaneously makes SOMs a useful tool for decision making.

Due to lack of comparable data, it was not possible to analyze all 24 Latin-American countries. A full comparison of all countries, together with the applica-

tion of other analyses methodologies will be the subject of our future work. As mentioned above, the results are not conclusive since they are conditioned by data quality and availability. It is vital, then, to improve data collection and count on statistical information drawn on common criteria for all the countries in the region. With more reliable information, further studies and valid comparisons will become viable.

7 Conclusions

Becoming an economy knowledge-based as a long-term strategic objective must be a goal for our region. It also would be a significant step towards achieving local growth and development. To that end, it is essential not only to increase the GDP (Gross Domestic Product) percentage assigned to R&D but also to use the budget available in an efficient way.

The aim of this paper was to analyze the effect of the varying levels of expenditure on the ability to innovate as measured by the number of patents, publications and invention indices. The results confirm that there is an imbalance among countries in the area- imbalance that should be redressed in order to increase the global efficiency of the region.

This analysis clearly shows that R&D must be improved in countries with low performance measures to increase their performance. Topic is beyond the scope of this paper and will be analyzed in future studies.

References

1. Andersen, P.Y., Petersen, N.C.: A Procedure for Ranking Efficient Units in Data Envelopment Analysis. Management Science 39, 1261–1264 (1993)
2. Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the Efficiency of Decision Making Units. European Journal of Operational Research 2, 429–444 (1978)
3. Council of the European Union: Presidency Conclusions Brussels European Council (2005), http://Www.Europarl.Eu.Int
4. Hwang, C.L., Yoon, K.: Multiple Attribute Decision Making: Methods and Applications. Springer, New York (1981)
5. Kohonen, T.: Self Organized Formation Of Topological Correct Feature Maps. Biol Cybernetics 43, 59–96 (1982)
6. Chuang, L.M., Liu, C.C., Chao, S.T.: Data envelopment analysis in measuring r&d efficiency of semiconductor industry's new product development in Taiwan, http://eco-science.net/f_articles/205-data-envelopment-analysis-in-measuring-rd.html
7. Amiri, M., Zandieh, M., Vahdani, B., Soltani, R., Roshanaei, V.: An integrated eigenvector–DEA–TOPSIS methodology for portfolio risk evaluation in the FOREX spot market. Expert Systems with Applications 37(1), 509–516 (2010)
8. Pölzlbauer, G.: Survey and Comparison of Quality Measures for Self-Organizing Maps. In: Proceedings of the Fifth Workshop on Data Analysis (Wda 2004), pp. 67–82 (2004)

9. Red de Indicadores de Ciencia y Tecnología Iberoamericana e Interamericana, http://www.ricyt.org/
10. Sexton, T.R.: The Methodology of DEA. In: Silkman, R.H. (ed.) Measuring Efficiency: An Assessment of DEA, pp. 73–104. Jossey-Bass, San Francisco (1986)
11. Seema, S., Thomas, V.J.: Inter-country R&D efficiency analysis: An application of data envelopment analysis. Scientometrics 76(3), 483–501 (2008)
12. Tasdemir, K., Merenyi, E.: Exploiting Data Topology in Visualization and Clustering Of Self-Organizing Maps. IEEE Transactions on Neural Networks 20(4), 549–562 (2009)
13. Vesanto, J.: Data Exploration Process Based on the Self-Organizing Map. Acta Polytechnica Scandinavica, Mathematics and Computing Series, No. 115, Espoo, p. 96. Published by the Finnish Academies of Technology (2002)

Tool Based Assessment of Electromobility in Urban Logistics

Tim Hoerstebrock, Axel Hahn, and Jürgen Sauer

Abstract. Compared to conventional vehicles with combustion engines, electric vehicles have several advantages concerning sustainability and efficiency. Unfortunately, these advantages are bound to low ranges of the vehicles and long charging times due to the battery as energy source. In addition, the expensive battery increase the investment cost of the vehicle. In case of private users, these costs cannot be amortized by the relatively low electricity price due to the low utilizations of the vehicle. Car sharing could be a possible answer to deploy electric cars in urban regions nevertheless. The objective of our research is to assess the feasibility of exchanging conventional vehicles through electric powered ones within a car sharing fleet. The goals of this analysis are to determine possible exchange rates of the vehicles, to specify the required charging infrastructure and to evaluate the effect on the quality of service in terms of availability of the vehicles. In order to achieve these goals, we developed a multi-agent framework that simulates vehicles with new drive systems in existing transportations systems in general and the potential of electromobility in existing road networks in particular. In this chapter, we explain our approach and evaluate the feasibility of electric vehicles in a particular car sharing fleet operating in the city of Oldenburg, Germany. We evaluate two customer patterns: working day and weekend. The results show that the weekend scenario leads to several fuel shortages – in contrast to the working day scenario. The findings indicate that a more intelligent booking system or a quantitative expansion of charging stations would lead to a higher reliability and user acceptance.

1 Introduction

Electric vehicles hold a high potential when it comes to sustainable transport because of its local zero emissions and the high efficiency of their engines.

Tim Hoerstebrock
OFFIS Oldenburg, Germany
e-mail: Tim.Hoerstebrock@offis.de

Axel Hahn · Jürgen Sauer
Carl von Ossietzky University of Oldenburg, Germany
e-mail: {hahn,sauer}@wi-ol.de

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*,
Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_25,

Compared to the conventional vehicles, the electric powered vehicle pays off in terms of running costs due to the relatively low electricity price. However, batteries of these vehicles are expensive and have a lower energy density than gasoline. That leads to high investment costs, short range and long charging times. Besides the loss of comfort, the vehicle user might not cover their mobility needs with the new technology. However, even when the vehicles meet the mobility need, the capacity utilization of the vehicles is likely to be so low that the higher investment cost of the electric vehicle cannot be amortized. In order to cope with these disadvantages, car sharing in urban regions could be one possible answer. Urban regions provide a good foundation to deploy electromobility due to the particular mobility patterns (short trips, long standing times) of its inhabitants. Furthermore, fleet owner have an advantage compared to private users because of the high occupancy of the vehicles. This utilization and the relatively low electricity price shorten the payback period of these vehicles. However, electromobility opens new questions for car-sharing companies due to the above-mentioned restrictions of the electric vehicles. They have to deal with questions which vehicles can be replaced by an electric car (technical requirements), which charging infrastructure is needed and whether the electric car fleet can handle the customers' requests with the same reliability and efficiency like conventional cars (quality of service). Current tools and methods of transportation planning do not offer an integrated solution for these questions.

Thus, our aim is to develop an evaluation methodology that helps us assessing new drive systems in transportation systems in general and the potential of electromobility in existing road networks in particular. The basis builds a multi-agent simulation tool that assesses the charging infrastructure layout of a given region with respect to the individual traffic patterns of the transportation system user. In this article, we explain our methodology and its closely coupled tool. For evaluation, we investigate the potential of electric vehicles in a car-sharing fleet in the city of Oldenburg, Germany. Therefore, we enhance the simulation tool to consider fleet specific movement patterns. Key measures in this investigation were breakdowns due to energy-shortage and waiting time due to the charging process. Two typical scenario sets of customer behaviour were applied: working day and weekend.

The remainder of this article is structured as follows: section 2 summarizes the related work on transportation engineering. Section 3 describes our approach, which is followed by implementation of the data model and simulation tool (section 4). Section 5 provides the investigated scenario to evaluate the potentials of electromobility within the car-sharing fleet in Oldenburg, Germany. The last section concludes the article and gives an outlook on future work.

2 Methods of Transportation Engineering

Traffic in terms of mobility patterns has a huge effect on the first deployment of electromobility. Especially its charging infrastructure needs to be placed at locations where users depend on it. In this chapter, we describe related approaches,

which are used to solve similar problems. We will focus on a micro- and mesoscopic models and give an overview of current research how to model users' mobility behaviour, since these works are important for our assessment.

Considering the microscopic view, there exist varieties of approaches that tackle all kinds of problems of transportation engineering. The tools usually cover only distinct problems and a limited catalogue of targets. The reduction of complexity is needed to handle this level of detail [27]. This also explains the high diversity of the methods and tools.

As one will notice in the following literature review, scientists often use multi-agent systems to tackle their problems. The reason this is done is that a transportation system is hard to describe with an analytic model because of the complex relations of its elements, the non-deterministic behaviour of its participants and the high dependence to external settings like weather conditions. The multi-agent approach is so appealing for transportation related problems because of it dynamic, large-scale and distributed elements [33]. Moreover, computing resources have become highly available and affordable which favour the use of this paradigm.

A main task before instantiating microscopic models is to determine the traffic behaviour of the participants in a transportation system and to map this behaviour to the desired model. Jochem gives an overview of the different methods and models that have been used to apply user demand to transportation models [14]. Usually this process is divided into four distinct steps: trip generation, trip distribution, modal choice and route assignment. This method of demand modelling and traffic assignment is trip based that means that every person acts in the same manner as long as they travel for the same reason (e.g. go to work). In order to model people's irrationality on a fine, microscopic level, this approach is complemented by other researchers who follow multi-agent-based approaches [12]. The idea is to transfer the whole process of activity planning, modal choice and route choice to the individual virtual traveller. This requires that each individual has access to his own schedule within the target system. With this approach the system designer gets a more realistic and flexible model of the individual behaviour within a transportation system. This allows the system designer, to model non-rational behaviour like people's choice to take the car although public transport offers a faster connection [12]. There are different ways to model this behaviour. Rosetti et. al [26], for example, use the BDI (beliefs, desires, intentions) architecture which characterizes the way the agent sees the world (beliefs), the goals they pursue (desires), and predefined (hierarchic) plans they can use to achieve their goals (intentions). The drawback of modelling every traveller on his own is the high requirement of data to be able to describe the decision process of the individual traveller. Usually this data is not available and must be obtained from general high-level data. Balmer et. al [3] introduced a demand-modelling framework that is capable of using different data inputs to generate these behaviours and export them to downstream tools for further usage. They evaluated the approach on two regions: the Kanton Zurich, Switzerland and Berlin/Brandenburg, Germany.

After modelling the behaviour of each individual entity, simulation runs are made to evaluate the impact of the combination of behaviours on the transport system. MATSim [5, 19], OPUS [30], AIMSUN [2, 4], VISSIM [24] and MITSIM [18, 34] are all microscopic simulation tools, which are capable to do so (not exclusively). They try to simulate the physical phenomena as close to reality as possible to get the most realistic picture of vehicle movement as possible [5]. The goal of these tools is to display temporal and spatial vehicle movement in order to investigate network capacity and to explain the occurrences of traffic jams and/or rush hours. The results of this analysis are time-based bottlenecks of the transportation system. After this determination, possible solutions are developed in order to resolve these bottlenecks. In literature, these solutions have a wide range. They reach from experimental road assignments [21, 28, 29] over optimizing the signal plans [6, 17, 25] to the introduction of intelligent infrastructure and management systems [1, 11, 15]. The reason for these approaches lie in the inflexible transportation system and the corresponding expensive modifications of road infrastructure [8, 26]. Whereas the first two areas are not relevant for our questions, intelligent infrastructure can be a promising measure to help the electric vehicle user to overcome the disadvantages of the new technology.

The basis of the approach is that the infrastructure pictures the real-time overview of the actual condition of the transportation system and that this information is passed to the user so that he can accordingly revise his route choice [7]. Therefore, investments are reduced in comparison to network extensions and the road information will lead to a better utilization of the transportation network. Unfortunately, the user does not always act rationally to the passed information so that congestion and other ineffective situations remain. To investigate this issue, scientist use the individual traveller model described above. They extend the traveller models by reasoning to get to a realistic route choice behaviour. This reasoning is implemented in different ways. Wahle et. al [31, 32], for example, use a simple approach with only a two-road system to determine, first, if information first leads to a better capacity utilization and, second, if an equilibrium is reached. Wahle et. al uses a cellular automaton proposed by Nagel and Schreckenberg [20]. The vehicles that leave the system announce their travel time. Entering vehicles can see on a board which road takes the shortest time. According to their strategy (static or dynamic), they choose their road randomly or by the information they receive. Wahle et al. conclude that this sort of information leads to oscillations in the road utilizations, which makes it hard to manage traffic. Klügl and Bazzan [16] investigate a similar scenario in which commuters have two road choices: main road or side road. The main road is shorter than the side road but accomplishment will take longer if congested by too many vehicles. The commuter has to choose between these two roads. Klügl and Bazzan developed heuristic that takes the experience of the driver into account. The driver remembers what reward he has gained with what decision in the past and acts more likely to his experience. This leads to equilibrium of travel time (reward) on the main and the side road. The authors come to the same conclusion as Wahle et al. after introducing traffic forecast into their scenario. If all drivers act to the information in the same way, the system

will get into an ineffective status (oscillation). Gringmuth et al. [10] extend the scenarios by informing the virtual driver not only on but also before his trip. Therefore, the decision making of the agent gets a lot more complex since the decision is not only reduced to which road the traveller takes but also when to start his trip, which mode to choose and which activity to undertake. They use the program mobiTopp, which characterizes the user's behaviour based on data from the German mobility panel [13]. Combined with the traffic simulation VISSIM and VISUM [24] they investigate the outcome of pre-trip information on the total street network.

In order to evaluate electromobility and its element, the multi-agent approach is promising since each individual user can be assessed and time-based bottlenecks and missing spatial coverage of charging infrastructure can be identified. Unfortunately, the above-mentioned simulation tools cannot be applied to the questions in this article due to the following reasons:

- The tools only focus on the road network. The requirement of our approach is to develop a tool-based methodology to assess new drive systems in transportation systems, generally. Therefore, our approach needs to be decoupled from any specific transportation system. This allows, for example, the investigation of new drive systems in the maritime domain.
- The tools mostly focus on a realistic physical representation of the car flows. The focus of our methodology is to assess new service infrastructure, technical features of the vehicle and quality of service. A realistic physical representation would overshoot our aims.
- The tools neglect the technical features of vehicles. Since gas stations are highly available and the tank of the cars are dimensioned in that way that ranges above 600 km are reached, fuel shortage was not a topic of traffic simulation. Thus, the technical features of vehicles were neglected. For our scenario, the modelling of these features is highly important since the ranges and consumption behaviour of the electric vehicles determine the applicability.

With respect to the last point, it should be noted that the user must be optimally supported in his mobility behaviour in order to cope with the negative effects of electric vehicles. This includes the installation of a charging infrastructure to provide energy to the user at many places. Therefore, the mobility patterns are major inputs for our assessment methodology and highly influence the location and technical equipment of the charging infrastructure. As stated above, this is bound to high data demand to display these patterns on a microscopic level. The mentioned methods and tools especially by Balmer et. al [3] to convert general data into individual schedules is a promising approach. In addition, the use of electric vehicles will be closely connected to new telematics applications [9] that allow the user to easily find and book the next available charging station. Furthermore, these systems shall lower the range anxiety of the user by real-time information about the remaining range of the vehicle. This opens the question of user behaviour in dependence of this information. Our desired solution should consequently be able to integrate the above methods to describe users' behaviour according to the information passed.

3 Approach

As stated above, not only the road traffic but also other transportation domains face strategic questions when it comes to adapting new technology. Therefore, our approach needs to cover all types of transportation systems (and their combination) like sea, railway or air traffic.

The challenge of our approach was to set up a highly generalized abstract data model, which can be extended to fit (almost) every research purpose in the context of introducing new propulsion technology within existing transportation systems. The researcher (user) is able to derive vehicles (e.g. ship, train, car, airplane), transhipment places (e.g. charging stations, gas stations, airports, ports) and domain specific route networks. The goal is to be as specific as needed concerning the domain problem, but generic enough to be adaptable to future changes inside/outside of the domain.

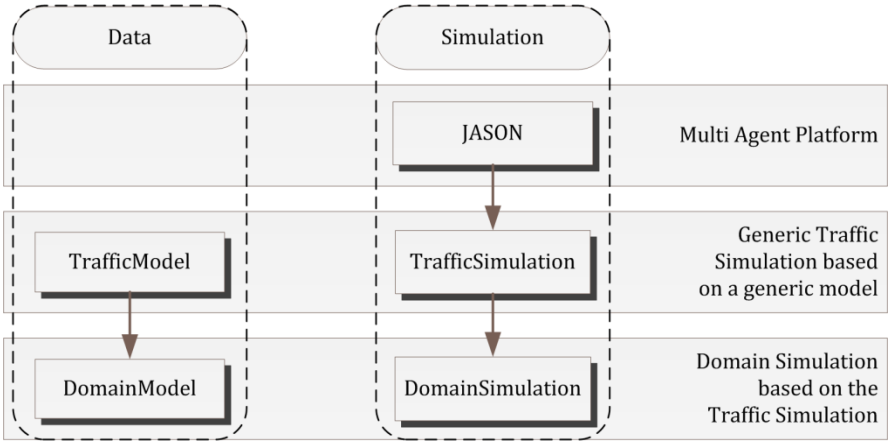


Fig. 1 Approach overview

Our approach incorporates the use of a MAS. Technically, the simulation is discrete event driven. The reason for choosing a discrete event simulation was to simulate only necessary time slots. This allows the user to simulate large timespans efficiently without wasting any computing power for simulating unnecessary simulation time. Additionally, we are confronted with a system that highly fits to the characteristics of an ideal application for multi-agent technology defined by Parunak [23]: modular, decentralised, changeable, ill structured and complex. Thus, we enabled the agents to act simultaneously and designed behaviour dependent communication between these agents, which might invoke further actions. When combing discreteness and communication between autonomous agents, it is essential to care about altering discrete time slots. For instance, while charging, an

EV-Agent might prolong the duration. Such an alteration would also cause the corresponding charging station to change charging plans due to the extended time span. Hence, alternation of the discrete time slots needs to be cascaded throughout every connected agent and its behaviours.

The various agents can follow any goal and therefore need to be customized to focus a specific user defined research purpose. The architecture consists of three levels (see fig. 1). The first level is the abstract implementation of a multi agent simulation platform called JASON. It contains a basic agent that can handle different behaviours simultaneously; a discrete scheduler for actions of various agents in parallel and the possibility to establish communication between agents (see section 4.1). The traffic simulation and the traffic model are positioned on the second layer. The traffic model was built to reflect all necessary abstract objects of the reality. These objects are any types of vehicles, stationary loading units and a network. Each vehicle and stationary loading unit is applied to an agent. This assignment serves as a connection of the traffic model to the traffic simulation. The traffic simulation is a derivate/extension of the JASON package. It describes how vehicles are able to move on the network and how they can load and unload any type of fuel or cargo. In order to receive a tool for the concrete research purpose, three steps need to be done: (1) derive a specific data model from the traffic model, (2) derive a specific simulation model from the traffic simulation and (3) link both derivations with each other. These steps lead to the third layer including the domain model and the domain simulation.

The next chapter deals with the concrete implementation of the three layers including their corresponding data and simulation models.

4 Implementation

We separated our implementation in two major tracks: model and simulation (see fig. 2). The traffic model within the model track was created using the Eclipse Modeling Framework (EMF) with its Ecore UML-dialect. The advantage of EMF within this approach is the model driven development, which implies automatic code generation of the model classes and the model editor classes. The generated editor enables the user to create or edit instances of the domain model over a graphical user interface making the creation of models over program code obsolete. This allows the user to easily setup his scenarios for evaluation.

Since the specific simulation mainly includes simulation logic, there is no advantage to use the EMF to support the simulation track also. The domain specific simulation packages (domain simulation) therefore are specialized in the conventional programmatic way. Further details about the concrete implementation are provided in the following sections.

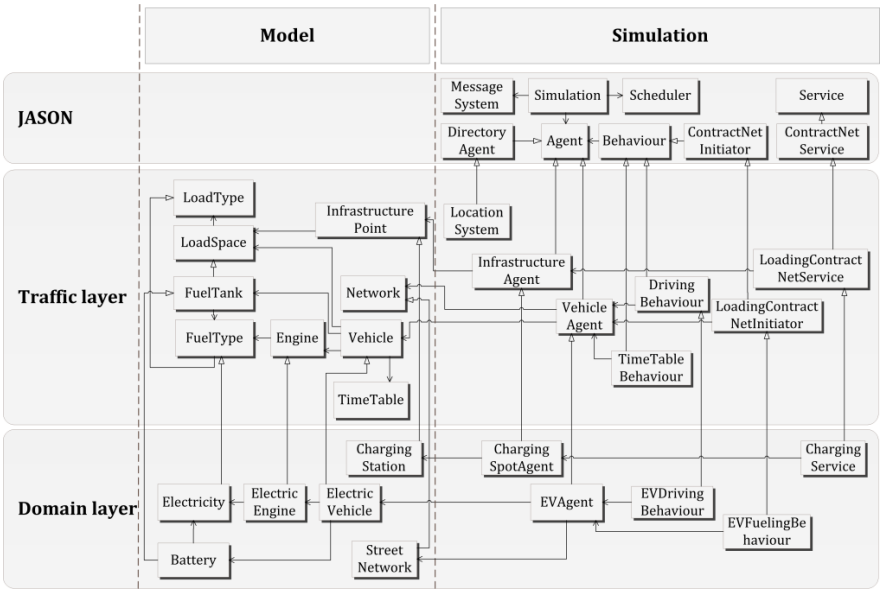


Fig. 2 Design of the three layers

4.1 JASON – Java Agent Simulation OldeNburg

The basic simulation framework JASON provides the necessary functionalities of a discrete and event based multi-agent simulation (see fig. 3). The AGENT class is the basic superclass for user-defined agents. Agents act inside a simulation and can interact through message-based communication. Every agent has an agent description with its unique name and supported services, e.g. fuelling. An agent acts through behaviours that are scheduled by the agent to perform actions at a specific simulation time. The agent provides basic functionality for its behaviours, e.g. sending and receiving messages, scheduling and un-scheduling of behaviours and the ability to add or remove provided services.

The class SIMULATION is the core of the Framework. All agents have to register with it. Registered agents are then able to schedule their behaviours into the SCHEDULER at a specific simulation time. Since the simulation is discrete, only time steps with scheduled behaviours will be executed. Thus, the simulation does not need to simulate every time step but only steps, when behaviours need to act.

The *Behaviour* class is the basic super-class for user-defined behaviours. Behaviour contains the logic of specific actions an agent can perform. They can schedule themselves or other behaviours within the agent. As long as an action is performed, it is guaranteed, that the current simulation time will not continue. Behaviours can perform actions that last over a specific period by rescheduling themselves. The *Scheduler* always calls the action method of a behaviour; therefore, a behaviour has to remember its own status and the corresponding tasks that are involved. Due to an agent's lock, it is guaranteed, that only one of the agent's behaviours runs at a time.

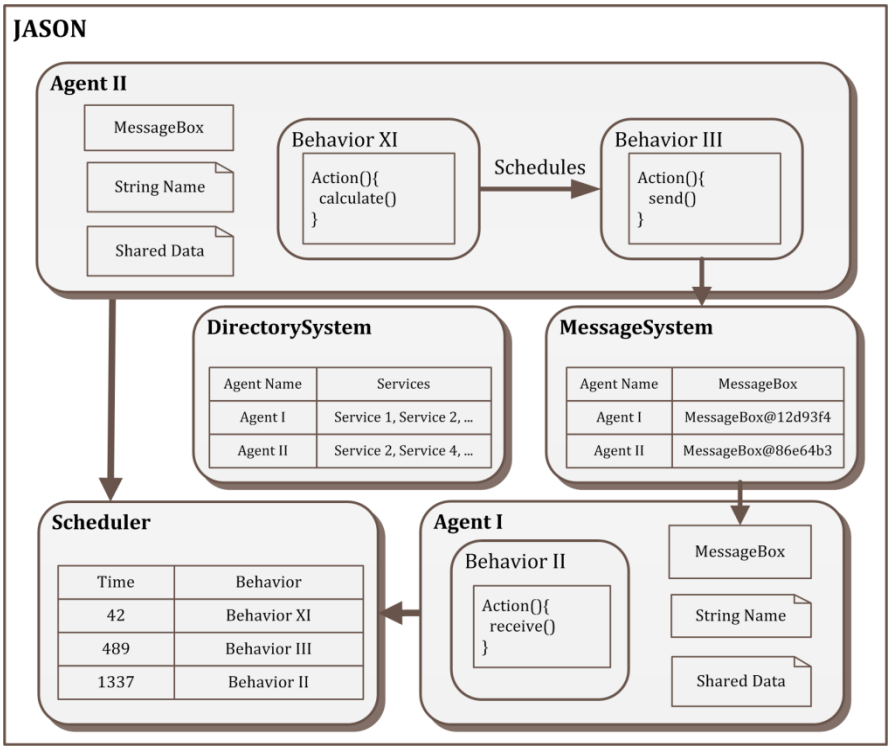


Fig. 3 Abstract design of JASON

Communication is performed through messages sent and received by the agent. For the communication of agents, a dedicated message system is provided. All agents registered with the simulation can be addressed. The agents provide an interface for behaviours to send messages through the agent. Thus, all behaviours are able to communicate in parallel. If an agent sends a message with a valid receiver, it will get an answer in any case, at least a *NotUnderstood* message. The simulation framework itself provides basic communication protocols in accordance with the FIPA protocols. Therefore, messages are designed aligned to the FIPA ACL Message Standard [22].

The implementation of the protocols provides basic functionality like creation of pre-designed messages. Users only have to set certain parameters to create a valid message inside a protocol. The main reason for implementing pre-designed protocols is to guarantee easy negotiations between agents, especially with 1:n relations. Already included in the framework are the contract net, subscribe, and register protocols. Every protocol is based on the initiator and responder classes that provide the necessary interfaces. This design allows a logical and easy to understand disjunction of the participating parties in the protocol.

Agents supporting different services need to register them with a Directory System Agent, which is acting as yellow pages. With this agent, other agents are enabled to search for supported services. The agents' services are registered with the system in an agent description, containing all supported services and the agent's name, to contact it directly if a service should be called on.

4.2 Traffic Layer

The traffic layer consists of two parts: traffic model and traffic simulation. The traffic model contains the general elements of traffic system like vehicles and infrastructure. They include technical specifications like capacity, loading/unloading patterns, maximum power of the engine, current and maximum speed. The traffic simulation covers the abstract agents and behaviours that are necessary to represent a transportation system with JASON. In the following, the traffic model and the traffic simulation will be explained in more detail.

Traffic Simulation. In an effort to create a generic simulation framework, an extension of JASON has been implemented. This extension enables the simulation of all kinds of transportation systems with only few extensions needed. For that purpose, the basic agents, behaviours and services from JASON are extended to represent the basic units of a transportation system. The two main agents are *InfrastructureAgent* and *VehicleAgent*. The *InfrastructureAgent* represents an entity that provides services to other agents in the system. In our case, this service is mainly loading and unloading some sort of goods. This function is covered by the *LoadingContractNetService* class, which extends the basic *ContractNetService* of JASON. The *InfrastructureAgent* registers the service to the *LocationSystem*, which extends the *DirectoryAgent*. This system allows other agents to find services on the network (e.g. gas stations).

The VehicleAgent is an entity that moves within the transportation system to pursue certain targets. They have certain behaviours to reach these goals. The three main behaviours in the traffic layer are TimeTableBehaviour, DrivingBehaviour and LoadingContractNetInitiator. The TimeTableBehaviour is supposed to be a part of the inner goal of an agent. It holds the information about the next steps an agent should perform. The basic goal in this abstract layer is to execute all defined action on schedule. The DrivingBehaviour is supposed to accomplish the driving on a network. For that purpose, it will check the generic cost of a connection. Within the concrete implementation, these costs can reflect, for instance emissions, fuel or time. The costs are calculated by a Dijkstra algorithm to find the cheapest route efficiently. If enough energy is available, the DrivingBehaviour will start driving and reschedule when the route will be finished. While an agent is driving, it is still able to receive messages and react on events. The LoadingContractNetInitiator behaviour initiates a negotiation between the VehicleAgent and the InfrastructureAgent that provides the appropriate LoadingContractNetService. After the two parties agreed on price and amount of the required good(s), the loading/unloading takes place.

Traffic Model. The traffic simulation from above provides the needed infrastructure to simulate the basic events in a transportation system. The traffic model complements this infrastructure by a suitable data structure to represent the agents' technical specifications. This model is needed to compare different sets of equipment and, therefore, to assess new technology in existing transportation systems. The main agents of the traffic simulation (*InfrastructureAgent* and *VehicleAgent*) are connected to their corresponding data models (*Vehicle* and *InfrastructurePoint*).

The technical elements of the *Vehicle* cover its *Engine(s)*, the corresponding *FuelType(s)* and one or more additional *LoadSpace(s)* if necessary. The *Engine(s)* contain several specifications about efficiency, maximum power, torque, etc. Every *Engine* is bound to one specific *FuelType*. Hence, programmatically it is prevented from consuming fuels that it is not designed for (e.g. gasoline in an electric engine). The *FuelType* provides criteria to evaluate its ecological and technological influence like CO_2, NOx, SOx, and energy content. In order to represent logistic scenarios, additional *LoadSpace(s)* can be assigned to the *Vehicle* to represent its transport capacity for other cargo. Each *LoadSpace* can store one or more *LoadTypes* each representing a certain type of cargo. These *LoadSpaces* can be limited by amount and weight.

On the top level, the *Vehicle* provides abstract methods to calculate the overall fuel consumption over a defined timespan and to calculate the optimal power input with respect to fuel efficiency. It also contains strategies about the use of its engine(s). This enables the user to define how the engine(s) should be handled during the simulation. In case of two engines, for instance, one engine could be used as main engine while the second engine is only used within defined situations like dealing with strong headwinds.

As stated in the traffic simulation, the *Vehicle* has a *TimeTableBehaviour* to represent its inner goals. Therefore, a concrete *TimeTable* is provided to specify the destinations, arrival times and the actions that need to be performed at the destinations.

In order to specify which transportation system a *VehicleAgent* can use, the traffic model holds one or more networks to represent roads, railways or seaways. These networks are directly assigned to the *VehicleAgent*.

The *InfrastructurePoint* simply has one or more *LoadSpaces* with certain *LoadType(s)*. It manages these *LoadSpace(s)* and holds the data about possible loading/unloading performance and certain restrictions about the *LoadSpace(s)*. It is assigned to one node of one of the networks of the traffic model.

4.3 Domain Layer

The traffic simulation is not concrete enough to investigate a specific transportation domain. Therefore, a domain simulation is needed which is the final extension. At this point, all abstract classes of the traffic simulation are implemented which leads to a runnable simulation tool. Additionally, domain specific data models are created by extending the traffic model to their final classes.

Domain Simulation. In order to represent our evaluation scenario, the basic functions from the traffic simulation are sufficient. Since the agent classes are abstract and, hence, do not work out of the box, they are extended by the *EVAgent* (extends *VehicleAgent*) and *ChargingSpotAgent* (extends *InfrastructureAgent*). They are complemented by certain setup routines to initiate the simulation.

Domain Model. Just like the traffic simulation, the traffic model needs another extension to represent the specific technical attributes of the EVs involved in our evaluation scenario. This extension can be done by programmatically or model-driven derivation. The model-driven approach includes the utilization of the Eclipse Modeling Framework. The user is forced to implement abstract methods to achieve the required domain specific functionality. This implementation can simply be a taking over the basic functionality of the traffic model or by extending it to fit to the user's needs. The fuel consumption behaviour of an engine, for example, can be modelled as either a linear function only depending on its current power or a multi-criteria function depending on load weight, additional electric consumer and weather conditions.

In order to constitute a runnable domain model, the user must extend at least five abstract classes of the traffic model: *Vehicle*, *Engine*, *Network* (plus the corresponding node class) and at least one action for the *TimeTable*. Additionally, all other classes may be extended. We additionally extended the *InfrastructurePoint*, to represent the *ChargingStation* and its corresponding technical specification like provided voltage, amperage and amount of phases. Another derivation, for example, could be the model of a battery switching station, which provides completely different attributes. After extension, these constructs can be seamlessly integrated into the domain simulation.

5 Evaluation

This chapter provides an evaluation scenario that shows the interaction of the domain model and the domain simulation for assessment. The scenario describes a commercial car-sharing fleet operated in Oldenburg located in the northwest of Germany. The fleet consists of 21 vehicles accessible to every signed up customer of the car-sharing company. This chapter shows how the domain model is instantiated and how the different scenarios are build. Finally, we analyse the potential exchange rate of electric vehicles by using the breakdown rate as the target function.

5.1 Model Instantiation

The model instantiation covers the three classes *StreetNetwork*, *ElectricVehicle* and *ChargingStation* that are explained in section 4.3. The generation of the actual timetables are discussed in section 5.2.

StreetNetwork. To analyse the potential of electromobility in the car-sharing fleet in Oldenburg we modelled the base locations of the different vehicles. In addition, potential destination were identified which consider different trip reasons of the customer. This trip reasons are connected to the destinations. Each position is represented by a node. To model the *StreetNetwork* we build average routes (edges) between the origins and the potential targets taking different types of streets into account (main street, highway, interstate (autobahn)).

ElectricVehicle. The electric car has a capacity of 17.25 kWh and a linear consumption of 0.14375 kWh/km. This leads to a range of 120 km. The charging time to recharge the battery completely is set to 8 hours with a power connection of 230V/16A. For this scenario, a linear charging progress is assumed.

ChargingStation. The *ChargingStation* contains the technical specifications (amperage, voltage, number of phases) and its placement within the region. In order to limit the access to certain usages, the driving reasons of the mobility pattern model can be assigned to the charging station (work, shopping, spare time, etc.). For the evaluation scenario, the infrastructure is placed at the base locations of the vehicles, so that very vehicle has its own charging station. The considered power connection is a conventional one with 230V, 16A and one phase. Charging infrastructure is not placed at the potential destinations.

5.2 Input Data

In order to specify the trips of the vehicles, two scenarios are developed: working day and weekend. Each scenario has subsets concerning different exchange rates. Following assumptions are made:

- The time horizon of the possible usage of the vehicles is set to18 hours.
- The starting time of the vehicle usage follows a normal distribution. The mean average is 8:00 am on working days and 9:00 am on weekends; the standard deviation is set to 50 minutes.
- The standing times at the destinations follow a normal distribution with following attributes: work (mean 8 hours; standard deviation 50 min); shopping (mean 1 hour; standard deviation 10 min); casual activities (mean 2 hours; standard deviation 30 min).

In this paper, the capacity utilization of the fleet is fixed to reduce the number of sub scenarios. It is set to 50%. For each scenario (working day/weekend), the exchange rate rises successively in steps of 10%. For each sub scenario, 10 sets are generated according to the defined procedure (fig. 4). After simulation, the mean averages of the breakdown rate from the 10 sets of each scenario are calculated. Table 1 summarizes the scenarios.

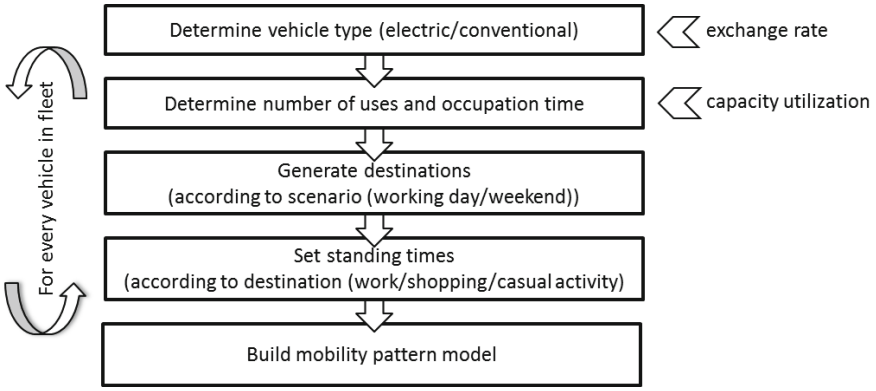


Fig. 4 Trip generation

Table 1 Scenario definition

Scenario	Working day / Weekend	Capacity utilization	Exchange rate	Simulation days
Scenario 1	Working day	50%	10%–100%; Steps of 10%	1
Scenario 2	Weekend	50%	10%–100%; Steps of 10%	1

5.3 Results and Discussion

The analysis deals with the two defined scenarios above. The working day scenario did not lead to any significant breakdown rates. Considering 50% capacity utilization, every exchange rate can deal with the mobility pattern of the working day. The situation is different with the weekend scenario. Fig. 5 shows the mean average of the breakdown rate of the whole fleet. As one can see, the step from 20% exchanged vehicles to 30% leads to a significant raise of the breakdown rate from 0% to 4.5%. From there it rises continuously up to 14.89%. The reason for the two courses of the breakdown rate can be seen in the different route choices and the corresponding times per trip. The weekend scenario provides longer trips with relatively shorter occupation times (see table 2). This leads to a higher number of trips. In the working day scenario, the destinations are closer to the base locations and the standing times at the destination is higher, so the actual distance driven per day is lower.

Considering the above made assumptions, it seems that electric vehicles are not suitable for weekend trips, especially for those trips with longer distances. The workday scenario promises suitable breakdown rates.

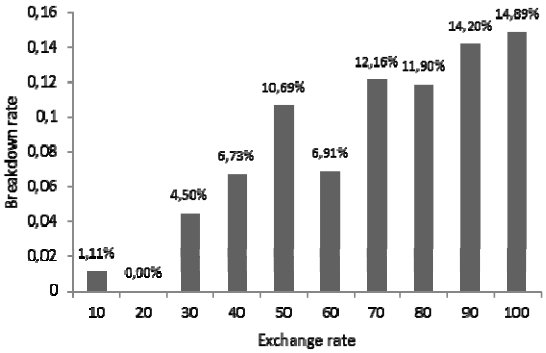


Fig. 5 Breakdown rate analysis (Weekend scenario)

Solutions to reduce the breakdown rate especially for the weekend scenario could be an intelligent booking system that plans the bookings of the vehicles according to the calculated energy capacity and the planned activity. Although this could lead to higher waiting times due to the forced charging times of the booking system. In addition, a quantitative expansion of the charging infrastructure can be added to the model so that charging spots are also available at the destination sites. Furthermore, the quality of the infrastructure can be enhanced to allow a faster charging of the vehicles (e.g. by modelling a fast charging station or a battery switching station). The effect of both expansions on the breakdown rate should be further evaluated.

Table 2 Distances driven by the vehicles (in km)

Scenario	Work	Shopping	Casual activities	Total
Scenario 1	53.80	2.81	0.23	56.84
Scenario 2	0.00	28.18	78.37	106.55

6 Conclusion

In this paper, a methodology was presented to assess electromobility, its required infrastructure and its feasibility to satisfy the mobility needs of car-sharing users. The methodology is supported by an agent-based simulation tool that simulates the vehicle movement and their interactions with the environment. We extended the tool and its corresponding model to be able to investigate the potential of electromobility within fleets. We provided an example scenario to show the elements (region, infrastructure, and vehicle), their instantiation and an evaluation of a car-sharing scenario in the city of Oldenburg, Germany.

Our future work will be devoted to integrating additional systems into the planning model (e.g. booking and planning systems) in order to evaluate the effect of those systems on the whole performance. In addition, we currently develop detailed models of batteries and vehicles to characterize the consumption and refill behaviour. Overall, the model and simulation track of our methodology is easily extensible so that the mentioned additions can be made with a relatively low amount of time and effort.

References

1. Adler, J.L., Blue, V.J.: A cooperative multi-agent transportation manage-ment and route guidance system. Transportation Research Part C: Emerging Technologies (2002), doi:10.1016/S0968-090X(02)00030-X.
2. AIMSUN, http://www.aimsun.com/wp/ (accessed December 15, 2011)
3. Balmer, M., Axhausen, K., Nagel, K.: Agent-Based Demand-Modeling Framework for Large-Scale Microsimulations. Transportation Research Record (2006), doi:10.3141/1985-14.
4. Barceló, J.: Parallelization of microscopic traffic simulation for ATT systems. In: Marcotte, P., Nguyen, S. (eds.) Equilibrium and advanced transportation modelling. Centre for Research on Transportation 25th Anniversary Series, 1971-1996, pp. 1–26. Kluwer Academic Publishers, Boston (1998)
5. Charypar, D., Axhausen, K., Nagel, K.: An event-driven queue-based mi-crosimulation of traffic flow. ETH, Eidgenössische Technische Hochschule Zürich (2006)
6. Choy, M.C., Srinivasan, D., Cheu, R.L.: Cooperative, hybrid agent architecture for real-time traffic signal control. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans (2003), doi:10.1109/TSMCA.2003.817394
7. Dia, H.: An agent-based approach to modelling driver route choice behaviour under the influence of real-time information. Transportation Research Part C: Emerging Technologies (2002), doi:10.1016/S0968-090X(02)00025-6
8. France, J., Ghorbani, A.: A multiagent system for optimizing urban traffic. In: IEEE/WIC International Conference on Intelligent Agent Technology, IAT 2003, pp. 411–414 (2003)
9. Frost, Sullivan: Strategic Market and Technology Assessment of Telematics Applications for Electric Vehicles: Summary of Frost & Sullivan Study (2010)
10. Gringmuth, C., Liedtke, G., Geweke, S., Rothengatter, W.: Impacts of intelli-gent information systems on transport and the economy - the micro based modelling system OVID. In: Advances in Modeling, Optimization and Management of Transportation Processes and Systems: Theory and Practice - 10th Meeting of the EURO Working Group Transportation, EWGT (2000)
11. Hernándes, J.Z., Ossowski, S., García-Serrano, A.: Multiagent architectures for intelligent traffic management systems. Transportation Research Part C: Emerging Technologies (2002), doi:10.1016/S0968-090X(02)00032-3
12. Hunecke, M., Schubert, S., Zinn, F.: Mobilitätsbedürfnisse und Verkehrsmit-telwahl im Nahverkehr. Ein einstellungsbasierter Zielgruppenansatz. Internationales Verkehrswesen 57(1/2), 26–33 (2004)
13. Institute for Transport Studies - Karlsruhe Institute of Technology (KIT). Deutsches Mobilitätspanel, http://mobilitaetspanel.ifv.uni-karlsruhe.de/en/links/index.html (accessed December 15, 2011)
14. Jochem, P.: A CO2 emission trading scheme for German road transport. Diss. Univ., Karlsruhe (2009)
15. van Katwijk, R., van Koningsbruggen, P.: Coordination of traffic manage-ment instruments using agent technology. Transportation Research Part C: Emerging Technologies (2002), doi:10.1016/S0968-090X(02)00034-7
16. Klügl, F., Bazzan, A.L.C.: Route Decision Behaviour in a Commuting Sce-nario: Simple Heuristics Adaptation and Effect of Traffic Forecast. Journal of Artificial Societies and Social Simulation 7, 1 (2004)

17. Kosonen, I.: Multi-agent fuzzy signal control based on real-time simulation. Transportation Research Part C: Emerging Technologies (2003), doi:10.1016/S0968-090X(03)00032-9.
18. Massachusetts Institute of Technology. MITSIMLab Intelligent Transportation Systems, http://mit.edu/its/mitsimlab.html (accessed December 15, 2011)
19. MATSim. Multi-Agent Transport Simulation, http://matsim.org/ (accessed December 15, 2011)
20. Nagel, K., Schreckenberg, M.: A cellular automaton model for freeway traffic. Journal de Physique I (1992), doi:10.1051/jp1:1992277
21. de Oliveira, D., Ferreira Jr., P.R., Bazzan, A.L.C., Klügl, F.: A Swarm-Based Approach for Selection of Signal Plans in Urban Scenarios. In: Dorigo, M., Birattari, M., Blum, C., Gambardella, L.M., Mondada, F., Stützle, T. (eds.) ANTS 2004. LNCS, vol. 3172, pp. 416–417. Springer, Heidelberg (2004)
22. Organisation for Economic Co-operation and Development. Impact of Transport Infrastructure Investment on Regional Development. OECD Publishing, Paris (2002)
23. Parunak, H.V.D.: Industrial and Practical Applications of DAI. In: Weiss, G. (ed.) Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, pp. 377–421. MIT Press, Cambridge (1999)
24. PTV AG. VISSIM - Multi-Modal Traffic Flow Modeling, http://www.ptvag.com/software/transportation-planning-traffic-engineering/software-system-solutions/vissim/ (accessed December 15, 2011)
25. Rochner, F., Prothmann, H., Branke, J., Müller-Schloer, C., Schmeck, H.: An Organic Architecture for Traffic Light Controllers. In: Hochberger, C., Liskowsky, R. (eds.) Informatik 2006 – Informatik für Menschen. Lecture Notes in Informatics (LNI), vol. 93, pp. 120–127. Köllen Verlag (2006)
26. Rossetti, R.J.F., Bordini, R.H., Bazzan, A.L.C., Bampi, S., Liu, R., van Vliet, D.: Using BDI agents to improve driver modelling in a commuter scenario. Transportation Research Part C: Emerging Technologies (2002), 10.1016/S0968-090X(02)00027-X
27. Steierwald, G.: Stadtverkehrsplanung: Grundlagen, Methoden, Ziele, 2nd edn. Springer, Berlin (2005)
28. Vasirani, M.: Vehicle-centric coordination for urban road traffic management: A market-based multiagent approach. Diss. Universidad Rey Juan Carlos, Madrid (2009), http://hdl.handle.net/10115/5146
29. Vasirani, M., Ossowski, S.: An artificial market for efficient allocation of road transport networks. In: Klügl, F., Ossowski, S. (eds.) MATES 2011. LNCS, vol. 6973, pp. 189–196. Springer, Heidelberg (2011)
30. Waddell, P., Borning, A., Sevcíková, Socha, D.: Opus (the Open Platform for Urban Simulation) and UrbanSim 4. In: Proceedings of the 2006 International Conference on Digital Government Research, pp. 360–361. ACM, San Diego (2006)
31. Wahle, J.: The impact of real-time information in a two-route scenario using agent-based simulation. Transportation Research Part C: Emerging Technologies (2002), doi:10.1016/S0968-090X(02)00031-1.
32. Wahle, J., Bazzan, A.L.C., Klügl, F., Schreckenberg, M.: Decision dynamics in a traffic scenario. Physica A: Statistical Mechanics and its Applications (2000), doi:10.1016/S0378-4371(00)00510-0.
33. Weiss, G.: Multiagent systems: A modern approach to distributed artificial intelligence. MIT Press, Cambridge (1999)
34. Yang, Q.: A Simulation Laboratory for Evaluation of Dynamic Traffic Manage-ment Systems. PhD thesis. Massachusetts Institute of Technology (1997)

An Evolutionary Multi-objective Algorithm for Inferring Parameters in Outranking-Based Decision Models: The Case of the ELECTRE III Method

Eduardo Fernández, Jorge Navarro, and Gustavo Mazcorro

Abstract. Methods based on fuzzy outranking relations constitute one of the main approaches to multiple criteria decision problems. In this paper we examine multi-criteria selection and ranking problems for which ELECTRE III is one of the most popular methods. ELECTRE III applications require the elicitation of a large number of parameters: weights and different thresholds; but direct eliciting is often an arduous task for the decision-maker. In this paper, an evolutionary-multiobjective-based indirect elicitation of the complete ELECTRE III model-parameter set is proposed. The proposal needs some kind of "preference knowledge", which is implicit in a set of reference examples illustrating the decision policy from a "decision-maker". This method performs very well in some illustrative examples; its generalization to other outranking methods is straightforward.

1 Introduction

Many real decisions can be modeled by the use of multicriteria decision analysis. Multicriteria methods entail a decision-maker (*DM*) reflecting his/her preferences in a pre-specified mathematical structure. The representation of the *DM* by

Eduardo Fernandez · Jorge Navarro
Autonomous University of Sinaloa, Mexico
e-mail: {eddyf,jnavarro}@uas.uasnet.mx

Gustavo Mazcorro
Instituto Politécnico Nacional (UPIICSA-IPN), Ciudad México, Mexico
e-mail: gmazcorro@ipn.mx

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*, Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_26,

preferential parameters is a crucial aspect in the construction of multicriteria decision models [5]. The development of these models can be based on direct or indirect elicitation procedures. In the first case, the *DM* must specify preferential parameters through an interactive process guided by a decision analyst [7]. Usually, the *DM*s reveal difficulties in defining parameter values whose meaning is barely clear for them. On the other hand, indirect procedures, which compose the so-called preference-disaggregation analysis (*PDA*), use regression-like methods for inferring the set of parameters from a battery of decision examples [7]. It is a way of managing "preference knowledge" which is implicit in a set of reference examples or decision statements. The object of this paper is to find a fitted model consistent with a decision policy embedded in a set of decision examples. According to Doumpos and Zopounidis [8], such examples may be provided by:

a) Former decisions made by the *DM*;
b) decisions on a limited set of fictitious actions for which the *DM* can easily express preferential judgments (decision policy);
c) decisions on a subset of actions under consideration, for which the *DM* is comfortable expressing a decision policy.

There are some early works related to *PDA* paradigm (e.g. [12]). Indeed, the process of assessing criterion weights in value-function and utility models (cf. [3]), may be considered an example of the *PDA* approach. In the framework of Multiple Criteria Decision Aid (*MCDA*), Jacquet-Lagreze and Siskos [11] pioneered the *UTA* method. According to Greco et al. [20], *MCDA* approaches based on disaggregation paradigms are of increasing interest because they imply relatively less cognitive effort from the *DM*. Marchant [13] maintains that the only valid preferential input-information is such arising from *DM*'s preferential judgments in pairwise comparisons. Our interest here is restricted to *PDA* in ELECTRE. In ELECTRE-based models, inferring all the parameters simultaneously requires solving a non-linear programming problem with non-convex constraints, which is usually difficult (cf. [6, 14]. According to Doumpos et al. [7], the relational form of these models and veto conditions may make it impossible to infer the model parameters in real-size date sets. Otherwise, in small data sets the non-linear problem may be ill-determined; there are many different parameter settlements that are compatible with preference information, but no mathematical programming technique is able to describe the whole compatible parameter settlement.

Two recent papers examine the problem of inferring outranking model parameters by evolutionary techniques, both in the context of multicriteria classification problems [7, 10]. In recent years, evolutionary algorithms have rendered powerful tools for solving difficult problems in a variety of fields, in particular, for the treatment of nonlinearity and global optimization in polynomial time (cf. [2]). Doumpos et al. [7] use a differential evolution algorithm for inferring parameter values in ELECTRE TRI method. Fernandez et al. [10] propose an evolutionary multiobjective algorithm for inferring parameters of a fuzzy indifference relation model for classificatory purposes. Compared with single-objective optimization,

the multiobjective approach is (though more complex) more flexible because it allows a richer modeling of preferences. The solution of the parameter inference problem must satisfy several constraints in the parameter space. The *DM* may be unable to establish the model parameters, but he/she may express subjective information about criterion importance and parameter value ranges. As constraints, these expressions may be accounted to reduce the search space and help to obtain more acceptable solutions. As an additional advantage, an evolutionary multiobjective algorithm is capable of generating many good compromise solutions in the associated parameter space. As a result of the evolutionary exploration process, a characterization of the complete set of different model parameter settlements is achieved. This information is then used to obtain a final parameter settlement.

This paper is the first of a paper series delineating an overall contribution. The main goal of the present paper is to develop an evolutionary multiobjective method for inferring the whole set of ELECTRE III model parameters. This approach combines the preference information contained in a reference-set with inter-criteria and intra-criteria soft information, both arising from the *DM*. Further, the proposal makes the exploration capacity of the evolutionary search useful to achieve a better characterization of the set of compatible parameter settlements. In a forthcoming paper the approach will be extended to achieve a more complex model to handling reinforced preferences, as proposed by Roy and Slowinski [18].

Aside of this introduction, the paper is organized as follows: notations and main assumptions are pointed-out in Section 2. The optimization model is outlined in Section 3. The method for inferring ELECTRE III parameters is detailed in Section 4. Experimental evidence is given in Section 5 through an illustrative example. Section 6 contains concluding remarks.

2 Assumptions and Notations

Let us consider a consistent family of criteria $G = \{g_1, \ldots, g_n\}$ defined on a decision set A.

In natural language, the statement "$\boldsymbol{x}$ outranks $\boldsymbol{y}$" (denoted $\boldsymbol{xSy}$) means that the *DM* is sufficiently confident with the statement "$\boldsymbol{x}$ is at least as good as $\boldsymbol{y}$".

Assumption 1: The *DM* can provide a reference set $T \subset A$ composed of action pairs *(a, b)* satisfying the following property: For each $(a, b) \in T$, one of the two statements below is true:

i. a outranks b;
ii. a does not outrank b.

Let $\sigma(\boldsymbol{x}, \boldsymbol{y})$ be a fuzzy outranking relation defined on A. $\sigma(\boldsymbol{x}, \boldsymbol{y})$ may be interpreted as the degree of credibility of the statement "$\boldsymbol{x}$ is at least as good as $\boldsymbol{y}$". σ may be assigned by ELECTRE III method, by PROMETHEE method, or by any other outranking-inspired approach. Here, we are interested in considering how the σ-image depends on a specific settlement of the model parameters (weights and thresholds). Let us denote by $\boldsymbol{P}$ the set of model parameters to be inferred.

Thus, the assessment of the degree of credibility for "x is at least as good as y" is a function $\sigma(x, y, P)$.

Assumption 2: The *DM* has additional information about criterion importance, symmetry, asymmetry, and acceptable parameter ranges. By using such information, the *DM* can express judgments of acceptability or preferences on different parameter settlements.

Now, we denote by $\boldsymbol{P}^*$ a specific settlement of model parameters. Let us consider a real value $\lambda > 0.5$ and the following crisp binary relations on T:

$(\boldsymbol{x}, \boldsymbol{y}) \in S(\lambda)$ iff $\sigma(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{P}^*) \geq \lambda$ (λ-outranking)

$(\boldsymbol{x}, \boldsymbol{y}) \in P(\lambda)$ iff $\sigma(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{P}^*) \geq \lambda \wedge \sigma(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{P}^*) < 0.5$ (λ-strict preference)

$(\boldsymbol{x}, \boldsymbol{y}) \in Q(\lambda)$ iff $\sigma(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{P}^*) \geq \lambda \wedge 0.5 \leq \sigma(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{P}^*) < \lambda$ (λ-weak preference)

$(\boldsymbol{x}, \boldsymbol{y}) \in I(\lambda)$ iff $\sigma(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{P}^*) \geq \lambda \wedge \sigma(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{P}^*) \geq \lambda$ (λ-indifference)

$(\boldsymbol{x}, \boldsymbol{y}) \in R(\lambda)$ iff $\sigma(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{P}^*) < \lambda \wedge \sigma(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{P}^*) < \lambda$ (λ-incomparability)

3 Parameter Inference by Using a Multicriteria Error Measure

A perfect consistency preference model-decision policy is reflected by the equivalence

$$\forall(\boldsymbol{x}, \boldsymbol{y}) \in T \quad \sigma(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{P}^*) \geq \lambda \Leftrightarrow \boldsymbol{x}S\boldsymbol{y} \tag{1}$$

Yet some effects of reinforced preferences (cf. [18]) could make $\boldsymbol{x}S\boldsymbol{y}$ true despite $\sigma(\boldsymbol{x},\boldsymbol{y},\boldsymbol{P}^*) < \lambda$ holds. Using the above λ-relations, (1) can be transformed into

$$\forall(\boldsymbol{x}, \boldsymbol{y}) \in T$$

$$\boldsymbol{x}P(\lambda)\boldsymbol{y} \Rightarrow \boldsymbol{x}S\boldsymbol{y} \tag{2.a}$$

$$\boldsymbol{x}Q(\lambda)\boldsymbol{y} \Rightarrow \boldsymbol{x}S\boldsymbol{y} \tag{2.b}$$

$$\boldsymbol{x}I(\lambda)\boldsymbol{y} \Rightarrow \boldsymbol{x}S\boldsymbol{y} \tag{2.c}$$

$$\boldsymbol{x}S\boldsymbol{y} \textit{ and no effect of intensity of preferences favoring } \boldsymbol{x} \Rightarrow \boldsymbol{x}S(\lambda)\boldsymbol{y} \tag{2.d}$$

Conditions

1. $(\boldsymbol{x}, \boldsymbol{y}) \in P(\lambda)$ with $\boldsymbol{x}.not\ S\boldsymbol{y}$
2. $(\boldsymbol{x}, \boldsymbol{y}) \in Q(\lambda)$ with $\boldsymbol{x}.not\ S\boldsymbol{y}$
3. $(\boldsymbol{x}, \boldsymbol{y}) \in I(\lambda)$ with $\boldsymbol{x}.not\ S\boldsymbol{y}$
4. $(\boldsymbol{x}, \boldsymbol{y}) \in S$ with $\boldsymbol{x}.not\ S(\lambda)\boldsymbol{y}$ *and no effect of intensity of preferences favoring* $\boldsymbol{x}$

are identified as inconsistencies with $P(\lambda)$, $Q(\lambda)$, $I(\lambda)$, and S, respectively. Such discrepancies can be interpreted as errors, deviations of $\sigma(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{P}^*)$ from a good model for the degree of truth of the predicate "$\boldsymbol{x}$ is at least as good as $\boldsymbol{y}$". Such inconsistencies can arise from inadequate assessments of some model parameters.

Let us define the following sets:

$D_P = \{(\boldsymbol{x}, \boldsymbol{y}) \in P(\lambda)$ with $\boldsymbol{x}.not\ S\boldsymbol{y}\}$

$D_Q = \{(\boldsymbol{x}, \boldsymbol{y}) \in Q(\lambda)$ with $\boldsymbol{x}.not\ S\boldsymbol{y}\}$

$D_I = \{(\boldsymbol{x}, \boldsymbol{y}) \in I(\lambda)$ with $\boldsymbol{x}.not\ S\boldsymbol{y}\}$

$D_S = \{(\boldsymbol{x}, \boldsymbol{y}) \in S$ with $\boldsymbol{x}.not\ S(\lambda)\boldsymbol{y}$ *and no effect of intensity of preferences favoring* $\boldsymbol{x}\}$

n_p, n_Q, n_I and n_S denote the respective cardinality of the above sets. Obviously, such values depend on $\boldsymbol{P}$.

Here, we propose to infer the model parameters from the best compromise solution to the multiobjective optimization problem:

$$\underset{P \in R_F}{\text{Minimize}}\left(n_p, n_Q, n_I, n_S\right) \tag{3}$$

where R_F is a feasible region in the parameter space. This region is determined by constraints that the *DM* imposes on the model parameters (Assumption 2). In the remainder of the paper we shall denote by $\left(n_p, n_Q, n_I, n_S\right)^*$ the best compromise solution to Problem 3 in the objective space.

We disregard the use of single objective minimization of some error function or any related criterion. Compared with single-objective optimization, a multiobjective approach is more flexible because it allows of modeling preferences on different objectives. The different inconsistency measures do not have the same importance and should not be merged into a single objective; inconsistencies with $P(\lambda)$ seem to be more relevant. The objective n_Q seems to be a little more important than n_I and n_S. However, since it is difficult to model the *DM*' priorities with respect to n_Q, n_I and n_S, we use a posterior modeling of preferences. Still, the complexity of solving (3) suggests the application of evolutionary algorithms. These are particularly convenient to solve multiobjective optimization problems, because they render approximations to the Pareto frontier in a single run, instead of performing many single-objective optimization processes as it is the case for non-heuristic conventional multiobjective optimization techniques [2, 10]. An evolutionary-based solution to Problem 3 for inferring the ELECTRE III model parameters is developed in the next section.

Beyond the mathematical complexities, several $\boldsymbol{P}^*$ may arise as a premap from $\left(n_p, n_Q, n_I, n_S\right)^*$. From such $\boldsymbol{P}^*$ the decision-maker should select a $\boldsymbol{P}_{best}{}^*$ as "the most appropriate settlement". Indeed, this choice is another selection problem whose solution demands the *DM* consider the following:

- how representative is the particular solution regarding the whole distribution of parameter settlements;
- how capable is such solution to restore additional preference-information not included in the reference set;

- to what extent such solution agrees with *DM's* additional information about criterion importance, symmetry, asymmetry, and acceptable parameter ranges.

These issues will be addressed in sub-section 4.4.

4 Inference of ELECTRE III Parameters by Evolutionary Multiobjective Optimization

4.1 Brief Outline of ELECTRE III

Proposition $\boldsymbol{xSy}$ ("$\boldsymbol{x}$ outranks $\boldsymbol{y}$") holds if and only if the coalition of the criteria in agreement with this proposition is strong enough, and there is no important coalition discordant with it [19]. It can be expressed by the following logical equivalence [17]:

$$\boldsymbol{xSy} \Leftrightarrow C(\boldsymbol{x},\boldsymbol{y}) \wedge\sim D(\boldsymbol{x},\boldsymbol{y})$$

where:
$C(\boldsymbol{x},\boldsymbol{y})$ is the predicate about the strength of the concordance coalition;
$D(\boldsymbol{x},\boldsymbol{y})$ is the predicate about the strength of the discordance coalition;
$\wedge$ and $\sim$ are logical connectives for conjunction and negation, respectively.

Let $c(\boldsymbol{x},\boldsymbol{y})$ denote the truth degree of predicate $C(\boldsymbol{x},\boldsymbol{y})$. In ELECTRE III the degree of credibility of $\boldsymbol{xSy}$ is calculated by

$$\sigma(\boldsymbol{x},\boldsymbol{y}) = c(\boldsymbol{x},\boldsymbol{y}).Nd(\boldsymbol{x},\boldsymbol{y}) \tag{4}$$

where $Nd(\boldsymbol{x},\boldsymbol{y})$ denotes the truth degree of the non-discordance predicate.

The concordance index $c(\boldsymbol{x},\boldsymbol{y})$ is defined as follows:

$$c(\boldsymbol{x},\boldsymbol{y}) = \sum\nolimits_{G} k_j c_j(\boldsymbol{x},\boldsymbol{y}) \tag{5}$$

where:
k_j is the weight of the *j*-th criterion $(k_1 + k_2 + \cdots + k_N = 1)$
$c_j(\boldsymbol{x},\boldsymbol{y})$ is the marginal (partial) concordance index for the *j*-th criterion.

Let us denote by p_j and q_j the preference and indifference thresholds for criterion j $(p_j \geq q_j \geq 0)$. The partial concordance index is calculated by:

$$c_j(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} 0 & \text{if } \mathrm{g}_j(\boldsymbol{y}) - \mathrm{g}_j(\boldsymbol{x}) \geq p_j \\ \left(\mathrm{g}_j(\boldsymbol{x}) - \mathrm{g}_j(\boldsymbol{y}) + p_j\right)/\left(p_j - q_j\right) & \text{if } q_j < \mathrm{g}_j(\boldsymbol{y}) - \mathrm{g}_j(\boldsymbol{x}) < p_j \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

Let $D_{x,y} = \left\{j \in G \text{ such that } g_j(\boldsymbol{y}) - g_j(\boldsymbol{x}) \geq p_j\right\}$ be the coalition discordant with $\boldsymbol{xSy}$. The partial discordance is measured in comparison with a veto threshold v_j, which is the maximum difference $g_j(\boldsymbol{y}) - g_j(\boldsymbol{x})$ compatible with $\sigma(\boldsymbol{x},\boldsymbol{y}) > 0$. Following Mousseau and Dias [15], we shall use a simplification of the original

formulation of the discordance indices in the ELECTRE III method, which is given by:

$$Nd(\boldsymbol{x},\boldsymbol{y}) = \min_{j \in D_{x,y}} \left[1 - d_j(\boldsymbol{x},\boldsymbol{y})\right] \tag{7}$$

$$d_j(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} 1 & \text{iff } \nabla_j \geq v_j \\ \left(\nabla_j - \mathrm{u}_j\right)/\left(v_j - u_j\right) & \text{iff } u_j < \nabla_j < v_j \\ 0 & \text{iff } \nabla_j \leq u_j \end{cases} \tag{8}$$

where $\nabla_j = \mathrm{g}_j(\boldsymbol{y}) - \mathrm{g}_j(\boldsymbol{x})$ and u_j is a discordance threshold (see Figure 1).

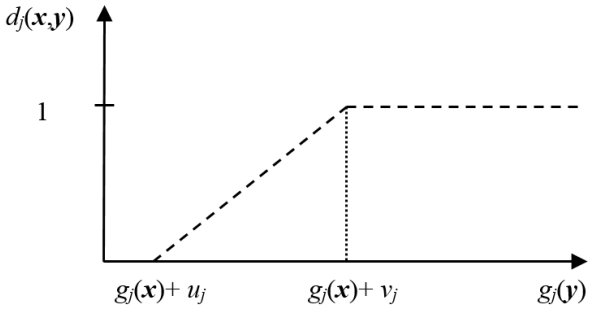


Fig. 1 Partial discordance relation $d_j(\boldsymbol{x},\boldsymbol{y})$

$\sigma(\boldsymbol{x},\boldsymbol{y})$ is calculated by combining Eq. 5-8 with Eq. 4. So, the following parameters must be specified:

1. The vector of weights;
2. The vector of indifference thresholds;
3. The vector of preference thresholds;
4. The vector of veto thresholds;
5. The vector of discordance thresholds (only when the simplification suggested by Mousseau and Dias [15] is used).

Additionally, if a crisp outranking relation is built on *A,* then a cutting level λ^* has to be specified.

4.2 Constraints in Problem (3)

As we stated above, the *DM* may reveal subjective information about criterion importance and parameter value ranges. The parameter settlement should be in correspondence with such beliefs. Otherwise, the *DM* would feel disappointed with a decision that contradicts his/her feelings.

The most obvious situation concerns "weights". Often, the *DM* has a clear idea about the importance of criteria, even though he/she is doubtful regarding weight values. The solution must agree with the order of importance in *DM*'s mind.

Fernandez et al. [9] state that the *DM* should be able to create the following binary relations on G:

$SI = \{(g_m, g_j) \in G \times G$ such that "criteria g_m and g_j have approximately equal importance"$\}$

$LI = \{(g_m, g_j) \in G \times G$ such that "criterion g_m is lightly more important than criterion g_j"$\}$

$MI = \{(g_m, g_j) \in G \times G$ such that "criterion g_m is clearly more important than criterion g_j"$\}$

MI, LI and SI should hold:

$$\begin{aligned} &\text{a. if } g_m MI g_j \text{ then } k_m - k_j \geq \beta \\ &\text{b. if } g_m LI g_j \text{ then } \beta > k_m - k_j \geq \beta/2 \\ &\text{c. if } g_m SI g_j \text{ then } |k_m - k_j| < \beta/2 \end{aligned} \tag{9}$$

where β is a threshold parameter for strict outranking that is proposed to be settled within the interval [1/(2*n*), 1/*n*] [9].

In order to obtain a preference-consistent model, the *DM* should impose other constraints:

$$\begin{aligned} &\text{For } j = 1, \dots, n \\ &0 \leq q_{j\,min} \leq q_j \leq q_{j\,max} \\ &p_{j\,min} \leq p_j \leq p_{j\,max} \ (q_{j\,max} < p_{j\,min}) \\ &u_{j\,min} \leq u_j \leq u_{j\,max} \ (p_{j\,max} < u_{j\,min}) \\ &v_{j\,min} \leq v_j \leq v_{j\,max} \ (u_{j\,max} < v_{j\,min}) \end{aligned} \tag{10}$$

If necessary, the *DM* may impose constraints specifying certain inter-criteria asymmetry conditions. For instance: $v_j \leq v_l \leq v_i$, $q_l \leq q_j$ or $p_l \leq p_j$.

If the *DM* felts that the discordance threshold u should be near to the middle point of the interval $[p, v]$, the constraints

$$|(v_j + p_j)/2 - u_i| \leq \varepsilon(v_j - p_j) \quad (0 < \varepsilon \ll 1) \ (i = 1, \dots, n) \tag{11}$$

could be added.

4.3 Description of the Evolutionary Approach for Inferring the Model Parameters

We shall use the Non-dominated Sorting Genetic Algorithm-II (*NSGA-II*) [4]. *NSGA-II* is one of the most efficient approaches in the literature on Evolutionary Multiobjective Optimization (cf. [2]). This method ranks every member of a *K'*-size population according to individual nondomination levels, applies evolutionary operators to build an offspring population, and combines parent and offspring populations in a new pool of *2K'* size. This combined population is sorted into nondominated classes. The next *K'*-size population is obtained by selecting the

best individuals of the parent-offspring combined pool. In order to keep diversity, a crowding distance (a density estimator) is associated with every individual.

For the selection of "parents", *NSGA-II* uses a special kind of binary tournament that is called "crowded tournament selection operator" [4]. It works as follows: Let *i, j* be two randomly selected solutions from the parent population. Solution *i* wins the tournament over *j* whenever one of the following conditions is true:

1. If solution *i* has a better rank than *j*.
2. If they have the same rank but solution *i* has a better crowding distance than *j* (that is, the crowding distance associated with *i* is greater than the associated to *j*).

Point 1 assures the winner lies on a better non-domination front. Point 2 solves possible ties between solutions, being on the same front, by deciding according to crowding distances. In this case, the winner resides in a less crowded region.

The pseudocode of *NSGA-II* is given below (cf. [2]):

```
Generate random population (size K')
Evaluate Objective Values
Generate non-dominated fronts
Assign to these fronts Rank Based on Pareto Dominance
Keep the best front (Rank) in the population memory
Generate Offspring Population
  Binary Tournament Selection
  Crossover and Mutation
For i = 1 to Number of Generations
  With Parent and Offspring Population
    Generate  non-dominated fronts
    Assign to these fronts Rank Based on Pareto Domin-
ance
    Loop (inside) by adding solutions to next genera-
tion
    starting from the best front until K' individuals
found
    calculate crowding distance between points on each
front
Update the population memory
Select points (elitist) on the better front (with bet-
ter Rank)
and which are outside a crowding distance
Form next generation
  Binary Tournament Selection
  Crossover and Mutation
  Increment generation index
End of Loop
```

Individuals are represented by a string composed of $5n + 1$ positions as shown in Figure 2.

p_1	u_1	v_1	…	p_n	u_n	v_n	k_1	k_2	…	k_n	q_1	q_2	…	q_n	λ

Fig. 2 Individual coding

We use one-point crossover. $2n + 1$ possible crossover points are defined on the individual (see Figure 3). Given two parents the specific crossover point is randomly generated.

			1		…				n				n+1	n+2	…		2n+1
p_1	u_1	v_1	…	…	p_n	u_n	v_n	k_1	k_2	…	k_n	q_1	q_2	…	q_n	λ	

Fig. 3 Possible crossover points

We use uniform mutation. The operator for mutation is implemented as follows:

Pick a random integer number $l \in [1; 4n + 2]$

1. If $l \in [1; n]$, then a random real number $a \in [q_{lmin}; q_{lmax}]$ is generated. Replace q_l with a.
2. If $l \in [n + 1; 2n]$, then a random real number $a \in [p_{lmin}; p_{lmax}]$ is generated. Replace p_l with a. To enforce restriction (11) a random real number $b \in B = \{x \in \Re \text{ such that } |(v_j + p_j)/2 - x| \leq \varepsilon(v_j - p_j)\}$ is generated. Replace u_l with b.
3. If $l \in [2n + 1; 3n]$, then a random real number $a \in B$ is generated; interval B is defined as in 2.
4. If $l \in [3n + 1; 4n]$, then a random real number $a \in [v_{lmin}; v_{lmax}]$ is generated. Replace v_l with a. To enforce restriction (11) a random real number $b \in$ ber $b \in \{x \in \Re, \text{such that } |(v_j + p_j)/2 - x| \leq \varepsilon(v_j - p_j)\}$ is generated; replace u_l with b.
5. If $l \in 4n + 1$, n real numbers $k_j \in (0; 1)$ are randomly generated. We use the approach of Butler et al. [1]. $n - 1$ uniform random numbers are generated in $(0; 1)$; further, these are ranked $0 < a_1 < a_2 < \cdots < a_{n-1} < 1$, and the weights are calculated as $k_i = a_i - a_{i-1}$. Thus, the normalization condition $(k_1 + k_2 + \cdots + k_n = 1)$ is satisfied and the random weights are uniformly distributed. This particular mutation is considered valid only if the constraints given by (9) hold. Otherwise, the random generation is repeated.
6. If $l \in 4n + 2$, then a random real number $a \in (0.5; 1]$ is generated. Replace λ with a.

Note that the above defined genetic operators keep feasibility with respect to constraints 9,10, 11. The initial population is generated from a feasible individual by reiterative mutations.

In a wide range, the algorithm performance was not very sensitive to the parameter settlement. Finally, the parameters of the evolutionary search were set to *Number of Generations*= 200, *Population size*= 100, *Crossover probability*= 0.8, *Mutation probability*= 0.05.

4.4 Final Formalization and Discussion

As we shall show through some examples, the above evolutionary methodology is able to find a good compromise solution $(n_p, n_Q, n_I, n_S)^*$ to Problem 3 in its objective space. This solution usually corresponds to many different points in the parameter space. Let us denote by $\{\boldsymbol{P}^*\}$ the set of points in the parameter space which are pre-image of $(n_p, n_Q, n_I, n_S)^*$. Each parameter settlement $\boldsymbol{P}_1^*$ is compatible with the preference information contained in the reference set under constraints 9-11. However, this does not mean that such compatible solution should be accepted by the *DM* as the correct parameter settlement. The *DM* probably has other beliefs and feelings, not contained in 9-11, which should be satisfied (this will be discussed in detail below). It is thus necessary to choose an element $\boldsymbol{P}_{best}{}^* \in \{\boldsymbol{P}^*\}$ as the final solution of the parameter inference problem. The *DM* may decide between two procedures: i) to select a particular $\boldsymbol{P}_1^* \in \{\boldsymbol{P}^*\}$ according to his/her own judgment, or ii) to use the information provided by their distribution in order to select a more acceptable setting. Let us discuss this issue thoroughly. There are two random factors that explain the deviation of $\boldsymbol{P}_1^*$ from more acceptable central points: 1) the reference set is a population sample. Different reference sets lead to different optimal parameter settlements, although the system of *DM*'s preferences is unique. 2) Even under the same reference set, the random nature of the evolutionary algorithm produces different solutions. Thus, $\boldsymbol{P}^*$ can be considered a random vector. There is a multivariate probability distribution function $\psi(\boldsymbol{P}_m{}^*, \boldsymbol{P}^*)$ that describes the random behavior of $\boldsymbol{P}^*$, being $\boldsymbol{P}_m{}^*$ its mean point. Let $\boldsymbol{P}_{near}{}^*$ be the nearest $\boldsymbol{P}^*$ to $\boldsymbol{P}_m{}^*$. $\boldsymbol{P}_{near}{}^*$ may be considered the most central point of $\psi(\boldsymbol{P}_m{}^*, \boldsymbol{P}^*)$, and from this view, its most representative point. $\boldsymbol{P}_{near}{}^*$ may coincide with $\boldsymbol{P}_m{}^*$ when this is capable of restoring the reference information.

As a final formalization, we propose the following steps:

First: The decision maker provides the preference information.

Second: The decision maker establishes constraints 9-11.

Third: Perform many runs of the evolutionary algorithm and obtain a good characterization of $\{\boldsymbol{P}^*\}$. Let us denote by $\{\boldsymbol{P}^*\}_{sample}$ this set and $\boldsymbol{P}_m$ its mean point ($\boldsymbol{P}_m$ is an estimator of $\boldsymbol{P}_m{}^*$).

Fourth: Provisional assignment of $\boldsymbol{P}_{best}{}^*$: Check if $\boldsymbol{P}_m{}^*$ belongs to $\{\boldsymbol{P}^*\}$ ($\boldsymbol{P}_m{}^*$ is able to restore the preference information from which $\{\boldsymbol{P}^*\}$ was derived). In the affirmative case, assign $\boldsymbol{P}_m{}^*$ to $\boldsymbol{P}_{best}{}^*$. Otherwise, find $\boldsymbol{P}_{near}{}^*$ in $\{\boldsymbol{P}^*\}_{sample}$ (the nearest $\boldsymbol{P}^*$ to $\boldsymbol{P}_m{}^*$), and assign it to $\boldsymbol{P}_{best}{}^*$.

Fifth: Check if the temporary $\boldsymbol{P}_{best}{}^*$ agrees with the whole system of preferences and beliefs from the *DM*. If possible, check its ability to restore some additional preference information not contained in the reference set from which $\{\boldsymbol{P}^*\}$ was derived. If these conditions are satisfactorily held, the current $\boldsymbol{P}_{best}{}^*$ can be accepted as the final solution of the inference problem.

Alternatively, the *DM* may 1) separate *T* in different subsets, thus obtaining several candidates to be $\boldsymbol{P}_{best}{}^*$, which will be judged according to the above conditions (fifth step); 2) select another element from $\{\boldsymbol{P}^*\}_{sample}$; 3) modify/add

some constraints thus including additional information and repeat from the third step; 4) use $\{\boldsymbol{P}^*\}_{sample}$ as starting basis for further refinement in an interactive *DM*-analyst decision support procedure as was proposed by Doumpos et al. [7].

5 An Illustrative Example

Let us consider the *R&D* project evaluation problem analyzed by Fernandez et al. [10]. In such example 81 projects were evaluated by a real decision-maker on four criteria, the results are shown in Table 1.

Table 1 Reference objects

Project	C1	C2	C3	C4	Global Impact	Project	C1	C2	C3	C4	Global Impact
1	7	4	7	7	Exceptional	42	2	7	6	3	Very High
2	6	6	6	6	Exceptional	43	3	6	2	3	High
3	4	4	4	4	Very High	44	6	3	1	7	Very High
4	2	4	4	4	High	45	1	1	2	3	Below Average
5	2	2	4	4	Above Average or High	46	3	2	1	2	Average
6	2	2	2	4	Above Average	47	5	5	6	3	Very High
7	2	2	2	2	Average	48	4	5	6	4	Very High
8	1	2	2	2	Low or Below Average	49	4	4	2	5	High
9	1	1	1	1	Very Low	50	3	2	2	2	Above Average
10	3	3	3	6	High	51	6	5	6	1	Very High
11	3	3	6	3	High	52	4	6	2	1	High
12	3	3	6	6	Very High	53	1	6	2	4	High
13	3	6	3	3	High	54	1	6	3	4	High
14	3	6	3	6	Very High	55	3	1	4	1	Average
15	3	6	6	3	Very High	56	5	2	3	6	High
16	3	6	6	6	Very High	57	6	3	7	2	Very High
17	6	3	3	3	High	58	3	4	2	6	High
18	6	6	3	6	Very High	59	6	6	7	1	Very High
19	6	3	6	3	Very High	60	4	2	4	3	High
20	6	3	6	6	Very High	61	2	1	4	5	Above Average or High
21	6	6	3	3	Very High	62	5	4	1	3	High
22	6	3	3	6	Very High	63	6	6	7	7	Exceptional
23	6	6	6	3	Very High	64	3	4	6	4	High
24	2	2	5	1	Average	65	4	3	2	7	Very High
25	5	1	2	2	Average	66	5	2	6	6	Very High
26	5	5	1	2	High	67	2	7	3	3	Very High

Table 2 (*continued*)

27	2	5	1	2	Average	68	6	5	1	6	Very High
28	1	5	1	3	Average	69	2	5	7	3	Very High
29	3	7	7	7	Exceptional	70	4	1	4	2	Average
30	7	7	3	7	Exceptional	71	4	7	3	1	Very High
31	7	7	7	3	Exceptional	72	1	6	5	6	Very High
32	5	5	3	1	High	73	4	3	6	2	High
33	7	2	5	3	Very High	74	6	1	6	4	Very High
34	1	1	4	4	Average	75	3	5	5	1	High
35	1	1	5	1	Average	76	2	4	3	5	High
36	1	3	6	1	High	77	5	3	2	2	High
37	1	1	1	6	Above Average	78	1	4	4	6	High
38	1	1	1	2	Very Low	79	6	5	2	3	High
39	1	1	1	7	High	80	2	2	5	2	Above Average or High
40	1	1	1	3	Low	81	1	7	5	6	Very High
41	1	1	1	4	Below Average						

The g criteria were assumed functions with domain [1, 7], although only integer values were considered. The *DM* stated that: i) full symmetry should exist in the criterion set; ii) there is no effect of intensity of preference when $\left|\mathrm{g}_j(\boldsymbol{x}) - \mathrm{g}_j(\boldsymbol{y})\right| < 2$; iii) there are remarkable effects of intensity of preferences when $\left|\mathrm{g}_j(\boldsymbol{x}) - \mathrm{g}_j(\boldsymbol{y})\right| \approx 3$; iv) the discordance threshold u_i should be not far from the middle point between strict preference and veto thresholds. In consequence, and according to 10-11, the *DM* imposed the following constraints:

$$\begin{gathered} \left|k_m - k_j\right| < 0.125 \; m = 1,2,3; j > m \\ \text{For } j = 1, \ldots, 4 \\ 0 < q_j \leq 0.3 \\ 0.5 \leq p_j \leq 0.9 \\ 1.5 \leq u_j \leq 2.4 \\ 2.5 \leq v_j \leq 5 \\ \left|(v_j + p_j)/2 - u_i\right| \leq 0.1(v_j - p_j) \end{gathered} \tag{12}$$

We performed five experiments (*A, B, C, D, E*): 41 objects were randomly generated and their pairs were chosen as the reference set. The evolutionary algorithm of Section 4 was run 10 times for solving Problem 3 subject to the constraints in (12). Coded in Turbo *C++* the run time was 65 seconds on a Pentium-4 3 *GHz* microprocessor, 2 *GB RAM,* and a 80 *GB* disk. In each execution an ideal solution (0,0,0,0) of Problem 3 was obtained. Moreover, there are many different points in the parameter space giving the same ideal solution. Considering this specific data set, it is easy to prove that if $\boldsymbol{p}^* = ({p_1}^*, {p_2}^*, {p_3}^*, {p_4}^*)$, $\boldsymbol{q}^* = ({q_1}^*, {q_2}^*, {q_3}^*, {q_4}^*)$ are components of an ideal solution $\boldsymbol{P}^*$ of Problem (3), any other solution $\boldsymbol{P}^{**}$

obtained from $\boldsymbol{P}^*$ by replacing $\boldsymbol{p}^{**} \leq \boldsymbol{p}^*$ instead of $\boldsymbol{p}^*$ and $\boldsymbol{q}^{**} \leq \boldsymbol{q}^*$ instead of $\boldsymbol{q}^*$ (but satisfying $0 < {q_j}^{**} \leq 0.3$ and $0.5 \leq {p_j}^{**} \leq 0.9$) is still an ideal solution of (3). This means that there are many ELECTRE-III models which are compatible with the decision policy embedded in the reference set. It should be noticed that an ideal solution to Problem (3) does not necessarily agree with *DM*'s view. Amongst the different parameter settlements, the *DM* should select the most compatible with his/her beliefs. In order to illustrate this point, let us consider one of the ideal solutions that was obtained from Instance A:
$\boldsymbol{q} = (0.05, 0.12, 0.029, 0.031)$; $\boldsymbol{p} = (0.525, 0.839, 0.898, 0.784)$;
$\boldsymbol{u} = (1.683, 2.489, 1.587, 1.717)$; $\boldsymbol{v} = (2.512, 4.403, 2.558, 2.53)$;
$\boldsymbol{k} = (0.254, 0.259, 0.216, 0.271)$.
Such solution should not be accepted by the *DM* because the above premises *i, ii* and *iii* are not held.

Applying the procedure of subsections 4.3 and 4.4 (with 10 replications) we obtained the results shown in Table 2. ${\boldsymbol{P}_{best}}^*$ were found as described by step fourth, subsection 4.4.

Table 3 ${\boldsymbol{P}_{best}}^*$

Instance	Optimal (n_p, n_Q, n_I, n_S)	K	Q	p	u	v	λ
A	0,0,0,0	0.258, 0.270, 0.219, 0.253	0.123, 0.117, 0.144, 0.124	0.684, 0.662, 0.721, 0.672	2.025, 2.011, 1.677, 2.019	3.372, 3.671, 2.804, 3.394	0.764
B	0,0,0,0	0.262, 0.220, 0.257, 0.261	0.123, 0.163, 0.134, 0.131	0.679, 0.640, 0.734, 0.719	1.693, 1.646, 2.268, 2.369	2.671, 2.765, 3.927, 4.389	0.768
C	0,0,0,0	0.257, 0.267, 0.257, 0.219	0.120, 0.186, 0.146, 0.135	0.729, 0.654, 0.672, 0.642	2.099, 1.958, 1.938, 1.611	3.649, 3.509, 3.453, 2.689	0.763
D	0,0,0,0	0.269, 0.220, 0.230, 0.280	0.201, 0.189, 0.155, 0.166	0.736, 0.715, 0.653, 0.706	2.256, 2.308, 2.426, 1.851	4.153, 4.230, 4.791, 3.166	0.759
E	0,0,0,0	0.218, 0.259, 0.266, 0.257	0.120, 0.166. 0.144, 0.164	0.676, 0.706, 0.671, 0.695	2.188, 2.236, 2.077, 2.058	3.903, 4.012, 3.598, 3.504	0.767
Most central point	0,0,0,0	0.252, 0.248, 0.246, 0.253	0.135, 0.163, 0.144, 0.143	0.699, 0.674, 0.692, 0.687	2.042, 2.019, 2.069, 1.994	3.519, 3.610. 3.685, 3.455	0.764
Deviation		0.022, 0.025, 0.023, 0.023	0.089, 0.091, 0.086, 0.089	0.113, 0.111, 0.111, 0.109	0.303, 0.380, 0.363, 0.369	0.754, 0.929, 0.899, 0.855	0.011

The central values shown in the seventh row solution seem to be the most compatible with the *DM's* premises listed in points *i-iii*. This central solution is compatible with the reference information, and strongly agrees with *i-iii*. In comparison with $\boldsymbol{P}_{best}{}^{*}$ from Instances A-E, the *DM* feels more comfortable accepting the identified central solution.

As a final test, the DM was questioned about 10 pairs $(\boldsymbol{x}, \boldsymbol{y})$ which had been not contained in the reference information. The elements $\boldsymbol{x}$ and $\boldsymbol{y}$ were chosen in such a way that no effects of intensity of preference prevailed between them, so $\boldsymbol{xSy}$ or $\boldsymbol{ySx}$ should be true. From the optimal solution identified in Table 2 the relation $S(\lambda)$ over the new pair set was then built. The result was $S(\lambda) \Leftrightarrow S$.

6 Concluding Remarks

This paper presented an evolutionary method for inferring appropriate ELECTRE-III parameters from preference statements. In a decision support framework, the final decision is conceived as the result of a cooperative effort between a DM and a decision analyst, performing an exhaustive exploration of the parameter space.

There are several necessary conditions to be an appropriate settlement of the ELECTRE-III model parameters:

i) To be a settlement that minimizes some error or inconsistency measure when predictions of the model are compared with the real decision policy; that is, to have capacity to restore the reference information.
ii) To satisfy the *DM* additional information about criterion importance, symmetry, asymmetry and acceptable parameter ranges; that is, the inferred values should be meaningful for the *DM;*
iii) To perform well when the inferred model is compared with preference statements which are not contained in the reference set. That is, to have capacity of explaining new decisions.

In this paper, point i) has been approached through evolutionary multiobjective optimization of several inconsistency measures. It yields the following advantages:

1. Compared with single-objective optimization, a multiobjective approach is more flexible because it allows to model preferences on different objectives. The different inconsistency measures do not have the same importance and should not be integrated into a single objective. For instance, inconsistencies with $P(\lambda)$ seem to be more relevant than other types.
2. Evolutionary multiobjective optimization techniques allow to perform a deep exploration of the set of satisfactory solutions; this is an important issue, because as it was shown by examples, there are many different solutions in the parameter space that satisfy the above Condition i), and the DM-analyst should have a wide representation of this set in order to select the most appropriate solution according to ii) and iii).
3. Evolutionary multiobjective optimization algorithms allows an easy handling of constraints (point ii).

The NSGA-II algorithm performed very well in some examples with only four criteria, identifying parameter settlements satisfying the necessary conditions. The whole method should be tested in examples with larger criterion set.

Acknowledgments. The authors acknowledge support from CONACyT project grant 57255.

References

1. Butler, J., Jia, J., Dyer, J.: Simulation techniques for the sensitivity analysis of multi-criteria decision models. European Journal of Operational Research 103, 531–546 (1997)
2. Coello, C.A., Lamont, G.B., Van Veldhuizen, D.A.: Evolutionary Algorithms for Solving Multi_Objective Problems, 2nd edn. Springer, New York (2007)
3. Pekelman, D., Sen, S.K.: Mathematical programming models for the determination of attribute weights. Management Science 20, 1217–1229 (1974)
4. Deb, K.: Multi-Objective Optimization using Evolutionary Algorithms. John Wiley & Sons, Chichester-New York-Weinheim-Brisbane-Singapore-Toronto (2001)
5. Dias, L., Mousseau, V., Figueira, J., Climaco, J.: An aggregation/disaggregation approach to obtain robust conclusions with ELECTRE-TRI. European Journal of Operational Research 138(2), 332–348 (2002)
6. Dias, L.C., Mousseau, V.: Inferring ELECTRE's veto-related parameters from outranking examples. European Journal of Operational Research 170(1), 172–191 (2006)
7. Doumpos, M., Marinakis, Y., Marimaki, M., Zopounidis, C.: An evolutionary approach to construction of outranking models for multicriteria classification: The case of ELECTRE TRI method. European Journal of Operational Research 199(2), 496–505 (2009)
8. Doumpos, M., Zopounidis, C.: Multicriteria decision aid classification methods. Kluwer Academic Publishers, Dordrecht (2002)
9. Fernandez, E., Navarro, J., Duarte, A.: Multicriteria sorting using a valued preference closeness relation. European Journal of Operational Research 185(2), 673–686 (2008)
10. Fernandez, E., Navarro, J., Bernal, S.: Multicriteria sorting using a valued indifference relation under a preference disaggregation paradigm. European Journal of Operational Research 198(2), 602–609 (2009)
11. Jacquet-Lagreze, E., Siskos, J.: Assesing a set of additive utility functions for multicriteria decision making: The UTA method. European Journal of Operational Research 10(2), 151–164 (1982)
12. Mangasarian, O.L.: Multisurface method for patter separation. IEEE Transactions on Information Theory 14(6), 801–807 (1968)
13. Marchant, T.: Debate on How to assign numerical values to different parameters that aim at differentiating the role that the criteria have to play in a comprehensive preference model? In: 71 Meeting of the Euro Working Group Multiple Criteria Decision Aiding, Torino, Italy (2010)
14. Mousseau, V., Slowinski, R.: Inferring an ELECTRE-TRI model from assignment examples. Journal of Global Optimization 12(2), 157–174 (1998)

15. Mousseau, V., Dias, L.C.: Valued outranking relations in ELECTRE providing manageable disaggregation procedures. European Journal of Operational Research 156(2), 467–482 (2004)
16. Mousseau, V., Figueira, J., Naux, J.P.: Using assignment examples to infer weights for ELECTRE TRI method: Some experimental results. European Journal of Operational Research 130(2), 263–275 (2001)
17. Perny, P.: Multicriteria filtering methods based on concordance and non-discordance principles. Annals of Operations Research 80, 137–165 (1998)
18. Roy, B., Slowinski, R.: Handling effects of reinforced preference and counter-veto in credibility of outranking. European Journal of Operational Research 188(1), 185–190 (2008)
19. Roy, B.: The Outranking Approach and the Foundations of ELECTRE methods. In: Bana e Costa, C.A. (ed.) Reading in Multiple Criteria Decision Aid, pp. 155–183. Springer, Berlin (1990)
20. Greco, S., Mousseau, V., Slowinski, R.: Ordinal regression revisited: Multiple criteria ranking with a set of additive value functions. European Journal of Operational Research 191, 415–435 (2008)

Customer Segmentation Based on Compensatory Fuzzy Logic within a Sustainability CRM for Intermodal Mobility

Benjamin Wagner vom Berg, Ariel Racet Valdés, Ammar Memari, Nasheda Barakat, and Jorge Marx Gómez

Abstract. Today Customer Relationship Management (CRM) is a major part of companies' strategies to increase consumption of customers with the goal of profit maximization. The integration of sustainability in CRM is in progress on different levels (e.g. Green Marketing), but sustainability is not integrated yet in a holistic approach in CRM strategies, processes and systems. A main contribution of such a Sustainable CRM can be to influence the behavior of customers to a more sustainable consumption. For identifying the right customers and applying effective marketing activities it is necessary to build customer segments. The data of customers in this context are various and often fuzzy. In this approach compensatory fuzzy logic is used for customer segmentation based on user preferences. The case study shows the appliance of this customer segmentation within a service for e-mobility for different means of transport (electric car, public transport, train, etc.) with the aim to lead costumers to a more sustainable mobility behavior.

1 Introduction

Nowadays Customer Relationship Management (CRM) is still a strategy, a method and a tool to build and maintain customer relations with the aim of profitability. The value of the customer relation is measured in long-term profits. Thus, CRM is a part of marketing [14].

Benjamin Wagner vom Berg · Ammar Memari · Nasheda Barakat · Jorge Marx Gómez
Carl von Ossietzky University of Oldenburg, Germany
email: benjamin.wagnervomberg@uni-oldenburg.de, memari@wi-ol.de, nasheda2006@yahoo.com, jorge.marx.gomez@wi-ol.de

Ariel Racet Valdés
"Jose Antonio Echeverria" Higher Technical Institute, Cuba
e-mail: aracet@ind.cujae.edu.cu

R.A. Espín Andrade et al. (eds.), *Soft Computing for Business Intelligence*, Studies in Computational Intelligence 537,
DOI: 10.1007/978-3-642-53737-0_27,

On the other side, companies, governments and the public are recognizing that resources are limited (e.g. oil) and other different problems are occurring at the economic, environmental and social dimension resulting from our economic and consumption behavior [1]. Due to this, sustainability is growing into a major topic and, according to the Brunlandt Commission [12], we have to integrate sustainability in our economic strategies and processes in order to maintain resources for future generations.

The integration of environmental aspects in the marketing area was already happening at end of the 80´s within Green Marketing, e.g. by eco-labeling, and is still in progress [10, 27]. Another newer important approach is sustainability marketing that is described as follows: "In other words, sustainability marketing represents an evolution of marketing that blends the mainstream economic and technical perspectives with the emerging concepts of relationship marketing and the social, ethical, environmental and intergenerational perspectives of the sustainable development agenda" [1]. So, there is a change only from economic and technical objectives towards sustainability objectives. An important brick within sustainability marketing is to shape consumer lifestyle and behavior in a way that would both expand the market for the offered products and to achieve sustainability goals [1].

These new approaches have also consequences for CRM as part of marketing, but these consequences and effects has not been revealed in all aspects. A holistic approach to use CRM as a strategy and an information system to support sustainability does not exist yet. One of the important goals of such a sustainable CRM has to be the modification of customer behavior [19] towards a more sustainable consumption according to the above mentioned approach of sustainability marketing. To influence the behavior in an effective way, it is crucial to identify customers who are open for sustainable offers and who are fitting to a certain offer. This leads to the need for special customer segmentation. There are already several approaches for customer segmentation in green marketing, e.g. by demographic or psychographic attributes [5]. In this paper, we present customer segmentation based on customer preferences.

Since customer information are not often sharp and the unambiguously correlation to one specific segment is not possible, there are several approaches to use Fuzzy methods for customer segmentation [17]. In this case, we will apply the Compensatory Fuzzy Logic (CFL) introduced by Espin et al. [8] that allows compensation and association of properties. These authors introduced logic with compensatory operators and decision-making-inspired properties. One important property in favor of the subtlety claim for good modeling through language is the non-associative property of the conjunction and disjunction.

Marketing, segmentation of markets and Customer Relationship Management, like all management activities, need this possibility of knowledge engineering modeling for shaping knowledge and the preferences of the costumers as well. The CFL properties allow reaching a model of customer preferences in accordance with actual customer behavior, as illustrated in the case study.

The here introduced case study is placed in the domain of mobility (people travelling from A to B). Through mobility it can be identified a huge potential for the support of sustainability by a changing behavior of customers [2]. The mobility behavior of people has already changed in many ways. On the one hand, the car doesn´t play such an important role for young people like it did 20 years ago [31]. On the other hand, we are getting more mobile and travelling more often and longer distances (e.g., by airplane). The problem is that we have different needs of transportation in our daily life for business trips, for vacation trips, etc., and we have different preferences according costs and comfort, in general and for a single trip. But we don´t have all information to find the best choice according to our preferences. And, especially, we don´t have enough information about the sustainability of the different options.

This case study presents a software project to develop an Agent-based Application that provides all these information to the customer via a smart phone app or/and a website. The customer gets information about different means of transport for an individual requested trip (destination, date, time) according to his preferences of comfort, costs, flexibility, sustainability, etc. Specific values that show the sustainability of each mean of transportation for the planned trip are included. The customer information generated by the usage of the application is transferred to a CRM-System. Then, the CRM-System analyzes these data and provides — based on customer segmentation according to customer preferences— sustainable offers for specific target groups to modify the mobility behavior towards a more sustainable moving.

2 CRM and Sustainability

Customer Relationship Management is one major part of customer based management of companies. Customer based management is shortly described with the words of Jack Welch, former CEO of the General Electric Company: “We want a company that focuses on nothing but serving the customers.” This customer based view is essential for the surviving of companies in the competitive global market, because the customer has a huge choice of different providers to fulfill his needs. So customer acquisition and customer loyalty are the main tasks nowadays for companies to be successful [14]. One most important goal is thereby to create economic success over influencing the quantity, quality and duration of customer relations [14]. A definition of CRM reads as follows: “CRM is a customer-oriented enterprise strategy which tries with help of modern information and communication technology, to develop and strengthen profitable customer relations in the long term by holistic and individual marketing, sales and service concepts.” [15] The term CRM names the strategy and the software (also eCRM or CRM-system). Within the customer-oriented strategy, CRM is a central component to reach the classical company targets enterprise maintenance and economic success [18]. The term of sustainability in combination with CRM is mostly used in the context of a sustainable customer relation in the sense of a long-term relation [16]. Long-term is usually interpreted as long-term profitable.

This interpretation of sustainability is only focused on the economical aspects, but the meaning of sustainability goes beyond the economic perspective.

The usually quoted definition for sustainability is the following one of the Brundlandt Commission from the year 1987: "Permanent development is development, which satisfies the needs of the present, without risking that future generations do not satisfy their own needs" [12]. This definition makes clear that sustainability means more than a long-term protection of profits and the enterprise maintenance. Sustainable management means to act resource-protective and to take further goals from the social and also the cultural range into focus. These goals can quite contradict profit-oriented goals [26]. The three pillars of sustainability according to the Lower German House of Parliament can be used to classify the dimensions of sustainability: The ecological dimension, the social dimension, and the financial (economical) dimension [4].

Green marketing is a very important approach for integrating sustainability into marketing. The “Green Marketing Manifesto” from Grant [10] provides a precise vision for Green Marketing. In the following, we intend to have a closer look at the three different kinds of activities proposed in the book to get a better understanding of this alternative marketing approach:

- *Green-setting new standards – communicate*: having commercial objectives only (where the product, brand or company is greener than alternatives, but the marketing is straightforward about establishing the difference).
- *Greener – sharing responsibility – collaborate*: having green objectives as well as commercial objectives (the marketing itself achieves green objectives, for instance changing the way people use the product).
- *Greenest – supporting innovation – cultural reshaping*: having cultural objectives as well (making new ways of life and new business models normal and acceptable).

These three alternatives are leading softly to a new understanding and a new aim of marketing. The classical aim of marketing is selling or purchasing in a market with the instruments of the 4P´s of the marketing-mix: Product, Price, Place, and Promotion. [22] At least Alternative b. and c. are showing activities that are going beyond just selling and purchasing products. They are also about influencing customer behavior and changing it to a more environmentally responsible behavior.

Green marketing is focusing on environmental aspects [1]. In contrast to that, sustainability marketing is focusing on ecological, environmental and economical aspects. The managerial approach of sustainability marketing contains six key elements according to Belz and Peattie: social-ecological problems, consumer behavior, sustainability marketing values and objectives, sustainability marketing strategies, sustainability marketing mix and sustainability marketing transformation. In this understanding of sustainability marketing the intersection of socio-ecological problems and consumer wants are building the context for all marketing activities [1].

The here described activities can be supported by CRM on strategic level and on system level because CRM is an essential part of marketing. The linkage between both Green marketing and Sustainability marketing and CRM is still missing attention and there are only a few works that address this linkage [21].

A holistic approach of CRM supporting sustainability on technical and strategic level is still missing and needs to be discovered.

3 Customer Segmentation

Customer segmentation is a fundamental instrument within customer-focused management [13]. Customer segmentation means segmentation of all potential and actual customers according to their market reactions in internal homogeneous and among each other heterogeneous subgroups (customer segments) plus the work in one or multiple customer segments [9]. Customer segmentation is crucial for any customer specific activities like marketing campaigns to address the right customers at the right time with the right offer. The more precise customer segmentation is, the bigger is the success of a campaign. There are different concepts of customer segmentation present in practice and theory. Segmentation concepts which are focusing on segments by characteristics of an existing costumer relation are customer lifetime, for example [29]. Another kind of segmentation method that is also used within green marketing is to focus on characteristics of the customer itself and to segment, e.g. by demographics like sex, marital status etc. [5] or by psychographics, e.g. to find correlations between ecological consumer behavior and altruism or liberalism [30].

In our case the goal is to change the customer consumption behavior. To provide a more fitting sustainable alternative to the customer this alternative has also to match with the customer preferences otherwise he would reject the offer. So we have to consider customer segmentation by customer behavior and customer preferences. Data for customer behavior and customer preferences are especially available for Online-customers [23]. There is a huge amount of data that can be used generated by the way a customer uses a website/online-application: by alternatives he is choosing and by data he is providing. Questions about his preferences can easily be integrated in the procedure of usage of the website/application and we don´t have to accomplish expensive surveys to know more about customer preferences.

Frequently, data contains qualitative and quantitative attributes. It is also necessary to deal with non-numeric attributes. Linguistic terms are sometimes more useful to describe classes, e.g. comfortable, medium comfortable, luxurious. These data are not sharp. Fuzzy logic, Fuzzy classification and compensatory fuzzy logic make it possible to deal with this fuzzy customer data. With fuzzy logic it is possible to treat each customer individually [23].

4 Fuzzy Logic and Modeling Human Preferences

One way to implement the "principle of gradualness" is the definition of logic where predicates are functions of the universe X in the interval [0,1] and the conjunction, disjunction, negation and involvement operation are defined in such a way that when restricted to the domain {0,1} we obtain Boolean Logic. The different ways of defining operations and their properties determine different multivalued logics that are part of the Fuzzy Logic Paradigm [6, 8].

The use of a set of different operators with properties that generalize the bivalent logic would seem to be the natural way to model human preferences from language. In fact, applications in the field of decision making – modeling and decision-maker preferences has been made basically from the operator concept, rather than multivalued logic [7]. However, this way to address decisions does not provide the best base to exploit capacity of Fuzzy Logic for knowledge transformation and decision maker preferences in logical formulas [8].

There are two main features that hinder the use of logic-based approaches in modeling human preferences:

- The associative property of conjunction and disjunction operators used
- No compensation among truth values of basic predicates when compound predicates veracity is calculated using the operators.

The associativity property of a large part of operators used for aggregation determines that objectives hierarchy trees, which represent different preferences, produce the same truth values of its compound predicates. Under the associativity property both trees in Fig. 1 represent the same preferences, not according with human preferences on the reality. It is obvious, for example, that preference x has greater relevance in the tree on the right than in the one on the left.

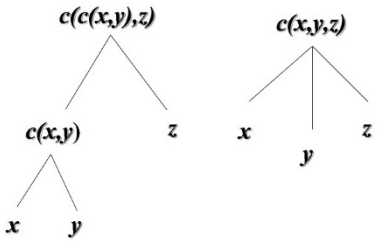


Fig. 1 Objectives hierarchy trees

The lack of compensation is an obstacle for a model that seeks to norm or to describe the reality of decision making or human preferences; the classic approaches of decision theory, the base of normative thinking, includes models such as the additive ones which accept compensation without limits. Descriptive

approaches accept partial compensation, which seems more akin to the reasoning of the actual agents. This way, the total lack of compensation and associativity are important limitations of operators frequently used for the addition of preferences.

The above suggests that it is desirable to use non-associative multivalued logic systems that facilitate the truth values compensation between basic predicates. Compensatory fuzzy Logic is a multivalued logic that meets these requirements [28].

5 Case Study: Sustainable Individual Mobility

Mobility in the meaning of mobility of persons who are moving from A to B has a strong connection to sustainability because, depending on the means of transport (e.g. train or car), the impact on resource consumption, environment and global climate is different. A definition for sustainable mobility is given by the Mobility Report 2001 prepared for the World Business Council for Sustainable Development: Sustainable mobility is "the ability to meet the needs of society to move freely, gain access, communicate, trade and establish relationships without sacrificing other essential human or ecological values today or in the future" [25].

The importance of this topic can be seen at the actual hype about electric vehicles (EV´s). The beginning shortage of oil and the impact of combustion engine cars on air pollution in cities and on global climate by CO2 emissions lead to the need of alternatives for conventional cars. Especially when we are looking at emerging markets like China or India, it is doubtable if our western usage patterns of cars are working in the long view. Also, it has to be considered if just using EV´s instead of conventional cars is solving these problems [3, 31].

Electric mobility offers the chance for introducing new business models which are able to break up old concepts of individual mobility both concerning the supply by the providers and the use by the customers. Here is a big opportunity to change mobility behavior under the criteria of sustainability. So far in practice not sufficiently accepted concepts (e.g., car sharing) can be integrated into a multimodal mobility service that is attractive to customers. These different services can be offered by one provider within one contract, e.g. similar to a contract for mobile communication [31] where the EV and the infrastructure is included. In the center of attention is no longer the supply of a vehicle, but the supply of services for mobility [20]. This creates possibilities for custom-made and sustainable offers to the customer for his individual mobility requirements. Through such new business models, the customer acceptance and the customer satisfaction are crucial for the market penetration. To enforce this acceptance and satisfaction also new information infrastructures and information services are needed [31].

6 Jinengo – Multi-modal Mobility

Jinengo is the name of a student group of Carl von Ossietzky University Oldenburg, Department of Business informatics. The name derives from the Chinese words “Jie neng” that means “saving energy”. Eleven Master students are working for one year on the creation of a software application for sustainable trip-planning that is named Jinengo as well. The application is connected to a standard CRM-System.

The planned application will offer the user a comfortable way to plan all his trips (daily trips, vacation trips, business trips, etc.) over his smart phone. The user will set his parameters for the trip (destination, date, time) and his preferences (costs, comfort, flexibility etc.). As a result, he gets back different possibilities with different means of transportation for the trip. The following means of transport are planned to be included: train (Deutsche Bahn), bus, car-pooling (e.g., car2gether), car-sharing (e.g., car2go) and the own car (electric or conventional car). For every single option the user gets information about costs, sustainability and further information. The sustainability could be expressed in CO2 emissions, CO2 equivalents and resource/energy consumption. The best way for the sustainability expression is still in consideration. Other types of information, e.g. the weather, will be integrated and considered in the suggested alternatives. The user is then able to choose the best alternative according to his preferences and according to sustainability. To provide the necessary information, the application will use the existing services for the means of transportation offered (e.g., from Deutsche Bahn) and CO2 calculators for car emissions (e.g. OPTIRESOURCE by Daimler AG). In a further step, a booking option can be also included. The adaptive application is based on a SOA-architecture with mobile software agents [24]. For providing the necessary information, the application will encapsulate and use already existing services for the means of transport offered. Keeping diversity will enable the application to reduce risk, e.g. of a biased evaluation of CO2 emission by averaging different sources.

Making the customer choice the more sustainable option is the essential aim. For Jinengo this means that customers should be influenced in the way that they choose a sustainable mean of transport or at least a more sustainable one that fits their requirements and preferences. First of all, we have to find out some more about sustainability of the different means of transport. CO2-emissions are only one important aspect according sustainability since some more aspects need to be considered like the resources for production of vehicles. To make a customer moving more sustainable, we have to identify an alternative means of transport (e.g. train) that is more sustainable to substitute the preferred mean of transport (e.g. car). There are some important variables that have to be considered in this context: actual preferred means of transport, availability of the alternative means of transport, travel distance, travel frequency, travel preferences (comfort, costs, flexibility). To apply different marketing activities, e.g. a mailing campaign by e-mail, we have to identify users with similar variables. Then, we can offer this group of users a specific offer according to their preferences. That is why; we need to segment our customers.

7 Customer Segmentation Based on User Preferences

Integrating CRM to store user data as mentioned above is important for customer segmentation as a step towards targeting the right segment with the right campaigns that will eventually alter behavior of the customer. We start by defining the segments we want to classify customers into by using natural language then transforming that into a Compensatory Fuzzy Logic predicate.

For example, we would like to identify customers who are good candidates to alter their behavior of heavily depending on their own car. Our target is to urge them to take the trains more often; so, we define a segment as: customers who have a good chance of changing their behavior from using their own car to using trains more often. We use the code name *C2T* to refer to this segment.

However, we still need to identify those customers based solely on their preferences that are discovered from their traveling patterns. We assume that a customer belongs to this segment if he should not be too keen on transportation qualities provided solely by his own car, like flexibility and comfort and, at the same time, has a high care for sustainability. The trips we would like the customer change are those which are frequent and repetitive. Moreover, the customer must have a high percentage of these trips. We name such a customer a *frequent traveller*.

All in all we can say: the customer belongs to this segment if he does not care so much about flexibility, he does not care so much about comfort, but he really cares about sustainability and he is a frequent traveller.

However, if the customer cares a lot about flexibility, but, at the same time, cares very much about sustainability, then compensation can take place and this customer also belongs to the segment.

The same can be said about comfort that can be compensated by a very high care for sustainability.

If the customer is a very frequent traveller, but he cares little about sustainability, then this customer is still a member of the aforementioned segment because of the high impact that a change in his behavior would have. So, a very high frequency of travelling can compensate for a little care about sustainability.

Definition of the predicate to be evaluated

C2T(X): The customer X is member of the segment C2T

Definition of the compound predicates

FT(X): The customer X is a frequent traveller
F(X): The customer X cares about flexibility
C(X): The customer X cares about comfort
S(X): The customer X cares about sustainability

Model of the compound predicate to be evaluated

$C2T(X) = [\neg F(X) \wedge \neg C(X) \wedge S^2(X) \wedge FT(X)] \wedge [F2(X) \rightarrow S^3(X)] \wedge [C^2(X) \rightarrow S^3(X)] \wedge [FT^2(X) \rightarrow S^{0.5}(X)]$

That is the required information for the segmentation, however, that is not the information we have about the customer. We must go more into details of the sub-predicates until we reach a level that can be populated directly with stored information.

Description of the compound predicates and their simple predicates

FT(X): The customer X is a frequent traveller

The more the customer chooses the same source and destination in his queries, the more a frequent traveller he is. The same measure can be applied at his choices. We call this measure similarity of source and destination. Another important factor for determining travel frequency is, of course, the number of travels the user makes. The more travels he makes, the more frequent traveller he is. If travels with similarity between source and destination are a little few, it should be compensated with very much travels.

SD(X): The customer X has travels with similarity between source and destination

NT(X): The customer X has many travels

$$FT(X) = [NT(X) \wedge SD(X)] \wedge [SD^{0.5}(X) \dashrightarrow NT^{2}(X)]$$

F(X): The customer X cares about flexibility

Flexible trips are ones that give the customer the freedom to change preferences without substantial effect on the overall travel time and costs. The user might want to change the starting time of the trip, change the destination slightly, and decide to change the route in order to pass through a certain station on the way. The more factors the plan is flexible to, the more flexible it is and at the same time, the wider the flexible range for a certain factor is, the more flexible the plan is.

For example, a day ticket is more flexible than a single-trip ticket since starting time can be any time within a day, source and destination are roughly defined within one city or state, and route is totally arbitrary as long as it is within the coverage area of the ticket.

If the customer is not that flexible in origin and destination locations, but is very flexible in time, he is still considered as a flexible traveller. A sample query might be: “I would like to travel from my house to my parents’ house sometime next week”.

The customer cares about flexibility if he chooses trips in which he can change the starting time and he can change the route without substantial effect on the overall travel time and costs. If the starting time is not flexible, that should be compensated by very flexible origin and destination locations.

TC(X): The customer X cares about travel time and costs

R(X): The customer X cares about flexible origin and destination locations

ST(X): The customer X cares about flexible starting time

$$F(X) = [ST(X) \wedge R(X) \wedge TC(X)] \wedge [\neg ST(X) \dashrightarrow R^{2}(X)]$$

C(X): The customer X cares about comfort

Comfort of the trip is based on an assumed order of transportation methods: most comfortable is the car (a passenger has more comfort than a driver) then train, airplane, bus, motorcycle, bike, and walking.

Comfort in the vehicle comes mainly from the comfort of seats, which is related as well to height of the ceiling (a bus is more comfortable than a minivan), available legroom and the ergonomic design.

Comfort is related to weather as well, and the last three transportation methods don't protect the customer from bad weather, so they are not as comfortable as the others.

Luggage capacity plays also a major role in defining comfort, the more luggages a transportation means can handle; the more comfortable it is for travellers with luggage. If a customer chooses high-luggage-capacity transportation means when he travels with luggage, then this customer prefers comfort. Changes of transportation vehicles along the journey is another issue related to comfort, the less the changes the more comfortable the trip is, especially when travelling with luggage.

CS(X): The customer X cares about good comfort of seats

HS(X): The height of the ceiling is good

LS(X): The legroom is good

ED(X): The ergonomic design is good

$$CS(X) = HS(X) \wedge LS(X) \wedge ED(X)$$

The customer cares about comfort if he chooses trips in which he has a good protection from bad weather, and the comfort of seats are acceptable and has capacity for luggage and a minimal number of changes between origin and destination, if the comfort of seats is somewhat bad it should be compensated with very low number of changes.

W(X): The customer X cares about good protection from bad weather conditions

L(X): The customer X cares for a good capacity for luggage

NC(X): The customer X cares for keeping the number of changes minimal

$$C(X) = [W(X) \wedge CS(X) \wedge L(X) \wedge NC(X)] \wedge [CS^{0.5}(X) \dashrightarrow NC^{2}(X)]$$

S(X): The customer X cares about sustainability

The customer can mark the checkbox "save the planet" in the query for his planned trip. If he uses this checkbox, he gets an ordered list where the more sustainable alternatives are shown on top of the list. If the customer is using this checkbox, we know that he takes care for sustainability. Also, it is possible that he is not picking the sustainable alternative from the list, although he marked the checkbox "save the planet".

Another case is when the customer does not use the checkbox, but picks (always) the more sustainable choice from the non-ordered list. Then we can assume that he also cares about sustainability. So not using the checkbox "save the planet" can be compensated by choosing the very good sustainable alternative repeatedly.

The Customer X cares about sustainability if he chooses trips where he cares about “save the planet” and he chooses the good sustainable alternative for the travel. If he doesn´t care about saving the planet that should be compensated by choosing a very good sustainable alternative from the non-ordered list.

SP(X): The customer X cares about “save the planet”

SA(X): The customer X choose a good sustainable alternative

$S(X) = [SP(X) \wedge SA(X)] \wedge [\neg SP(X) \dashrightarrow SA^2(X)]$

So far, it has been obtained the model for defining a segment of customers who have a good chance of changing their behavior from using their own car to using trains more often. The logic tree that gathers all the predicates is shown in Figure 2.

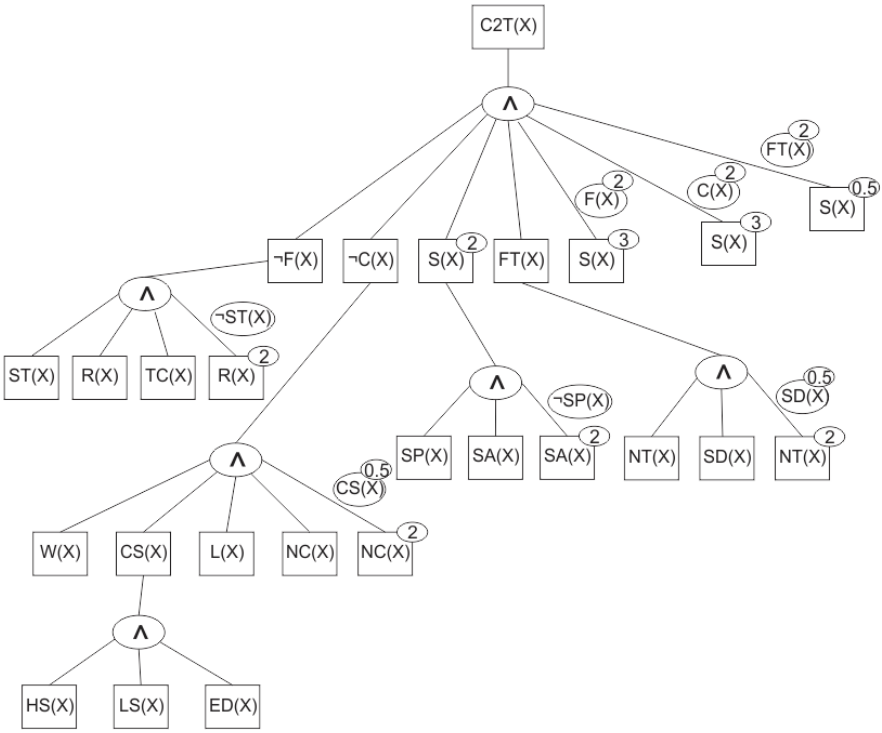


Fig. 2 Logic tree for C2T customer segmentation

With the application of this model, we will get the ordering of the customers from the one with the highest membership degree till the lowest. It is relevant information to take into account when it comes to make decisions related to marketing campaigns. By using this information, it can be decided which kind of marketing campaign could be use on each customer to make him move from using car to using train (more often).

8 Sustainable Offers Based on Customer Segmentation

Based on the customer segmentation we are now able to start marketing activities to take influence on the moving behavior of the customer. We even can use the preferences to develop individual campaigns. For instance, there is a segment of customers who rated “comfort” very high and should be convinced to use the train instead of a car. Then, the campaign can focus on comfort characteristics of trains like sleeping in the train, working in the train, no stress because of traffic jams, etc. For cost-sensitive customers we can give discounts for train tickets and so on.

We have also the choice between different ways of marketing activities on the side of the output. We can do classical campaigns like e-mail campaigns, but we also can post individual banners or snippets within the application. We can integrate a recommender system that directly makes suggestions to the user. A further interesting option is to modify the order of the result list that the user gets for a request of a trip. More sustainable choices could be highlighted or ranked higher. Services like Google or Amazon use different algorithms and techniques like this very successful in a profit-oriented way. Jinengo can easily adapt these techniques to use it in a sustainable-oriented way.

9 Conclusion and Outlook

In this article, it has been presented some new ideas on CRM and the use of fuzzy logic that can be summarized as followed:

- A Sustainable CRM might be an important contribution to support a more sustainable acting within our economy.
- There are existing techniques used in CRM which can be modified according to the aim to focus on sustainability.
- Influencing the choice of the customer can be one important way to change economic acting towards more sustainability by CRM.
- A customer segmentation based on customer preferences helps to apply the fitting marketing activities to the right customers to modify their behavior.
- Properties of compensatory fuzzy logic allow a very useful representation of customer preferences using the natural language and their translation to the language of predicate calculus.
- A model of the customer preferences with compensatory fuzzy logic was presented, and will be used for projecting better marketing strategies to our customers

These ideas are still not evaluated, but the presented case study in the area of multi-modal mobility provides a useful framework for the evaluation of all of the presented ideas. For this purpose, the application will be brought to market within one of the described business models to get real customer data. Then the ideas about a Sustainable CRM on a strategic and on a system level and the use of customer segmentation based on compensatory fuzzy logic can be evaluated.

References

1. Belz, F.-M., Peattie, K.: Sustainability Marketing - a global perspective. Wiley & Sons, Chichester (2009)
2. Black, W.R.: Sustainable Transportation: Problems and Solutions. Guilford Publications Inc., New York (2010)
3. Brake, M.: Mobilität im regenerativen Zeitalter - Was bewegt uns nach dem Öl? Heise Verlag, Hannover (2009)
4. Bundestag, D.: Konzept Nachhaltigkeit - Vom Leitbild zur Umsetzung, Abschlussbericht der Enquête-Kommission. In: Schutz des Menschen und der Umwelt“ of 13th Deutschen Bundestages, 1st edn., Universitäts-Buchdruckerei, Bonn (1998)
5. Diamantopoulos, A., et al.: Can socio-demographics still play a role in profiling green consumers? - A review of the evidence and an empirical investigation. Journal of Business Research 56(2003), 465–480 (2003)
6. Dubois, D., Prade, H.: Fuzzy Sets and Systems: Theory and Applications. Academic Press Inc. (1980)
7. Dubois, D., Prade, H.: A review of fuzzy set aggregation connectives. Information Sciences 36, 85–121 (1985)
8. Espin, R., Fernández, E., Mazcorro, G., Marx, J., Lecich, M.I.: Compensatory Logic: A fuzzy normative model for decision making. Investigación Operativa 27(2), 188–197 (2006)
9. Freter, H.: Marktsegmentierung. Kohlhammer Verlag, Stuttgart (1983)
10. Grant, J.: The Green Marketing Manifesto. John Wiley & Sons Ltd., England (2007)
11. Greenpeace. Make it Green: Cloud Computing and its Contribution to Climate Change. Greenpeace International, Amsterdam (2010)
12. Hauff, V. (ed.): Unsere gemeinsame Zukunft: Der Brundtland-Bericht der Weltkommission für Umwelt und Entwicklung. Eggenkamp Verlag, Greven (1987)
13. Hinterhuber, H.H., Matzler, K. (eds.): Kundenorientierte Unternehmensführung: Kundenorientierung – Kundenzufriedenheit – Kundenbindung, 6th edn. Gabler Verlag, Wiesbaden (2009)
14. Hippner, H., Wilde, K.D. (eds.): Grundlagen des CRM: Konzepte und Gestaltung, 2nd edn. Gabler Verlag, Wiesbaden (2006)
15. Hippner, H., Wilde, K.D.: CRM – ein Überblick. In: Helmke, S., Dangelmaier, W. (eds.) Effektives Customer Relationship Management, 2nd edn., Gabler Verlag, Wiesbaden (2006)
16. Homburg, C. (ed.): Kundenzufriedenheit: Konzepte – Methoden – Erfahrungen, 7th edn. Gabler Verlag, Wiesbaden (2008)
17. Hruschka, H.: Market Definition and Segmentation Using Fuzzy Clustering Methods. International Journal of Research in Marketing 3(2), 117–135 (1986)
18. Kaiser, M.-O.: Erfolgsfaktor Kundenzufriedenheit: Dimensionen und Messmöglichkeiten, 2nd edn. Erich Schmidt Verlag, Berlin (2005)
19. Kantsperger, R.: Modifikation von Kundenverhalten als Kernaufgabe von CRM. In: Hippner, H., Wilde, K.D. (eds.) Grundlagen des CRM: Konzepte und Gestaltung, pp. 291–304. Gabler Verlag, Wiesbaden (2006)
20. Landmann, R., et al.: Winning the powertrain race: The frontline role of marketing and sales. Roland Berger Strategy consultants (2009)

21. Landua, I.: Gaining Competitive Advantage through Customer Satisfaction, trust and Confidence in Consideration of the influence of Green Marketing. Unpublished master thesis. University of Gävle – Department of Business Administration, Gävle (2008)
22. McCarthy, E.J., Perrault, W.D.: Basic marketing: a managerial approach. Irwin, Homewood (1975)
23. Meier, A., Werro, N.: A Fuzzy Classification Model for Online Customers. Informatica 31, 175–182 (2007)
24. Memari, A., Heyen, C., Marx-Gómez: A Component-based Framework for Adaptive Applications. In: Modeling of Business Information Systems (MoBIS), Dresden, Germany (2010)
25. MIT & CRA (Massachusetts Institute of Technology and Charles River Associates Incorpo- rated) Mobility 2001: World Mobility at the End of the Twentieth Century and Its Sustainability. Prepared for the World Business Council for Sustainable Development. MIT Press, Cambridge (2001)
26. Müller-Christ, G., Hülsmann, M. (eds.): Nachhaltigkeit und Widersprüche: Eine Managementperspektive. LIT Verlag, Hamburg (2007)
27. Peattie, K.: Environmental Marketing management: Meeting the green challenge. Pitman Publishing, London (1995)
28. Racet, A., Espin, R., Marx Gómez, J.: Compensatory Fuzzy Ontology. In: Davcev, D., Gómez, J.M. (eds.) ICT Innovations 2009, pp. 35–44. Springer, Berlin (2009)
29. Stauss, B.: Perspektivenwandel, Vom Produkt-Lebenszyklus zum Kundenbeziehungs-Lebenszyklus. Thexis 17(2), 15–18 (2000)
30. Straughan, R.D., Roberts, J.A.: Environmental segmentation alternatives: a look at green consumer behavior in the new millennium. Journal of Consumer Marketing 16(2), 558–575 (1999)
31. Weidlich, A.: Geschäftsmodelle Elektromobilität. Paper presented at the Fünfzehntes Kasseler Symposium Energie-Systemtechnik 2010. Fraunhofer IWES, Kassel (2010)

Author Index