FSCS 2006
Symposium on Fuzzy Systems in Computer Science 2006

Otto-von-Guericke-Universität Magdeburg, Germany

September 27th and 28th, 2006

Edited by

Eyke Hüllermeier
Rudolf Kruse
Andreas Nürnberger
Jens Strackeljan

FSCS 2006
Symposium on Fuzzy Systems in Computer Science 2006

Otto-von-Guericke-Universität Magdeburg, Germany

September 27th and 28th, 2006

Edited by

Eyke Hüllermeier
Rudolf Kruse
Andreas Nürnberger
Jens Strackeljan

FSCS 2006 – Symposium on Fuzzy Systems in Computer Science 2006
Otto-von-Guericke-Universität Magdeburg, Germany, 27-28 September 2006
Editors: Eyke Hüllermeier, Rudolf Kruse, Andreas Nürnberger, Jens Strackeljan

Printed by
Otto-von-Guericke-University Magdeburg
Faculty of Computer Science
Universitätsplatz 2
39106 Magdeburg
Germany

Preface

On the occasion of the 10th anniversary of the Neuro-Fuzzy research group in the Computer Science department in Magdeburg, the symposium "Fuzzy Systems in Computer Science 2006" was jointly organized by the the North German Softcomputing Association (AFN), the European Society for Fuzzy Logic and Technology (EUSFLAT), and several fuzzy-oriented research groups within the Otto-von-Guericke University of Magdeburg. The aim of this event was to provide an international forum for reporting recent advances in the research area of fuzzy systems. Special emphasis was put on the applicability of the methods to real world problems.

The scientific program includes a track on Fuzzy Methods in Learning and Data Mining, a special session about the applications of Fuzzy Methods in Intelligent Data Analysis, and the Annual Meeting of the AFN. The keynote address was given by the well-known UC Berkeley Professor Lotfi Zadeh, the inventor of fuzzy logic. The program was complemented by an invited lecture of Rudolf Seising about the history of fuzzy systems.

We like to thank all authors for their inspiring contributions and all those persons who assisted in the organization of this event.

Eyke Hüllermeier, Rudolf Kruse, Andreas Nürnberger, Jens Strackeljan

Magdeburg, September 2006

Workshop Organisation

Organisers:

Eyke Hüllermeier
Faculty of Computer Science
Data and Knowledge Engineering Group
e-mail: eyke.huellermeier@iti.cs.uni-magdeburg.de
Tel.: + 49 391 67 18842

Rudolf Kruse
Faculty of Computer Science
Neuro-Fuzzy Group
e-mail: kruse@iws.cs.uni-magdeburg.de
Tel.: + 49 391 67 18706

Andreas Nürnberger
Faculty of Computer Science
Information Retrieval Group
e-mail: nuernb@iws.cs.uni-magdeburg.de
Tel.: + 49 391 67 18718

Jens Strackeljan
Faculty of Mechanical Engineering
Research Group Technische Dynamik
e-mail: jens.strackeljan@mb.uni-magdeburg.de
Tel.: + 49 391 67 11379

Local Organization Team

Melanie Baumgarten
Kerstin Giesswein
Faculty of Computer Science
Sabine Laube
Frank Rügheimer
Faculty of Computer Science, Neuro-Fuzzy Group

Contents

Invited Talks

Fuzzy Data Mining

Classification

Intelligent Data Analysis

Visual Data Analysis

Applications of Computational Intelligence

FSCS 2006
Symposium on Fuzzy Systems in Computer Science 2006

Proceedings

A New Frontier in Computation – Computation with Information Described in Natural Language

Lotfi A. Zadeh*

Department of EECS, University of California
Berkeley, CA 94720-1776, USA
E-Mail: zadeh@eecs.berkeley.edu
Telephone: 510-642-4959; Fax: 510-642-1712

Extended Abstract

What is meant by Computation with Information Described in Natural Language, or NL-Computation, for short? Does NL-Computation constitute a new frontier in computation? Do existing bivalent-logic-based approaches to natural language processing provide a basis for NL-Computation? What are the basic concepts and ideas which underlie NL-Computation? These are some of the issues which are addressed in the following.

What is computation with information described in natural language? Here are simple examples. I am planning to drive from Berkeley to Santa Barbara, with stopover for lunch in Monterey. It is about 10 am. It will probably take me about two hours to get to Monterey and about an hour to have lunch. From Monterey, it will probably take me about five hours to get to Santa Barbara. What is the probability that I will arrive in Santa Barbara before about six pm? Another simple example: A box contains about twenty balls of various sizes. Most are large. What is the number of small balls? What is the probability that a ball drawn at random is neither small nor large? Another example: A function, *f*, from reals to reals is described as: If *X* is small then *Y* is small; if *X* is medium then *Y* is large; if *X* is large then *Y* is small. What is the maximum of *f*? Another example: Usually the temperature is not very low, and usually the temperature is not very high. What is the average temperature? Another example: Usually most United Airlines flights from San Francisco leave on time. What is the probability that my flight will be delayed?

Computation with information described in natural language is closely related to Computing with Words. NL-Computation is of intrinsic importance because much of human knowledge is described in natural language. This is particularly true in such fields as economics, data mining, systems engineering, risk assessment and emergency management. It is safe to predict that as we move further into the age of machine intelligence and mechanized decision-making, NL-Computation will grow in visibility and importance.

Computation with information described in natural language cannot be dealt with through the use of machinery of natural language processing. The problem is semantic

* Research supported in part by ONR N00014-02-1-0294, BT Grant CT1080028046, Omron Grant, Tekes Grant, Chevron Texaco Grant and the BISC Program of UC Berkeley.

imprecision of natural languages. More specifically, a natural language is basically a system for describing perceptions. Perceptions are intrinsically imprecise, reflecting the bounded ability of sensory organs, and ultimately the brain, to resolve detail and store information. Semantic imprecision of natural languages is a concomitant of imprecision of perceptions.

Our approach to NL-Computation centers on what is referred to as generalized-constraint-based computation, or GC-Computation for short. A fundamental thesis which underlies NL-Computation is that information may be interpreted as a generalized constraint. A generalized constraint is expressed as X isr R, where X is the constrained variable, R is a constraining relation and r is an indexical variable which defines the way in which R constrains X. The principal constraints are possibilistic, veristic, probabilistic, usuality, random set, fuzzy graph and group. Generalized constraints may be combined, qualified, propagated, and counter propagated, generating what is called the Generalized Constraint Language, GCL. The key underlying idea is that information conveyed by a proposition may be represented as a generalized constraint, that is, as an element of GCL.

In our approach, NL-Computation involves three modules: (a) Precisiation module; (b) Protoform module; and (c) Computation module. The meaning of an element of a natural language, NL, is precisiated through translation into GCL and is expressed as a generalized constraint. An object of precisiation, p, is referred to as precisiend, and the result of precisiation, p^*, is called a precisiand. Usually, a precisiend is a proposition, a system of propositions or a concept. A precisiend may have many precisiands. Definition is a form of precisiation. A precisiand may be viewed as a model of meaning. The degree to which the intension (attribute-based meaning) of p^* approximates to that of p is referred to as cointension. A precisiand, p^*, is cointensive if its cointension with p is high, that is, if p^* is a good model of meaning of p.

The Protoform module serves as an interface between Precisiation and Computation modules. Basically, its function is that of abstraction and summarization.

The Computation module serves to deduce an answer to a query, q. The first step is precisiation of q, with precisiated query, q^*, expressed as a function of n variables $u_1, \ldots, u_n$. The second step involves precisiation of query-relevant information, leading to a precisiand which is expressed as a generalized constraint on $u_1, \ldots, u_n$. The third step involves an application of the extension principle, which has the effect of propagating the generalized constraint on $u_1, \ldots, u_n$ to a generalized constraint on the precisiated query, q^*. Finally, the constrained q^* is interpreted as the answer to the query and is retranslated into natural language.

The generalized-constraint-based computational approach to NL-Computation opens the door to a wide-ranging enlargement of the role of natural languages in scientific theories. Particularly important application areas are decision-making with information described in natural language, economics, systems engineering, risk assessment, qualitative systems analysis, search, question-answering and theories of evidence.

Fuzzy Sets and Systems –
Their History and Future in Science

Rudolf Seising

Medical University of Vienna
Core Unit for Medical Statistics and Informatics
Spitalgasse 2, A-1090 Vienna, Austria

e-mail: Rudolf.seising@meduniwien.ac.at

Abstract. In the history of science, new theories have always been necessary in order for existing scientific theories to progress and this will continue to be true in the future. Two examples of essentially different mathematical theories that deal with the concept of uncertainty are probability theory und the theory of fuzzy sets. Whereas probability theory has a history of around 350 years, the theory of fuzzy sets is little more than 41 years old. This paper is, first of all, a review of innovations during these more than four decades in the work of Professor Lotfi Zadeh, whose new scientific thoughts and concepts included contributions from circuit theory, filter theory, and system theory prior to his formulation of the theory of fuzzy sets and systems. Since the 1960s fuzzy methods have entered the scientific and technological world, good theoretical progress (e.g., fuzzy logic, fuzzy probability theory, fuzzy topology, fuzzy algebra) has been made, and there have been technical advances in various areas (e.g., fuzzy control, fuzzy expert systems, fuzzy clustering and data mining). In short, fuzzy technology has now become "normal." In the second part of this contribution, the future prospects of fuzzy set theory in science and its philosophy are discussed. Considering the situation in basic and very successful scientific theory – i. e., quantum mechanics – we see 1) that uncertainty is an essential and integral concept in this fundamental theory of science and 2) that this concept is not satisfying when modeled on classical probability theory. Thus in the last half of the 20th century new approaches were proposed to introduce "quantum logics" and "quantum probabilities," and with the progress of computer sciences, information theory, and artificial intelligence, new theories of uncertainty appeared – one of which is the theory of fuzzy sets. It would be instrumental for the understanding of scientific handling in our times to establish a kind of order among uncertainty theories. To this end, the "structuralist program" in the philosophy of science could be helpful. This has been developed in recent decades as a representation scheme for scientific knowledge. It makes it possible to embed theories into a network of theories and to show their interrelationships in a clear way. In the last section of this contribution some programmatic ideas are set out to guide the construction of a general network of "uncertainty theories" that represents probability theory and fuzzy set theory, and perhaps some additional "uncertainty theories," and to explain the interrelationships of these theories. This approach is presented as a step toward the Generalized Theory of Uncertainty (GTU) that Lotfi Zadeh has targeted and outlined in recent publications.

Introduction

Now that fuzzy sets and systems have a history of more than 40 years, it is time to document their origins and development. This has been the author's project during recent years as a historian of science and technology. Based on his original research work (studies of scientific articles, newspapers, letters, etc., and interviews with Lotfi Zadeh and many other pioneers), the facts of the history of fuzzy sets and fuzzy systems have been assembled[1].

The genesis of fuzzy sets is not a story from set theory or symbolic logic or the philosophy of mathematics, but rather the result of fundamental research work carried out by the mathematically oriented electrical engineer and system theorist Lotfi A. Zadeh. In the 1960s system theory was a new field of research in the field of electrical engineering and Zadeh was a protagonist of this interdisciplinary field and one of the authors of the book *Linear System Theory – The State Space Approach* [5], where he introduced the concept of "state" in system theory. In the next sections the reader will find a reconstruction of the theory of fuzzy sets and fuzzy systems as an integral part of the development of system theory in the 1960s.

Part I: From Circuit Theory to Filter Theory to System Theory to Fuzzy Set Theory

The heading of this section alludes to Zadeh's trend-setting article *From Circuit Theory to System Theory* in the *Proceedings of the IRE* in May 1962 [6]. There, he described the subordination process of the classical theory of circuits as a special sector in the much wider scientific discipline of system theory. "Thus, whether a system is electrical, mechanical or chemical in nature does not matter to a system theorist. What matters are the mathematical relations between the variables in terms of which the behavior of the system is described" ([6], p. 856).

It this article Zadeh used the word "fuzzy" for the very first time to characterize his vision of new mathematics: "In fact, there is a fairly wide gap between what might be regarded as 'animate' system theorists and 'inanimate' system theorists at the present time, and it is not at all certain that this gap will be narrowed, much less closed, in the near future. There are some who feel that this gap reflects the fundamental inadequacy of the conventional mathematics – the mathematics of precisely-defined points, functions, sets, probability measures, etc. – for coping with the analysis of biological systems, and that to deal effectively with such systems, which are generally orders of magnitude more complex than man-made systems, we need a radically different kind of mathematics, the mathematics of fuzzy or cloudy quantities which are not describable in terms of probability distributions." ([6], p. 857).

---

[1] The first part of this contribution is a very short abridgement of my book on the origins of fuzzy set theory and its initial applications that appeared in German in 2005 [1] – the English version will be published in the first half of 2007 [2]. The second part concerns connections of the theory of fuzzy sets and philosophy of science. For this project in progress see [3, 4]

Fig. 1. Left to right: Claude E. Shannon, Norbert Wiener, John R. Ragazzini, and Lotfi A. Zadeh (all photographs were taken in the 1950s or 1960s).

When Zadeh published these ideas, he did not know what the mathematics of fuzzy quantities would look like. He was then a professor of Electrical Engineering at the University of California at Berkeley and could look back on a very successful career in electrical engineering and system theory – the creation of the theory of fuzzy sets lay in his immediate future. In the next subsections we will delineate the scientific path Zadeh took that brought him to this point.

From Signals to Filters

Lotfi Zadeh was born in 1921 as the son of an Azerbaijani father and a Russian mother in Baku, Soviet Azerbaijan. He spent his youth and school years in Teheran, Iran, where he attended an American Presbyterian school, learning English in the process. Later, he studied at the University of Tehran, from which he received the B.S. degree in Electrical Engineering in 1942. He came to the USA. in 1944 and after a while continued his studies at the Massachusetts Institute of Technology (MIT), where he took courses in circuit theory and network theory. He was involved with the theory and practice of relays, antennas, and electrical filters and he already mentioned that real systems and idealized systems are different. In 1946 he received the S.M. degree in Electrical Engineering from MIT.

Zadeh then moved to New York, where he joined the faculty of Columbia University as an instructor. In 1949 he wrote his Ph. D. thesis on *Frequency Analysis of Variable Networks*, under the supervision of Professor John Ralph Ragazzini, and in 1950 he was appointed to be an assistant professor. At that time Zadeh was intrigued by Norbert Wiener's cybernetics and Claude E. Shannon's information theory, and he was interested in the theory of ideal and optimal filtering. Along with Ragazzini, he published *An Extension of Wiener's Theory of Prediction* [7] in 1950, an article that was a milestone in the development of network synthesis. The mathematical techniques of this theory of prediction and filtering have been commonly employed in mathematical physics, particularly in quantum mechanics, e.g., multidimensional Euclidean spaces and Hilbert space representation. In his many papers on prediction and filtering, linear and nonlinear systems, time-varying networks, etc. at the beginning of the 1950s, Zadeh used the mathematical calculus of functionals and operators. In February 1952, he presented *Some Basic Problems in Communication of Information* [8] at the meeting of the Section of Mathematics and Engineering of the New

York Academy of Sciences, showing that it is useful to apply function space techniques in communication theory. The following discussion outlines these problems. The first one deals with the *recovery process* of transmitted signals:

"Let $X=\{x(t)\}$ be a set of signals. An arbitrarily selected member of this set, say $x(t)$, is transmitted through a noisy channel Γ and is received as $y(t)$. As a result of the noise and distortion introduced by Γ, the received signal $y(t)$ is, in general, quite different from $x(t)$. Nevertheless, under certain conditions it is possible to recover $x(t)$ – or rather a time-delayed replica of it – from the received signal $y(t)$." In this paper, Zadeh did not examine the case where $\{x(t)\}$ is an ensemble; restricting his view to the problem of recovering $x(t)$ from $y(t)$ "irrespective of the statistical character of $\{x(t)\}$" ([8], p. 201). Corresponding to the relation $y = \Gamma x$ between $x(t)$ and $y(t)$, he represented the recovery process of $y(t)$ from $x(t)$ by $x = \Gamma^{-1}y$, where $\Gamma^{-1}y$ is the inverse of Γ, if it exists, over $\{y(t)\}$.

Zadeh represented signals as ordered pairs of points in a signal space Σ, which is embedded in a function space with a delta-function basis, and to measure the disparity between $x(t)$ and $y(t)$ he attached a distance function $d(x, y)$ with the usual properties of a metric. Then he considered the special case in which it is possible to achieve a perfect recovery of the transmitted signal $x(t)$ from the received signal $y(t)$. He supposed that "$X = \{x(t)\}$ consist of a finite number of discrete signals $x_1(t)$, $x_2(t)$, …, $x_n(t)$, which play the roles of symbols or sequences of symbols. The replicas of all these signals are assumed to be available at the receiving end of the system. Suppose that a transmitted signal x_k is received as y. To recover the transmitted signal from y, the receiver evaluates the 'distance' between y and all possible transmitted signals x_1, x_2, …, x_n, by the use of a suitable distance function $d(x, y)$, and then selects that signal which is 'nearest' to y in terms of this distance function (fig. 2). In other words, the transmitted signal is taken to be the one that results in the smallest value of $d(x, y)$. This in brief, is the basis of the reception process." ([8], p. 201) In this process the received signal x_k is always "nearer" – in terms of the distance functional $d(x, y)$ – to the transmitted signal $y(t)$ than to any other possible signal x_i, i.e.,

$$d(x_k, y) < d(x_i, y), \qquad i \neq k, \qquad \text{for all } k \text{ and } i. \tag{1}$$

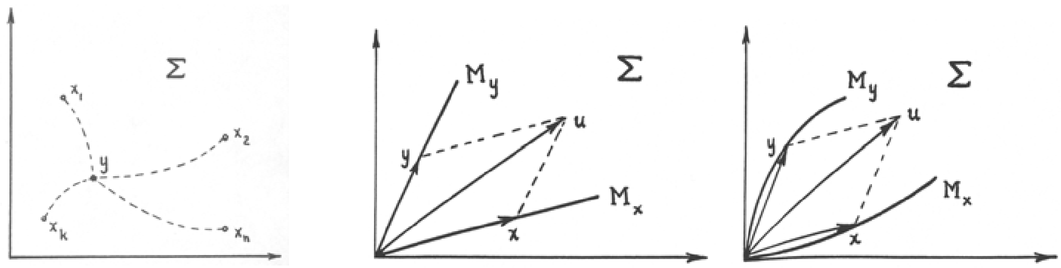


Fig. 2. Recovery of the input signal by means of a comparison of the distances between the received signal *y* and all possible transmitted signals (left). Geometrical representation of filtering (nonlinear: middle, linear: right). ([8], p. 202 f)

Zadeh conceded that "in many practical situations it is inconvenient, or even impossible, to define a quantitative measure, such as a distance function, of the disparity between two signals. In such cases we may use instead the concept of neighborhood, which is basic to the theory of topological spaces" ([8], p. 202). – About 15 years later he proposed another "concept of neighborhood," which is now basic to the theory of fuzzy systems!

Another "basic problem in communication of information" that Zadeh presented to the New York Academy of Sciences was the *multiplex transmission of two or more signals*. In the interests of simplicity he assumed that the system has two channels and the sets of signals assigned to their respective channels are $X = \{x(t)\}$ and $Y = \{y(t)\}$. "At the receiving end, one is given the sum signal $u(t) = x(t) + y(t)$, and is required to extract $x(t)$ and $y(t)$ from $u(t)$. The problem is essentially to find two filters N_1 and N_2 such, that, for any x in X and any y in Y,

$$N_1(x+y) = x \quad \text{and} \quad N_2(x+y) = y \tag{2}$$

irrespective of the probabilities of x and y. A filter such as N_1, which can extract any signal belonging to a set X from the sum of x and a signal y belonging to a set Y, is called an ideal filter" ([8], p. 203).

Zadeh represented signals $x(t)$ and $y(t)$ as vectors x and y in an n-dimensional signal space Σ. Then, to the sets of signals, X and Y, there correspond two manifolds M_x and M_y in Σ, which are characterized by k and l relations respectively:

$$f_i(x_1, x_2, ..., x_n) = 0, \quad g_j(y_1, y_2, ..., y_n) = 0, \qquad i = 1, 2, \ldots, k; \; j = 1, 2, \ldots l, \tag{3}$$

where $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are the coordinates of x and y respectively, and f_i and g_j are specified functions of them. The relations between the coordinates of the vectors x, y and u,

$$x_\nu + y_\nu := u_\nu, \qquad \nu = 1, \ldots, n. \tag{4}$$

yield the sum $u(t) = x(t) + y(t)$. Therefore, we have $k + l + n$ equations in $2n$ unknowns $x_1, x_2, \;\; , x_n, \; y_1, y_2, \;\; , y_n$.

At that time Zadeh was familiar with the "electronic brains" that were developed in the late 1940s. He wrote: "If $k + l = n$, these equations can, in general, be solved for the x_ν and y_ν by the use of machine computation or other means. Needless to say, the functions f_i and g_j should be such as to result in unique real values for the x_ν and y_ν." ([8], p. 204). He asserted that the solution of the system of equations in (4) yields the expressions for the coordinates of the signal vector $x(t)$ in terms of the coordinates of the signal vector $u(t)$ in the following form (when the above conditions are fulfilled):

$$x_\nu = H_\nu(u_1, u_2, \;\; , u_n) \qquad \nu = 1, 2, \ldots n \tag{5}$$

Symbolically, he wrote $x \;\; = H\,(u)$, where H (or equivalently, the components H_ν) provides the desired characterization of the ideal filter.

At the end of his talk, Zadeh mentioned the case of linearity of the equations for f_i and g_j. "In this case the manifolds M_x and M_y are linear, and the operation performed by the ideal filter is essentially that of projecting the signal space Σ on M_x along M_y" ([8], p. 204). He illustrated both the nonlinear and the linear modes of ideal filtering in figure 2 (middle and left) in terms of two-dimensional signal space. This analogy between projection in a function space and filtration with an ideal filter led Zadeh in the early 1950s to a functional symbolism of filters [9]. Thus, $N = N_1 + N_2$ represents a filter consisting of two filters connected by addition, whereas $N = N_1N_2$ represents their tandem combination and $N = N_1|N_2$ the separation process (fig. 3).

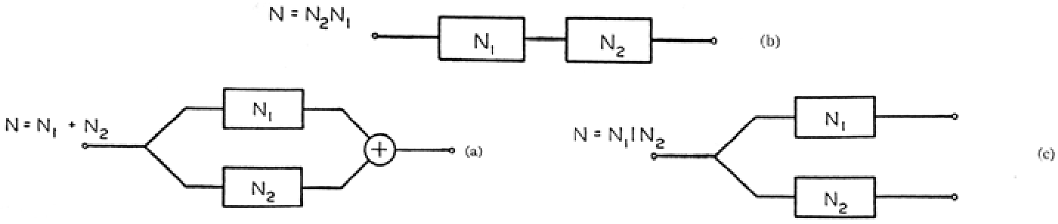


Fig. 3. Functional symbolism of ideal filters, ([9], p. 225).

Later, e.g., in [10], Zadeh introduced the concept of optimal filters contrary to ideal filters. Ideal filters are defined as filters which achieve a perfect separation of signal and noise, but in reality there are no such ideal filters. He knew from experience that characteristics of electrical filters do not show an exact step at the limiting frequency, but instead show smooth functions (fig. 4). Their shapes are similar to the membership functions of fuzzy sets that are well known today, but in the 1950s the time was not ripe for this new mathematical theory. Zadeh regarded optimal filters to be those that give the "best approximation" of a signal and he noted that the "best approximations" depend on reasonable criteria. At that time he formulated these criteria in statistical terms.

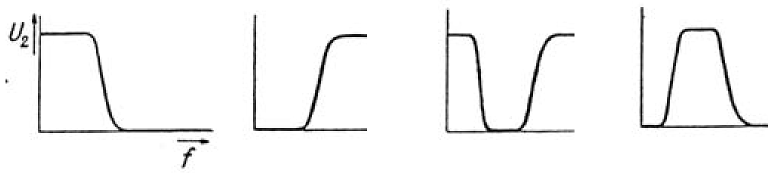


Fig. 4. Characteristics of high-pass, low-pass, band-pass filters.

From Filters to Systems

From Circuit Theory to System Theory [6] includes some paragraphs from Zadeh's eight years older article *System Theory*, published in the *Columbia Engineering Quarterly* [11], where he characterized a system as a "black box" (fig. 5) with inputs $u_1, \ldots, u_m$ and outputs $v_1, \ldots, v_n$, $(m, n \in \mathbb{N})$, and in the case that these inputs and outputs are describable as time dependent functions, then the dynamic behavior of the system can be studied mathematically, and the input-output-relationship of the system is

$$(v_1, \ldots, v_n) = f(u_1, \ldots, u_m). \qquad \textbf{(6)}$$

In the early 1950s, system theory was a scientific discipline of rising importance for "the study of systems per se, regardless of their physical structure." Engineers at that time were, in general, not really trained to think in abstract terms, but nonetheless Zadeh believed that it was only a matter of time before system theory would gain acceptance. It turned out that he was right: Eight years later, when he wrote *From Cir-*

cuit Theory to System Theory, he could describe problems and applications of system theory and its relations to network theory, control theory, and information theory. Furthermore, he pointed out: "… that the same abstract 'systems' notions are operating in various guises in many unrelated fields of science is a relatively recent development. It has been brought about, largely within the past two decades, by the great progress in our understanding of the behavior of both inanimate and animate systems – progress which resulted on the one hand from a vast expansion in the scientific and technological activities directed toward the development of highly complex systems for such purposes as automatic control, pattern recognition, data-processing, communication, and machine computation, and, on the other hand, by attempts at quantitative analyses of the extremely complex animate and man-machine systems which are encountered in biology, neurophysiology, econometrics, operations research and other fields" ([6], p. 856f.).

In his 1954 article *System Theory* Zadeh represented a system as a *block diagram*, i.e., a graphical description of the interrelationships between the variables associated with a system (fig. 6). Thus the block diagram presents in a graphical form the same information about a system as is conveyed by writing the input-output relationship (5). In Zadeh's later system theory papers a more sophisticated treatment of these interrelationships is apparent, but in this contribution it is not possible to cover Zadeh's detailed "input-output analysis," which was a very ambitious work in advanced mathematics; for details the reader is referred to chapter 4 in [1, 2]. In the next section Zadeh's path toward establishing an alternative is traced.

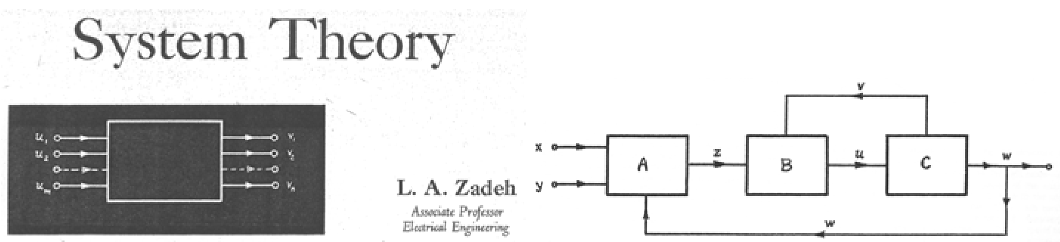


Fig. 5. Left: System with inputs and outputs. Frontispiece of the article [11];
Right: Combination of systems with inputs and outputs ([11], p. 32).

From Systems to Fuzzy Systems

In April 1963, Zadeh participated in the *Second Systems Symposium at Case Institute of Technology* in Cleveland, Ohio, where the organizers brought together, on the one hand, systems scientists concerned with general systems theory and cybernetics and, on the other, technical system scientists. The proceedings were published by Mihaljo D. Mesarović, under the title *Views on General Systems Theory* [12]. In fact, this book contains some very different views and approaches, and Mesarović emphasized in the preface: "Finally, it was expressed that a broad-enough collection of powerful methods for the synthesis (design) of systems of diverse kinds should be considered as constituting the sought-for theory and any further integration was unnecessary." ([12], p. xiv). Kenneth E. Boulding, the well-known economist, philosopher, and founding member of the *Society for General Systems Research* was inspired to write little poems aimed at some of the talks. These poems are printed in the proceedings as introductions to the contributions of the authors. The heading of Zadeh's texts was *The Concept of State in System Theory* [13] and for this presentation Boulding composed the following poem ([12], p. 39), which shows that Zadeh's introduction of a system's state was very clearly impressive:

A system is a big black box
Of which we can't unlock the locks,
And all we can find out about
Is what goes in and what goes out.
Perceiving input-output pairs,
Related by parameters,
Permits us, sometimes, to relate
An input, output, and a state.
If this relation's good and stable
Then to predict we may be able,
But if this fails us – heaven forbid!
We'll be compelled to force the lid! K. B.

Zadeh's starting points to establish a general notion of state in system theory were the fields of dynamical systems and of automata. As an example, he had already presented the Turing machine in [6]: "Roughly speaking, a Turing machine is a discrete time (t = 0, 1, 2, …) system with a finite number of states or internal configurations, which is subjected to an input having the form of a sequence of symbols (drawn from a finite alphabet) printed on a tape which can move in both directions along its length. The output of the machine at time t is an instruction to print a particular symbol in the square scanned by the machine at time t and to move in one or the other direction by one square. A key feature of the machine is that the output at time t+1 and the state at time t+1 are determined by the state and the input at time t." ([6], p. 858).

If s_t, u_t, and y_t denote the *state*, *input*, and *output* of the Turing machine at time t, respectively, and if f and g are functions on pairs of s_t and u_t, then the machine-operation is characterized by the following set of state equations:

$$s_{t+1} = f(s_t, u_t), \qquad y_t = g(s_t, u_t), \qquad t = 0,1,2,\ldots, \tag{7}$$

If the system is a differential system instead of a discrete-state system, the *state*, *input*, and *output* of the system are represented by vectors $s(t)$, $y(t)$, and $u(t)$, respectively. With $\dot{s}(t) = d/dt\, s(t)$ state equations in (1) assume the forms

$$\dot{s}(t) =) f(s(t), u(t)), \qquad y(t) = g(s(t), u(t)). \tag{8}$$

In the 1940s and 1950s some mathematicians and control theorists in the Soviet Union used these state equations earlier than western scientists, and Lotfi Zadeh took notice of the scientific progress in the Soviet Union after he emigrated to the United States. He referred to the fact that "in the United States, the introduction of the notion of state and related techniques into the theory of optimization of linear as well as nonlinear systems is due primarily to Richard Ernest Bellman, whose invention of dynamic programming has contributed by far the most powerful tool since the inception of the variational calculus to the solution of a whole gamut of maximization and minimization problems." ([6], p. 858.)

Richard Bellman and Lotfi Zadeh had been friends since the late 1950s and in the summer of 1964 they intended to collaborate for some weeks at RAND in Santa Monica. In the preceeding years, Zadeh had tried to find precise mathematical definitions of some systems' properties, for instance, optimality, adaptivity, and linearity. He came up with new and more precise definitions, but this research project proved to be very difficult. Gradually Zadeh began to realize that he had lost! In sum, he saw that conventional mathematics had gone too far away from real world problems. Moreover, he perceived that it would not be possible to compute all the system equations of any real system. Even the new digital computers of the 1940s and 1950s could not help obtain exact knowledge of what happens in real world systems. To compute, to describe, or to control processes in complex systems – in particular in large-scale systems – there are too many state equations and, as a result, too many systems of differential equations. Computers cannot solve that many differential equations in a limited time.

In the summer of 1964 Zadeh planned to go to Santa Monico to visit his friend Richard Bellman, after he had given a talk on pattern recognition at a conference at the Wright-Patterson Air Force Base in Dayton, Ohio. During this time Zadeh started thinking about the possibility of using grades of membership in the pattern separation problem. The separation of patterns or sets of points in an n-dimensional Euclidean space was a basic problem for pattern recognition and Zadeh thought that grades of membership might be a good tool.

A few weeks later, in Santa Monica, Zadeh described a preliminary version of these thoughts to Bellman. It was not too difficult to develop these ideas and to extend them to build a mathematical theory. After some discussion, Zadeh wrote the manuscript "Fuzzy Sets" and submitted it to the editors of the journal *Information and Control* in November 1964 [14].

Fig. 6. Left to right: Richard E. Bellman, Robert E. Kalaba, and Lotfi A. Zadeh (all photographs were taken in the 1960s).

A preprint version of this text appeared as a report of the Electronics Research Laboratory (ERL) of the University of California at Berkeley in November 1964 [15]. In addition, Zadeh sent an elaboration of their discussions in Santa Monica to Bellman, who was then the editor of the *Journal of Mathematical Analysis and Applications*on.. Bellman decided to publish this paper in the journal, but it did not appear until 1966 under the title "Abstraction and Pattern Classification" by the authors Richard Bellman, Robert Kalaba and Lotfi Zadeh [16]. The text of this article is identical to that of the memorandum RM-4307-PR of the RAND Corporation, which had already appeared in October 1964 [17]. This memo was also a "preliminary paper in which the authors discuss a general framework for the treatment of pattern recognition problems." It was written by Lotfi Zadeh, who here introduced the concept of a fuzzy set as "a notion which extends the concept of membership in a set to situations in which there are many, possibly a continuum of, grades of membership." ([17], p. 1)

In the spring of 1965 the Symposium on Systems Theory took place at the Polytechnic Institute in Brooklyn (April 20-22, 1965) and Zadeh gave a talk to an audience primarily composed of non-technical systems scientists. The title of his talk was "A New View on System Theory," but he only treated one subject, namely, fuzzy sets and systems.[2] His "new view" dealt with the concepts of fuzzy sets "which provide a way of treating fuzziness in a quantitative manner" ([18], p. 29). In this talk he defined "fuzzy systems" to a large audience for the first time:

DEFINITION: A system S is a fuzzy system if (input) $u(t)$, output $y(t)$, or state $x(t)$ of S or any combination of them ranges over fuzzy sets. ([18], p. 33)

Zadeh explained that "these concepts relate to situations in which the source of imprecision is not a random variable or a stochastic process but rather a class or classes which do not possess sharply defined boundaries." ([18], p. 29) His examples were "the 'class' of real numbers which are much larger than, say, 10" and "the 'class' of bald men". He also considered the case of pattern classification: "For example, suppose that we are concerned with devising a test for differentiating between handwritten letters O and D. One approach to this problem would be to give a set of handwritten letters and indicate their grades of membership in the fuzzy sets O and D. On performing abstraction on these samples, one obtains the estimates $\tilde{\mu}_O$ and $\tilde{\mu}_D$ of μ_O and μ_D respectively. Then given a letter x which is not one of the given samples, one can calculate its grades of membership in O and D, and, if O and D have no overlap, classify x in O or D." ([18], p. 29) Here and in his later article, "Fuzzy Sets" Zadeh noted that "Such classes are not classes or sets in the usual sense of these terms, since they do not dichotomize all objects into those that belong to the class and those that do not."[3]

"Fuzzy Sets"

In April 1965, when Zadeh gave his talk in Brooklyn [18], his article "Fuzzy Sets" was already in press, and he anticipated his substance, i.e., a new "way of dealing with classes in which there may be intermediate grades of membership." He introduced "the concept of a fuzzy set," that is, a class in which there may be a continuous infinity of grades of membership, with the grade of membership of an object x in a fuzzy set A represented by a number $\mu_A(x)$ in the interval [0,1]." Zadeh maintained that these new concepts provided a "convenient way of defining abstraction — a process which plays a basic role in human thinking and communication."

---

[2] In the Proceedings of this symposium there is a shortened manuscript version of the talk with the heading *Fuzzy Sets and Systems*: [18]

[3] Zadeh used quotation marks to indicate the difference between usual classes or sets and his new (fuzzy) sets.([18], p. 29).

To generalize various concepts of ordinary set theory, Zadeh defined *equality*, *containment*, *complementation*, *intersection,* and *union* relating to fuzzy sets A, B in any universe of discourse X as follows (for all $x \in X$):

- $A = B$ if and only if $\mu_A(x) = \mu_B(x)$,
- $A \subseteq B$ if and only if $\mu_A(x) \leq \mu_B(x)$,
- $\neg A$ is the complement of A if and only if $\mu_{\neg A}(x) = 1 - \mu_A(x)$,
- $A \cup B$ if and only if $\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))$,
- $A \cap B$ if and only if $\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))$.

For his interpretation of fuzzy unions and intersections, he had a separate paragraph presenting a very important analogy with sieves. Zadeh wrote: "Specifically, let $f_i(x)$, $i = 1,\ldots, n$, denote the value of the membership function of A_i at x. Associate with $f_i(x)$ a sieve $S_i(x)$ whose meshes are of size $f_i(x)$. Then, $f_i(x) \vee f_j(x)$ and $f_i(x) \wedge f_j(x)$ correspond, respectively, to parallel and series combinations of $S_i(x)$ and $S_j(x)$ as shown in figure 7.[4].

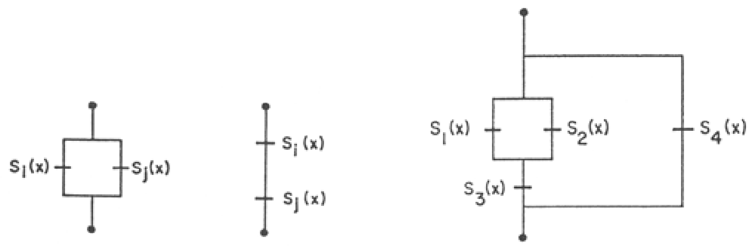


Fig. 7. Figures in Zadeh's paper [17] and their original captions: Left: Parallel and serial connection of sieves simultating $\cup$and $\cap$. Right: A network of sieves simultating $\{[f_1(x) \vee f_2(x)] \wedge f_3(x)\} \vee f_4(x)$.

If one takes into account that the term "sieve" connotes the meaning of a filter, then the analogy with fuzzy sets and electrical filters as discussed in the first section can be seen. From this paper's perspective on the history of the theory of fuzzy sets and systems, it can be concluded that the genesis of the concepts of fuzzy sets and fuzzy systems was an integral part of system theory in the USA in the 1960s.

In 1973 in "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," Zadeh introduced "linguistic variables" that are variables whose values may be sentences in a specific natural or artificial language. [19] To illustrate: the values of the linguistic variable "age" might be expressible as *young*, *very young*, *not very young*, *somewhat old*, *more or less young*. These values are formed with the label *old*, the negation *not*, and the hedges *very*, *somewhat*, and *more or less*. In this sense the variable "age" is a linguistic variable (fig. 8).

Linguistic variables became a proper tool for reasoning without exact values. Since in many cases, it is either impossible or too time-consuming (and therefore too expensive) to measure or compute exact values, the concept of linguistic variables has been successful in many fuzzy application systems, e.g., in control and decision making.

---

[4] These illustration shows very clear the connection of fuzzy union (maximum) and intersection (min) with Zadeh's training in the theory of electrical filters (sieves).

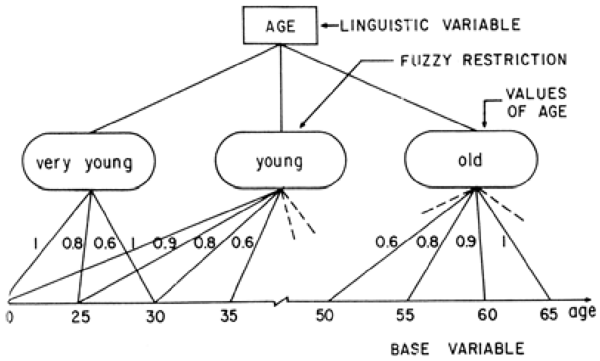


Fig. 8. Example of the linguistic variable “Age” [19].

Part II: The Philosophy of Science and Fuzzy Set Theory

The aim of this section is to propose some ideas regarding what directions further research on the theory of fuzzy sets should take. Fuzzy sets are a new concept in mathematics and also a new concept in science – a concept that forgoes precision. This can be regarded as an advantage – especially in connection with nonclassical scientific theories. In this part attention will focus on considerations pertaining to quantum mechanics and fuzzy sets and on a generalized approach to theories of uncertainty comprising probability theory and fuzzy set theory.

Classical Science

In classical science we examine real systems, we observe variables (observables) assigned to these systems, and we measure their values. Consequently, we have to keep in mind that scientific research is a mixture of theory and practice: The assignment of variables to objects or systems is a theoretical act, but the measurement of their values in experiments is empirical. In order to investigate the interrelationships of scientific systems, we combine their observables into equations that describe or refer to natural laws. To verify these laws, we have to measure these time-dependent variables in nature or in laboratories. For example, we consider two physical systems (fig. 9): First we have a mechanical system: variable v_2 represents the effective force at particle M, whereas variable v_1 represents the velocity of M. The second system is an electrical network: here, variable v_1 represents the input-voltage and variable v_2 the electric current through the network.

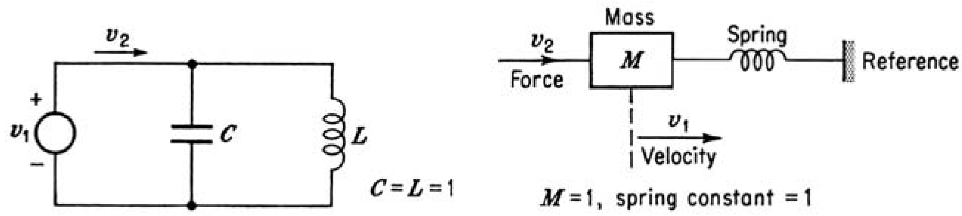


Fig. 9. Mechanical system and electrical network ([5], p. 14).

Both of these systems may be governed by identical mathematical equations, e.g., a mathematical construct with two terminal variables v_1 and v_2, yields the equation

$$\frac{d^2 v_2}{dt^2} = \frac{d^2 v_1}{dt^2} + v_1 \tag{3}$$

This example from Lotfi A. Zadeh's part of the textbook *Linear System Theory* [5] shows that scientific disciplines are concerned with physical and theoretical systems as mathematical constructs and the system-theoretic approach in engineering allows us "to study systems per se, regardless of their physical structure" ([11], p. 16).

To investigate classical physical systems, we have to observe characteristic variables, at best a set of observables that characterize the system completely. This complete set of observables represents the state of the system. For example, in Newtonian mechanics the state of a system (a particle with mass m) is given by the pair of values of the system's position vector $\boldsymbol{x}$ and its momentum vector $\boldsymbol{p}$. These two values implicate all other properties of the system that are relevant in the Newtonian theory.

To generalize, we can formulate that the state of a physical system is the collection of all the system's properties. In order to represent these properties in terms of the physical theory, we must determine the formally possible observables in this mechanical theory, and in order to know the system's properties at a given point in time t, we must measure the values of these observables. Thus, the representation of the "state of a classical system" is related to the measurement or the perception process of the observer. Due to the possible errors of measurement and the systematic errors occurring in every experiment, we can attribute their probability of this being the real value to all measured values of observables. Thus, the state of a system in Newtonian mechanics is given by the pair of the values of position $\boldsymbol{x}$ and momentum $\boldsymbol{p}$ and their probability distributions.

Quantum Mechanics

Due to the scientific revolution brought about by the discovery of quantum mechanics in the first third of the 20th century, a basic change took place in the relationship between the exact scientific theory of physics and the phenomena observed in basic experiments. Systems of quantum mechanics do not behave like systems of classical theories in physics – they are not particles and they are not waves, they are different. This change led to a new mathematical conceptual fundament in physics. Werner Heisenberg, Niels Bohr, and others introduced new theoretical terms in the new quantum mechanics theory that differ significantly from those of classical physics. Their properties are completely new and are not comparable to those of observables in classical theories. The theory of quantum mechanics is completely abstract: it is a theory of mathematical state functions that have no exact counterpart in reality.

Why do we need this new concept of a state function in this non-classical theory? – In experimental physics we can observe only classical variables, e.g., position and momentum. Therefore, physicists subject subatomic objects of reality to classical experiments. It is much more difficult to determine the state of a quantum mechanical system than it is to determine that of classical systems, as we cannot measure sharp values for both variables simultaneously. This is the meaning of Heisenberg's uncertainty principle.

But upon what is Heisenberg's uncertainty principle based? We can experiment with quantum mechanical objects in order to measure a position value, and we can also experiment with these objects in order to measure their momentum value. However, we cannot conduct both experiments simultaneously and thus are not able to get both values for the same point in time respectively. But we can predict these values as outcomes of experiments at this point in time. Since predictions are targeted on future events, we cannot valuate them with the logical values "true" or "false", but use probabilities.

To determine the state of a quantum mechanical system, we have to modify the classical concept. Analogous to the state of classical systems, the state of a quantum mechanical system consists of all the probability distributions of all the object's properties that are formally possible in this physical theory. In classical mechanics the probability distributions for the observables position and momentum and all the other formally possible properties are marginal probability distributions of the unique probability distribution of the system's state, but in quantum mechanics we have no probability distribution that would be the joint probability distribution of all observables: in accordance with the uncertainty principle, not all of them are compatible.

Approaches to Deal with Uncertainty in Quantum Mechanics

In 1926 Max Born proposed an interpretation of this non-classical peculiarity of quantum mechanics – the quantum mechanical wave function is a "probability-amplitude" [20, 21]: The absolute square of its value equals the probability of it having a certain position or a certain momentum if we measure the position or momentum respectively. In 1932, John von Neumann published the *Mathematical Foundations of Quantum Mechanics* [22], in which he defined the quantum mechanical wave function as a one-dimensional subspace of an abstract Hilbert space, which is defined as the state function of a quantum mechanical system or object. Its absolute square equals the probability density function of its having a certain position or a certain momentum in the position or momentum representation of the wave function respectively. Unfortunately there is no joint probability distribution for events in which both variables have a certain value simultaneously, as there is no classical probability space that comprises these events. Such pairs would describe classical states. Thus, the quantum mechanical object's state function embodies the probabilities of all properties of the object, but it delivers no joint probability distribution for all these properties. Thus: We need a radically different kind of uncertainty theory that is not describable in terms of probability distributions.

In 1936 Garett Birkhoff and John von Neumann proposed the introduction of a new "quantum logic," as the lattice of quantum mechanical propositions is not distributive, and therefore not Boolean [23]. In 1963 George Whitelaw Mackey attempted to provide a set of axioms for the propositional system of predictions of experiments' outcomes. He was able to show that this system is an orthocomplemented partially ordered set. [24] In this logico-algebraic approach, the "probabilities" of evaluating the predictions of the properties of a quantum mechanical object do not satisfy Kolmogoroff's well-known axioms. The double-slit experiment shows that they are not additive and together with their non-distributivity, it is indicated that the probabilistic structure of quantum mechanics is more complicated than that of the classical probability space as it was defined by Kolmogoroff. Already in the 1960s, the philosopher Patrick Suppes discussed the "probabilistic argument for a non-classical logic of quantum mechanics" [25, 26]. He introduced the concept of a "quantum mechanical σ-field" as an "orthomodular partial ordered set" covering the classical σ-fields as substructures. Later, in the 1980s, a "quantum probability theory" was proposed and developed by Gudder and Pitowski [27, 28]. The quantum mechanical lattice of predictions is Suppes' "quantum mechanical σ-field," which can be restricted to a Boolean lattice corresponding to a given observable. The quantum probabilities became classical probabilities again, only applying to predictions of compatible observables. During this period the question of whether it was beneficial to use fuzzy sets instead of probabilities or "quantum probabilities" to interpret quantum mechanics had already arisen. However, this was not successful at that time. This disappointment may have been due to the fact that fuzzy set theory was not as widely accepted as a mathematical tool then as it has become in recent decades and because theoretical physicists showed no interest in using the new theory of fuzzy sets. Moreover, there has also been a lack of interest in fuzzy set theory in the history and philosophy of science until the present day. Now, more than 40 years after its creation, the theory of fuzzy sets is broadly approved and attracts attention in the history of science [1]. In physics, the results of new experiments, such tests by Alain Aspect and his co-workers test of Bells inequality in 1982 [29, 30] and experiments by Anton Zeilinger and his group on quantum teleportation since 1997 [31], have sparked a new debate on the interpretation of quantum mechanics.

Quantum Mechanics and Fuzzy Sets

In a manner similar to Zadeh's extension of systems to fuzzy systems, the definition of a linguistic variable operating on a fuzzy set, and assignment of membership degrees and elements of the term set of the linguistic variable, we propose to define the "fuzzy state" of a physical system as a vector of linguistic instead of numerical variables.

By now linguistic variables have become very successful for reasoning with fuzzy values in cases where it is not possible to measure or compute exact existing values. In quantum mechanics, however, the situation is different: exact values of the classical observables do not exist, but outcomes of a physicist's experiments have to be values of observables, i.e., an observing physicist assigns a sharp value or a probability distribution of an observable (e.g., its position) to a quantum theoretical object. This value or probability distribution is not sufficient to determine the quantum object's state. It is only one representation of the state that is not complete.

A concrete system a has a certain number of properties P_i, $i \in \{1, \ldots, n\}$ and a linguistic variable LV_i representing the property P_i can be found. These linguistic variables operate on fuzzy sets and assign membership degrees and elements of a term set, for example:

$$T_{LVi} = \{\text{very small, small, big, very big, … etc .}\}$$

We can imagine the *n*-tuple $\boldsymbol{LV} = \langle LV_1, LV_2, \ldots, LV_n\rangle$ to be a vector in an *n*-dimensional Cartesian space. The value $\boldsymbol{s_f} = \boldsymbol{LV}(a, t)$ for a system *a* at time point *t* is called the "fuzzy state" of this system at this time. During this time, the state $\boldsymbol{s_f}$ moves in the state space $\Sigma(\mathrm{a}) = \{\boldsymbol{LV}(\mathrm{a}, t) \mid t \in T\}$ of the system.

In the case of a classical particle in Newtonian mechanics, the "fuzzy state" is a pair $\boldsymbol{s_f} = (LV_{\boldsymbol{x}}, LV_{\boldsymbol{p}})$ of the two linguistic variables – position $LV_{\boldsymbol{x}}$ and momentum $LV_{\boldsymbol{p}}$.

Usually, the shape of the fuzzy sets' membership functions is subjectively chosen or dependent on the problem at hand. In one particular case, the membership function may have the shape of the Gaussian distribution over one of its values (linguistic terms) and thus the fuzzy state variable yields the probabilities of measuring the observables position $\boldsymbol{x}$ and momentum $\boldsymbol{p}$ due to the calculation of errors. Thus, this is the case if the system under consideration is a system of classical physics.

However, in general, membership functions of fuzzy sets do not represent probability distributions of measurement errors or randomness, but more general uncertainties that are deeply rooted in the absence of the theoretical concept's strict boundaries. Classical concepts such as position and momentum, which have strict boundaries in Newtonian mechanics, do not have such boundaries in the new theory of quantum mechanics; this is the result of Heisenberg's uncertainty relations. Therefore, this pair of classical concepts does not match the quantum mechanical state variable – in the represent of the quantum mechanical "state," there is some uncertainty regarding position and momentum.

This concept of uncertainty is often misleadingly represented by classical probability as well, but in a strict sense quantum mechanical uncertainty is different from the concept of classical probability. Therefore, this approach can be extended to include the assumption that the classical theoretical concepts are not the right concepts, but that we have no better concepts to interpret the outcomes of classical experiments. Thus, we can use fuzzy sets and linguistic variables to convert classical observables to objects of quantum mechanics.

In the case of a quantum mechanical system, the "fuzzy state" is a vector $\boldsymbol{s_f} = \boldsymbol{LV(a,t)}$ in the abstract Hilbert space, with an infinite tuple $\boldsymbol{LV} = \langle LV_1, LV_2, \ldots\rangle$ of linguistic variables $LV_{i.}$, and not all linguistic variables $LV_{i.}$ and $LV_{j.}$ are compatible, e.g., $LV_{\boldsymbol{x}}$ (the position observable) and LV_p (the momentum observable). We can measure one of these LV_i and this measurement reduces the membership function to a numerical value – or more realistically – a probability distribution of values of this observable. This is an effect that is known in usual quantum mechanics as the "collapse" of the quantum mechanical state function, but in the fuzzy approach we have no collapse of the fuzzy state. In fact, we see that the fuzzy state of a quantum mechanical system is a vague concept.

Uncertainty Theories in the Philosophy of Science

As we have seen in the previous sections, two examples of new theories arose in the 20th century, quantum mechanics and fuzzy set theory. Both deal with the concept of uncertainty and both use mathematical tools that are essentially different from classical probabilities, but these mathematical concepts can be restricted (or specialized) to this classical case. In both theories, the introduction of new concepts (theoretical terms) was necessary due to phenomena in reality that were not explainable using existing mathematical concepts at the time. The older mathematical theories were not able to represent the phenomena and processes that had been observed in reality.

Two trends in obtaining systematic rational reconstructions of empirical theories can be found in the philosophy of science in the latter half of the 20th century: the *Carnap approach* and the *Suppes approach*. In both approaches, the first step consists of an axiomatization that intends to determine the mathematical structure of the theory in question. However, whereas in the Carnap approach the theory is to be axiomatized within a formal language, the Suppes approach uses informal set theory. Thus, in the Suppes approach, one is able to axiomatize real physical theories in a precise way without recourse to formal languages. This approach traces back to the proposal of the philosopher Patrick Suppes in the 1950s to include the axiomatization of empirical theories of science in the metamathematical program of the French group "Bourbaki" [32]. Later, in the 1970s, Joseph D. Sneed developed informal semantics meant to consider not only mathematical aspects, but also application subjects of scientific theories in this framework, based on this method.

In his book [33], Sneed presents this view as stating that all empirical claims of physical theories have the form "*a* is an *S*" where "is an *S*" is a set-theoretical predicate (e.g., "*a* is a classical particle mechanics"). Every physical system that fulfills this predicate is called a model of the theory. To give concrete examples, the class M of a theory's models is characterized by empirical laws that consist of conditions governing the connection of the components of physical systems. Therefore, we have models of a scientific theory, and by removing their empirical laws, we get the class M_p of so-called potential models of the theory. Potential models of an empirical theory consist of *theoretical terms*, i.e., observables with values that can be measured in accordance with the theory. This connection between theory and empiricism is the basis of the philosophical "problem of theoretical terms."

If we remove the theoretical terms of a theory in its potential models, we get structures that are to be treated on a purely empirical layer; we call the class M_{pp} of these structures of a scientific theory its "partial potential models." Finally, every physical theory has a class I of intended applications and, of course, the different applications of a theory partially overlap. This means that there is a class C of constraints that produces cross connections between the overlapping applications.

In brief, this structuralist view of scientific theories regards the core K of a theory as a quadruple $K = \langle M_p, M_{pp}, M, C\rangle$. This core can be supplemented by the class I of intended applications of the theory $T = \langle K, I\rangle$. To make it clear that this concept reflects both sides of scientific theories, these classes of K and I are shown in figure 10.

Later, Sneed, Wolfgang Stegmüller, C. Ulises Moulines, Wolfgang Balzer, and others developed this view into a framework intended to consider networks of theories and evolutions of theories [34]. In such networks the theories are linked by set-theoretical relations, called "specialization," "theoretization," "reduction," etc. Numerous theories for different intended cross-connected applications can be studied in this environment: within the structuralist view, commonalities and differences are expressible as intertheoretic relations (specialization, generalization, and theoretization, as well as strict and approximative reduction). With these concepts of networking in the space of scientific theories, a dynamic – and therefore historical – aspect of these philosophical considerations is possible.

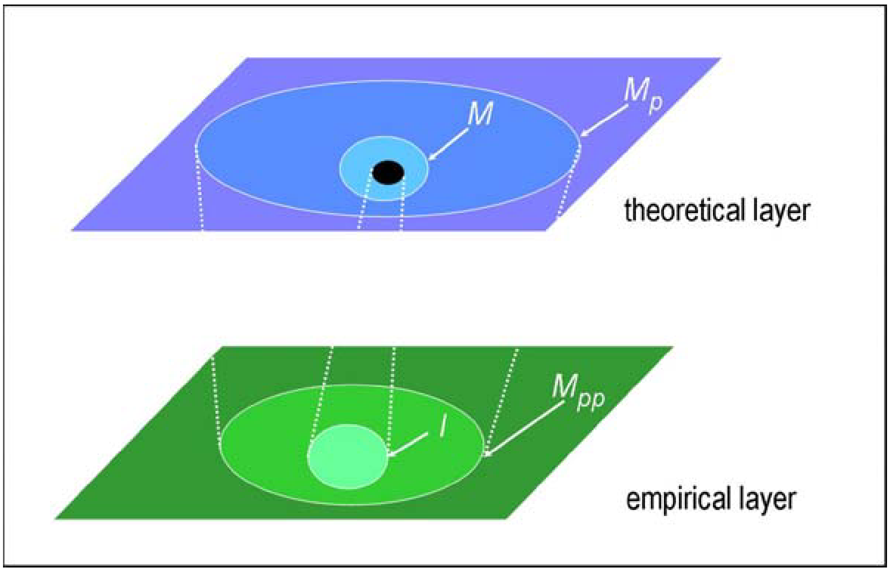


Fig. 5. Empirical and theoretical layer of structures in scientific theories.

Conclusion: Constructing a Theory-Net for Uncertainty

Fuzzy set theory is an uncertainty theory of "quantities which are not describable in terms of probability distributions"; quantum mechanics is a physical theory that cannot be modeled mathematically with classical probabilities; thus quantum probability theory is a new theory as well. In his Ph.D. dissertation on "probabilistic structures in quantum mechanics" [35], the author has already shown that probability theory and quantum probability theory are expressible in the framework of the structuralist view of scientific theories and, moreover, that these theories are both theories in one network of structures of a "general probability theory," GPT. The addition of fuzzy set theory to this network as a supplementary theory requires changing the network core from GPT to a "basic uncertainty theory," BUT.

It is a new view of fuzzy set theory, quantum mechanics, probability theory, and perhaps even additional theories dealing with uncertainty, to embed them in one network of "uncertainty theories" and to explain their differences and their commonalities as intertheoretical relationships. This new view suggests a research project analyzing and reconstructing uncertainty theories – in the first instance, probability theory, quantum probability theory, and fuzzy set theory. Perhaps we can add the theories of Dempster-Shafer, certainty factors, belief functions, etc. This project would make a significant contribution to the understanding of many concepts of uncertainty in science and technology and of their relation to reality. Moreover, we can expect to get an overview of the intended applications of these network elements and their cross connections.

Last year Lotfi Zadeh published a paper called “Toward a Generalized Theory of Uncertainty (GTU) – an Outline” [36]. He began with the following words: “It is a deep-seated tradition in science to view uncertainty as a province of probability theory. The Generalized Theory of Uncertainty (GTU) which is outlined in this paper breaks with this tradition and views uncertainty in a broader perspective.” ([36], p. 1) In Zadeh’s favored GTU, probabilistic and “statistical information is just one – albeit an important one – of many forms of information” ([36], p. 2).

Zadeh’s approach to achieving a generalized theory of uncertainty is very different from the approach introduced in the present paper, which is to find a structuralist view of uncertainty theories, their basic elements, and many of their interrelationships. Nevertheless, it is desirable to proceed in both directions in order to encompass the entire system of classical and non-classical sciences.

Acknowledgements

I wish to thank Lotfi A. Zadeh (Berkeley) for his generous support of my historical project on the theory of fuzzy sets and systems during the past few years. I also want to thank Rudolf Kruse (Magdeburg) for his encouragement to continue this project. Many thanks go to Wolfgang Balzer (Munich), my former supervisor in the philosophy of science for his introduction to the stucturalist view of theories and to Dieter Straub (Munich) for important and enjoyable discussions. Finally, I would like to thank Witold Pedrycz (Edmonton) for his encouragement to make a shift toward this interdisciplinary and science-oriented research project.

References

[1] R. Seising, Die Fuzzifizierung der Systeme. Die Entstehung der Fuzzy Set Theorie und ihrer ersten Anwendungen – Ihre Entwicklung bis in die 70er Jahre des 20. Jahrhunderts. Stuttgart: Steiner 2005.

[2] R. Seising, The Fuzzification of Systems The Origins of Fuzzy Set Theory and Its Initial Applications – Its Development to the 1970s. Berlin: Springer-Verlag 2007.

[3] R. Seising, Proposals for Future Developments in Fuzzy Set Technology. 2006 IEEE World Congress on Computational Intelligence, Vancouver, BC, Canada. July 16-21, 2006 (CD, ISBN: 0-7803-9489-5).

[4] R. Seising, Can Fuzzy Sets be Useful to (re)interpret Uncertainty in Quantum Mechanics? K. Demirli and A. Akgunduz (eds.): Proceedings of the 2006l Conference of the North American Fuzzy Information Processing Society (NAFIPS-06), June 3-6, 2006, Montréal, Canada.(CD, ISBN: 0-7803-9188-8).

[5] L. A. Zadeh, C. A. Desoer, Linear System Theory: The State Space Approach. New York 1963.

[6] L. A. Zadeh, From Circuit Theory to System Theory, Proceedings of the IRE, vol. 50,, 1962, pp. 856-865.

[7] L. A. Zadeh, J. R. Ragazzini, An Extension of Wiener’s Theory of Prediction, Journal of Applied Physics, Vol 21 (1950), pp. 645-655.

[8] L. A. Zadeh: Some Basic Problems in Communication of Information, in: The New York Academy of Sciences, 1952, Ser. II, Vol.14, No. 5, pp. 201-204.

[9] L. A.Zadeh, Theory of Filtering. Journal of the Society for Industrial and Applied Mathematics, Vol. 1, 1953, pp. 35-51.
[10] L. A. Zadeh, General Theory of Linear Signal Transmission Systems. Journal of the Franklin Institute, April, 1952, pp. 293-312.
[11] L. A. Zadeh, System Theory, Columbia Engineering Quarterly, Nov. 1954, pp. 16-19, 34.
[12] M. D. Mesarovic, Views on General Systems Theory. Proc. of the Second Systems Symposium at Case Institute of Technology. Huntington, New York: R. E. Krieger Publ. 1964.
[13] L. A. Zadeh, The Concept of State in System Theory, in: [12], pp. 39-50.
[14] L. A. Zadeh, Fuzzy Sets, Information and Control, 8, 1965, pp. 338-353.
[15] L. A. Zadeh, Fuzzy Sets, Electronic Research Laboratory (ERL), University of California Berkeley, Report No. 64-44, November 16, 1964.
[16] R. E. Bellman, R. Kalaba, L. A Zadeh, Abstraction and Pattern Classification. Journal of Mathematical Analysis and Applications 13, 1966, pp. 1-7.
[17] R. E. Bellman, R. Kalaba, L. A. Zadeh, Abstraction and Pattern Classification. Memorandum RM- 4307-PR, Santa Monica, California: The RAND Corporation, October 1964.
[18] L. A. Zadeh, Fuzzy Sets and Systems. In: Fox, J. (ed.): System Theory. Microwave Research Institute Symp. Series XV, Brooklyn, New York: Polytechnic Press 1965, p. 29-37.
[19] L. A. Zadeh, On the Analysis of Large-Scale Systems. Memo No. ERL-M418, Electronic Research Laboratory, College of Engineering, Univ. of California, Berkeley, Jan. 8, 1974.
[20] M. Born, Zur Quantenmechanik der Stoßvorgänge, Zeitschrift für Physik, vol. 37, pp. 863-867, 1926.
[21] M. Born, Das Adiabatenprinzip in der Quantenmechanik, Zeitschrift für Physik, vol. 40, pp. 167-191, 1926.
[22] J. von Neumann, Mathematical Foundations of Quantum Mechanics, Princeton Univ. Press, Princeton, New Jersey, 1955.
[23] G. Birkhoff, J. von Neumann, The Logic of Quantum Mechanics, Annals of Mathematics, Series 2, vol. 37, p. 823 ff, 1936.
[24] G. W. Mackey, Mathematical Foundations of Quantum Mechanics, New York: W. A. Benjamin, 1963.
[25] P. Suppes, Probability Concepts in Quantum Mechanics, Philosophy of Science, vol. 28, pp. 378-389, 1961.
[26] P. Suppes, The Probabilistic Argument for a Non-Classical Logic of Quantum Mechanics, Philosophy of Science, vol. 33, pp. 14-21, 1966.
[27] S. P. Gudder, Quantum Probability, Academic Press: San Diego, 1988.
[28] I. Pitowski, Quantum Probability, Quantum Logic, Lecture Notes in Physics, p. 321, Berlin: 1989.
[29] A. Aspect, P. Grangier, and G. Roger, "Experimental Realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A New Violation of Bell's Inequalities," Physical Review Letters, vol. 49, no. 2, pp. 91-94, 1982.
[30] A. Aspect, Dalibard, G. Roger, "Experimental test of Bell's inequalities using time-varying analyzers," Physical Review Letters, vol. 49, pp. 1804-1807, 1982.
[31] Dik Bouwmeester, Jan-Wei Pan, Klaus Mattle, Manfred Eibl, Harald Weinfurter, Anton Zeilinger, "Experimental Quantum Teleportation," Nature, vol. 390, p. 575 ff, 1997.
[32] N. Bourbaki, (pseudo.): Elements of Mathematics: Theory of Sets, Addison-Wesley, Reading, Mass. 1968.
[33] J. D. Sneed, The Logical Structure of Mathematical Physics, Dordrecht: Reidel 1971.
[34] W. Balzer, C. U. Moulines, J. D. Sneed, An Architectonic for Science. The Structuralist Program. Dordrecht: Reidel 1987.
[35] R. Seising, Probabilistische Strukturen der Quantenmechanik, Frankfurt: Lang 1996.
[36] L. A. Zadeh, "Toward a generalized theory of uncertainty (GTU) – an outline", Information Sciences, vol. 172, Issue 1-2, pp. 1-40, 2005.

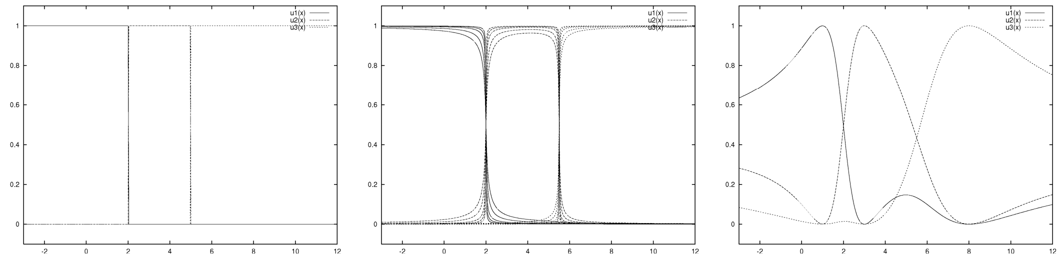
Improved Fuzzy Partitions for Clustering

Frank Höppner

Department of Information Systems
University of Applied Sciences Braunschweig /Wolfenbüttel
Robert Koch Platz 10-14
38440 Wolfsburg, Germany
f.hoeppner@fh-wolfenbuettel.de

When building fuzzy systems automatically from data, we are in need of procedures that automatically divide up the input space in fuzzy granules. These granules are the building blocks for the fuzzy rules. When modeling an input/output relationship, the membership functions of these rules play the same role as basis functions in conventional function approximation tasks. To keep interpretability we usually require that the fuzzy sets are specified in *local regions*, that is, the membership functions have bounded support or decay rapidly. If this requirement is not fulfilled, many rules must be applied and aggregated simultaneously, such that the final result becomes more difficult to grasp – one is not allowed to interpret a fuzzy system *rule by rule* any longer. A second requirement is that the fuzzy sets of the primitive linguistic values should be simple and unimodal. It would be counterintuitive if the membership of the linguistic term "young", which is high for for "17 years", would be higher for "23 years" than for "21 years".

To gain such fuzzy granules clustering algorithms can be used. Especially fuzzy clustering algorithms seem well suited, because they provide the user with a fuzzy membership function which could be used directly for the linguistic terms. Unfortunately, the family of the fuzzy c-means (FCM) clustering algorithms [1] and derivatives produce membership functions that do not fulfill the above-mentioned requirements [6]. Figure 1(c) shows an example for FCM membership functions for a partition of the real line with cluster representatives $c_1 = 1$, $c_2 = 3$ and $c_3 = 8$. We can observe that the support of the membership functions is unbounded for all clusters, in particular for the cluster whose center is located at $c_2 = 3$. While for $c_1 = 1$ and $c_3 = 8$ one allows even in the context of fuzzy systems for an unbounded support if $x < 1$ and $x > 8$ respectively, but at least the membership function for $c_2 = 3$ should be defined locally. Furthermore, we can observe that the membership degree for the cluster at $c_1 = 1$ increases near 5, the FCM membership functions are not unimodal. These undesired properties can be reduced by tuning a parameter of the FCM algorithm, the so-called fuzzifier, however, then we also decrease the fuzziness of the partition and finally end up with crisp indicator functions as shown in Fig. 1(a). The problem of unimodality can be solved by using possibilistic memberships [3], but the possibilistic c-means is not a partitional but a mode-seeking algorithm. In [6] the objective function has been completely abandoned to allow user-defined membership functions, thereby also loosing the partitional property. For further literature about different aspects of interpretability in fuzzy systems, see for instance [2].



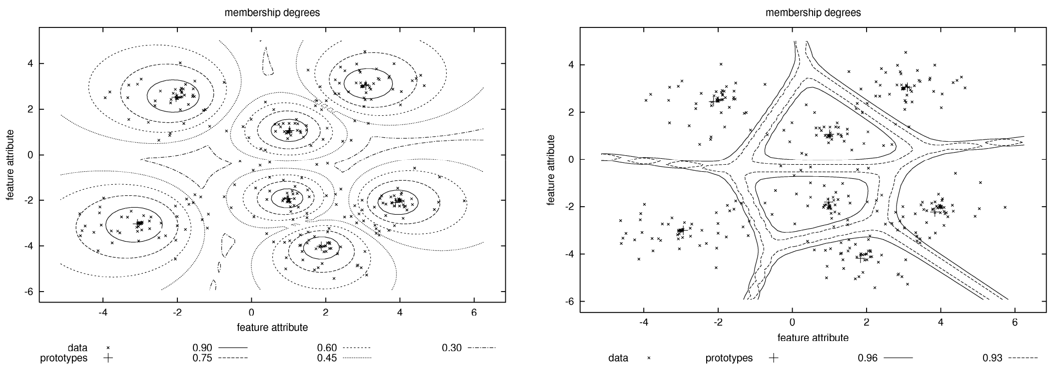
(a) Indicator functions of crisp partition.

(b) Intuitively fuzzified partitions of fig. 1(a).

(c) FCM membership functions ($m = 2.0$).

Fig. 1. Different kinds of membership functions.

We discuss an alternative approach to influence the fuzziness of the final partition [5]. A "reward" term for membership degrees near 0 and 1 is considered in order to force a more crisp assignment. If we choose an (in some sense) maximal reward, we arrive at fuzzy membership functions which are identical to those that we would obtain by using a (scaled) distance to the Voronoi cell that represents the cluster instead of the Euclidean distance to the clusters center. Furthermore, the membership functions – as a whole – can be interpreted as a *fuzzified minimum function* [4], for which we give an estimation of the error we make when substituting a crisp minimum function by its fuzzy version. Figure 2 illustrates the differences when these membership functions are utilized in a clustering algorithm: The iso-membership lines indicate the slow decay and unsatisfying locality in case of FCM in Fig. 2(a) and the Voronoi-like membership degrees in Fig. 2(b) for an artificial 2-dimensional data set.



(a) FCM partition.

(b) Voronoi-like partition.

Fig. 2. Effect of modification on the resulting partition.

Bibliography

[1] J. C. Bezdek. A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2(1):1–8, Jan. 1980.

[2] J. Casillas, O. Cordón, F. Herrera, and L. Magdalena, editors. *Trade-off between Accuracy and Interpretability in Fuzzy Rule-Based Modelling*. Studies in Fuzziness and Soft Computing. Physica, 2003.

[3] R. N. Davé and R. Krishnapuram. Robust clustering methods: A unified view. *IEEE Trans. on Fuzzy Systems*, 5(2):270–293, May 1997.

[4] F. Höppner. Fuzzy shell clustering algorithms in image processing: Fuzzy c-rectangular and 2-rectangular shells. *IEEE Trans. on Fuzzy Systems*, 5(4):599–613, Nov. 1997.

[5] F. Höppner and F. Klawonn. Improved fuzzy partitions for fuzzy regression models. *International Journal of Approximate Reasoning*, 32:85–102, 2003.

[6] T. A. Runkler and J. C. Bezdek. Alternating cluster estimation: A new tool for clustering and function approximation. *IEEE Trans. on Fuzzy Systems*, 7(4):377–393, Aug. 1999.

Visualization of Fuzzy Rule Classifiers for Flight Duration Forecast

Frank Rehm[1], Frank Klawonn[2] and Rudolf Kruse[3]

[1] German Aerospace Center, Braunschweig, Germany
frank.rehm@dlr.de
[2] University of Applied Sciences Braunschweig/Wolfenbüttel, Germany
f.klawonn@fh-wolfenbuettel.de
[3] Otto-von-Guericke-University of Magdeburg, Germany
kruse@iws.cs.uni-magdeburg.de

Abstract

The impact of the weather on the flight duration of aircraft has been analysed in various studies. The complex aspects of the weather, which are accordingly reflected in the weather data, demand sophisticated techniques to visualize the analytical results. In this paper we present an approach to visualize fuzzy rules describing high-dimensional data. By means of this method, the rules, as well as the classified data, can be presented on an arbitrary low-dimensional space. We will demonstrate the efficiency of this technique on some benchmark examples and on real weather data set that is used to predict aircraft flight duration on a European hub airport.

References

1. Berthold, M.R.: Mixed Fuzzy Rule Formation. International Journal of Approximate Reasoning (IJAR), 32, 67–84, Elsevier, 2003.
2. Borg, I., Groenen, P.: Modern Multidimensional Scaling : Theory and Applications. Springer, Berlin, 1997.
3. Rehm, F., Klawonn, F., Kruse, R.: POLARMAP - Efficient Visualisation of High Dimensional Data. IEEE Proceedings of the 10th International Conference on Information Visualisation, London, 2006.
4. Rehm, F., Klawonn, F., Kruse, R.: Rule Classification Visualization of High-Dimensional Data. Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU), Paris, 1944–1948, 2006.
5. Lesot, M-J., Rehm, F., Klawonn, F., Kruse, R.: Prediction of Aircraft Flight Duration. Proceedings of the 11th IFAC Symposium on Control in Transportation Systems, Delft, 107–112, 2006.

On the role of Interpretability in Fuzzy Data Mining

Corrado Mencar, Giovanna Castellano and Anna Maria Fanelli

Department of Informatics, University of Bari, v. E. Orabona, 4, 70126, Bari, Italy
{mencar, castellano, fanelli}@di.uniba.it

Abstract. Fuzzy techniques have been introduced in the realm of Data Mining with the objective of providing for an added value to the “interestingness” of mined patterns. However, a fuzzy model is able to provide interesting patterns of data only when it is not only accurate but also interpretable. This note highlights the main motivations of introducing interpretability and outlines the basic interpretability issues to be addressed in Fuzzy Data Mining.

Keywords: Data Mining, Fuzzy Logic, Interpretability.

1 Introduction

In recent years we are witnessing a massive production of data from the industrial world, which require sophisticated processing techniques in order to be somehow useful to their users. Data Mining can be considered as an umbrella term that covers those learning and statistical techniques which have evolved so as to be applied to such highly dimensional, often noisy, large volumes of data produced by the real world. The aim of Data Mining techniques is to find "interesting" patterns (e.g. regularities, dependencies, etc.) in data, which can be used by users for decision making, strategy planning, and everything else that can be useful to make some profit from available data.

Fuzzy techniques have been introduced in the realm of Data Mining with the objective of providing for an added value to the "interestingness" of mined patterns. The key feature of fuzzy techniques is, indeed, the ability of formally representing properties –such as vagueness, imprecision, preference, similarity– whose semantics is matter of degree. Without fuzzy techniques, such properties could be represented in two alternative ways. The first one concerns the representation of such properties with mathematical logic. Properties are labelled by symbols whose semantics is crisp, or threshold-based. This kind of representation allows mathematical deduction with standard – widely accepted – inference rules. On the other hand, symbolic representation of properties whose semantics is matter of degree diverges from the smooth nature of human perceptions. This divergence may result in difficult interpretation of the results of a data mining process. In other words, the discovered patterns may loose "interestingness".

The second form of representation of properties whose semantics is matter of degree is by using purely numerical methods. Using such methods it is possible to capture the quantitative nature of such properties, hence numerical information is not lost, but the results of computation are not representable in any symbolic form, which is important when knowledge has to be communicated to users, especially to those users that are not skilled (nor interested) in sophisticated mathematical representations. Even in this case, patterns discovered by numerical methods may lose a part of "interestingness".

2 Fuzzy Logic and Data Mining

Fuzzy Logic (in its broader sense) helps data miners in bridging the gap between numerical precision and symbolic representation. The key of its success is its ability of representing properties through symbols which have gradual semantics. This possibility, which gave rise to important developments, such the Theory of Information Granulation, the Generalized Constraint Theory, the Precisiated Natural Language, etc., can be exploited with success to improve the "interestingness" of patterns discovered through data mining. Properties are indeed represented by symbols, possibly natural language symbols, so as to offer a direct communication of knowledge to users. The semantic of such symbols is numeric (graded), so as to catch the smooth nature of properties in a way that is deemed affine to human perception of observed phenomena.

Fuzzy Logic was deemed the panacea until the eighties, where everything was fuzzified sometimes without a matured understanding of the effective meaning of the resulting models. We are still paying for this superficial modus operandi, often in relation to the objections of fuzzy logic detractors. Fortunately, the research in the last two decades has evolved toward more rigorous directions. The development of Fuzzy Logic in the narrow sense (also called Mathematical Fuzzy Logic), the Possibility Theory and other studies helped (and help) modelers to use Fuzzy Logic with more understanding [3].

The cause of possible misunderstandings of fuzzy models is due to the extreme flexibility of Fuzzy Logic representations. As an example, the same formalism can be used to represent properties with completely different (often incompatible) semantics, like degrees of truth and degrees of possibility. Furthermore, the uncontrolled adoption of data driven methods to generate fuzzy models may lead to generate models that are fuzzy only on a purely formal level, but can be considered as pure "black-box" models (at the same level of neural networks) when they have to be understood by the users. Without a proper care, fuzzy models (especially those automatically derived from data) do not provide for "interesting" patterns of data, and therefore the key feature of Fuzzy Logic in data mining is lost.

3 Interpretability issues in Fuzzy Data Mining

A fuzzy model is able to provide "interesting" patterns of data when it meets both the requirements of accuracy and interpretability. While studies on accuracy provided well established theoretical and methodological results, the research on interpretability is still flourishing and not yet matured (see [1] for recent developments in the field). Interestingly, research on interpretability is still in development in the field of Machine Learning and Artificial Intelligence, which surely have a longer tradition than Fuzzy Logic [2]. From the scholars of these two fields we have maybe the most exhaustive definition of the notion interpretability (called "Comprehensibility Postulate" in A.I.) [6]:

The results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single "chunks" of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion.

This definition, even though coming from the A.I. tradition, evidently promotes Fuzzy Logic as a promising tool to meet the comprehensibility postulate, though with the proper care. Fuzzy Logic, and the Theory of Fuzzy Information Granulation in particular, is able to represent "chunks" (granules) of information, which could be directly interpretable in natural language since the gradual semantics of fuzzy information granules is affine to the gradual semantics of natural language terms. Moreover, because of the symbolic/numeric bridging property of Fuzzy Logic, fuzzy information granules have the potential to relate quantitative and qualitative concepts in an integrated fashion.

However, interpretability is not given for grant when fuzzy models are adopted. Interpretability must be formalized, and translated in a number of features (both mathematical or fuzzy) that the fuzzy model must possess in order to be judged as interpretable. To formalize interpretability, a number of issues have to be addresses, such as [5]:

- Who needs interpretability?
- Why and when interpretability is needed?
- What should be interpretable?
- How interpretability can be achieved?
- How interpretability can be assessed?

Each of such issues casts a direction for scientific investigation. Some of these directions have already been undertaken, but none of these can be retained exhausted. As an example, the issue concerning the achievement of interpretability is now addressed by defining a number of interpretability constraints, i.e. formal (or sometimes informal) properties to be imposed during the design of a fuzzy model. Several studies belong to such development line (see, e.g. [4]) However, there is often disagreement on which interpretability constraints have to be used. This disagreement is maybe consequence of the lack of results concerning the assessment of interpretability, which calls for interdisciplinary studies involving psychological, philosophical and mathematical contributions.

Data Mining –especially through fuzzy techniques– cannot exclude the interpretability issue, which is complementary but not less important than the accuracy issue, even though it is the least studied and assessed in literature. Without interpretability, patterns discovered through fuzzy data mining processes might be not as "interesting" as desired. Considering the higher computational complexity of fuzzy methods w.r.t. classical counterparts, fuzzy techniques in data mining that do not take into account interpretability would not be fully motivated.

References

1. Casillas, J. , Cordón, O., Herrera, F., Magdalena L., (eds.): Interpretability Issues in Fuzzy Modeling, Springer (2003)
2. Giboin, A.: ML Comprehensibility and KBS explanation: stating the problem collaboratively, Proceedings of IJCAI'95 Workshop on Machine Learning and Comprehensibility (1995) 1-11
3. Kerre, E.E., Mordeson, J.M.: A historical overview of fuzzy mathematics, New Mathematics and Natural Computation, vol. 1, no. 5, World Scientific (2005) 1-26
4. Mencar, C.: Theory of Fuzzy Information Granulation: Contributions to Interpretability Issues, Ph.D. Thesis, University of Bari, Italy, 2005, available at http://www.di.uniba.it/~mencar/phdthesis.htm
5. Mencar, C., Castellano, G., Fanelli, A.M.: Some Fundamental Interpretability Issues in Fuzzy Modeling, in Proceedings of EUSFLAT 2005
6. Michalski, R.S.: A theory and methodology of inductive learning, Artificial Intelligence, Vol. 20 (1983) 111–161

Mining Fuzzy Rules with Different Semantics

F. Berzal[a], J.C. Cubero[a], D. Sánchez[a,1], J.M. Serrano[b], M.A. Vila[a]

[a] Dep. Computer Science and A.I., University of Granada, Spain
e-mail: {fberzal,JC.Cubero,daniel,vila}@decsai.ugr.es
[b] Dep. Computer Science, University of Jaén, Spain
e-mail: jmserrano@ujaen.es

Extended Abstract

There is a large amount of work devoted to mining fuzzy rules, covering aspects as measures, algorithms and applications [4, 3]. Most of this work is about mining fuzzy association rules under different definitions.

Mining fuzzy rules allows us to deal with fuzzy data, either because data is fuzzy itself, like fuzzy transactions obtained from texts for text mining, or because additional fuzzy information is employed to transform the original data. The latter case includes (among others) using linguistic labels to change the domain of attributes in relational databases, introducing fuzzy resemblance relations between values, or a combination of them.

In these fuzzy datasets it is possible to discover different kinds of rules, according to different semantics. For example, fuzzy association rules of the form $A \Rightarrow C$ in a set of transactions have the meaning "every transaction containing A also contains C", i.e., $t(A) \leq t(C)$ $\forall t \in T$, where T is the set of transactions and $t(X)$ is de inclusion degree of the set of items X in transaction t.

There are several ways in which semantics of rules can be different. For example, we have employed fuzzy association rules with the usual semantics to mine for fuzzy approximate dependencies (rules with a different semantics) by transforming the set of tuples of a certain table into a set of transactions in a different way than usual. The semantics of a dependence like $V \rightarrow W$ in fuzzy relational tables is "whenever two tuples agree in attribute V, they agree also in W", where the agreement and even the values of the attributes can be fuzzy.

A large number of models of fuzzy rules and their corresponding semantics is shown in [2], including the following:

- Implication-based models
  - Certainty rules, "the more x is A, the more certain y is in B"
  - Gradual rules, "the more x is A, the more y is B"
  - Impossibility rules, "the more x is A, the less possible y is not B"
- Conjunction-based models
  - Possibility rules, "the more x is A, the more the values in B are possible for y"

[1] Corresponding author.

- Antigradual rules, “the more x is A, the larger the set of possible values for y around the core of B”

among others. We propose to employ techniques introduced in [1, 3] to model these semantics and to develop algorithms to mine them in fuzzy databases.

We also employ these techniques to mine for a kind of gradual rules that apply on pairs of values of attributes, instead than on particular values, by using a fuzzy relation. These are rules of the kind “If (x_1, x_2) are R_X then (y_1, y_2) are R_Y”, where X and Y are attributes, R_X and R_Y are relations defined on the domain of X and Y, and x_i and y_j are values of X and Y, respectively. Fuzzy approximate dependencies are a particular case of this kind of rules.

As an important contribution, we describe how to adapt existing algorithms to mine for fuzzy rules in order to discover the previous kinds of rules without increasing time and space complexity. Finally, our results are applied in real databases in order to show their usefulness in real applications.

References

1. F. Berzal, I. Blanco, D. Sánchez, J.M. Serrano, and M.A. Vila. A definition for fuzzy approximate dependencies. *Fuzzy Sets and Systems*, 149(1):105–129, 2005.
2. B. Bouchon-Meunier, D. Dubois, LL. Godó, and H. Prade. Fuzzy sets and possibility theory in approximate and plausible reasoning. In J.C. Bezdek, D. Dubois, and H. Prade, editors, *Fuzzy Sets in Approximate Reasoning and Information Systems*, chapter 1, pages 15–190. Kluwer Academic Publishers, 1999.
3. M. Delgado, N. Marín, D. Sánchez, and M.A. Vila. Fuzzy association rules: General model and applications. *IEEE Transactions on Fuzzy Systems*, 11(2):214–225, 2003.
4. E. Hüllermeier, editor. *Fuzzy Sets and Systems 149(1). Special Issue on Fuzzy Sets in Knowledge Discovery.* Elsevier B.V., 2005.

Interpreting Data Mining Quantifiers in Mathematical Fuzzy Logic

Emilia Ylirinne, Esko Turunen*

Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland

June 13, 2006

Abstract: We give a fuzzy logic interpretation for several GUHA style data mining quantifiers, define some new ones and show their basic properties.

Key words: Data mining, BL–algebra, MV–algebra, continuous t–norm.

1 Introduction

In [4, 5, 6], Ivánek showed, in GUHA style data mining framework, that relations between two Boolean attributes derived from data can be quantified by $[0, 1]$–valued functions defined on four–fold tables corresponding to pairs of attributes. In other words, each relation on a data has a $[0, 1]$–valued *strength* that can be interpreted as a *degree of truth* of this relation on the data. Moreover, two such relations can be combined by fuzzy logic connectives, i.e. continuous t–norms, and obtain a new relation. As an example, Ivánek established the following

THEOREM 1 *Let $\Rightarrow^*$ be an implicational quantifier and $\odot$ be a t–norm. Then the quantifier $\Leftrightarrow^*$ constructed for all four–fold tables $\langle a, b, c, d\rangle$ from $\Rightarrow^*$ by the formula*

$$\Leftrightarrow^* (a, b, c) =_{def.} [\Rightarrow^* (a, b)] \odot [\Rightarrow^* (a, c)]$$

is a double implicational quantifier satisfying the property

$$\Leftrightarrow^* (a, b, c) \leq min\{[\Rightarrow^* (a, b)], [\Rightarrow^* (a, c)]\}.$$

Our aim is to take one step forward and show that several GUHA style quantifiers have a natural interpretation in the language of mathematical fuzzy logic. Our approach generates new quantifiers, too. Moreover, our intention is to give a user–friendly description of several quantifiers. To this end we recall some necessary definitions and results.

*Corresponding author. E-mail: esko.turunen@tut.fi

1.1 Quantifiers in data minining

The *GUHA method* for data mining was introduced in [3]. Assuming we have a data file composed of m columns and n rows, consider two Boolean attributes ϕ and ψ. A *four–fold table* $\langle a, b, c, d\rangle$ related to these attributes is composed from numbers of objects in the data satisfying four different binary combinations of these attributes:

	ψ	$\neg\psi$
ϕ	a	b
$\neg\phi$	c	d

where
- a is the number of objects satisfying both ϕ and ψ,
- b is the number of objects satisfying ϕ but not ψ,
- c is the number of objects not satisfying ϕ but satisfying ψ,
- d is the number of objects not satisfying ϕ nor ψ.

Various relations between ϕ and ψ can be measured in the data by different *four–fold table quantifiers*, denoted by $\phi \sim \psi$, $\sim (a, b, c, d)$, $\sim (a, b, c)$ or $\sim (a, b)$ depending on context, which here are understood as functions with values in the real unit interval $[0, 1]$. The most well-known is the *basic implicational quantifier* characterized by an equation

$$\Rightarrow_{\phi} (a, b) = \frac{a}{a+b},$$

see Chapter 2. An intuitive ideal behind this quantifier is that a proposition 'ϕ *implies* ψ' is accepted to be *true* (or, *supported by the data*) if, in the set of objects satisfying ϕ, the amount of objects satisfying ψ, too, is large enough or, equivalently, the number of objects *not* satifying ψ is small enough. In the opposite case the proposition 'ϕ *implies* ψ is *false* (*not supported by the data*). – To give later a user friendly interpretation of a general quantifier, we direct reader's attention to two matters in this example. To begin with, not all the data is important, indeed, only a subset (of number $a+b$) where ϕ is satisfied is relevant to determine the truth of the proposition 'ϕ *implies* ψ'. In the second place, a kind of *complement* of 'enough large a'– in a logical sense – is 'enough small b'.

There are well-known quantifiers that have a clear *statistical* justification, e.g. Fisher quantifier and χ^2 quantifier (cf. [3]) are such quantifiers. Recently Burian [1] gave a *probability theoretical* interpretation of *above average quantifier*. Our intention is to show that there are quantifiers that have a *fuzzy logical* justification.

1.2 Mathematical fuzzy logic

An essential difference between classical Boolean logic and Hájek's *Basic fuzzy logics* introduced in [2] is that Boolean logic is two–valued while, in the latter, propositions have truth values on the real unit interval $[0, 1]$. Moreover, Basic fuzzy logics permit several new logical connectives. In fact, there are different Basic fuzzy logics depending on the choice of the connectives. In Basic fuzzy logics, the role of Boolean algebras is replaced by BL–algebras; a *BL–algebra* is an algebra $L = \langle L, \wedge, \vee, \odot, \to, \mathbf{0}, \mathbf{1}\rangle$ with four binary operations $\wedge, \vee, \odot, \to$ and two different constants $\mathbf{0}$, $\mathbf{1}$ such that, for each $x, y, z \in L$, by setting $x \leq y$ iff $x \wedge y = x$ hold:

$$\langle L, \wedge, \vee, \mathbf{0}, \mathbf{1}\rangle \text{ is a distributive lattice with universal bounds } \mathbf{0} \text{ and } \mathbf{1}, \tag{1}$$

$$\odot \text{ is an associative, commutative and isotone operation and } x \odot \mathbf{1} = x, \tag{2}$$

$$x \odot y \leq z \text{ iff } x \leq y \to z, \tag{3}$$

$$x \wedge y = x \odot (x \to y), \tag{4}$$

$$(x \to y) \vee (y \to x) = \mathbf{1}. \tag{5}$$

Simple examples of BL–algebras are t–algebras $\langle [0,1], \wedge, \vee, \odot_t, \to_t, 0, 1\rangle$, where $\langle [0,1], \wedge, \vee, 0, 1\rangle$ is the usual lattice on the real unit interval $[0,1]$ and $\odot_t$ is a continuous t–norm, whereas $\to_t$ is the corresponding residuum. The most known t–algebras are the following

Gödel algebra: $x \odot_t y = \min\{x, y\}$, $x \to_t y = \begin{cases} 1 & \text{if } x \leq y \\ y & \text{otherwise} \end{cases}$

Product algebra: $x \odot_t y = xy$, $x \to_t y = \begin{cases} 1 & \text{if } x \leq y \\ y/x & \text{otherwise} \end{cases}$

Lukasiewicz algebra: $x \odot_t y = \max\{0, x + y - 1\}$, $x \to_t y = \min\{1, 1 - x + y\}$.

These three examples are fundamental as they characterize all continuous t–norms (for details, see [2]). In this paper we will consider only *Lukasiewicz logic*, the algebraic counterpart of this logic is an *MV–algebra*, a BL–algebra L fulfilling an addtional *double negation law*

$$\text{for each } x \in L, x = x^{**}, \tag{6}$$

where $x^* = x \to \mathbf{0}$.

As mentioned above, several logical connectives can be defined in Basic fuzzy logics. Below we list only those that we will use later. Here they are listed together with their standard semantics in $[0,1]$.

Connective	Semantics	Name
$\neg\phi$	$x^* = 1 - x$	Lukasiewicz negation
$\phi \overline{\wedge} \psi$	$x \odot y = \max\{0, x + y - 1\}$	Lukasiewicz conjunction
$\phi \underline{\vee} \psi$	$x \oplus y = \min\{1, x + y\}$	Lukasiewicz disjunction
$\phi \Leftrightarrow \psi$	$x \leftrightarrow y = 1 - \lvert x - y\rvert$	Lukasiewicz equivalence

It is easy to see that Lukasiewicz negation, disjunction and conjunction satisfy a *De Morgan Law*, i.e. for all $x, y \in [0,1]$,

$$(x \oplus y)^* = x^* \odot y^*.$$

2 Quantifiers and mathematical fuzzy logic

From a general methodological point of view it is of great importance to recognize that fuzziness and probability are different concepts and, therefore, are to be treated with different tools. If, in a set A of n elements, m of the elements have a property ϕ, then a probability theoretical statement is *'Probability of a random element in the set A to have property ϕ is $\frac{m}{n}$'*. A fuzzy approach is to say *'The truth value of set A with respect to property ϕ^* is $\frac{m}{n}$'*, where ϕ^* refers to the elements of A with property ϕ.

Having this formal difference in mind we call a data mining quantifier a *fuzzy logic quantifier* if the condition for it to be true can be expressed by propositions formed by Basic fuzzy logic connectives. For example, the following are Lukasiewicz logic quantifiers.

1. Basic implicational quantifier A proposition connecting two attributes ϕ and ψ by *basic implicational quantifier* is defined to be true in a given data if

$$a \geq n \text{ and } \frac{a}{a+b} \geq p,$$

where $n \in \mathcal{N}$ and $p \in [0,1]$ are parameters given by user.

A fuzzy logic interpretation of this quantifier is the following

Given a data, the determining subset A is formed of cases that satisfy ϕ; there must be enough of them. The data supports a relation 'ϕ implies ψ' if there are few cases in A not satisfying ψ.

We recognize that a proposition *Cases in A not satisfying ψ* has a truth value $\frac{b}{a+b}$ which should be low, therefore, its complement – in terms of Lukasiewicz logic – should be large enough, i.e.

$$\left(\frac{b}{a+b}\right)^* = 1 - \frac{b}{a+b} = \frac{a+b-b}{a+b} = \frac{a}{a+b} \geq p.$$

2. Basic double implicational quantifier A proposition connecting two attributes ϕ and ψ by *basic double implicational quantifier* is defined to be true in a given data if

$$a \geq n \text{ and } \frac{a}{a+b+c} \geq p,$$

where $n \in \mathcal{N}$ and $p \in [0,1]$ are parameters given by user.

A fuzzy logic interpretation of this quantifier is now the following

Given a data, the determining subset A is formed of cases that satisfy ϕ or ψ; there must be enough cases satisfying both of them. The data supports a relation 'ϕ implies ψ and ψ implies ϕ' if there are few cases in A not satisfying ψ or few cases in A not satisfying ϕ.

It is easy to see that a proposition *Cases in A not satisfying* ψ has a truth value $\frac{b}{a+b+c}$ and a proposition *Cases in A not satisfying* ϕ has a truth value $\frac{c}{a+b+c}$, therefore a proposition *Cases in A not satisfying* ψ *or not satisfying* ϕ is related to a Lukasiewicz logic truth value $\frac{b}{a+b+c} \oplus \frac{c}{a+b+c}$ which should be low enough, therefore, its complement should be high enough, i.e.

$$\begin{aligned}
\left(\frac{b}{a+b+c} \oplus \frac{c}{a+b+c}\right)^* &= \left(\frac{b}{a+b+c}\right)^* \odot \left(\frac{c}{a+b+c}\right)^* \\
&= \left(1 - \frac{b}{a+b+c}\right) \odot \left(1 - \frac{c}{a+b+c}\right) \\
&= \max\left\{\left(1 - \frac{b}{a+b+c}\right) + \left(1 - \frac{c}{a+b+c}\right) - 1, 0\right\} \\
&= \frac{a+b+c-b-c}{a+b+c} \\
&= \frac{a}{a+b+c} \geq p.
\end{aligned}$$

3. Basic equivalence quantifier A proposition connecting two attributes ϕ and ψ by *basic equivalence quantifier* is defined to be true in a given data if

$$\frac{a+d}{a+b+c+d} \geq p,$$

where $p \in [0, 1]$ is given by user.

A fuzzy logic interpretation of this quantifier is the following

The determining subset A is the whole data; this data supports a relation 'ϕ and ψ are logically equivalent' if there are plenty of cases satisfying ψ and ϕ or plenty of cases not satisfying ψ nor ϕ.

Obviously, *plenty of cases satisfying* ψ *and* ϕ or *plenty of cases not satisfying* ψ *nor* ϕ requires that $\frac{a}{a+b+c+d}$ is high enough or $\frac{d}{a+b+c+h}$ is high enough. In Lukasiewicz logic this fact can be expressed by saying that

$$\begin{aligned}
\frac{a}{a+b+c+d} \oplus \frac{d}{a+b+c+d} &= \min\{\frac{a}{a+b+c+d} + \frac{d}{a+b+c+d}, 1\} \\
&= \frac{a+d}{a+b+c+d} \geq p.
\end{aligned}$$

The above studied quantifiers are well–known and implemented e.g. in the system LISp–Miner [7]. To avoid misunderstanding, it should be noticed that, in this setting, a proposition

'ϕ implies ψ and ψ implies ϕ'

has, in general, different value of truth than a proposition

'ϕ and ψ are logically equivalent'.

This is due the fact that, in the former, the determining subset A is formed of cases that satisfy ϕ or ψ while, in the latter, the determining subset A is the whole data.

Now we introduce a new Lukasiewicz logic quantifier, namely

4. Quantifier of simultaneous existence A proposition connecting two attributes ϕ and ψ by *quantifier of simultaneous existence* is defined to be true in a given data if

$$\frac{(a+d)-(b+c)}{a+b+c+d} \geq p,$$

where $p \in (0,1]$ is given by user.

A fuzzy logic interpretation of this quantifier is the following

The determining subset A is the whole data. If there are sufficiently more simultaneous occurences or simultaneous absences of ϕ and ψ than non coincident occurences then a proposition connecting the attributes ϕ and ψ by quantifier of simultaneos existence *is true.*

It is evident that a high degree of simultaneous occurences or simultaneous absences of ϕ and ψ is related to truth values

$$\frac{a}{a+b+c+d}, \frac{d}{a+b+c+d},$$

one or the other should be high. Thus, in terms of Lukasiewicz logic, a value

$$\frac{a}{a+b+c+d} \oplus \frac{d}{a+b+c+d} = \frac{a+d}{a+b+c+d} = X$$

should be high. In a same manner the degree of non coincident occurences of ϕ and ψ is related to a truth value

$$\frac{b+c}{a+b+c+d} = Y.$$

Moreover, that X and Y are *sufficently different* means that they are *not similar*, i.e. that a degree $(X \leftrightarrow Y)^*$ is high enough or, equivalently,

$$\begin{aligned}
\left(\frac{a+d}{a+b+c+d} \leftrightarrow \frac{b+c}{a+b+c+d}\right)^* &= 1-\left(1-\left|\frac{(a+d)-(b+c)}{a+b+c+d}\right|\right) \\
&= \left|\frac{(a+d)-(b+c)}{a+b+c+d}\right| \\
&= \frac{(a+d)-(b+c)}{a+b+c+d} \geq p.
\end{aligned}$$

Next we establish some basic properties of the quantifier of simultaneous existence, denoted in the sequal by $\phi \Leftrightarrow^* \psi$.

REMARK 2 *Evidently, whenever $\phi \Leftrightarrow^* \psi$ is true in a data, then $\phi \equiv \psi$, too, is true in the data, where $\phi \equiv \psi$ refers to quantifier of basic equivalence. Thus, the following deduction rules are correct*

$$\frac{\phi \Leftrightarrow^* \psi}{\phi \equiv \psi}, \frac{\phi \Leftrightarrow^* \psi}{\psi \Leftrightarrow^* \phi}, \frac{\phi \Leftrightarrow^* \psi}{\neg\phi \Leftrightarrow^* \neg\psi}.$$

Ivánek [6] called a quantifier $\sim$ Σ*–equivalence* if, having two data M_1, M_2 with he following four–fold tables

$$\begin{array}{r|c|c} M_1 & \psi & \neg\psi \\ \hline \phi & a & b \\ \hline \neg\phi & c & d \end{array} \quad \text{and} \quad \begin{array}{r|c|c} M_2 & \psi & \neg\psi \\ \hline \phi & a' & b' \\ \hline \neg\phi & c' & d' \end{array}$$

such that $a' + d' \geq a + d > 0$ and $b' + c' \leq b + c$, then if $\phi \sim \psi$ is true in data M_1 then $\phi \sim \psi$ is true in data M_2, too.

THEOREM 3 *Quantifier of simultaneous existence is Σ-Equivalence.*

Proof.
We show that

$$\frac{a' + d' - (b' + c')}{a' + b' + c' + d'} \geq \frac{a + d - (b + c)}{a + b + c + d}. \tag{7}$$

To this end, let us denote $x = a + d, x' = a' + d', y = b + c, y' = b' + c'$. Now, (7) holds iff

$$\begin{array}{rcll} \frac{x'-y'}{x'+y'} & \geq & \frac{x-y}{x+y} & \text{iff} \\ (x+y)(x'-y') & \geq & (x-y)(x'+y') & \text{iff} \\ xx' - xy' + yx' - yy' & \geq & xx' + xy' - yx' - yy' & \text{iff} \\ -xy' + yx' & \geq & xy' - yx' & \text{iff} \\ 2yx' & \geq & 2xy' & \text{iff} \\ yx' & \geq & xy', & \end{array}$$

which holds true as $x' \geq x$ and $y' \leq y$. The proof is complete.

Ivánek [6] has proved that a Σ-double implicational quantifier $\sim^*$ can be constructed from a Σ-equivalence quantifier $\equiv^*$ by stipulating

$$\sim^* (a, b, c) =_{def.} \equiv^* (a, b, c, 0).$$

We follow Ivánek's idea and construct a new quantifier from a quantifier of simultaneous existence via

$$\sim^* (a, b, c) =_{def.} \Leftrightarrow^* (a, b, c, 0) = \frac{a - (b + c)}{a + b + c} \geq p,$$

where $p \in (0, 1]$. We call it a *quantifier of co–existence.* The fuzzy logic interpretation of this quantifier is analogous to the interpretation of the quantifier of simultaneous existence:

The determining subset A is formed of cases that satisfy ϕ or ψ. If there are sufficiently more simultaneous occurences of ϕ and ψ than non–coincident occurences then a proposition connecting the attributes ϕ and ψ by quantifier of co–existence *is true.*

A value

$$X = \frac{a}{a+b+c}$$

should be high and a value

$$Y = \frac{b+c}{a+b+c}$$

should be low. Moreover, X and Y should be sufficiently different (not similar) i.e. a degree $(X \leftrightarrow Y)^*$ should be high enough or, equivalently,

$$\begin{aligned}
\left(\frac{a}{a+b+c} \leftrightarrow \frac{b+c}{a+b+c}\right)^* &= 1 - \left(1 - \left|\frac{a-(b+c)}{a+b+c}\right|\right) \\
&= \left|\frac{a-(b+c)}{a+b+c}\right| \\
&= \frac{a-(b+c)}{a+b+c} \geq p.
\end{aligned}$$

Notice also that if we join a basic double implicational quantifier

$$\sim (a,b,c) = \frac{a}{a+b+c}$$

to itself by Lukasiewicz conjunction, then we obtain a formula

$$[\phi \textit{ implies } \psi \textit{ and } \psi \textit{ implies } \phi] \textit{ and } [\phi \textit{ implies } \psi \textit{ and } \psi \textit{ implies } \phi].$$

Intuitively, Lukasiewicz conjunction diminishes truth value i.e. ϕ has higher truth value than ϕ *and* ϕ. Moreover,

$$\begin{aligned}
\frac{a}{a+b+c} \odot \frac{a}{a+b+c} &= \max\left\{0, \frac{a}{a+b+c} + \frac{a}{a+b+c} - 1\right\} \\
&= \max\left\{0, \frac{2a-a-b-c}{a+b+c}\right\} \\
&= \max\left\{0, \frac{a-(b+c)}{a+b+c}\right\} \\
&= \frac{a-(b+c)}{a+b+c} \geq p.
\end{aligned}$$

3 Conclusion and future work

We have deepened the connection between fuzzy logic and data mining quantifiers familiar in GUHA method framework introduced in [4]. Basic implicational quantifier, basic double implicational quantifier and basic equivalence quantifier, for example, have a natural interpretation in Lukasiewicz fuzzy logic. Conversely, the dissimilarity relation of Lukasiewicz fuzzy logic generates a (new) quantifier of simultaneous existence. Such a connection clears the road for future investigation into common features of fuzzy logic and data mining.

References

[1] Burian, J.: Datamining a AA (Above Average) kvantifikátor. Manuscrip (in Czech).

[2] Hájek, P.: Metamathematics of Fuzzy Logic, Kluwer (1998).

[3] Hájek, P. and Havránek, T.: Mechanising Hypothesis Formation. Mathematical Foundations for a General Theory. Springer-Verlag, Berlin (1978), 396 p.

[4] Ivánek, J.: Construction of implicational quantifiers from fuzzy implication. Fuzzy Sets and Systems **151** (2005) 381–391.

[5] Ivánek, J.: Combining Implicational Quantifiers for Equivalence Ones by Fuzzy Connectives. Int. J. Intelligent Systems **21** (2006) 325–334.

[6] Ivánek, J: On the Correspondence between Classes of Implicational and Equivalence Quantifier, in Zytkow, J.M. and Rauch, J. (Eds.) PKDD'99, LNAI 1704, pp. 116-124, 1999. Springer-Verlag (1999).

[7] Rauch, J and Simunek, M. Mining 4ft–rules, in Arikawa, S and Morishita, S. (Eds.) Discovery Science, Springer, Berlin, 2000, pp.268–272.

[8] Turunen, E.: Mathematics behind Fuzzy Logic. Springer-Verlag (1999).

Fuzzy Functional Dependencies in Mulitiargument Relationships

Krzysztof Myszkorowski

Technical University of Lodz, Institute of Computer Science
Wolczanska 215, 93-005 Lodz, Poland
email:kamysz@ics.p.lodz.pl

Abstract. The subject of this paper is analysis of functional dependencies in multiargument relationships $\boldsymbol{R}(X_1, X_2, ..., X_n)$. Apart from dependencies between all n attributes there may be also dependencies describing relationships of fewer attributes. However, there is no complete arbitrariness. The $(n$-1)-ary relationships embedded in n-ary relationship must not be contrary to it. The paper formulates the rules to which they must be subordinated. Included is also the possibility of an incomplete knowledge about a modeled fragment of reality, which entails fuzzy dependencies and relationships.

Keywords: Fuzzy databases, n-ary relationships, fuzzy functional dependencies, fuzzy normal forms

1 Introduction

Conventional database systems are designed with the assumption of precision of information collected in them. The problem becomes more complex if our knowledge of the fragment of reality to be modeled is imperfect [14]. In such cases one has to apply tools for describing uncertain or imprecise information [5, 13, 15]. One of them is the fuzzy set theory [6, 19]. So far, a great deal of effort has been devoted to the development of fuzzy data models [2, 7, 9, 17, 18]. Bordogna et al. [1] presented a fuzzy extension of a graph-based data model in which fuzzy and uncertain information occurs in conceptual scheme graphs. Ma et al. [12] extended certain major notions in object-oriented environment, investigated fuzzy inheritance hierarchies and proposed a generic fuzzy object-oriented model. Numerous works discuss how uncertainty existing in databases should be handled. Some authors proposed incorporating fuzzy logic into data modeling techniques. In [10, 11] an entity relationship model extended by a fuzzy logic was described. The main elements of the model, entities, relationships and attributes were presented in the context of fuzzy sets.

In database models, usually binary relationships occur between entity sets. When designing, it may be necessary to define n-ary relationships. Furthermore, within such connections there may exist relationships comprising fewer than n sets. However, there is no complete arbitrariness. The relationships “embedded” in n-ary relationships are subjected to certain restrictions. This issue for ternary relationships was presented in [4, 8]. For various types of ternary relationships the authors formulated the rules to which the binary relationships between pairs of sets are subjected. This analysis may be also carried out using the theory of functional dependencies which reflect integrity constraints and should be studied during the design process [18].

The subject of this paper is an analysis of multiargument relationships. Included is also the possibility of occurrence of fuzzy values. Section 2 discusses n-ary relationships in conventional databases. Section 3 contains the definition of a fuzzy functional dependency, extended Armstrong's rules and extended normal forms. These notions are used in section 4 in analysis of fuzzy multiargument relationships.

2 Multiargument Relationships

A multiargument relationship may be formally presented using the relational notation: $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$, where $\boldsymbol{R}$ is the name of the relationship, and attributes X_i denote keys of sets which participate in it. The n number is called the relationship degree. The relationship's cardinality may be presented as follows:

$$Q\,(X_1, X_2, \ldots, X_n) = M_1\colon M_2\colon \ldots\colon M_n\,, \tag{1}$$

where M_i denotes the number of X_i values that can occur for each value of $\{X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$.

For ternary relationships (n=3), analysis of four possible cases is necessary: 1:1:1, M:1:1, M:N:1, M:N:P. The article [8] formulates the rules determining the possibility of the occurrence of binary relationships within a ternary relationship. Pursuant to these, cardinalities of imposed binary relationships cannot be lower than the cardinality of the ternary relationship. Therefore, for the first case, the binary relationships of any cardinalities can be imposed. In the fourth case the binary relationships may be only of the many-to-many type.

The above analysis may be also carried out using the theory of functional dependencies. For respective cases the dependencies occurring in the ternary relationship $\boldsymbol{R}(X, Y, Z)$ are as follows:

1. $XY \rightarrow Z,\ XZ \rightarrow Y,\ YZ \rightarrow X$ – relationship 1:1:1
2. $XY \rightarrow Z,\ XZ \rightarrow Y$ – relationship M:1:1
3. $XY \rightarrow Z$ – relationship M:N:1
4. There are no functional dependencies between attributes – relationship $M{:}N{:}P$.

Let us notice that the first case comprises all possible dependencies between three attributes. In the other cases their quantity is consecutively decreased by 1. The presented dependencies describe the integrity constraints and must not be infringed. They constitute a restriction for binary relationships.

Example 1. Let us consider the possibility to impose the dependency $Z \rightarrow Y$ in each of the described cases.

1. Relationship 1:1:1

Based on Armstrong's rules we obtain: $Z \rightarrow Y \Rightarrow XZ \rightarrow Y$ (rules of augmentation and decomposition). The obtained dependency exists in the description of the relationship.

The integrity constraints have not been infringed. Therefore, dependency $Z \rightarrow Y$ is admissible. We can notice that imposing of this dependency implies also dependency $Z \rightarrow X$. For we have: $Z \rightarrow Y \wedge YZ \rightarrow X \Rightarrow Z \rightarrow X$ (pseudotransitivity rule).

2. Relationship M:1:1.

Similarly as in case 1 we obtain – by virtue of augmentation and decomposition rules - dependency $XZ \rightarrow Y$. Therefore, dependency $Z \rightarrow Y$ is admissible.

3. Relationship M:N:1

Dependency $XZ \rightarrow Y$, which is a consequence of the introduced dependency, does not occur in description of the relationship. Its occurrence means infringement of integrity constraints. Therefore, dependency $Z \rightarrow Y$ is inadmissible.

4. Relationship M:N:P

Basing on augmentation and decomposition rules we obtain dependency $XZ \rightarrow Y$. This constitutes an infringement of the integrity constraint of the lack of functional dependencies between attributes. Dependency $Z \rightarrow Y$ is inadmissible.

Table 1 presents admissible functional dependencies which may be imposed onto binary relationships existing between pairs of entity sets participating in the ternary relationship.

Table 1. Binary relationships within the ternary relationship

Cardinality of relationship	Admissible dependencies
1:1:1	Every dependency
M:1:1	$X \rightarrow Z, X \rightarrow Y, Y \rightarrow Z, Z \rightarrow Y$
M:N:1	$X \rightarrow Z,\ Y \rightarrow Z$
M:N:P	No dependencies

The above result can be generalized for n-ary relationships $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$. Let us denote by U the set of all attributes: $U = \{X_1, X_2, \ldots, X_n\}$. The functional dependencies describing n-ary relationships may be presented as:

$$U - \{X_i\} \rightarrow X_i \ , \qquad i = 1, 2, \ldots, n \ . \tag{2}$$

Their quantity is equal to the quantity of number one in relationship's cardinality. For the relationship having cardinalities 1:1: … :1 this number is n. If $M_i > 1$ for every i, the dependencies determined by formula (2) do not occur at all.

Functional dependencies describing the relationships between (n-1) attributes "embedded" in n-ary relationship may be presented as follows:

$$U - \{X_i, X_j\} \rightarrow X_i \ , \qquad \text{where} \quad i \neq j, \ i, j = 1, 2, \ldots, n \ . \tag{3}$$

There are, however, some restrictions. Let us denote by F the set of functional dependencies determined by formula (2). Attributes can be divided into two disjoint sets: $\mathcal{L}$ and $\mathcal{B}$, such that:

$\mathcal{L}$ – contains the attributes occurring only on the left side of dependencies (2).

$\mathcal{B}$ – contains the other attributes.

Cardinality $|F|$ of set F reaches n maximum. Then all attributes belong to $\mathcal{B}$. If F is an empty set, all attributes belong to $\mathcal{L}$. The values of attributes belonging to $\mathcal{L}$ determine the values of other attributes. Thus they must belong to every candidate key of relation scheme $\boldsymbol{R}$ [16]. Imposing the (n-1)-ary relationship described by dependency (3), in which $X_i \in \mathcal{L}$, would infringe the data integrity. Such a relationship is possible if attribute X_i does not belong at least to one candidate key.

Theorem 1. In n-ary $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ relationship with F set of functional dependencies between n attributes, determined by formula (2), there may exist a functional dependency $U - \{X_i, X_j\} \rightarrow X_i$, where $i \neq j$, $i, j = 1, 2, \ldots, n$, if attribute X_i does not belong to set $\mathcal{L}$.

Proof. The dependency concerned is admissible, if it does not cause any infringement of integrity constraints. This may occur when a consequence of its introduction will be the occurrence of a new functional dependency in a form determined by formula (2), which does not belong to F. The sets $\mathcal{L}$ and $\mathcal{B}$ would be also changed then. Let us check admissibility of the existence of dependency: $U - \{X_i, X_j\} \rightarrow X_i$. Basing on the rules of augmentation and decomposition, we obtain: $U - \{X_i, X_j\} \rightarrow X_i \Rightarrow U - \{X_i\} \rightarrow X_i$. If $X_i \in \mathcal{B}$, then $U - \{X_i, X_j\} \rightarrow X_i \in F$ and only in this case the considered dependency $U - \{X_i, X_j\} \rightarrow X_i$ may hold in $\boldsymbol{R}$. At $X_i \in \mathcal{L}$ it is inadmissible because $U - \{X_i\} \rightarrow X_i \notin F$.

3 Functional Dependencies in Fuzzy Databases

The functional dependency $X \rightarrow Y$ between attributes X and Y in conventional databases means that every value of attribute X corresponds exactly to one value of attribute Y. This corresponds to the assumption that the equality of attributes values may be evaluated formally, using a two-valued logic. In fuzzy databases it is possible to evaluate the degree of the closeness of compared values. Therefore, the notion of the functional dependency has to be modified [3]. The existence of such a dependency means that close values of attribute X correspond to close values of attribute Y.

Definition 1. Let $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ be a relation scheme and let X and Y be subsets of the set of attributes: $X, Y \subseteq U$, where $U = \{X_1, X_2, \ldots, X_n\}$. Y is functionally dependent on X in θ degree, $X \rightarrow_\theta Y$, if and only if for every relation R of $\boldsymbol{R}$ the following condition is met:

$$\min_{t,t' \in R} (I (t(X) =_c t'(X), t(Y) =_c t'(Y)) \geq \theta \ , \tag{4}$$

where $\theta \in [0,1]$, $=_c : [0,1] \times [0,1] \rightarrow [0,1]$ is a closeness measure and $I : [0,1] \times [0,1] \rightarrow [0,1]$ is a fuzzy implicator.

The above definition is very general. To be specified, it requires establishing a fuzzy implicator and a closeness measure. In further considerations the following implicator will be used:

$$I(a, b) = 1, \quad \text{for } a \le b, \qquad I(a, b) = b, \quad \text{for } a > b \quad . \tag{5}$$

Let us assume that attribute values are given by means of possibility distributions. The closeness measure of possibility distributions $\pi_1(x)$ and $\pi_2(x)$ is defined by the formula:

$$=_c (\pi_1, \pi_2) = \sup_x \min(\pi_1(x), \pi_2(x)) \quad . \tag{6}$$

Table 2. Relation *PES* – example 2

	P	E	S
t_1	P1	E1	low
t_2	P2	E2	9 000
t_3	P3	1/E3, 1/E4	8 000
t_4	1/P2, 1/P3	1/E2, 0.8/E3	high
t_5	P1	E2	3 000
t_6	P4	E1	1 000

Example 2. Relation *PES* (table 2) presents a relationship between the post held (*P*), education (*E*) and salary (*S*). The membership functions for fuzzy sets - *low* (μ_l) and *high* (μ_h) are presented in Figure 1.

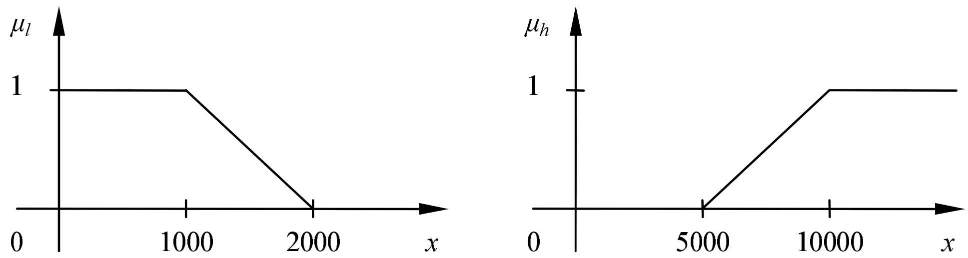


Fig. 1. Membership functions for fuzzy sets – *low* and *high*

Attributes of relation *PES* satisfy the dependency $PE \rightarrow_{0.6} S$. For we have:
1. for tuples t_2 and t_4: $=_c(t_2(PE), t_4(PE)) = 1$ and $=_c(t_2(S), t_4(S)) = 0.8$
2. for tuples t_3 and t_4: $=_c(t_3(PE), t_4(PE)) = 0.8$ and $=_c(t_3(S), t_4(S)) = 0.6$

Basing on formula (5) we obtain $I(1, 0.8) = 0.8$ and $I(0.8, 0.6) = 0.6$.

If the fuzzy implicator I satisfies the following conditions [3]:
C1: $a \le b \Rightarrow I(a, b) = 1$,
C2: $I(a, b) \ge \theta \Rightarrow I(a', b') \ge \theta$, where $a' = \min(a, c)$, $b' = \min(b, c)$,
C3: $I(a, b) \ge \alpha \wedge I(b, c) \ge \beta \Rightarrow I(a, c) \ge \gamma$, where $\gamma = \min(\alpha, \beta)$,
then the extended Armstrong's axioms for fuzzy functional dependencies are satisfied:
A1: $Y \subseteq X \Rightarrow X \to_\theta Y$ for every θ (reflexivity),
A2: $X \to_\theta Y \Rightarrow XZ \to_\theta YZ$ (augmentation),
A3: $X \to_\alpha Y \wedge Y \to_\beta Z \Rightarrow X \to_\gamma Z$, where $\gamma = \min(\alpha, \beta)$ (transitivity).

Conditions C1, C2 and C3 are satisfied for implicators which can be presented in the form:

$$I(a, b) = 1, \text{ for } a \le b, \qquad I(a, b) = f(a, b) \text{ for } a > b \ , \tag{7}$$

where $f(a, b)$ is a-nonincreasing and b-nondecreasing function.
The following rules result from Armstrong's axioms:
D1: $X \to_\alpha Y \wedge X \to_\beta Z \Rightarrow X \to_\gamma YZ$, where $\gamma = \min(\alpha, \beta)$ (union),
D2: $X \to_\alpha Y \wedge WY \to_\beta Z \Rightarrow XW \to_\gamma Z$, where $\gamma = \min(\alpha, \beta)$ (pseudotransitivity),
D3: $X \to_\alpha Y \wedge Z \subseteq Y \Rightarrow X \to_\alpha Z$ (decomposition),
D4: $X \to_\alpha Y \Rightarrow X \to_\beta Y$ for $\beta \le \alpha$.

Inclusion of fuzzy functional dependencies requires extension of the notion of relation key. Let G denote a set of fuzzy functional dependencies for scheme $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$. Its closure (the set of functional dependencies that are logically implied by G) shall be denoted by G^+. Subset K of the set of attributes U ($K \subseteq U$) is a θ-key of $\boldsymbol{R}$, if dependency $K \to_\theta U$ belongs to G^+ and there is no set $K' \subset K$, such that $K' \to_\theta U \in G^+$ (U is fully functionally dependent on K in θ degree). Attributes belonging to θ-key are called θ-prime-attributes. The other ones are called θ-non-prime-attributes. For scheme ***PES*** (example 2) subset PE is the 0.6-key.

Definitions of normal forms have also been extended. They applied the notions of θ-key and θ-prime-attribute.

Definition 2. Scheme $R(X_1, X_2, \ldots, X_n)$ is in the θ-fuzzy second normal form (θ-F2NF), if every θ-nonprime-attribute is fully functionally dependent on θ-key in α degree, where $\alpha > 0$.

Example 3. Scheme ***PES*** from the previous example is in 0.6-F2NF. Let us augment it by attribute A determining age and assume that its values are connected with values of attribute P. Let us assume that the relationship between the post and age is expressed by dependency $P \to_\varphi A$. In result of such modification the θ-key is PE with $\theta = \min(\varphi, 0.6)$. However, because of the introduced dependency the modified scheme is not in 0.6-F2NF. There is a non-prime-attribute dependent on a part of θ-key.

Definition 2 excludes the possibility of the occurrence of θ-key subset, which would determine a θ-nonprime-attribute.

Definition 3. Scheme $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ is in the θ-fuzzy third normal form (θ-F3NF), if for every functional dependency $X \to_\phi Y$, where $X \subseteq U$, $U = \{X_1, X_2, \ldots, X_n\}$ and $Y \notin X$, X contains the θ-key of $\boldsymbol{R}$ or Y is a θ-prime-attribute.

Definition of the θ-fuzzy third normal form eliminates the possible occurrence of transitive dependencies between relation attributes.

Example 4. Let us consider the relation *EES* with attributes *Em* – employee, *E* – education and *S* – salary. Let us assume that between its attributes there are dependencies: $Em \to_{0.8} E$ and $E \to_{0.9} S$. Basing on transitivity axiom they yield the dependency $Em \to_{0.8} S$. The key of *EES* is *Em*. It is the 0.8-key. Attributes *E* and *S* disturb the conditions of definition 3, because *E* is not a θ-key and *S* is not a θ-prime attribute. The θ-fuzzy third normal form is obtained in result of the decomposition into relations with schemes: $\boldsymbol{EE}(Em, E)$ – 0.8-F3NF and $\boldsymbol{ES}(E, S)$ – 0.9-F3NF. This decomposition maintains the dependencies.

Eliminating from definition 3 the possibility that attribute Y in dependency $X \to_\theta Y$ is θ-prime leads to a stronger definition. This is a definition of θ-fuzzy Boyce-Codd normal form.

Definition 4. Scheme $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ is in the θ-fuzzy Boyce-Codd normal form (θ-FBCNF), if for every functional dependency $X \to_\phi Y$, where $X \subseteq U$, $U = \{X_1, X_2, \ldots, X_n\}$ and $Y \notin X$, X contains a θ-key of $\boldsymbol{R}$.

4 Fuzzy Multiargument Relationships

Let us consider the ternary relationship $\boldsymbol{R}(X,Y,Z)$ with the following fuzzy functional dependencies:

$$XY \to_\alpha Z, \quad XZ \to_\beta Y, \quad YZ \to_\gamma X, \quad \text{where } \alpha > 0, \beta > 0, \gamma > 0 \quad . \tag{8}$$

Let us assume that $\alpha > \beta > \gamma$. Scheme $\boldsymbol{R}$ has three candidate keys at the $\theta \le \gamma$ level, two at the $\theta \le \beta$ level and one at the $\theta \le \alpha$ level. The sets $\mathcal{L}$ and $\mathcal{B}$ defined in section 2 are fuzzy sets:

$$\mathcal{L} = \{(1 - \gamma) / X, (1 - \beta) / Y, (1 - \alpha) / Z\} \ , \tag{9}$$

$$\mathcal{B} = \{\gamma / X, \beta / Y, \alpha / Z\} \ . \tag{10}$$

The possibility to impose fuzzy binary relationships is limited by values of α, β and γ. The values determining attributes membership in sets $\mathcal{L}$ and $\mathcal{B}$ cannot be changed due to the introduced new functional dependency. Its level cannot be arbitrary. Let us check the possible existence of the dependency $Y \to_\phi X$. The rules of augmentation and decomposition yield the dependency $YZ \to_\phi X$. Integrity constraints will not be disturbed if $\phi \le \gamma$, because then $YZ \rightarrow_\gamma X \Rightarrow YZ \rightarrow_\phi X$ (rule D4). Hence, the following theorem can be formulated:

Theorem 2. In the fuzzy ternary relationship $\boldsymbol{R}(X, Y, Z)$ with functional dependencies: $XY \rightarrow_\alpha Z$, $XZ \rightarrow_\beta Y$, $YZ \rightarrow_\gamma X$ there may exist binary relationships determined by dependencies of form $V \rightarrow_\phi W$ where $V, W \in \{X, Y, Z\}$, if $\phi \le \alpha$ for $W = Z$, $\phi \le \beta$ for $W = Y$ and $\phi \le \gamma$ for $W = X$.

Remark 1. Let us notice that the dependency $Y \rightarrow_\phi X$, where $\phi \le \alpha$, implies dependency $Y \rightarrow_\phi Z$. For we have: $Y \rightarrow_\phi X \wedge XY \rightarrow_\alpha Z \Rightarrow Y \rightarrow_\phi Z$.

Example 5. Let us return to relation *PES* from example 2. In view of the dependency $PE \rightarrow_{0.6} S$ the sets $\mathcal{L}$ and $\mathcal{B}$ for this relation are as follows: $\mathcal{L} = \{1/P, 1/E, 0.4/S\}$ and $\mathcal{B} = \{0.6/S\}$. Let us assume that between attributes P and S there should be dependency $P \rightarrow_\phi S$, where $\phi = 0.4$. This dependency is disturbed by values of attribute S in tuples t_1 and t_5. To impose it, it is sufficient to change $t_5(S)$ into a value lower or equal to 1600. The other values do not have to be modified. Dependency $PE \rightarrow_{0.6} S$ is not disturbed. Sets $\mathcal{L}$ and $\mathcal{B}$ are not changed. If, however, $\phi = 0.9$, the disturbance of dependency $P \rightarrow_\phi S$ is much higher. Analysed should be the values in the first five tuples. To impose the dependency $P \rightarrow_{0.9} S$, the following changes should be made: $t_2(S) \ge 9500$, $t_3(S) \ge 9500$ and $t_5(S) \le 1100$. The dependency $PE \rightarrow_{0.9} S$ is then forced, which is contradictory to the assumption. Changed are sets $\mathcal{L}$ and $\mathcal{B}$: $\mathcal{L} = \{1/P, 1/E, 0.1/S\}$ and $\mathcal{B} = \{0.9/S\}$. The same situation would occur when dependency $E \rightarrow_\phi S$ were imposed. Noteworthily, other functional dependencies cannot exist for any $\phi > 0$.

The obtained result can be generalized for fuzzy n-ary relationships. Let there exist in relationship $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ the following dependencies:

$$U - \{X_i\} \rightarrow_{\alpha_i} X_i , \quad \text{where } \alpha_i > 0,\ i = 1, 2, \ldots, n \ . \tag{11}$$

Scheme $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ has n θ-keys in the form $U - \{X_i\}$. From dependencies (11) the following sets $\mathcal{L}$ and $\mathcal{B}$ result:

$$\mathcal{L} = \{(1 - \alpha_1) / X_1, (1 - \alpha_2) / X_2, \ldots , (1 - \alpha_n) / X_n\}, \tag{12}$$

$$\mathcal{B} = \{ \alpha_1 / X_1, \alpha_2 / X_2, \ldots , \alpha_n / X_n\} \ . \tag{13}$$

Levels γ_i of functional dependencies:

$$U - \{X_i, X_j\} \rightarrow_{\gamma_i} X_i, \quad \text{where } i \ne j,\ i, j = 1, 2, \ldots, n \tag{14}$$

determining $(n$-1)-ary relationships cannot exceed relevant values of α_i because of the disturbance of integrity constraints determined by dependencies (11). A consequence of existence of any dependency (14) is one of the existing functional dependencies (11) with a changed level. For due to the augmentation and decomposition rules we have: $U - \{X_i, X_j\} \to_{\gamma_i} X_i \Rightarrow U - \{X_i\} \to_{\gamma_i} X_i$. This dependency is not contradictory to assumptions if $\gamma_i \leq \alpha_i$. This results from rule D4: if $\gamma_i \leq \alpha_i$, then $U - \{X_i\} \to_{\alpha_i} X_i \Rightarrow U - \{X_i\} \to_{\gamma_i} X_i$. Otherwise, the obtained dependency is inadmissible. Its consequence is a change of the degree of attributes membership to sets $\mathcal{L}$ and $\mathcal{B}$.

Theorem 3. In the fuzzy n-ary relationship $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ with functional dependencies in the form determined by formula (11), there may exist $(n\text{-}1)$-ary relationships determined by functional dependencies (14), in which $\gamma_i \leq \alpha_i$.

Remark 2. If the number of dependencies in set F is lower than n, there are attributes fully belonging to $\mathcal{L}$. They cannot occur on the right side of the functional dependency (14) of any positive level.

Scheme $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ with functional dependencies between n attributes in the form determined by formula (11) occurs in θ-FBCNF, where $\theta = \min_i (\alpha_i)$. After having introduced the functional dependencies (14) the conditions of definition 4 may be disturbed. Let us denote by m the number of functional dependencies (11) occurring in relationship $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ and assume that $m > 1$. These dependencies correspond to the following sets $\mathcal{L}$ and $\mathcal{B}$:

$$\mathcal{L} = \{ (1 - \alpha_1) / X_1, (1 - \alpha_2) / X_2, \ldots, (1 - \alpha_m) / X_m, 1/ X_{m+1}, 1/ X_{m+2}, \ldots, 1/ X_n\}, \quad (15)$$

$$\mathcal{B} = \{ \alpha_1 / X_1, \alpha_2 / X_2, \ldots, \alpha_m / X_m\} . \quad (16)$$

Attributes $X_{m+1}, X_{m+2}, \ldots, X_n$ fully belong to $\mathcal{L}$. Their number equals $n - m$.

Let us examine the consequences of imposing the functional dependency $U - \{X_i, X_j\} \to_{\gamma_i} X_i$, where $i \neq j$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$ and $\gamma_i \leq \alpha_i$. If $X_j \in \mathcal{B}$, i.e. $j \leq m$, from dependency $U - \{X_j\} \to_{\alpha_j} X_j$ the attribute X_i can be eliminated. Basing on rule D2 we obtain: $U - \{X_i, X_j\} \to_{\gamma_i} X_i \wedge U - \{X_j\} \to_{\alpha_j} X_j \Rightarrow U - \{X_i, X_j\} \to_{\lambda_{i,j}} X_j$, where $\lambda_{i,j} = \min (\gamma_i, \alpha_j)$. A new key arises, at the level not higher than $\theta' = \max_i (\alpha_i)$. This is formed by attributes occurring on the left side of the introduced dependency. The conditions of the definition of θ-FBCNF have not been disturbed. If in the introduced dependency $X_j \in \mathcal{L}$, i.e. $j > m$, no new key will be formed. Scheme $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ will not occur in θ-FBCNF. The left side of the dependency $U - \{X_i, X_j\} \to_{\gamma_i} X_i$ does not contain the key. However, it will remain in θ-F3NF, because attribute X_i is θ-prime.

With one functional dependency (11) ($m = 1$) there is only one candidate key. Having introduced a new dependency, scheme $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ will not occur in θ-F3NF, because the left side does not contain the θ-key, and attribute X_i is not θ-prime. Furthermore, a partial dependency of attribute X_i on the θ-key will arise, which means a disturbance in the conditions defining the θ-fuzzy second normal form (definition 2).

5 Conclusion

The paper analyses multiargument relationships, using functional dependencies. The starting point are dependencies existing between all attributes of relationship $\boldsymbol{R}(X_1, X_2, \ldots, X_n)$ in the following form: $U - \{X_i\} \rightarrow X_i$, where $U = \{X_1, X_2, \ldots, X_n\}$ and $i = 1, 2, \ldots, n$. They determine the integrity constraints. The subject of considerations was the possible occurrence – within the n-ary relationship – of functional dependencies describing the relationships between (n-1) attributes. They cannot disturb the integrity constraints. Such dependencies are admissible if their right side attributes do not belong to all candidate keys. In fuzzy databases the level of the considered dependency must not exceed the levels of relevant dependencies determined by formula (11). The dependencies between (n-1) attributes may disturb the normal form of scheme $\boldsymbol{R}$.

References

1. Bordogna G., Pasi G., Lucarella D.: A Fuzzy Object-Oriented Data Model for Managing Vague and Uncertain Information. International Journal of Intelligent Systems 14 (1999) 623-651
2. de Caluwe R., Vandenberghe R., van Gyseghem N., van Schooten A.: Integrating fuzziness in database models. In Studies in Fuzziness: Fuzziness in Database Management Systems. P. Bosc and J. Kacprzyk, (Eds.), Physica Verlag, Heidelberg (1995)
3. Chen G.: Fuzzy Logic in Data Modeling – semantics, constraints and database design. Kluwer Academic Publishers, Boston (1998)
4. Dullea J., Song I., Lamprou I.: An Analysis of Structural Validity in Entity-Relationship Modeling. Data & Knowledge Engineering 47 (2001) 167-205
5. Dey D., Sarkar S.: Generalized Normal Forms for Probabilistic Relational Data. IEEE Transactions on Knowledge and Data Engineering, 14 (3) (2002) 485-497
6. Fedrizzi M., Kacprzyk J.: A brief introduction to fuzzy sets. In Studies in Fuzziness: Fuzziness in Database Management Systems. P. Bosc and J. Kacprzyk, (Eds.), Physica Verlag, Heidelberg (1995)
7. van Gyseghem N., de Caluwe R.: Imprecision and uncertainty in UFO database model,.Journal of the American Society for Information Science 49(3) (1998) 236-252
8. Jones T., Song I.: Analysis of Binary/Ternary Relationships Cardinality Combinations in Entity-Relationship Modeling. Data & Knowledge Engineering 19 (1996) 39-64
9. Kerre E., Chen G.: An overview of fuzzy data models. In Studies in Fuzziness: Fuzziness in Database Management Systems. P. Bosc and J. Kacprzyk, (Eds.), Physica Verlag, Heidelberg (1995)
10. Kerre E., Chen G.: Fuzzy Data Modeling at a Conceptual Level: Extending ER/EER Concepts. In Studies in Fuzziness and Soft Computing: Knowledge Management in Fuzzy Databases. O. Pons, M. Vila, J. Kacprzyk, (Eds.), Physica Verlag, Heidelberg and New York (2000)
11. Ma Z.M., Zhang W.J., Ma W.Y., Chen G.Q: Conceptual Design of Fuzzy Object-Oriented Databases Using Extended Entity-Relationship Model. International Journal of Intelligent Systems 16 (2001) 697-711
12. Ma Z.M., Zhang W.J., Ma W.Y.: Extending object-oriented databases for fuzzy information modeling. Information Systems 29 (2004) 421-435

13. Motro A.: Accommodating Imprecision in Database Systems: Issues and Solutions. SIGMOD RECORD, 19(4) (1990) 69-74
14. Motro A.: Imprecision and uncertainty in database systems. In Studies in Fuzziness: Fuzziness in Database Management Systems. P. Bosc and J. Kacprzyk, (Eds.), Physica Verlag, Heidelberg (1995)
15. Petry F.: Fuzzy Databases: Principles and Applications. Kluwer Academic Publishers, (1996)
16. Saiedian H., Spencer K.: An Efficient Algorithm to Compute the Candidate Key of a Relational Database Schema. The Computer Journal, 39(2) (1996) 124 – 132
17. de Tre G., de Caluwe R., Van der Cruyssen B.: A Generalized Object-Oriented Database Model. In Studies in Fuzziness and Soft Computing: Recent Issues on Fuzzy Databases. Bordogna G. and Pasi G., (Eds.), Physica Verlag, Heidelberg and New York (2000)
18. Yazici A, George R.: Fuzzy Database Modeling. Physica-Verlag, Heidelberg and New York (1999)
19. Zadeh L. A.: Fuzzy sets. Information and Control 8 (1965) 338-353

Obtaining Compact and still Accurate Linguistic Fuzzy Rule-Based Systems by Using Multi-objective Genetic Algorithms*

R. Alcalá, J. Alcalá-Fdez, M. J. Gacto, and F. Herrera

University of Granada, Dept. Computer Science and A.I., E-18071 Granada, Spain,
{alcala, jalcala, herrera}@decsai.ugr.es, mjgacto@ugr.es

Abstract. In this work, we propose the use of Multi-Objective Genetic Algorithms to obtain Fuzzy Rule-Based Systems with a good trade-off between interpretability and accuracy. To do that, we present a new post-processing method that by considering selection of rules together with tuning of membership functions gets solutions only in the Pareto zone with the best trade-off, i.e., that containing solutions with the least number of possible rules but still presenting high accuracy. This method is based on the well-known SPEA2 algorithm, applying appropriate genetic operators and including some modifications to concentrate the search in the desired Pareto zone.

1 Introduction

One of the aims that focus the research in the Linguistic Fuzzy Modeling (LFM)[1] area in recent years is the trade-off between interpretability and accuracy. Of course, the ideal thing would be to satisfy both criteria to a high degree, but since they are contradictory issues generally it is not possible.

A widely-used approach to improve the accuracy of linguistic Fuzzy Rule-Based Systems is the *tuning* of Membership Functions (MFs) [1, 2], which refines a previous definition of the Data Base (DB) once the rule base has been obtained. Although tuning usually improves the system performance, sometimes a large number of rules is used to reach an acceptable degree of accuracy. In this case, some works [1, 2] consider the selection of rules together with the tuning of MFs by only considering accuracy criteria.

In this contribution, we focus on this problem by using Genetic Algorithms as a tool for evolving the MFs parameters and rule base size and by coding all of them (rules and parameters) in the same chromosome. Since the problem presents multi-objective nature we consider the use of Multi-Objective Genetic Algorithms (MOGAs)[3] to obtain a set of solutions with different degrees of accuracy and number of rules by using both characteristics as objectives.

Our main interest is to design an appropriate MOGA for this problem due to standard MOGAs can present some problems. Generally, MOGAs are based

* Supported by the Spanish Project TIN-2005-08386-C05-01.

on obtaining a set of non-dominated solutions. However, in this case, there are solutions that are not interesting although they are in the Pareto frontier. For example, non-dominated solutions with a small number of rules and high error are not interesting since they have not the desired trade-off between accuracy and interpretability. Furthermore, the existence of these kinds of solutions favors the selection of solutions with very different number of rules and accuracy to apply the crossover operator, which gives results with poor accuracy (the tuning parameters would be very different and the crossover would not have any sense except for exploring new combinations of rules).

In our proposal, we concentrate the search in the Pareto zone with still accurate solutions trying to obtain the least number of possible rules. To do that, we propose a modification of the well-known SPEA2 [11] algorithm that considering the rule selection together with the tuning of MFs concentrates the search in the Pareto zone having accurate solutions with the least number of possible rules. Besides, we have performed the same modification and experiments with NSGA-II [6], showing that this approach is not the most adequate for this problem.

In the next section, we present a study of the Pareto frontier for this problem. SPEA2 algorithm is introduced in Section 3 together with the modifications proposed and the genetic operators considered. Section 4 shows an experimental study of the proposed methods in a real-world problem. Finally, Section 5 gives some conclusions.

2 Pareto Frontier in the Problem of the Interpretability-Accuracy Trade-off

In this section, we present a study of what kind of solutions we could find in the optimal Pareto frontier when the system error and the number of rules (both considered as objectives) are optimized by tuning the MFs and selecting the most promising rules. In this way, we can obtain an approximation of the optimal Pareto in order to determine the desired Pareto zone.

Tuning of MFs usually needs of an initial model with large number of rules to get an appropriate level of accuracy. Generally, to obtain a good number of initial rules, methods ensuring covering levels higher than needed are used. In this way, we could obtain rules that being needed at first could be unnecessary once the tuning is applied or rules that could impede the tuning of the remaining ones. Thus, we can find the following types of rules: *Bad Rules* (erroneous or conflicting rules) that degrade the system performance; *Redundant or Irrelevant Rules* that do not significantly improve the system performance; *Complementary Rules* that complement to some others slightly improving the system performance; and *Important Rules* that should not be removed to obtain a reasonable system performance.

Taking into account the possible existence of these kinds of rules, different rule configurations and different tuning parameters, we can distinguish the following zones in the space of the objectives:

– Zone with Bad Rules, which contains solutions with bad rules. In this zone, the Pareto front does not exist since by removing these kinds of rules would improve the accuracy and these solutions would be dominated by others.
– Zone with Redundant or Irrelevant Rules, which is comprised of solutions without bad rules but still maintaining redundant or irrelevant rules. By deleting these kinds of rules the accuracy would be practically the same.
– Zone with Complementary Rules, comprised of solutions without any bad or redundant rule. By removing these rules the accuracy would be slightly decreased.
– Zone with Important Rules, which contains solutions only comprised of essential rules. By removing these kinds of rules the accuracy is really affected.

In Figure 1, we can find an approximation of the optimal Pareto in the problem of tuning and rule selection with the double objective of simplicity and accuracy. This figure shows the different zones in the space of the objectives together with the desired Pareto zone to find solutions with good trade-off. This zone corresponds with the zone of complementary rules, i.e., we would like to delete all the possible rules but without affecting the accuracy of the model finally obtained seriously.

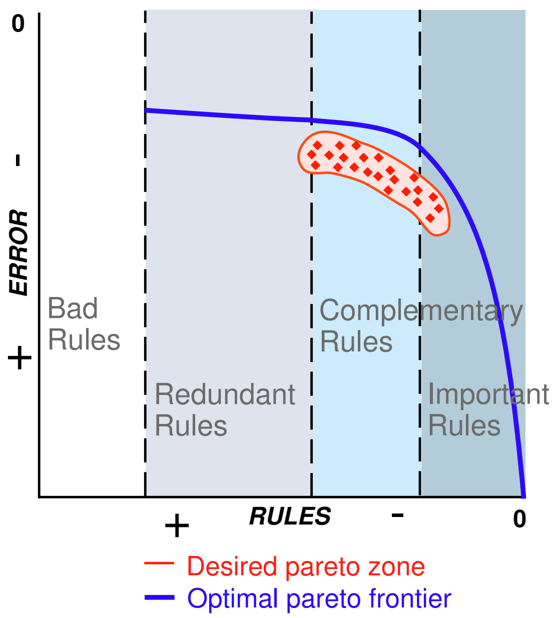


Fig. 1. Pareto Frontier

Taking into account what we previously exposed, the MOGA should not obtain all the Pareto front since it is difficult to obtain accurate solutions by favoring the crossing of solutions with very different rule configurations (those in the Pareto), which try to obtain the optimum by learning very different parameters for the membership functions. In the next section, we present a modification of SPEA2 [11] with the main aim of guiding the search towards the desired zone.

3 Accuracy-Oriented SPEA2 Algorithm

In this section, we firstly introduce the basis of SPEA2 [11]. Then we describe the coding scheme and the genetic operators used to implement the proposed algorithm. Finally, we explain the needed changes for guiding the search towards the desired zone.

3.1 SPEA2

The SPEA2 algorithm [11] (*Strength Pareto Evolutionary Algorithm for multi-objetive optimization*) is one of the most used techniques for solving problems with multi-objective nature. This algorithm differs from other MOGAs in several aspects although there are two with a especial interest:

- This uses a fine fitness assignment scheme which, for each individual, takes into account how many individuals it dominates and how many individuals dominates it.
- This includes a nearest neighbor density estimation technique which allows a more efficient way to guide the search.

According to the descriptions of the authors in [11], the SPEA2 algorithm consists of the next steps:

Input: N (population size),
$\overline{N}$ (external population size),
T (maximum number of generations).
Output: A (non-dominated set).

1. Generate an initial population P_0 and create the empty external population $\overline{P}_0 = \emptyset$.
2. Calculate fitness values of individuals in P_t and $\overline{P}_t$.
3. Copy all non-dominated individuals in $P_t \cup \overline{P}_t$ to $\overline{P}_{t+1}$. If $|\overline{P}_{t+1}| > \overline{N}$ apply truncation operator. If $|\overline{P}_{t+1}| < \overline{N}$ fill with dominated in $P_t \cup \overline{P}_t$.
4. If $t \geq T$, return A and stop.
5. Perform binary tournament selection with replacement on $\overline{P}_{t+1}$ in order to fill the mating pool.
6. Apply recombination and mutation operators to the mating pool and set P_{t+1} to the resulting population. Go to step 2 with $t = t + 1$.

3.2 Coding Scheme and Initial Gene Pool

A double coding scheme for both *rule selection* (C_S) and *tuning* (C_T) is used:

$$C^p = C_S^p C_T^p$$

- For the C_S part, the coding scheme consists of binary-coded strings with size m (with m being the number of initial rules). Depending on whether a rule is selected or not, values '1' or '0' are respectively assigned to the corresponding gene.

$$C_S^p = (c_{S1}, \ldots, c_{Sm}) \mid c_{Si} \in \{0, 1\} \ .$$

– For the C_T part, a real coding is considered, being m^i the number of labels of each of the n variables comprising the DB.

$$C_i = (a_1^i, b_1^i, c_1^i, \ldots, a_{m^i}^i, b_{m^i}^i, c_{m^i}^i), \ \ i = 1, \ldots, n \ ,$$
$$C_T^p = C_1 C_2 \ldots C_n \ .$$

The initial population is obtained in the following way:

– For the C_T part the initial DB is included as first individual. The remaining individuals are generated at random within the corresponding variation intervals. Such intervals are calculated from the initial DB. For each MF $C_i^j = (a^j, b^j, c^j)$, the variation intervals are calculated in the following way:

$$[I_{a^j}^l, I_{a^j}^r] = [a^j - (b^j - a^j)/2, a^j + (b^j - a^j)/2]$$
$$[I_{b^j}^l, I_{b^j}^r] = [b^j - (b^j - a^j)/2, b^j + (c^j - b^j)/2]$$
$$[I_{c^j}^l, I_{c^j}^r] = [c^j - (c^j - b^j)/2, c^j + (c^j - b^j)/2] \ .$$

– For the C_S part all genes take value '1' in all the individuals of the initial population.

3.3 Crossover and Mutation Operators

The crossover operator depends on the chromosome part where it is applied:

– In the C_T part, the BLX-0.5 [8] crossover is used.
– In the C_S part, the HUX [7] crossover is used.

Finally, four offspring are generated by combining the two from the C_S part with the two from the C_T part (the best two replace to their parent). The mutation operator changes a gene value at random in the C_S and C_T parts (one in each part) with probability P_m.

3.4 Modifications Applied on SPEA2

In order to focus the search on the desired Pareto zone, high accuracy with least possible number of rules, we propose two main changes with the aim of giving more selective pressure to those solutions that have a high accuracy. The proposed changes are described in the next:

– A restarting operator is applied exactly at the mid of the algorithm, by maintaining the most accurate individual as the sole individual in the external population ($\overline{P}_{t+1}$ with size 1) and obtaining the remaining individuals in the population (P_{t+1}) with the same rule configuration of the best individual and tuning parameters generated at random within the corresponding variation intervals. This operation is performed in step 4 then returning to step 2 with $t = t + 1$. In this way, we concentrate the search only in the desired pareto zone (similar solutions in a zone with high accuracy).

– In each stage of the algorithm (before and after restarting), the number of solutions in the external population ($\overline{P}_{t+1}$) considered to form the mating pool is progressively reduced, by focusing only on those with the best accuracy. To do that, the solutions are sorted from the best to the worst (considering accuracy as sorting criterion) and the number of solutions considered for selection is reduced progressively from 100% at the beginning to 50% at the end of each stage.

Besides, we have to highlight that the way to create the C_S part in the solutions of the initial population favors a progressive extraction of bad rules (those that do not improve with the tuning of parameters), only by means of the mutation at the beginning and then by means of the crossover.

Table 1. Methods Considered for Comparison

Ref.	Méthod	Description
[10]	**WM**	Wang & Mendel algorithm
[2]	**WM+T**	Tuning of Parameters
[2]	**WM+S**	Rule Selection
[2]	**WM+TS**	Tuning and Rule Selection
[11]	**SPEA2**	SPEA2 Algorithm
—	**SPEA2**$_{ACC}$	Accuracy-Oriented SPEA2
[6]	**NSGAII**	NSGA-II algorithm
—	**NSGAII**$_{ACC}$	Accuracy-Oriented NSGA-II

4 Experiments

To evaluate the usefulness of the method proposed, we have considered a real-world problem [5] with 4 input variables that consists of estimating the maintenance costs of medium voltage lines in a town. The methods considered for the experiments are briefly described in Table 1. WM [10] method is considered to obtain the initial rule base to be tuned. T and S methods perform the tuning of parameters and rule selection respectively. TS indicates tuning together with rule selection in the same algorithm. All of them consider the accuracy of the model as the sole objective. The remaining are MOGAs with and without the proposed modifications (all of them perform rule selection with tuning of parameters considering two objectives, accuracy and number of rules).

The linguistic partitions are comprised by *five linguistic terms* with triangular shape. The *center of gravity weighted by the matching* strategy acts as defuzzification operator and the fuzzy reasoning method is the *minimum t-norm* playing the role of implication and conjunctive operators. The values of the input parameters considered by the MOGAs studied are presented as follows: population size of 200, external population size of 61 (in the case of SPEA2 and SPEA2$_{ACC}$), 50000 evaluations and 0.2 as mutation probability per chromosome.

4.1 Problem Description

Estimating the maintenance costs of the medium voltage electrical network in a town [5] is a complex but interesting problem. Since a direct measure is very difficult to obtain, it is useful to consider models. These estimations allow electrical companies to justify their expenses. Moreover, the model must be able to explain how a specific value is computed for a certain town. Our objective will be to relate the *maintenance costs of the medium voltage lines* with the following four variables: *sum of the lengths of all streets in the town*, *total area of the town*, *area that is occupied by buildings*, and *energy supply to the town*. We will deal with estimations of minimum maintenance costs based on a model of the optimal electrical network for a town in a sample of 1,059 towns.

To develop the different experiments, we consider a *5-folder cross-validation model*, i.e., 5 random partitions of data each with 20%, and the combination of 4 of them (80%) as training and the remaining one as test.

4.2 Results

For each one of the 5 data partitions, the tuning methods have been run 6 times, showing for each problem the averaged results of a total of 30 runs. In the case of methods with multi-objective approach (the last four), the averaged values have been calculated considering the most accurate solution from each Pareto obtained. The proposed algorithm has been compared with several single objective based methods and with the widely-known NSGA-II [6] MOGA (with the same two objectives previously mentioned).

The results obtained by the analyzed methods are shown in table 2, where $\#R$ stands for the number of rules, MSE$_{tra}$ and MSE$_{tst}$ respectively for the averaged error obtained over the training and test data, σ for the standard deviation and *t-test* for the results of applying a *test t-student* (with 95 percent confidence) in order to ascertain whether differences in the performance of the proposed approach are significant when compared with that of the other algorithms in the table. The interpretation of this column is:

- $\star$ represents the best averaged result.
- $+$ means that the best result has better performance than that of the corresponding row.

Analysing the results showed in table 2 we can highlight the following facts:

- SPEA2$_{ACC}$ gets an important reduction of the mean square error respect with that obtained by the classic methods and NSGA-II. Furthermore, this algorithm improves the results obtained by SPEA2 with only 1.5 more rules.
- The models obtained by SPEA2$_{ACC}$ seem to show very good trade-off between interpretability and accuracy.

Table 2. Results obtained by the studied methods

Method	**#R**	**MSE$_{tra}$**	σ_{tra}	**t-test**	**MSE$_{tst}$**	σ_{tst}	**t-test**
WM	65	57605	2841	+	57934	4733	+
WM+T	65	18602	1211	+	22666	3386	+
WM+S	40.8	41086	1322	+	59942	4931	+
WM+TS	41.9	14987	391	+	18973	3772	+
SPEA2	**33**	13272	1265	+	17533	3226	+
SPEA2$_{ACC}$	34.5	**11081**	1186	*	**14161**	2191	*
NSGAII	41.0	14488	965	+	18419	3054	+
NSGAII$_{ACC}$	48.1	16321	1636	+	20423	3138	+

– NSGAII and NSGAII$_{ACC}$ present a not so good performance in this particular problem because of the crowding operator makes very difficult to concentrate the search in the desired Pareto zone.

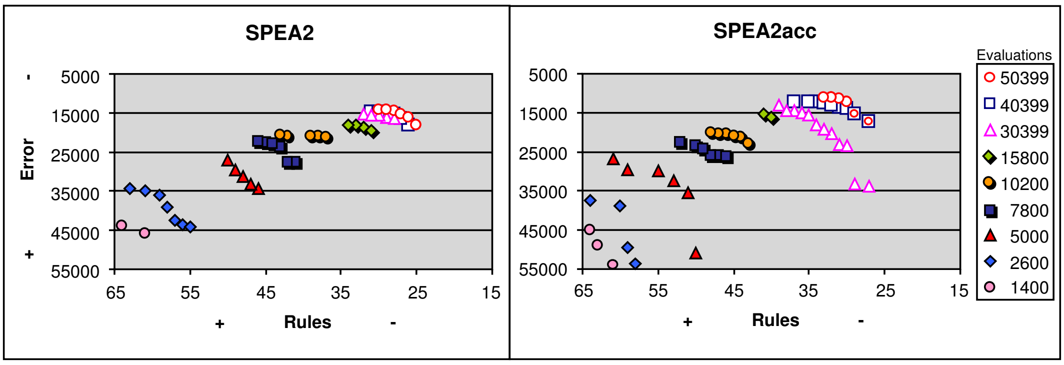


Fig. 2. Pareto fronts of $SPEA2$ and $SPEA2_{acc}$.

In figures 2 and 3, we can see the Pareto evolution for each algorithm. In figure 2, we can observe that SPEA2$_{ACC}$ mainly explores in the mid part of the evolution (before applying the restarting operator) in order to finally focusing on a specific zone of the Pareto. After restarting, the Pareto is extended in order to continue concentrating the search on the Pareto zone presenting solutions with less number of rules but still accurate.

In the remaining methods, figures 2 and 3, we can see as the Pareto moves along without having a big extension, which does not allow to obtain very good results even in the case of NSGA-II.

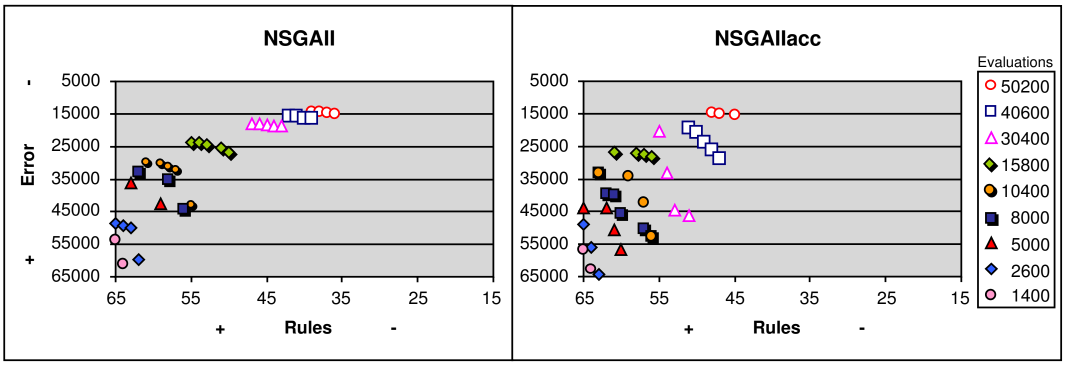


Fig. 3. Pareto fronts of $NSGAII$ and $NSGAII_{acc}$.

5 Concluding Remarks

Taking into account the showed in the previous section, we can conclude that the models obtained by the proposed method present a better trade-off between interpretability and accuracy. By searching for a good configuration of rules (only removing rules with little importance) and by tuning the parameter for a small set of rules, the proposed algorithm has obtained models even with a better accuracy than those obtained by methods only guided by measures of accuracy.

On the other hand, the proposed algorithm (SPEA2$_{ACC}$) could be of interest in problems that, although presenting a multi-objective nature, need as solution not all the Pareto frontier but only a specific area of it.

References

1. Casillas J., Cordón O., Herrera F., Magdalena L.(Eds): Accuracy improvements in linguistic fuzzy modeling. Springer-Verlag (2003)
2. Casillas J. , Cordón O., del Jesus M.J., Herrera F.: Genetic tuning of fuzzy rule deep structures preserving interpretability and its interaction with fuzzy rule set reduction. IEEE Trans. on Fuzzy Systems **13:1** (2005) 13–29
3. Coello Coello C.A., Van Veldhuizen D.A., Lamont G.B.: Evolutionary algorithms for solving multi-objective problems. Kluwer Academic Publishers (2002)
4. Cordón O., Herrera F.: A three-stage evolutionary process for learning descriptive and approximate fuzzy logic controller knowledge bases from examples. Int. J. of Approximate Reasoning **17:4** (1997) 369–407
5. Cordón O., Herrera F., Sánchez L.: Solving electrical distribution problems using hybrid evolutionary data analysis techniques. Applied Intelligence **10** (1999) 5–24
6. Deb K., Pratap A., Agarwal S., Meyarivan T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. on Evolutionary Computation **6:2** (2002) 182–197
7. Eshelman L.J.: The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination. Foundations of genetic Algorithms **1** (1991) 265–283
8. Eshelman L.J., Schaffer J.D.: Real-coded genetic algorithms and interval-schemata. Foundations of Genetic Algorithms **2** (1993) 187–202

9. Narukawa K., Nojima Y., Ishibuchi H.: Modification of Evolutionary Multiobjective Optimization Algorithms for Multiobjective Design of Fuzzy Rule-Based Classification Systems. IEEE Int. Conf. on Fuzzy Systems (2005) 809–814
10. Wang L.X., Mendel J.M.: Generating fuzzy rules by learning from examples. IEEE Trans. on Systems, Man, and Cybernetics **22:6** (1992) 1414–1427
11. Zitzler E., Laumanns M., Thiele L.: SPEA2: Improving the strength pareto evolutionary algorithm for multiObjetive optimization. Evolutionary Methods for Design, Optimization and Control with Applications to Industrial Problems (EUROGEN'01) (2001) 95–100
100–106

A First Study on the Use of Fuzzy Rule Based Classification Systems for Problems with Imbalanced Data Sets ⋆

María José del Jesus[1], Alberto Fernández[2], Salvador García[2], and Francisco Herrera[2]

[1] Dept. of Computer Science, University of Jaén, Spain,
mjjesus@ujaen.es
WWW home page: http://wwwdi.ujaen.es

[2] Dept. of Computer Science and A.I., University of Granada,
alfh@ugr.es,{salvagl,herrera}@decsai.ugr.es
WWW home page: http://sci2s.ugr.es

Abstract. In this work a preliminary study on the use of classification systems based on fuzzy reasoning in classification problems with non-balanced classes is carried out. The objective of this study is to evaluate the cooperation with pre-processing mechanisms of instances and the use of different granularity levels (5 and 7 labels) in the fuzzy partition considered. To do so, we will use simple fuzzy rule based models obtained with the Chi (and co-authors') method that extends the well-known Wang and Mendel method to classification problems.
The results obtained show that the previous step of instance selection and/or over sampling is needed. We have observed that a high over-fitting exists when we use 7 labels per variable. We will analyze this fact and we will discuss some proposals on the subject.

Key words: Fuzzy Rule Based Classification Systems, Instance Selection, Over-sampling, Imbalanced Data-sets.

1 Introduction

The design of a classification system, from the point of view of supervised learning, consists in the establishment of a decision rule that enables to determine the class of a new example in a set of known classes. When this knowledge extraction process uses as a representation tool fuzzy rules, the classification system obtained is called fuzzy rule-based classification system (FRBCS) [7].

In the classification problem field, we often encounter the presence of classes with a very different percentage of patterns between them: classes with a high pattern percentage and classes with a low pattern percentage. These problems receive the name of "classification problems with imbalanced data sets" and recently they are being studied in the machine learning field [5].

⋆ Supported by the Spanish Project TIN-2005-08386-C05-01 and TIC-2005- 08386-C05-03.

Learning systems can have difficulties in the learning of the concept related to the minority class, so in the specialized literature it is common to use pre-processing techniques to adjust the databases to a more balanced format [4].

Studying specialized literature, we have found only a few works [10,11,12] that study the use of fuzzy classifiers for this problem, and all of them from the point of view of approximate fuzzy systems, not from the descriptive fuzzy systems ones that are the ones used in this work.

In this work our aim is to analyze the behaviour of descriptive FRBCSs applied to data-bases with non-balanced classes. We want to evaluate the pre-processing mechanism of instances that are commonly used in the field in co-operation with the FRBCS, and to study the importance of the granularity of fuzzy partitions in these problems.

To do that, this paper is organized as follows. In Section 2 we introduce the components of an FRBCS and the inductive learning algorithm used. Section 3 presents the pre-processing techniques considered in this work . In Section 4 we introduce the way to evaluate the classification systems in domains with imbalanced data-sets. Section 5 shows the experimental study carried out with seven different data-sets. Finally, in Section 6 we present some conclusions about the study done.

2 Fuzzy rule based classification systems

An FRBCS is composed of a Knowledge Base (KB) and a Fuzzy Reasoning Method (FRM) that, using the information of the KB, it determines the class for any pattern of data admissible that comes to the system.

The power of the approximate reasoning consists in the possibility to obtain a result (a classification) even when we have not an exact compatibility (with degree 1) between the example and the antecedent of the rules.

2.1 Knowledge base

In the KB two different components are distinguised:

- The *Data Base* (DB), that contains the definition of the fuzzy sets associated to the linguistic terms used in the Rule Base.
- The *Rule Base* (RB), composed of a set of classification rules

$$R = \{R_1, ..., R_L\} \tag{1}$$

There are different types of fuzzy rules in the specialized literature but in our case we will use the following one:

- Fuzzy rules with a class and a certainty degree associated to the classification for this class in the consequent

$$\begin{array}{ll} R_k: & \text{If } X_1 \text{ is } A_1^k \text{ and } \ldots \text{ and } X_N \text{ is } A_N^k \\ & \text{then } Y \text{ is } C_j \text{ with degree } r_k \end{array} \tag{2}$$

where $X_1, \ldots, X_N$ are features considered in the problem, $A_1^k, \ldots, A_N^k$ are linguistic labels employed to represent the values of the variables and r_k is the certainty degree associated to the classification of the class C_j for the examples that belong to the fuzzy subspace delimited by the antecedent of the rule.

2.2 Fuzzy reasoning method

The FRM is an inference procedure that uses the information of the KB to predict a class from an unclassified example. Usually, in the specialized literature [8] the FRM of the maximum has been used, also named classic FRM or the winning rule, that considers the class indicated by only one rule having account the association degree of the consequent of the rule over the example. Other FRMs combine the information contributed for all the rules that represent the knowledge of the area of which the example belongs [8]. In this work we will use, besides the classic FRM, the FRM of additive combination among rules classification degree per class.

Next we present the general model of fuzzy reasoning that combines the information given by the fuzzy rules compatibles with the example.

In the classification process of the example $e = (e_1, \ldots, e_N)$, the steps of the general model of a FRM are the following:

1. Computing the compatibility degree of the example with the antecedent of the rules.
2. Computing the association degree of the example to the consequent class of each rules by means of an aggregation function between the compatibility degree and the certainty degree of the rule with the class associated.
3. Setting the association degree of the example with the different classes.
4. Classification. Applying a decision function F over the association degree of the example with the classes which will determine, on base to the criterion of the maximum, the label of the class v with the greatest value.

At point (3) we distinguish the two methods used in this study, that is, using the function of the maximum to select the rule with the greatest association degree for each class, and using the additive function over the association degrees of the rules associated with each class.

2.3 Chi et al. Algorithm

For our experimentation we will use simple rule base models obtained with the method proposed in [7] that extends the well-known Wang and Mendel method [13] to classification problems. This FRBCS desing method establishes the relationship between the variables of the problem and sets an association between the space of the features and the space of the classes by means of the following steps:

1. *Establishment of the linguistic partitions.* Once determined the domain of variation of each feature X_i, the fuzzy partitions are computed.
2. *Generation of a fuzzy rule for each example* $e^h = (e_1^h, \ldots, e_N^h, C_h)$. To do this is necessary:
   2.1 To compute the matching degree of the example e^h to the different fuzzy regions.
   2.2 To assign the example e^h to the fuzzy region with the greatest membership degree.
   2.3 To generate a rule for the example, which antecedent is determined by the selected fuzzy region and with the label of class of the example in the consequent.
   2.4 To compute the certainty degree. In order to do that the ratio S_j/S is determined, where S_j is the sum of the matching degree for the class C_j patterns belonging to this fuzzy region delimited by the antecedent, and S the sum of the matching degrees for all the patterns belonging to this fuzzy subspace, regardless its associated class.

3 Preprocessing imbalanced datasets.

In this work we evaluate different instance selection and oversampling techniques to adjust the class distribution in training data. We have chosen the following ones[4]:

- Undersampling methods:
  - **Condensed Nearest Neighbor Rule (CNN)**. This technique is used to find a consistent subset of examples. A subset $\subseteq$ E is consistent with E if using a 1-nearest neighbor, correctly classifies the examples in E.
  - **Tomek links** This method works as follows: given two examples e_i and e_j belonging to different classes, the distance between e_i y e_j (d(e_i, e_j)) is determined. A (e_i,e_j) pair is called a Tomek link if there is not an example e_l, such that d(e_i,e_l) ¡ d(e_i,e_j) or d(e_j ,e_l) ¡ d(e_i,e_j). If two examples form a Tomek link, then either one of these examples is noise or both examples are borderline.
  - **One-sided selection (OSS)** is an under-sampling method resulting from the application of Tomek links followed by the application of CNN. Tomek links are used as an under-sampling method and removes noisy and borderline majority class examples. CNN aims to remove examples from the majority class that are distant from the decision border.
  - **CNN + Tomek links** It is similar to the one-sided selection, but the method to find the consistent subset is applied before the Tomek links.
  - **Neighborhood Cleaning Rule (NCL)** uses the Wilson‘s Edited Nearest Neighbor Rule (ENN) [15] to remove majority class examples. ENN removes any example whose class label differs from the class of at least two of its three nearest neighbors. NCL modifies the ENN in order to increase the data cleaning.

    - **Random under-sampling** is a non-heuristic method that aims to balance class distribution through the random elimination of majority class examples.
- Oversampling methods:
    - **Random over-sampling** is a non-heuristic method that aims to balance class distribution through the random replication of minority class examples.
    - **Smote Synthetic Minority Over-sampling Technique (Smote)**[6] is an over-sampling method which form new minority class examples by interpolating between several minority class examples that lie together. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.
- Hybrid methods: Oversampling + Undersampling:
    - **Smote + Tomek links**. In order to create better-defined class clusters, it could be applied Tomek links to the over-sampled training set as a data cleaning method. Thus, instead of removing only the majority class examples that form Tomek links, examples from both classes are removed.
    - **Smote + ENN**. The motivation behind this method is similar to Smote + Tomek links. ENN tends to remove more examples than the Tomek links does, so it is expected that it will provide a more in depth data cleaning.

4 Evaluation of FRBCS for imbalanced data sets

In this section we introduce our experimentation framework. First of all we present the metric we will use to compare the different methods considered. Then we will describe the data sets we have chosen for this work and all the parameters used.

4.1 Measuring error: geometric mean on positive and negative examples

Weiss and Hirsh [14] show that the error rate of the classification of the rules of the minority class is 2 or 3 time greater than the rules that identify the examples of the majority class and that the examples of the minority class are less probable to be predict than the examples of the majority one.

The most straightforward way to evaluate the performance of classifiers is based on the confusion matrix analysis. ¿From a confusion matrix for a two class problem it is possible to extract a number of widely used metrics for measuring the performance of learning systems, such as Error Rate, defined as $Err = \frac{FP+FN}{TP+FN+FP+TN}$ and Accuracy, defined as $Acc = \frac{TP+TN}{TP+FN+FP+TN} = 1 - Err$.

Instead of using the error rate (or accuracy), in the ambit of imbalanced problems more correct metrics are considered. Specifically, it is possible to derive four performance metrics that directly measure the classification performance on positive and negative classes independently:

False negative rate $FN_{rate} = \frac{FN}{TP+FN}$ is the percentage of positive cases misclassified as belonging to the negative class;
False positive rate $FP_{rate} = \frac{FP}{FP+TN}$ is the percentage of negative cases misclassified as belonging to the positive class;.
True negative rate $TN_{rate} = \frac{TN}{FP+TN}$ is the percentage of negative cases correctly classified as belonging to the negative class;
True positive rate $TP_{rate} = \frac{TP}{TP+FN}$ is the percentage of positive cases correctly classified as belonging to the positive class.

These four performance measures have the advantage of being independent of class costs and prior probabilities. The aim of a classifier is to minimize the false positive and negative rates or, similarly, to maximize the true negative and positive rates.

The metric used in this work is the geometric mean [3], which can be defined as $g = \sqrt{a^+ \cdot a^-}$, where a^+ means the accuracy in the positive examples (TP_{rate}) and a^- is the accuracy in the negative examples (TN_{rate}). This metric tries to maximize the accuracy of each one of the two classes with a good balance. It is a performance metric that links both objectives.

4.2 Data sets and parameters

In this study we have considered seven data sets from UCI which have different degrees of imbalance. Table 1 summarizes the data employed in this study and shows, for each data set the number of examples (#Examples), number of attributes (#Attributes), class name of each class (majority and minority) and class attribute distribution. All attributes are qualitative.

Table 1. Data sets summary descriptions.

Data set	**#Examples**	**#Attributes**	**Class (min., maj.)**	**%Class(min.,maj.)**
Glass	214	9	(Ve-win-float-proc, remainder)	(7'94,92'06)
Pima	768	8	(1,0)	(34'77,66'23)
Yeast	1486	8	(mit,remainder)	(16'49,83'51)
Ecoli	336	7	(iMU, remainder)	(10'42,89'58)
Haberman	306	3	(Die, Survive)	(26'47,73'53)
New-thyroid	215	5	(hypo,remainder)	(16'28,83'72)
Vehicle	846	18	(van,remainder)	(23'52,76'48)

In order to realize a comparative study, we use a ten folder cross validation approach We consider the following parameters and functions:

- Number of labels per fuzzy partition: 5 and 7 labels.
- Computation of the compatibility degree: Min t-norm.
- Combination of the compatibility degree and the certain rule degree: Min t-norm.

– Inference method: Classic method (winning rule) and additive combination among rules classification degree per class (addition) [8].

In table 2 we show the percentages of examples for each class after balancing.

Table 2. Average of class percentage after balancing.

Balance Method	% Positives (minority class)	% Negatives (majority class)
CNN_TomekLinks	63.23	36.77
CNNRb	81.29	18.71
NCL	25.52	74.48
OSS	34.56	65.44
RandomOS	50.00	50.00
RandomUS	50.00	50.00
SMOTE	50.00	50.00
SMOTE_ENN	52.85	47.15
SMOTE_TomekLinks	54.35	45.65
TomekLinks	23.84	76.16

5 Analysis of experiments

We have divided our study into three parts: the analysis of the use of preprocessing for imbalanced problems, the study of the effect of the FRM and finally the analysis of the influence of the granularity applied to the linguistic partitions together with the inference method.

Tables 3 and 4 show the global results (in training and test sets) for all the data-sets used in the experimental study, showing the behaviour of the FRBCSs. Each column represents the following:

– the FRM used (WR for the Winning Rule and AC for Additive Combination) and the number of labels employed (5-7),
– the balancing method employed, where "none" means that the original data set is maintained for training,
– the accuracy per class (a^- y a^+) where the subindex indicates if it refers to training (tr) or test (tst). It also shows the geometric mean (GM) for training (TR) and test (TST).

1. **The effect of the preprocessing methods**: Our results show that in all the cases pre-processing is a necessity to improve the behaviour of the learning algorithms.
   Specifically it is noticed that the over-sampling methods provide very good results in practice. We found a kind of mechanism (the SMOTE pre-process family) that are very good as pre-process technique, both individually and

Table 3. Global Results WMWR.

Classifier	Balancing Method	a^-_{tr}	a^+_{tr}	GM_{TR}	a^-_{tst}	a^+_{tst}	GM_{TST}
FRBCS-WR5	CNN_TomekLinks	23.86	*98.59*	45.04	22.49	**91.14**	40.9
FRBCS-WR5	CNNRb	70.15	73.84	68.64	65.84	63.41	60.01
FRBCS-WR5	NCL	90.87	67.23	74.54	87.26	56.13	64.26
FRBCS-WR5	None	*98.68*	52.74	68.61	*94.51*	39.78	55.01
FRBCS-WR5	OSS	86.28	62.01	71.46	83.82	52.45	63.54
FRBCS-WR5	RandomOS	82.33	88.31	*84.77*	76.9	72.88	74.46
FRBCS-WR5	RandomUS	72.28	87.53	78.11	68.06	77.59	70.61
FRBCS-WR5	SMOTE	81.19	88.32	84.13	75.91	74.86	**75.11**
FRBCS-WR5	SMOTE_ENN	74.41	90.7	81.56	70.01	80.06	74.29
FRBCS-WR5	SMOTE_TomekLinks	71.94	94.22	81.69	67.94	83.51	74.8
FRBCS-WR5	TomekLinks	93.88	63.62	73.79	90.2	51.29	62.35
FRBCS-WR7	CNN_TomekLinks	30.21	**99.1**	52.31	26.85	*80.1*	43.81
FRBCS-WR7	CNNRb	65.04	80.25	70.08	58.17	53.87	51.77
FRBCS-WR7	NCL	89.13	80.81	83.82	79.02	55.34	60.89
FRBCS-WR7	None	**99.02**	66.8	79.22	*87.13*	42.9	55.68
FRBCS-WR7	OSS	74.83	65.48	69.69	68.91	46.38	55.11
FRBCS-WR7	RandomOS	89.54	91.19	**90.23**	76.54	63.36	69.33
FRBCS-WR7	RandomUS	67.23	92.14	77.38	59.51	69.5	63.08
FRBCS-WR7	SMOTE	86.7	92.19	89.23	74.04	66.64	69.95
FRBCS-WR7	SMOTE_ENN	80.68	92.02	85.95	70.46	70.3	70.04
FRBCS-WR7	SMOTE_TomekLinks	78.94	94.99	86.35	68.7	73.47	*70.87*
FRBCS-WR7	TomekLinks	93.16	75.46	82.46	83.17	50.88	59.73

the hybrid ones. In this way, for FRBCSs we have highly competitive models. Nevertheless, this over-sampling can introduce an additional computation cost if the dataset is relatively large.
Also we may stress that the results in the case of no preprocess method is employed are very high for the negative class (majority) but quite low for the positive one (minority); hence the clear necessity of the preprocess methods.
2. **The reasoning method**: Analyzing the tables we find that there are no great differences between the type of FRM.
3. **Granularity analysis**: It is empirically shown that a big number of labels produces over-fitting, the training results are significantly better than the test ones when 7 labels per variable are used. This situation is evident in table 5. Besides, we must note that we are using relatively small databases and with few attributes, which stresses more this undesirable behaviour.

6 Concluding remarks.

In this work we analyze the behaviour of the FRBCSs applied to classification problems with imbalanced data sets and the cooperation with pre-processing methods of instances.

Table 4. Global Results FRBCS-AC.

Classifier	**Balancing Method**	a^-_{tr}	a^+_{tr}	GM_{TR}	a^-_{tst}	a^+_{tst}	GM_{TST}
FRBCS-AC5	CNN_TomekLinks	25.81	*96.88*	44.7	24.9	*89.71*	41.26
FRBCS-AC5	CNNRb	69.47	71.54	66.12	66.04	62.05	59.15
FRBCS-AC5	NCL	90.85	63.55	72.02	87.09	54.41	62.86
FRBCS-AC5	None	*98.42*	46.18	63.7	**94.65**	36.04	52.2
FRBCS-AC5	OSS	86.74	57.6	68.52	84.98	50.74	62.47
FRBCS-AC5	RandomOS	90.74	73.81	81.03	86.23	61.85	71.39
FRBCS-AC5	RandomUS	70.79	88.23	77.49	67.17	81.36	71.83
FRBCS-AC5	SMOTE	87.35	78.84	82.34	83.08	66.62	72.57
FRBCS-AC5	SMOTE_ENN	80.28	85.84	*82.55*	76.71	73.47	*74.24*
FRBCS-AC5	SMOTE_TomekLinks	77.33	88.56	81.9	73.53	75.28	72.66
FRBCS-AC5	TomekLinks	93.99	58.55	70.48	90.92	49.03	60.94
FRBCS-AC7	CNN_TomekLinks	29.15	*98.04*	50.63	26.58	*80.03*	43.12
FRBCS-AC7	CNNRb	64.77	77.46	67.73	58.76	55.38	50.37
FRBCS-AC7	NCL	89.54	77.48	81.72	79.45	54.16	60.34
FRBCS-AC7	None	*98.82*	62.14	75.74	*87.38*	40.91	54.36
FRBCS-AC7	OSS	75.92	62.24	68.17	70.17	43.17	50.91
FRBCS-AC7	RandomOS	94.06	78.71	85.3	81.54	53.9	63.62
FRBCS-AC7	RandomUS	67.33	91.2	77.28	60.46	69.28	63.79
FRBCS-AC7	SMOTE	90.66	84.94	*87.5*	78.79	58.31	65.2
FRBCS-AC7	SMOTE_ENN	84.38	87.81	85.91	74.94	63.49	*68.24*
FRBCS-AC7	SMOTE_TomekLinks	82.3	91.13	86.23	72.71	65.57	67.62
FRBCS-AC7	TomekLinks	93.06	72.83	80.42	83.7	50.34	59.53

Table 5. FRBCS with 5 labels opposite 7 labels.

FRM	**Balancing Method**	GM_{TR} 5	GM_{TR} 7	GM_{TST} 5	GM_{TST} 7
Winning Rule	RandomOS	84.77	90.23	74.46	69.33
Winning Rule	SMOTE	84.13	89.23	75.11	69.95
Winning Rule	SMOTE_TL	81.69	86.35	74.8	70.87
Additive Comb.	SMOTE	82.34	87.5	72.57	65.2
Additive Comb.	SMOTE_ENN	82.55	85.91	74.24	68.24
Additive Comb.	SMOTE_TL	81.9	86.23	72.66	67.62

The main conclusions of our analysis are: the necessity of using pre-processing instances methods to improve the balance between classes before the use of the FRBCS method, the similar behaviour of the two fuzzy reasoning methods analyzed, and the over-fitting produced when we use a high number of labels per variable.

We must point out that FRBCSs with 5 labels do not reach high classification percentages in training. It seems that classes with very few examples may need labels with a low support that enables to obtain the information associated to the class, but without including examples from the other class. It seems interesting to post-process the rule base by means of tuning methods and/or the integration of labels in a different granularity level to gather all the possible information.

Following this idea, our future work will deal with this problem. We want to use a post-processing 2-tuples and 3-tuples tuning, two methods that have shown a good behaviour adjusting the support of the membership functions for regression problems [1,2].

References

1. R. Alcalá, J. Alcalá-Fdez. and F. Herrera. A proposal for the Genetic Lateral Tuning of Linguistic Fuzzy Systems and its Interaction with Rule Selection. *IEEE Transactions on Fuzzy Systems*, 2006, to appears.
2. R. Alcalá, J. Alcalá-Fdez. M.J. Gacto and F. Herrera. Rule Base Reduction and Genetic Tuning of Fuzzy Systems based on the Linguistic 3-Tuples Representation. *Soft Computing*, 2006, to appears.
3. R. Barandela, J.S. Sánchez, V. García, E. Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36:3 849–851, 2003.
4. G.Batista, R.C.Prati, M.C.Monard. A study of the behaviour of several methods for balancing machine learning training data. *SIGKDD Explorations* 6:1 20–29, 2004.
5. N.V. Chawla, N. Japkowicz, A. Kolcz. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations*, 6:1 1–6, 2004.
6. N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *JAIR* 16, 321–357, 2002.
7. Z. Chi, H. Yan, and T. Pham. Fuzzy algorithms with applications to image processing and pattern recognition. *World Scientific*, 1996.
8. O. Cordón, M.J. del Jesus, and F. Herrera. A proposal on reasoning methods in fuzzy rule-based classification systems. *International Journal of Approximate Reasoning*, 20:1 21–45, 1999.
9. J. Laurikkala. Improving Identification of Difficult Small Classes by Balancing Class Distribution. *Tech. Rep. A-2001-2*, University of Tampere, 2001.
10. S. Visa, and A. Ralescu. Learning Imbalanced and Overlapping Clases using Fuzzy Sets *Workshop on Learning from Imbalanced Datasets II*, Washington DC., 2003
11. S. Visa, and A. Ralescu. Fuzzy Classifiers for Imbalanced, Complex Classes of Varying Size. *IPMU*, Perugia (Italy), 393–400, 2004.
12. S. Visa, and A. Ralescu. The Effect of Imbalanced Data Class Distribution on Fuzzy Classifiers - Experimental Study. *IEEE International Conference on Fuzzy Systems*, 749–754, 2005
13. L.X. Wang and J.M. Mendel. Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man, and Cybernetics*, 25:2 353–361, 1992.
14. G. M. Weiss, and H. Hirsh. A quantitative study of small disjuncts. *Seventeenth National Conference on Artificial Inteligence*, AAAI Press 665–670, 2000.
15. D.L. Wilson. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Communications* 2:3 408–421, 1972.

Classification via Fuzzy Preference Learning

Eyke Hüllermeier

Fakultät für Informatik
Otto-von-Guericke-Universität Magdeburg, Germany
eyke.huellermeier@iti.cs.uni-magdeburg.de

We introduce a new approach to classification which combines pairwise decomposition techniques with ideas and tools from fuzzy preference modeling. More specifically, our approach first decomposes a polychotomous classification problem involving m classes $\{c_1 \dots c_m\}$ into an ensemble of binary problems, one for each pair of classes (c_i, c_j). The corresponding classifiers $\mathcal{M}_{ij}$ are trained on the relevant subsets of the original training data. Given a new instance x to be classified, this instance is submitted to every binary learner, the output of which is assumed to be a score $p_{ij} \in [0, 1]$. The latter is interpreted as a fuzzy degree of preference for class c_i in comparison with class c_j. By combining the outputs of all classifiers, one thus obtains a fuzzy preference matrix which is taken as a point of departure for the final classification decision. In other words, the problem of classification has now been reduced to a problem of decision making based on a fuzzy preference relation and, hence, can take advantage of techniques that have been developed in the latter field. It will be shown that this approach allows for quantifying different types of uncertainty in classification and is particularly interesting in the context of reliable classification.

Next Generation Data Analysis for Business Applications

Extended Abstract

Detlef Nauck, Mirko Böttcher, and Martin Spott

BT Group, Chief Technology Office,
Intelligent Systems Research Centre,
Ipswich IP5 3RE, United Kingdom
detlef.nauck@bt.com

Abstract. Businesses collect and keep large volumes of customer and process data as part of their processes. Analysis of this data by business users often leads to discovery of valuable patterns and trends that otherwise would go unnoticed and that can lead to prioritization of decisions on future investments. The majority of tools currently available to business users are typically limited to computing summary statistics, simple visualization and reporting of data. More complex tools that could offer possible explanations for observations, discover knowledge, or allow making predictions are usually aimed at an academic audience or at users who are highly trained in analytics. However, it is business users with little experience in analytics who require access to tools that allow them to easily model customer behavior and build future scenarios. In this paper we look at a selection of next generation data analysis systems we have developed to support business users in performing advanced analytics.

1 Introduction

Typically, data analysis software is called "intelligent" if it uses advanced analytics derived from artificial intelligence, computational intelligence or machine learning. However, from a user perspective software is only intelligent if it can work – at least to some extent – on behalf of the user, solve problems automatically and present relevant results in an intuitive and comprehensible manner. That means the *intelligence* level of a software platform is not only based on the cleverness of some algorithms or the complexity of the knowledge that it can represent, but largely also on its usability for a particular user domain. In the business intelligence domain we encounter mainly business users who are experts in their domain, but have typically very little understanding of data analysis issues. We need to provide intelligent automated tools to these users in order to empower them and to reduce their dependence on analysts who are rare and expensive. Users should not be bothered with model parameters and model building beyond the selection of the data that they want to use.

2 Automatic Data Exploration and Pattern Discovery

Nowadays, enterprises are confronted with a rapidly changing business environment. As markets, innovations and customers are changing faster than ever before the key to survival for businesses is the ability to detect, assess and respond to changing conditions rapidly and intelligently. Discovering changes and reacting to or acting upon them before others do has therefore become a strategical issue for many companies.

Many businesses collect huge volumes of data over long periods. This data reflects changes in processes, markets or customer behaviour and it is crucial for businesses to detect such changes early and precisely. Existing data analysis techniques are insufficient for this task because they assume the domain under consideration is stable over time and that the user knows what to look for and which analysis approach he should use. The widely used method of defining key performance indicators (KPI) is too weak to detect changes early enough or to identify new business drivers that suddenly have become important.

We have developed a framework for intelligent data exploration (IDEAL) that detects changes within a data set at virtually any level of granularity [1]. IDEAL does association rule mining on a data set at different points in time and subsequently analyses how detected rules have changed over time. Based on four different trend measures and partial contradictions between rules IDEAL assesses how interesting a detected change or stability potentially is. It then uses a fuzzy rule base to combine the different measures into a single interestingness score for each rule and presents a list of association rules ordered by their scores.

IDEAL uses methods for pruning temporally redundant association rules [2], and can learn from user feedback to discard rules over irrelevant variables and to focus on an area of particular interest. IDEAL is domain-independent and the user does not need to instruct IDEAL what to search for. IDEAL can potentially discover any combination of attribute values that display an interesting temporal development. Thus it can serve as an fully automatic early warning system that discovers threats and opportunities well before threshold-oriented KPI models could react. The detected patterns can trigger an in-depth analyses and potentially lead to predictive models that can be used in business applications.

3 Automatic Model Building

In order to support their planning activities and to facilitate pro-active decision making more and more businesses use models derived from data to predict changes in processes, markets or customer behaviour. Data analysis tools that are commercially available today and support predictive modelling are still very much a collection of data analysis methods that require analysis experts as users. The user not only needs domain knowledge about the data, but he also needs to know which data analysis methods are applicable to his problem, which ones meet some special requirements for the solution, how data needs to be prepared for the chosen method and finally, how the method needs to be configured.

On the other hand, many business users are keen to employ data analysis to make use of data that is being collected. Although a great proportion of typical analysis problems look quite simple to a data analysis expert, business users are overwhelmed by the sheer knowledge that is required to use current tools. We consider this one of the reasons for the fact that modern machine learning techniques like decision trees, neural networks, fuzzy techniques, support vector machines etc. are still not industry standard.

Business users require a much more user- or problem-oriented approach to data analysis. Rather than knowing analysis methods, they are experts in the data domain and they know what they want to achieve with data analysis. If they only knew how. They might know, for example, that they want to classify insurance claims as fraudulent or non-fraudulent, given historic information of the customer and the current case. They might want to understand, how the analysis method actually classifies customers (e.g. with a rule set), they might require a certain classification accuracy and that the algorithm is so simple that it can be implemented as an SQL query. Ideally, such users would simply like to feed all these high-level requirements and the data into a tool that would then automatically find the best algorithm in terms of requirements, configure it, run it and create a software module that can be plugged into the business application.

Based on these ideas we developed SPIDA (Soft computing Platform for Intelligent Data Analysis) [6] and equipped it with a Wizard that, to certain extent, does most of the things mentioned above. SPIDA uses a fuzzy expert system for selecting the most appropriate data analysis algorithm given a problem definition, a set of requirements and a data file. Once an algorithm has been selected SPIDA can determine appropriate parameters for the model creation (learning) process, automatically build (several versions of) different models in parallel and determine how well they match the user requirements. Finally, a successful model is turned into a Java software module that can be plugged into application software.

4 Customer Analytics – An Application

Many large businesses regularly conduct customer surveys in order to understand different aspects of customer attitudes and expectations. Typically, survey results are analysed statistically to highlight, for example, customer satisfaction in different areas. Such reports can pinpoint areas where customer satisfaction should be improved, but they do not provide information about dependencies between different influence factors and they cannot be used as a planning tool.

We have developed iCSat – a platform for intelligent customer satisfaction modelling [4, 5]. The approach of the software is to analyse the dependencies between all satisfaction indicators and to automatically learn a Bayesian network to represent a probabilistic model of customer responses. This model can then be used to plan changes in selected satisfaction indicators and to understand the influence on other indicators. This allows users to plan projects for improving customer satisfaction and to tackle the most promising indicators for improvement.

iCSat is a Java-based client/server application. The user is provided with an intuitive client that connects to a server that manages data in the user's database schema and carries out the automatic model building process. For learning the structure of a Bayesian network we use a combination of the K2 algorithm by Cooper and Herskovitz [3] and the algorithm by Singh and Valtorta [7]. The parameter learning is handled by the Java library of the commercial Bayesian tool Netica (www.norsys.com) that is also used to represent and run the model. iCSat allows users to determine the main influences for any variable and run different types of what-if scenarios.

References

1. Böttcher, M., Nauck, D., Borgelt, C. and Kruse, R.: A framework for discovering interesting business changes from data. BT Technology Journal **24**(2), Springer, Dordrecht (2006) 219–228
2. Böttcher, M., Spott, M. and Nauck, D. Detecting temporally redundant association rules. In *Proceedings of 4th ICMLA*, pp. 397–403, Los Angeles, USA, 2005. IEEE Computer Society.
3. Cooper, G.F., Herskovitz E.H.: A Bayesian method for the induction of probabilistic networks from data. Machine Learning **9** (1992) 309–347
4. Nauck, D., Ruta, D., Spott, M., and Azvine, B.: Being proactive – analytics for predicting customer actions. BT Technology Journal **24** (1), Springer, Dordrecht (2006) 17–26
5. Nauck, D., Ruta, D., Spott, M., and Azvine, B.: A Tool for Intelligent Customer Analytics. In *Proceedings of the 3rd IEEE Conference On Intelligent Systems*, London, UK, 2006 (to appear).
6. Nauck, D., Spott, M., Azvine, B.: SPIDA – a novel data analysis tool. BT Technology Journal **21**(4), Kluwer, Dordrecht (2003) 104–112
7. Singh, M. and Valtorta, M.: Construction of Bayesian Network Structures from Data: a Brief Survey and an Efficient Algorithm. International Journal of Approximate Reasoning **12**, Elsevier Science, New York, NY (1995) 111–131.

Item Planning with Graphical Models (Extended Abstract)

Jörg Gebhardt

Intelligent Systems Consulting (ISC Gebhardt)
Leonhardtstraße 19, D-29227 Celle, Germany
e-Mail: info@ISC-Gebhardt.de

In the last decade graphical models have become one of the most popular tools to structure uncertain knowledge about high-dimensional domains in order to make reasoning in such domains feasible. Their most prominent representatives are Bayesian networks and Markov networks. For both types of networks several clear, correct, and efficient propagation methods have been developed, with join tree propagation and bucket elimination being among the most widely known. With these methods it is possible to *condition* the probability distribution represented by a graphical model on given evidence, i.e. on observed values for some of the variables, a reasoning process that is also called *focusing*. Efficient and user-friendly commercial tools for this task, like HUGIN and NETICA, are widely available. In practice, however, the need also arises to support a variety of additional knowledge-based operations on graphical models, where *revision*, *updating*, the *fusion* of networks with relational rule systems, network *approximation*, and *learning* from data samples are some of the most important ones. Furthermore, it is essential to provide software tools in order to make interactive planning, reasoning, and decision making feasible, even in complex networks of real world applications.

The research to be reported about here was mainly triggered by ISC's consulting of the automobile manufacturer Volkswagen Group, where Markov networks are now established for item planning and capacity management. In opposite to many competing car manufacturers, Volkswagen Group favours a marketing policy that provides a maximum degree of freedom in choosing individual specifications of vehicles. That is, considering personal preferences, a customer may select from a large variety of options, each of which is taken from a so-called item family that characterizes a certain line of equipment. Typical examples include body variants, engines, gearshifts, door layouts, seat coverings, radios, and navigation systems. In case of the VW Golf - being Volkswagen's most popular car class - there are about 200 families with typically 4 to 8 values each, and a total range of cardinalities from 2 up to 150. In applications of such complexity, the importance of performance aspects is obvious, since it turns out that item planning, even when reduced to a single vehicle class within a single planning week, requires handling Markov networks of about 150 cliques and maximum dimensionalities of at least 12 for individual cliques. As a consequence, domains with a cardinality of more than 100,000,000 elements have to be considered, and in total we have to deal with about 2,000 of such networks.

The planning problem is addressed by a project called EPL (Eigenschaftsplanung, item planning) that was initiated in 2001 by Corporate IT, Sales, and Logistics of the Volkswagen Group. The aim was to develop an item planning system common to all trademarks that reflects a modelling approach based on Markov networks. System design and most of the implementation work of EPL is currently done by Corporate IT. The mathematical modelling, theoretical problem solving, and the development of efficient algorithms, extended by the implementation of a new software library called MARNEJ (MARkov Networks in Java) for the representation of Markov networks and the above-mentioned functionalities have been entirely provided by ISC Gebhardt. The world-wide rollout of the system EPL to all trademarks of the Volkswagen Group reflects the current state of the project. Up to 15 system developers implemented a client-server architecture in Java. The planned configuration uses 6 to 8 Hewlett Packard machines with 4 AMD Opteron 64-Bit CPUs and 24 GB of main memory each, and a terabyte storage device. The system is running Linux and an Oracle database system.

The presentation refers to new theoretical and algorithmic results as well as some details on the Volkswagen application.

References

1. Borgelt, C., Kruse, R.: Graphical Models—Methods for Data Analysis and Mining. J. Wiley & Sons, Chichester (2002)
2. Gebhardt, J., Detmer, H., Madsen, A.: Predicting parts demand in the automotive industry – an application of probabilistic graphical models. In: Proc. Int. Joint Conf. on Uncertainty in Artificial Intelligence (UAI'03, Acapulco, Mexico), Bayesian Modelling Applications Workshop. (2003)
3. Gebhardt, J., Borgelt, C., Kruse, R., Detmer, H.: Knowledge revision in markov networks. Mathware and Soft Computing, Special Issue "From Modelling to Knowledge Extraction" **11: 2-3** (2004) 93–107
4. Gebhardt, J., Kruse, R.: Knowledge-based operations for graphical models in planning. In Godo, L., ed.: Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU'05). Lecture Notes in Artificial Intelligence 3571. Springer, Berlin (2005) 3–14
5. Gebhardt, J., Klose, A., Detmer, H., Rügheimer, F., Kruse., R.: Graphical models for industrial planning on complex domains. In Riccia, G.D., ed.: Intelligent Agents: Decision Support and Planning. Lecture Notes in Artificial Intelligence. Springer, Berlin (2006)
6. Gebhardt, J., Kruse, R.: Parallel combination of information sources. In Gabbay, D.M., Smets, P., eds.: Handbook of Defeasible Reasoning and Uncertainty Management Systems. Volume 3: Belief Change. Kluwer Academic Publishers, Dordrecht (1998) 393–439

Subgroup Discovery Based on Ideas from Fuzzy Cluster Analysis

Frank Klawonn

Department of Computer Science
University of Applied Sciences Braunschweig/Wolfenbuettel
Salzdahlumer Str. 46/48
Wolfenbuettel, Germany
f.klawonn@fh-wolfenbuettel.de

Abstract. This extended abstract provides a brief outline on a new algorithm for subgroup discovery, exploiting some principles from fuzzy cluster analysis.

1 Introduction

Subgroup discovery refers to the following supervised inductive learning task [6, 9]. Given a population of individuals and a specific property of individuals, the aim is to find subgroups in the population that are statistically significant or interesting. Supervised classification means to predict the value (class) of a categorical attribute based on other attributes that can be categorical, but also numerical. Supervised classification tries to predict the correct class for any possible input. In certain applications it is, however, impossible to achieve good predictions for all possible inputs. Nevertheless, it can be of interest, to identify subsets for which reliable predictions can be made. For instance, when a company wants to initiate an advertisement campaign to specific customers who have a high potential to buy the advertised product, it would be perfect, to have a classifier that can predict for all customers whether they might have a high interest in buying the product. But in most cases, it is impossible to make a reliable prediction for all types of customers. Nevertheless, there might be possible to identify subgroups of customers with a high potential to buy the product. Only these customers would be addressed in the campaign, although this might mean that a significant proportion of the potential buyers would be excluded.

In this paper, we address a specific problem in subgroup discovery. We assume that we have one categorical attribute and a specific class of interest and want to find subgroups in a data set where all or at least an unusually high fraction of individuals belongs to the class of interest.

2 Characterisation of Subgroups

As mentioned in the introduction, the subgroups, we are interest in, belong to a specific class. There are many ways to characterise a subgroup. Here, we only consider numerical attributes to describe the subgroup. We are furthermore interested in a description of the subgroup in terms of a compact, connected subset in the form of hyper-spheres, hyper-ellipsoids or hyperboxes. The advantage of hyperboxes is that they are easy to understand, since the decision, whether an object belongs to a hyperbox or not, can be made by just looking at each of the attributes individually. In this way, the assignment of objects to the subgroup is very plausible even for high-dimensional data sets.

3 Cluster Analysis and Subgroup Discovery

The intention of cluster analysis is mainly unsupervised classification. This means that the task is to partition a given data set into subsets such that the objects within a subset are very similar, whereas objects from different subsets should be well distinguishable. In objective function-based clustering (for an overview see for instance [1, 3]) clusters are usually described by a prototype and in addition an optional modified distance function enabling the detection of clusters of different shapes.

Cluster analysis has been extended to supervised classification, for instance in [7] where each cluster is intended to cover one class and elements from other classes have a repulsing influence on the prototype of the cluster. In [8] clusters are punished, when they contain data from other classes. Although these ideas might seem appealing, these extensions of cluster analysis try to solve the full supervised classification task and have their own specific quality measures that are difficult to adapt to subgroup discovery.

In [2] the concept of noise clustering is described. An additional noise clusters is introduced for clustering. This noise cluster is supposed to cover those data objects that do not fit well to other clusters. Based on the principle of noise clustering, algorithms to identify single clusters in data sets were developed [4, 5]. It is no longer necessary to partition the data set into meaningful clusters, which is very often impossible. It is sufficient to identify a single good cluster and to assign even the majority of the data to the noise cluster. Based on these concepts an approach to subgroup discovery will be described in the following section.

4 A New Approach

In order to identify a subgroup, it is first necessary to specify the class of interest to which most or all of the objects in the subgroup should belong. Then, an error measure has to be defined that indicates the quality of the subgroup. Usually, two criteria are considered for this quality measure. The subgroup should be as large as possible and it should be as pure as possible, i.e. it should contain a high rate of objects from the class of interest. A high rate can also mean a high rate only in comparison to the overall rate of the class of interest in the data set. For instance, in fault detection, situations where faults occur might be quite rare with a rate far below than 1% and it might be of high value to identify subgroups with a rate of 10% of faulty individuals in order to take steps of fault prevention in the corresponding situations.

As a first step, we identify a single good cluster considering only the data from the class of interest based on techniques as they are proposed in [4, 5]. The cluster should cover a certain fraction of the data from the desired class. The fraction depends on the chosen quality measure. If the quality measure demands a high number of objects from the class of interest, then the cluster will be chosen larger. In this way, we obtain a cluster prototype and a radius of the cluster.

After the single good cluster has been identified ignoring the data from the other classes, these data objects are also included and the quality measure for the cluster is computed taking not only the coverage of the class of interest into account, but also the error rate in the subgroup, i.e. how many objects from other classes are contained in the subgroup. If the desired coverage or error rate is not achieved, we adjust the distance function and the prototype of the cluster in order to minimize the specified quality message. Here we apply heuristic and partly greedy search strategies for adjusting these parameters. Since we always require an improvement of the quality of the subgroup, when the parameters of the subgroup are changed, we can at least guarantee convergence for our algorithm, even so it might get stuck in a local minimum. If the subgroup does not satisfy the required quality, we might initiate the clustering of the first step again. In this case, we apply a modified clustering technique with forbidden regions in order to make sure that we do not identify the same cluster again.

5 An Example

For illustration purposes, we consider an extremely simple data set, the well known iris data set, in figure 1. In order to identify a subgroup four numerical attributes are available. The figure shows only two out of these four dimensions.

The rectangle indicates the two-dimensional projection of the four-dimensional hyperbox that characterises the subgroup. The specified subgroup does not cover all objects from the class of interest, but a very large fraction. The point out of the rectangle marked by a circle is an example for an object that belongs to the class of interest, but not to the subgroup. There is only one element, that belongs to the subgroup, but not to the class of interest. It is also marked by a circle lying within the rectangle.

6 Conclusions

In this paper, we have outlined a method to identify subgroups in terms of compact shapes in the form of hyper-spheres, hyper-ellipsoids or hyperboxes. The basic strategy is to first identify a single cluster in the class of interest for the subgroup and then adjust this cluster in order to reduce the fraction of objects from other classes in the subgroup. Different quality measures combing the required minimum size of the subgroup as well as the error rate, i.e. the

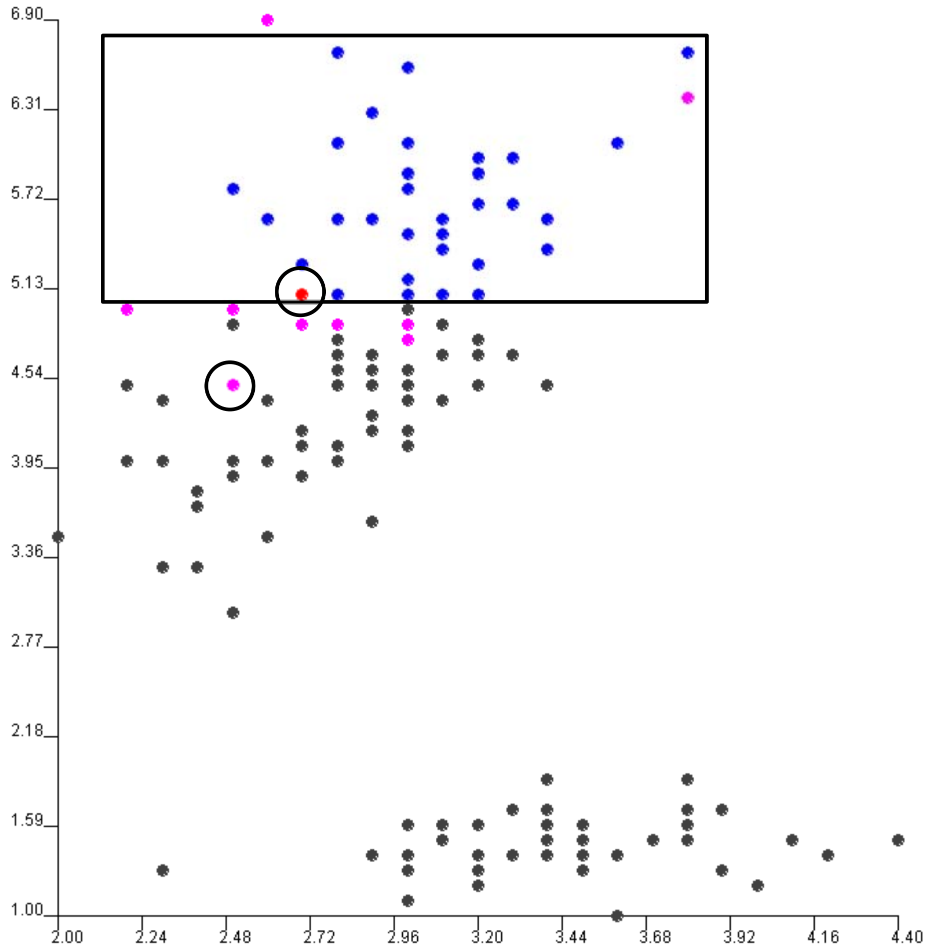


Fig. 1. An example data set

number of objects from other classes in the subgroup, can be used depending on the underlying data set and problem.

References

1. Bezdek, J.C., Keller, J., Krishnapuram, R., Pal, N.R.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.Kluwer, Boston, 1999.
2. Davé, R.N.: Characterization and Detection of Noise in Clustering. Pattern Recognition Letters, 12, pp. 657-664, 1991.
3. Höppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis, John Wiley & Sons, Chichester, 1999.
4. Klawonn, F., Georgieva, O.: Identifying Single Clusters in Large Data Sets. In: Wang, J. (ed.): Encyclopedia of Data Warehousing and Mining. Idea Group, Hershey, pp. 582–585, 2006.
5. Klawonn, F.: Identifying Single Good Clusters in Data Sets. In: Zheng, N., Jiang, X., Lan, X. (eds.): Advances in Machine Vision, Image Processing, and Pattern Analysis. Springer, Berlin pp. 160–167, 2006.

6. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, Calif., pp. 249–271, 1996.
7. Timm, H.: Fuzzy Cluster Analysis of Classified Data. In: Proc. Joint 9th IFSAWorld Congress and 20th International Conf. of the North American Fuzzy Information Processing Society (NAFIPS01). Vancouver, Canada, 2001.
8. Timm, H., Klawonn, F., Kruse, R.: An Extension of Partially Supervised Fuzzy Cluster Analysis. In: Proc. 21th Intern. Conf. of the North American Fuzzy Information Processing Society – NAFIPS. IEEE, Piscataway, NJ, pp. 63–68, 2002.
9. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Komorowski, J., Zytkow, J. (eds): Principles Of Data Mining And Knowledge Discovery. Springer, Berlin, pp. 78–87, 1997.

Resampling for Fuzzy Clustering

Christian Borgelt

European Center for Soft Computing
c/ Gonzalo Gutiérrez Quirós s/n, 33600 Mieres, Spain
christian.borgelt@softcomputing.es

Abstract. Resampling methods are among the best approaches to determine the number of clusters in prototype-based clustering. The core idea is that with the right choice for the number of clusters basically the same cluster structures should be obtained from subsamples of the given data set, while a wrong choice should produce considerably varying cluster structures. In this paper I give a brief overview how such resampling approaches can be transferred to fuzzy and probabilistic clustering.

1 Introduction

A core problem of prototype-based clustering algorithms—like classical c-means [12, 17], its fuzzy counterpart (fuzzy c-means) [2, 13], or expectation maximization for mixtures of Gaussians [5, 7]—is that they require the number of clusters to be known in advance. A common approach to tackle this problem is to cluster the data set several times, each time with a different number of clusters from a user-specified range, and then to choose the number of clusters yielding the best evaluation (see, for example, [2, 13, 4] for overviews of evaluation measures).

In this paper I study an alternative approach that has recently attracted a lot of attention in crisp and probabilistic clustering. The core idea is that if we cluster subsamples of the given data set with the "right" number of clusters, we should end up with basically the same cluster structure in each run. With a "wrong" number of clusters, however, the clustering result should be unstable, showing considerable variation between different subsamples. Thus, by measuring the stability of the clustering result w.r.t. subsampling (similarity of results from different runs), one may be able to determine the "best" number of clusters: it is the one for which the clustering results are most stable.

Intuitively, one may think of this as follows: if the "true" number of clusters is c and we try to find $c+1$ clusters, one cluster has to be split. If we try to find $c-1$ clusters, some pair of clusters has to be merged. As it depends on particular properties of the subsample which cluster is split or which clusters are merged, we should get somewhat differing structures in each run. By measuring how well the clustering results coincide, we can thus discover such situations and choose the number of clusters based on this information.

In addition to a general discussion of this highly promising approach, I study experimentally how the choice of t-norms in the needed relative cluster evaluation measures (to combine membership degrees) affects the quality and clarity of the results, that is, how well the "best" number of clusters can be determined.

	$u^{(2)}_{kj} = 1$	$u^{(2)}_{kj} = 0$	Σ
$u^{(1)}_{ij} = 1$	$n^{(i,k)}_{11}$	$n^{(i,k)}_{10}$	$n^{(i,k)}_{1.}$
$u^{(1)}_{ij} = 0$	$n^{(i,k)}_{01}$	$n^{(i,k)}_{00}$	$n^{(i,k)}_{0.}$
Σ	$n^{(i,k)}_{.1}$	$n^{(i,k)}_{.0}$	n

Table 1. Contingency table comparing rows of two (crisp) partition matrices (i and k are the cluster indices).

2 Relative Cluster Evaluation Measures

Relative cluster evaluation measures compare two partitions of given data, each of which can be described by a $c \times n$ partition matrix $\mathbf{U} = (u_{ij})_{1 \le i \le c, 1 \le j \le n}$, where c is the number of clusters and n the number of data points. An element u_{ij} of such a matrix states, in the crisp case, whether the j-th data point belongs to the i-th cluster ($u_{ij} = 1$) or not ($u_{ij} = 0$). In the fuzzy case, u_{ij} is the degree of membership to which the j-th data point belongs to the i-th cluster (usually satisfying the constraint $\forall j; 1 \le j \le n : \sum_{i=1}^{c} u_{ij} = 1$).

The main problem of the comparison is how to relate the clusters of one partition to the clusters of the other. There are basically three solutions: (1) for each cluster in the one partition we determine the *best fitting* cluster in the other, (2) we find the *best row permutation*, that is, the best one-to-one mapping of the clusters, or (3) we compare indirectly by first setting up a *coincidence matrix* for each partition matrix, which records for each pair of data points whether they are assigned to the same cluster or not, and then compare these matrices. Here I confine myself to the second and the third alternative.

2.1 Comparing Partition Matrices

To compare two $c \times n$ partition matrices $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ directly, we need a measure that compares two rows, one from each matrix. Such measures can be derived from measures comparing binary classifications, like, for example, the *accuracy* or the F_1-*measure* [19]. Formally, we set up a 2×2 contingency table for each pair of rows, one from each matrix (cf. Table 1). That is, for each pair $(i, k) \in \{1, \dots, c\}^2$ and each row-column pair $(a, b) \in \{0, 1\}^2$ we compute

$$n^{(i,k)}_{ab}\big(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}\big) = \sum_{j=1}^{n} \Big((1-a) + (2a-1)\, u^{(1)}_{ij}\Big) \cdot \Big((1-b) + (2b-1)\, u^{(2)}_{kj}\Big).$$

(In the following I generally drop the arguments $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ to make the formulae easier to read.) These numbers may also be computed from fuzzy membership degrees, where they have a fairly natural interpretation: in the crisp case, n_{11} is the number of data points that are assigned to the i-th cluster of the first partition *and* to the k-th cluster of the second partition, where the *and* is formally expressed by a product. Allowing membership degrees from $[0, 1]$ and drawing on the theory of fuzzy logic, we see that this is only a special case of a t-norm that combines the two statements. Hence, in the general case, we may replace the product by an arbitrary t-norm. Analogously, the expressions $1 - u_{ij}$ (for $a = 0$ or $b = 0$) can be seen as resulting from an application of the standard fuzzy negation, and indeed: they refer to negated statements "The j-th data point does *not* belong to the i-th cluster." In this way we achieve a straightforward generalization of all following measures to fuzzy clustering results.

From the numbers $n_{ab}^{(i,k)}$ computed above we may now compute any measure for evaluating a binary classification, maximizing the result over all row permutations.[1] An example is the (averaged) F_1 *measure* [19]

$$F_1\big(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}\big) = \max_{\varsigma \in \Pi(c)} \frac{1}{c} \sum_{i=1}^{c} \frac{2\,\pi_{i,\varsigma(i)}\rho_{i,\varsigma(i)}}{\pi_{i,\varsigma(i)} + \rho_{i,\varsigma(i)}},$$

where $\Pi(c)$ is the set of all permutations of the c numbers $1, \dots, c$ and cluster-specific precision and recall are

$$\pi_{i,k} = \frac{n_{11}^{(i,k)}}{n_{01}^{(i,k)} + n_{11}^{(i,k)}} \qquad \text{and} \qquad \rho_{i,k} = \frac{n_{11}^{(i,k)}}{n_{10}^{(i,k)} + n_{11}^{(i,k)}}.$$

Another example is *(cross-classification) accuracy*, averaged over all columns:

$$Q_{\text{acc}}\big(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}\big) = \max_{\varsigma \in \Pi(c)} \frac{1}{cn} \sum_{i=1}^{c} \left(n_{00}^{(i,\varsigma(i))} + n_{11}^{(i,\varsigma(i))}\right).$$

Two partition matrices $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ are the more similar, the higher the values of the (averaged) F_1 measure or the (cross-classification) accuracy. An alternative is a simple mean squared difference comparison of the partition matrices (which, at least to my knowledge, has not been used before). That is, we compute

$$Q_{\text{diff}}\big(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}\big) = \min_{\varsigma \in \Pi(c)} \frac{1}{cn} \sum_{i=1}^{c} \sum_{j=1}^{n} \left(u_{ij}^{(1)} - u_{\varsigma(i)j}^{(2)}\right)^2.$$

The smaller this measure, the more similar are the partitions.

2.2 Comparing Coincidence Matrices

As an alternative to comparing partition matrices directly, one may first compute from each of them an $n \times n$ *coincidence matrix*, also called a *cluster connectivity matrix* [16], which states for each pair of data points whether they are assigned to the same cluster or not. Formally, a coincidence matrix $\mathbf{\Psi} = (\psi_{jl})_{1 \le j,l \le n}$ can be computed from a partition matrix $\mathbf{U} = (u_{ij})_{1 \le i \le c, 1 \le j \le n}$ by

$$\psi_{jl} = \sum_{i=1}^{c} u_{ij} u_{il}.$$

These values may also be computed from fuzzy membership degrees, possibly replacing the product (which represents a conjunction) by some other t-norm.

[1] Note that with the so-called *Hungarian method* for solving optimum weighted bipartite matching problems [18] the time complexity of finding the maximum over all permutations for given pairwise column comparison values is $O(c^3)$ and not $O(c!)$.

	$\psi^{(2)}_{jl} = 1$	$\psi^{(2)}_{jl} = 0$	Σ
$\psi^{(1)}_{jl} = 1$	N_{11}	N_{10}	$N_{1.}$
$\psi^{(1)}_{jl} = 0$	N_{01}	N_{00}	$N_{0.}$
Σ	$N_{.1}$	$N_{.0}$	$N_{..}$

Table 2. Contingency table for comparing (crisp) coincidence matrices (the indices 1 and 0 mean same and different cluster, respectively).

Such matrices are compared by computing statistics of the number of data point pairs that are in the same group in both partitions, in the same group in one, but in different groups in the other, or in different groups in both. Formally, we compute a 2×2 contingency table (cf. Table 2) containing the numbers (which are basically counts of the different pairs $\big(\psi^{(1)}_{jl}, \psi^{(2)}_{jl}\big)$)

$$N_{ab}\big(\boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}\big) = \sum_{j=2}^{n} \sum_{l=1}^{j-1} \Big((1-a) + (2a-1)\,\psi^{(1)}_{jl}\Big) \Big((1-b) + (2b-1)\,\psi^{(2)}_{jl}\Big),$$

where an index $a, b = 1$ stands for “same group” and an index $a, b = 0$ stands for “different groups”. (The arguments $\boldsymbol{\Psi}^{(1)}$ and $\boldsymbol{\Psi}^{(2)}$ are dropped in the following.) Again the product may be replaced by any t-norm (note that $\psi_{jl} \in [0, 1]$, since fuzzy clustering satisfies $\forall j; 1 \le j \le n : \sum_{i=1}^{c} u_{ij} = 1$). From these numbers a large variety of measures may be computed, including the *Rand statistic*

$$Q_{\text{Rand}}\big(\boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}\big) = \frac{N_{11} + N_{00}}{N_{..}},$$

which is a simple ratio of the number of data point pairs treated the same in both partitions to all data point pairs, and the *Jaccard coefficient*

$$Q_{\text{Jaccard}}\big(\boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}\big) = \frac{N_{11}}{N_{11} + N_{10} + N_{01}},$$

which ignores negative information, that is, pairs that are assigned to different groups in both partitions. Both measures are to be maximized. Another frequently encountered measure is the *Folkes–Mallows index*

$$Q_{\text{Folkes-Mallows}}\big(\boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}\big) = \frac{N_{11}}{\sqrt{(N_{11} + N_{10})(N_{11} + N_{01})}},$$

which can be interpreted as a cosine similarity measure and thus is also to be maximized. A final example is the *Hubert index*

$$Q_{\text{Hubert}}\big(\boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}\big) = \frac{N_{..}N_{11} - N_{1.}N_{.1}}{\sqrt{N_{1.}N_{.1}N_{0.}N_{.0}}},$$

which may either be interpreted as a product-moment correlation or as the square root of the (normalized) χ^2 measure. It should be clear that this list does not exhaust all possibilities. Basically all of the abundance of measures, by which (binary) vectors and matrices can be compared, are applicable.

3 Resampling

Resampling methods can be found with basically two sampling strategies. In the first place, one may use *subsampling* [8], that is, the samples are drawn without replacement from the given data set, so that each data point appears in at most one data subset. This strategy is usually applied in a cross validation style, that is, the given data set is split into a certain number of disjoint subsets (with two subsets being the most common choice). The alternative is *bootstrapping* [6], in which samples are drawn with replacement, so that a data point may appear multiple times in the same data subset. There are good arguments in favor and against both approaches, but the results often do not differ much.

Resampling is used for cluster validation and model selection as follows: a cluster model can usually be applied as a classifier, thus enabling us to assign data points, which have not been used to build the cluster model, to the clusters. In this way we obtain, with the same algorithm, two different groupings of the same set of data points. For example, one may be obtained by clustering the data set, the other by applying a cluster model that was built on another data set. These two groupings can be compared using, for example, one of the measures discussed in the preceding section. By repeating such comparisons with several samples drawn from the original data set, one can obtain an assessment of the variability of the cluster structure (or, more precisely, an assessment of the variability of the evaluation measure for the similarity of partitions). Such an approach may be applied to select the most appropriate cluster model—and in particular, the "best" number of clusters—by executing the above algorithm for different parameterizations of the clustering algorithm and then to select the one showing the lowest variability. Specific algorithms following this general scheme have been proposed in [16, 20, 15], which differ in the exact resampling strategies and the evaluation measures used. All indicate that this approach is very robust and a fairly reliable way of choosing the number of crisp clusters.

4 Experiments

I carried out several experiments by applying a resampling approach for fuzzy clustering based on the above explanations to five data sets. The first three are artifical two-dimensional data sets of 400 data points each with three, four, and six clusters, respectively. They are shown in Figure 1. The fourth data set is an artificial three-dimensional data set of 400 data points with five equally populated, but ellipsoidal clusters. It is shown on the left in Figure 2. The last data set is the well-known wine data set from the UCI machine learning repository [3], a view of which is shown on the right in Figure 2. It comprises three classes of Italian wines and thus one can expect to find three clusters. I used attributes 7, 10, and 13, which are the most informative w.r.t. the class.

Before clustering all datasets were normalized in all dimensions to mean 0 and standard deviation 1 to rule out scaling effects. The experiments were carried out with the following resampling scheme: first the whole data set was clustered.

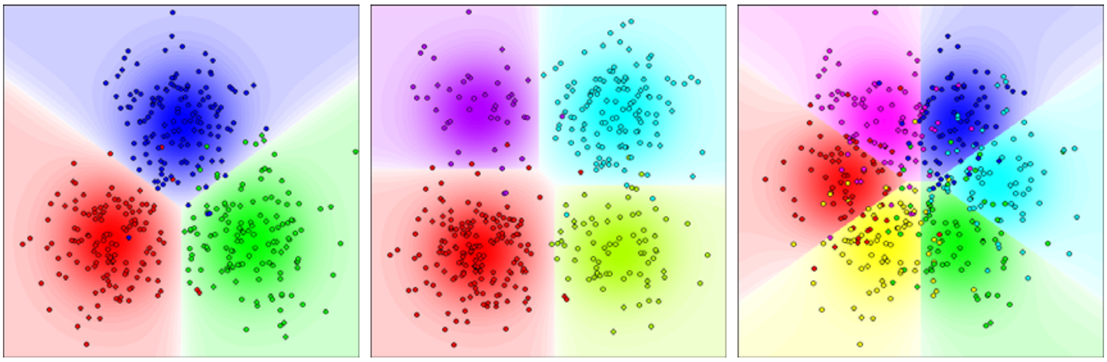


Fig. 1. Artificial data sets with 3 (equally populated), 4 (differently populated), 6 (equally populated) spherical clusters.

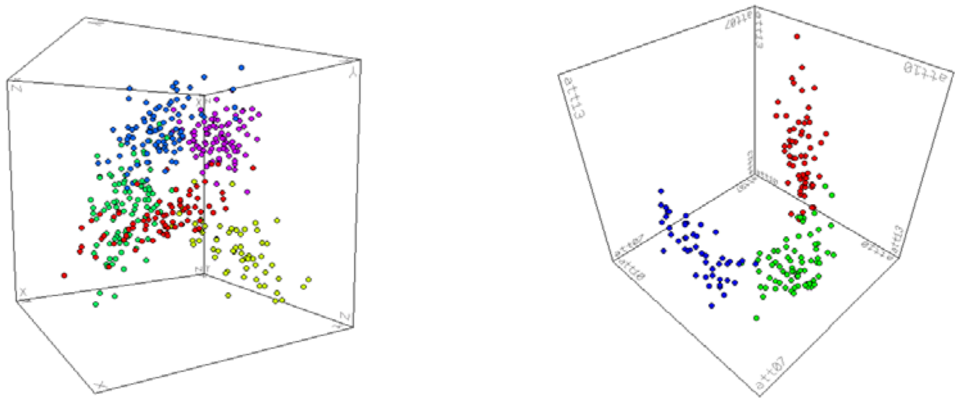


Fig. 2. An artificial data set with 5 (equally populated) ellipsoidal clusters and a view of the wine data set (attributes 7, 10, and 13).

Then 100 random samples (without replacement) were drawn from the data set, each of which comprised about half of the data points. (The data set was split into two equal parts, one of which was used). Each sample was clustered with the same number of clusters as the full data set and then the two cluster structures (one obtained from the full data set and one from the sample) were compared on the full data set using the measures described in Section 2. The evaluation results were averaged over the 100 samples, thus yielding a stability measure.

In the measures I used four different t-norms to combine membership degrees and the coincidence matrix entries (see Figure 3 for illustrations):

$$\begin{aligned} \top_{\text{min}}(a,b) &= \min\{a,b\}, & \top_{\text{minnp}}(a,b) &= \min\{a,b\} \text{ if } a+b \geq 1,\ 0 \text{ otherwise}, \\ \top_{\text{prod}}(a,b) &= a \cdot b, & \top_{\text{Luka}}(a,b) &= \max\{0, a+b-1\}, \end{aligned}$$

where $\top_{\text{minnp}}$ is the so-called *nil-potent minimum*. Since there are two places where a t-norm is needed in the measures based on comparing coincidence matrices, I tried all pairs of t-norms to explore their interactions. As it turns out, they cannot be combined freely: some combinations do not work well.

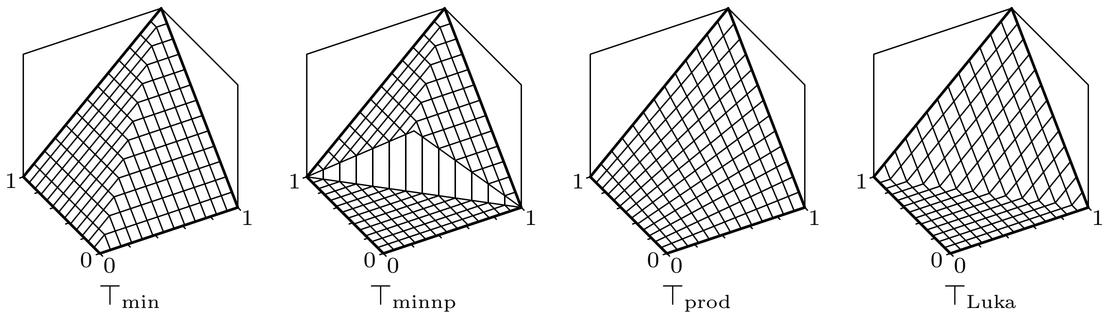


Fig. 3. The different t-norms used in the experiments.

data	diff	accuracy				F_1			
		min	mnp	prd	Luk	min	mnp	prd	Luk
art. 3	4	4	1	1	1	4	5	4	4
art. 4	3	2	4	0	0	3	3	1	3
art. 6	6	6	6	6	6	6	6	6	6
wine	4	4	0	1	1	4	4	0	4
art. 5	4	4	0	1	1	1	1	0	6
wine	5	4	4	0	0	0	0	0	1

Table 3. Overview of the results of comparing partition matrices with different measures and t-norms on the different data sets. Fuzzy c-means clustering was used for the first four rows, Gustafson–Kessel clustering for the last two.

Since it is not possible to show all individual results in this paper (there are simply too many different experiments), I try to give an impression of the performance of the different measures (in combination with different selections of t-norms) by providing a rough overview and reporting some individual results. The overview is shown in Tables 3 and 4 and uses grades to assess the performance of the different measures, with the following meanings:

6: clear global optimum at the correct cluster number,
no local optimum at any other cluster number
5: clear global optimum at the correct cluster number,
but there is a (weak) local optimum at another cluster number
4: only weak global optimum at the correct cluster number,
or a competing local optimum at another cluster number
3: clear local optimum at the correct cluster number,
but global optimum is at another cluster number
2: only weak local optimum at the correct cluster number,
or global optimum is significantly higher than local optimum
1: only a discernable step at the correct cluster number,
but not even a weak local optimum
0: no discernable characteristics at the correct cluster number

With grades 6 and 5, maybe also 4, the measure is usable for fully automatic selection, with grades 4, 3 and 2 for semi-automatic processing (with user interaction). With grades 1 and 0 a measure fails to find the correct cluster number.

Rand	min				minnp				prod				Luka			
data	min	mnp	prd	Luk	min	mnp	prd	Luk	min	mnp	prd	Luk	min	mnp	prd	Luk
art. 3	4	1	1	0	4	1	1	1	4	1	1	0	4	4	4	4
art. 4	4	1	1	0	4	0	1	1	2	1	1	0	3	3	2	2
art. 6	3	6	6	2	3	6	6	6	3	6	1	3	6	6	6	6
wine	4	0	1	0	4	1	1	1	4	0	0	0	4	0	0	0
art. 5	4	0	0	0	4	1	0	0	4	0	0	0	4	4	4	4
wine	4	1	0	0	4	0	0	1	0	1	1	0	0	4	4	0

Jaccard	min				minnp				prod				Luka			
data	min	mnp	prd	Luk	min	mnp	prd	Luk	min	mnp	prd	Luk	min	mnp	prd	Luk
art. 3	0	5	4	5	0	5	6	5	0	5	1	5	5	5	6	5
art. 4	1	3	1	3	1	3	3	3	0	3	0	3	3	3	3	3
art. 6	3	6	6	6	3	6	6	6	3	6	6	6	3	6	6	6
wine	0	4	1	4	0	0	0	5	0	0	0	4	0	0	0	5
art. 5	0	6	0	6	0	6	0	1	0	6	0	1	6	6	1	6
wine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Folkes	min				minnp				prod				Luka			
data	min	mnp	prd	Luk	min	mnp	prd	Luk	min	mnp	prd	Luk	min	mnp	prd	Luk
art. 3	1	4	4	4	1	5	6	5	0	5	1	5	4	5	6	5
art. 4	1	3	1	3	1	3	3	3	0	3	0	3	3	3	3	3
art. 6	3	6	6	6	3	6	6	6	3	6	6	6	3	6	6	6
wine	0	4	0	4	0	0	0	4	0	0	0	4	0	0	0	4
art. 5	0	6	0	6	0	6	0	1	0	6	0	1	6	6	1	6
wine	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

Hubert	min				minnp				prod				Luka			
data	min	mnp	prd	Luk	min	mnp	prd	Luk	min	mnp	prd	Luk	min	mnp	prd	Luk
art. 3	4	4	4	4	5	5	6	5	5	4	6	5	5	5	5	5
art. 4	6	4	4	4	3	6	6	6	3	6	6	5	3	3	3	3
art. 6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
wine	4	0	0	4	4	4	4	4	6	4	4	4	4	4	4	4
art. 5	1	6	1	4	0	6	1	6	0	6	1	6	6	6	6	6
wine	6	4	4	6	6	6	1	6	1	6	1	6	1	1	1	1

Table 4. Overview of the results of comparing coincidence matrices with different measures and t-norms on the different data sets. In each table the upper header row shows the t-norm for combining coincidence matrix entries, the lower header row the t-norm for combining membership degres. Fuzzy c-means clustering was used for the first four rows, Gustafson–Kessel clustering for the last two rows of each table.

	partition matrix			coincidence matrix			
#	diff	acc	F1	Rand	Jacc.	Folkes	Hubert
2	**.0001**	**.9918**	**.7606**	.7028	**.5769**	**.7317**	.3987
3	.0157	.9521	.6799	.6859	.4988	.6650	.3696
4	**.0009**	**.9850**	**.6851**	**.7123**	.4903	.6579	**.4097**
5	.0203	.9365	.5885	.6741	.4156	.5870	.3178
6	.0150	.9461	.5691	.6741	.3926	.5636	.3036
7	.0132	.9520	.5542	.6756	.3769	.5473	.2946
8	.0159	.9470	.5213	.6767	.3633	.5329	.2858

Table 5. Fuzzy clustering results on second artifical data set (4 clusters). All measures were computed with the minimum for the t-norm(s).

	partition matrix			coincidence matrix			
#	diff	acc	F1	Rand	Jacc.	Folkes	Hubert
2	.1135	.4729	.6971	.5968	.2925	.4470	.1299
3	.0337	.6125	.7659	.7745	.2580	.4057	.2667
4	.0066	.7224	.8618	.8709	**.3140**	**.4768**	.4033
5	**.0022**	.7636	**.8781**	.9076	.3081	.4709	**.4203**
6	.0109	.7663	.7036	.9299	.2365	.3820	.3449
7	.0122	.7838	.6049	.9477	.2008	.3340	.3068
8	.0103	**.8030**	.5652	**.9602**	.1786	.3024	.2820

Table 6. Fuzzy clustering results on fourth artifical data set (5 clusters). All measures were computed with the Łukasiewicz t-norm to combine the membership degrees and the product to combine the coincidence matrix entries.

These result tables show that one has to be very careful when choosing the measure and the t-norm(s), since a lot of combinations fail miserably. However, there are also a lot of combinations that work very nicely. Especially the Hubert index, which appears to be fairly robust w.r.t. the choice of the t-norms yields excellent results if either the Łukasiewicz t-norm or the nil-potent minimum are chosen to combine the membership degrees. (The t-norm used to combine the membership degrees is stated in the second header row.) This behavior is almost independent of the t-norm that is used to combine the coincidence matrix entries. All other coincidence matrix based measures seem to have problems with the wine data set (see below for a possible explanation).

Among the partition matrix based measures the newly introduced simple mean squared difference comparison performs fairly reliably, followed by the accuracy computed with the minimum as the t-norm. However, none of these measures quite reaches the performance of the properly parameterized Hubert index. Therefore the Hubert index seems to be the best choice.

To give an impression of individual results, Tables 5 to 8 show detailed tables for two artificial data sets and the wine data set. The results in Tables 6 and 8 are based on Gustafson–Kessel clustering [9], the rest on fuzzy c-means clustering. The used t-norms are indicated in the table captions. For each column the global and, if it exists, a relevant local optimum are highlighted.

The results on the wine data set (Table 7) indicate that maybe five clusters are an alternative to the number of classes (three). However, this may also be explained by ellipsoidal cluster shapes. The results shown in Table 8 make this likely, as here no local optima can be observed for five clusters.

	partition matrix			coincidence matrix			
#	diff	acc	F1	Rand	Jacc.	Folkes	Hubert
2	.0102	.7747	**.7668**	.7007	**.5566**	**.7139**	.4009
3	**.0013**	.8539	.7689	.8176	.5489	.7091	.5770
4	.0244	.8180	.6032	.8232	.4200	.5878	.4761
5	**.0056**	**.8669**	**.6409**	.8753	**.4345**	**.6049**	**.5313**
6	.0125	.8655	.5556	.8921	.3463	.5129	.4525
7	.0115	.8760	.5039	.9124	.3174	.4813	.4337
8	.0133	**.8837**	.4510	**.9244**	.2874	.4463	.4060

Table 7. Fuzzy clustering results on the wine data set (3 classes), processed with fuzzy c-means clustering. All measures were computed with the nil-potent minimum for the t-norm(s).

	partition matrix			coincidence matrix			
#	diff	acc	F1	Rand	Jacc.	Folkes	Hubert
2	.0054	.7395	**.9581**	.7321	**.5419**	**.7023**	.4589
3	**.0037**	**.7695**	.9305	.8109	.4561	.6262	**.4997**
4	.0260	.7153	.7099	.8430	.2819	.4388	.3477
5	.0231	.7471	.6421	.8794	.2344	.3789	.3123
6	.0254	.7730	.5537	.9046	.2078	.3433	.2921
7	.0279	.7891	.4587	.9225	.1683	.2883	.2477
8	.0285	.8092	.4075	**.9328**	.1442	.2534	.2187

Table 8. Fuzzy clustering results on the wine data set (3 classes), processed with Gustafson–Kessel clustering. All measures were computed with the nil-potent minimum for the t-norm(s).

5 Conclusions

In this paper I transferred resampling ideas that have been used in classical crisp clustering to fuzzy clustering and introduced the mean squared difference as a simple, but effective measure for comparing fuzzy and probabilistic partition matrices. In addition, I explored the influence of different t-norms, which can be used to combine membership degrees and coincidence matrix entries. As the experiments show, the resampling approach is applicable to fuzzy clustering, but one has to be careful which relative cluster evaluation measure to choose and how to parameterize it: not all measures that work with crisp clustering also work with fuzzy clustering. The best results are obtained with the Hubert index, parameterized with either the nil-potent minimum or the Łukasiewicz t-norm to combine the membership degrees. A close competitor, which has the advantage of being simple and straightforward, is a direct comparison of the partition matrices based on the mean squared difference.

References

1. J.C. Bezdek and N. Pal. *Fuzzy Models for Pattern Recognition.* IEEE Press, New York, NY, USA 1992
2. J.C. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.* Kluwer, Dordrecht, Netherlands 1999

3. C.L. Blake and C.J. Merz. *UCI Repository of Machine Learning Databases.* University of California, Irvine, CA, USA 1998
http://www.ics.uci.edu/˜mlearn/MLRepository.html
4. C. Borgelt. *Prototype-based Classification and Clustering.* Habilitation thesis, University of Magdeburg, Germany 2005
5. A.P. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society (Series B)* 39:1–38. Blackwell, Oxford, United Kingdom 1977
6. B. Efron and R.J. Tibshirani. An Introduction to the Boostrap. Chapman & Hall, London, United Kingdom 2003
7. B.S. Everitt and D.J. Hand. *Finite Mixture Distributions.* Chapman & Hall, London, United Kingdom 1981
8. P. Good. *Resampling Methods.* Springer-Verlag, New York, NY, USA 1999
9. E.E. Gustafson and W.C. Kessel. Fuzzy Clustering with a Fuzzy Covariance Matrix. *Proc. of the IEEE Conf. on Decision and Control (CDC 1979, San Diego, CA)*, 761–766. IEEE Press, Piscataway, NJ, USA 1979. Reprinted in [1], 117–122
10. M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering Validity Checking Methods: Part I. *ACM SIGMOD Record* 31(2):40–45. ACM Press, New York, NY, USA 2002
11. M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering Validity Checking Methods: Part II. *ACM SIGMOD Record* 31(3):19–27. ACM Press, New York, NY, USA 2002
12. J.A. Hartigan and M.A. Wong. A k-means Clustering Algorithm. *Applied Statistics* 28:100–108. Blackwell, Oxford, United Kingdom 1979
13. F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis.* J. Wiley & Sons, Chichester, England 1999
14. A.K. Jain and J. Moreau. Bootstrap Technique in Cluster Analysis. *Pattern Recognition* 20:547–569. Pergamon Press, Oxford, United Kingdom 1986
15. M.H.C. Law and A.K. Jain. *Cluster Validity by Bootstrapping Partitions.* Technical Report MSU-CSE-03-5, Dept. of Computer Science and Engineering, Michigan State University, Michigan, , USA 2003
16. E. Levine and E. Domany. Resampling Method for Unsupervised Estimation of Cluster Validity. *Neural Computation* 13:2573–2593. MIT Press, Cambridge, MA, USA 2001
17. S. Lloyd. Least Squares Quantization in PCM. *IEEE Trans. on Information Theory* 28:129–137. IEEE Press, Piscataway, NJ, USA 1982
18. C.H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization, Algorithms and Complexity.* Prentice-Hall, Englewood Cliffs, NJ, USA 1982
19. C.J. van Rijsbergen. *Information Retrieval.* Butterworth, London, United Kingdom 1979
20. V. Roth, T. Lange, M. Braun, and J.M. Buhmann. A Resampling Approach to Cluster Validation. *Proc. Computational Statistics (CompStat'02, Berlin, Germany)*, 123–128. Springer-Verlag, Heidelberg, Germany 2002

On the Visualization of the Soft Computing Knowledge Domain*

Nees Jan van Eck, Ludo Waltman, Jan van den Berg, and Uzay Kaymak

Faculty of Economics, Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
{nvaneck,lwaltman,jvandenberg,kaymak}@few.eur.nl

1 Introduction

Knowledge domain visualization is concerned with the construction of maps that visualize the structure and the evolution of a field of science. In this paper, we apply a knowledge domain visualization approach to the field of soft computing (SC). The focus of the paper is on so-called concept maps. These maps visualize the associations between concepts in a scientific field.

2 Analysis

We constructed and analyzed a concept map that is based on the abstracts of the papers that have been accepted for presentation at the 2006 IEEE World Congress on Computational Intelligence (WCCI 2006). The WCCI 2006 is a joint conference of the 2006 International Joint Conference on Neural Networks (IJCNN 2006), the 2006 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2006), and the 2006 IEEE Congress on Evolutionary Computation (CEC 2006). The concept map is shown in Fig. 1 and can be examined more closely using the concept map viewer that we have made available online (see http://people.few.eur.nl/nvaneck/wcci2006/). In the concept map, concepts are located in such a way that the distance between two concepts reflects the strength of their association as accurately as possible. The importance of a concept is indicated by the size of its label, and the distribution of the interest in a concept over the subconferences of the WCCI is indicated by the color of the concept label (red = FUZZ-IEEE, green = IJCNN, blue = CEC). In addition to the concept map, a concept density map was constructed. In the concept density map, colors are used to indicate the density of concepts (blue = low density, red = high density). The concept density map is shown in Fig. 2.

Structure of the SC Field. Based on Fig. 1 and 2, a number of observations concerning the SC field can be made. The well-known division of the field into the neural networks, fuzzy systems, and evolutionary computation subfields indeed appears in the concept map. Especially when concept densities are taken into

* The full paper on this topic has been published elsewhere [1].

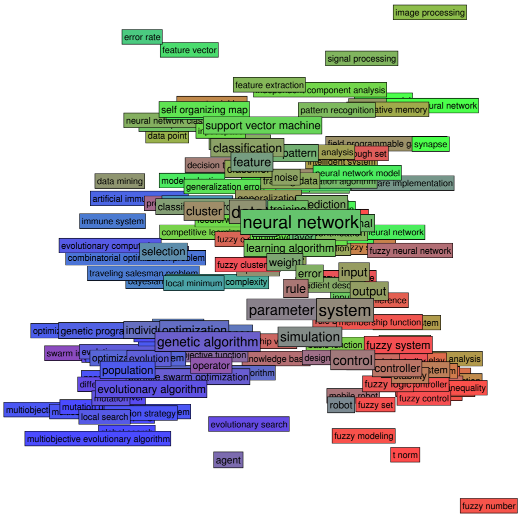


Fig. 1. Concept map calculated based on the WCCI 2006 abstracts.

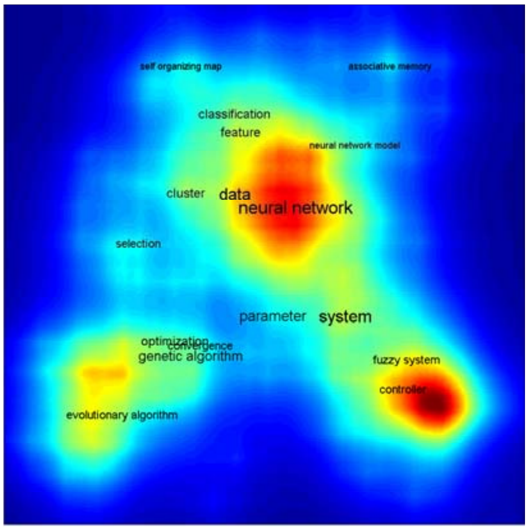


Fig. 2. Concept density map calculated based on the WCCI 2006 abstracts.

account (Fig. 2), the three subfields are clearly visible. This observation may be regarded as a justification for the existence of three separate IEEE conferences in the SC field. The concept map further shows that the neural networks and fuzzy systems subfields are less separated from each other than from the evolutionary computation subfield. So, the evolutionary computation subfield appears to have a relatively independent position within the SC field.

Developments in the SC Field during the Last Years. We also analyzed the main developments in the SC field during the last four years. Comparing the concept maps of the WCCI 2006 and the WCCI 2002 (not shown here), it turns out that the relations between the main concepts in the SC field have been fairly stable during the last years. However, the evolutionary computation subfield has become more independent from the rest of the SC field. Based on a comparison of the abstracts of the WCCI 2006 and the WCCI 2002, the four most significant recent developments in the SC field have been identified. These developments are the increased interest in the areas of differential evolution, multiobjective evolutionary computation, particle swarm optimization, and support vector machines. Three out of the four developments have taken place in the evolutionary computation subfield. In a certain sense, the evolutionary computation subfield therefore appears to have been the most innovative subfield during the last years.

References

[1] van Eck, N.J., Waltman, L., van den Berg, J., Kaymak, U.: Visualizing the WCCI 2006 knowledge domain. In: Proceedings of the 2006 IEEE International Conference on Fuzzy Systems. (2006) 7862–7869

Fuzzy Clustering Techniques in Autonomous Sensor-Based Landing Systems

Bernd Korn
German Aerospace Center (DLR)
Institute of Flight Guidance, Braunschweig
email: bernd.korn@dlr.de

Abstract. Enhanced Vision Systems (EVS) are currently developed with the goal to alleviate restrictions in airspace and airport capacity in low-visibility conditions. The demand for all-weather flight operations becomes most important for the approach and landing phase of a flight, when safety concerns resulting from low visibility require the nominal airport capacity to be significantly reduced. EVS relies on weather- penetrating forward-looking sensors that augment the naturally existing visual cues in the environment and provide a real-time image of prominent topographical objects that may be identified by the pilot. Infra-red (IR) and millimetre-wave (MMW) sensors are currently envisaged as the most promising EVS support of pilot vision in low visibility. The recently released final rule of the FAA for Enhanced Flight Vision Systems (EFVS) clearly acknowledges the operational benefits of such a technology by stating the following: Use of an EFVS with a head-up display (HUD) may improve the level of safety by improving position awareness, providing visual cues to maintain a stabilized approach, and minimizing missed approach situations. Although MMW sensors being remarkably better in penetration of bad weather than IR sensors the MMW data are difficult to interpret by the human being and pilots are in general not able to derive the necessary navigation information such as the aircraft's position relatively to the runway directly from the radar image.

In this contribution a system is presented which provides such navigation information primarily based on the analysis of millimetre wave (MMW) radar data. The core part of the presented system is a fuzzy rule based inference machine which controls the data analysis based on the uncertainty in the actual knowledge in combination with a-priori knowledge. The quality of MMW images is rather poor and data is highly corrupted with noise and clutter. Therefore, one main task of the inference machine is to handle uncertainties as well as ambiguities and inconsistencies to draw the right conclusions. The output of different sensor data analysis processes are fused and evaluated within a fuzzy/possibilistic clustering algorithm whose results serve as input to the inference machine. The only a-priori knowledge used in the presented approach is the same pilots already know from airport charts which are available of almost every airport. The performance of the approach is demonstrated with real data acquired during extensive flight tests to several airports in Germany.

References

[1] B. Korn, Autonomous Sensor based Landing Systems: Fusion of Vague and Incomplete Information by Application of Fuzzy Clustering Techniques, *In: M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger and W. Gaul (eds.), From Data and Information Analysis to Knowledge Engineering - Proc. of 29th Ann. Conf. of the Gesellschaft für Klassifikation*, pp. 454-461, Springer-Verlag, Berlin, 2006

Interactivity Closes the Gap
Lessons Learned in an Automotive Industry Application

Axel Blumenstock[1], Jochen Hipp[2], Carsten Lanquillon[2], Rüdiger Wirth[2]

[1] Department of Applied Information Processing, University of Ulm, Germany
axel.blumenstock@uni-ulm.de
[2] DaimlerChrysler Group Research, Ulm, Germany
{jochen.hipp, carsten.lanquillon, ruediger.wirth}@dcx.com

Abstract. After nearly two decades of data mining research there are many commercial mining tools available, and a wide range of algorithms can be found in the literature. One might think there is a solution to most of the problems practitioners face. In our application of descriptive induction on warranty data, however, we found a considerable gap between many standard solutions and our practical needs. Confronted with challenging data, and requirements such as understandability and support of existing work flows, we tried many things that did not work, ending up in simple solutions that do. We feel that the problems we faced are not so uncommon, and would like to advocate that it's better to focus on inclusion of domain expert knowledge rather than on complex algorithms. Interactivity and simplicity turned out to be key features to success.

1 Introduction

An air bellow bursts: This happens on one truck, on another it does not. Is this random coincidence, or the result of some systematic weakness?

Questions like these have ever been keeping experts busy at DaimlerChrysler's After Sales Services. Recently, they have attracted even more attention, when Chrysler's CEO LaSorda introduced the so-called tag process: a rigorous quality enhancement initiative that once more stresses the enormous business relevance of fast problem resolution [1].

This primary goal of quality enhancement entails several tasks to be solved:

- detecting upcoming quality issues as early as possible
- explaining *why* some kind of quality issue occurs and feeding this information back into engineering
- isolating groups of vehicles that might suffer a certain defect in the future, so as to make service actions more targeted and effective.

Our research group picks up common data mining methods and adapts them to the practical needs of our engineers and domain experts. This contribution reports on the lessons learned. In particular, we elaborate on our experience that the right answer to domain complexity need not be algorithmic complexity—but rather simplicity. Simplicity opens ways to create an interactive setup which involves experts without overwhelming them. And if truly involved, an expert will understand the results and turn them into action.

We will outline the problem setting in Section 2. The subsequent sections respectively discuss the theoretical aspects, tool selection and model building methods, each answering the questions of what we tried and what finally worked.

2 Domain and Requirements

2.1 The Data

Most of the data at hand is warranty data, providing information about diagnostics and repairs at the garage. Further data is about vehicle production, configuration and usage. All these sources are heterogeneous, and the data was not collected for the purpose of quality mining. This raises questions about reliability, appropriateness of scale, and level of detail. Apart from these concerns, our data has some properties that make it hard to analyze, including

Imbalanced classes: The class of interest, made up of all instances for which a certain problem was reported, is very small compared to its contrast set. A typical proportion is 0.1 %.

Multiple causes: A single kind of problem report can sometimes be traced back to different causes that produced the same phenomenon. Partitioning the positive class makes it even more sparse.

Semi-labeledness: The counterpart of the positives is not truly negative. If there is a warranty entry for some vehicle, it is (almost) sure that it indeed suffered the problem reported on. For any non-positive example, however, it is unclear whether it carries problematic properties and may fall defective in near future.

High-dimensional space of influence variables (1000s)

Influence variables interact strongly: Some quality issues do not occur until several influences coincide. And, if an influence exists in the data, many other non-causal variables follow by showing positive statistical dependence with the class as well.

True causes not in data: By chance, they are concludable from other, influenced variables.

2.2 The Domain Experts and Their Tasks

Our users are experts in the field of vehicle engineering, specialized on various subdomains such as engine or electrical equipment. They keep track of what goes on in the field, mainly by analyzing warranty data, and try to discover upcoming quality issues as early as possible. If they recognize a problem, they strive for finding out the root causes in order to address it most accurately.

They have been doing these investigations successfully over years. Now, data mining can help them to better meet the demands of fast reaction, well-founded insight and targeted service. But any analysis support must fit into the users' mindset, their language, and their work flow.

The structure of the problems to be analyzed varies substantially. This task requires inspection, exploration and understanding for every case anew. Ideally, the engineers should be enabled to apply various exploration and analysis methods from a rich repository. And it is important that they do it themselves, because no one else could decide quickly enough whether a certain clue is relevant and should be pursued, and ask the proper questions. Finding out reasons of strange phenomena requires both comprehensive and detailed background knowledge.

Yet, the engineers are not data mining experts. They could make use of data mining tools out of the box, but common data mining suites already require deeper understanding of the methods. Further, the users are reluctant to accept any system-generated hypothesis if the system cannot give exact details that justify this hypothesis. The bottom line is that penetrability and, again, interactivity are almost indispensable features of any mining system in our field.

3 Understanding the Task

Let us first have a theoretical look at the problem. It is noteworthy that we will meet the following arguments again when we investigate individual methods.

3.1 What we tried

A great portion of the task can be seen as a classification problem. We would like to separate the good from the bad. It may be possible to tell for any vehicle whether it might encounter problems in the future. And if we choose a symbolic method, we can use the model to explain the problem.

But as stated above, data is semi-labeled, and the problem behind the positive class may have multiple causes. These properties act as if there were a strong inherent noise that changes the class variable in either direction. While classifier induction tries to separate the classes in the best possible way but returns unpredictable, arbitrary results when noise increases, it suffices for our application to grab the most explainable part of the positives and leave the rest for later investigation or, finally, ascribe it to randomness. In other words, we experienced that anything beyond *partial* description is seldom adequate.

So we came up with subgroup discovery. It means to identify *in any way* subsets of the entire object set which show some unusual distribution with respect to a property of interest. In our case, this property is the binary variable that a certain quality issue occurred.

Results from subgroup discovery approaches need not be restricted to "explaining" a class, but can be re-used for picking out objects of interest. This is the partial classification we want, where a statement about the contrast set is not adequate or required.

Still, data properties make subgroup discovery results unusable most of the time. There are many candidate influences, and they interact strongly. Therefore, even if the cause could be described by a sole variable, it would be hard to find it among the set of variables influenced by it otherwise. All these variables, including the causal one, would refer to roughly the same subset of vehicles with an increased proportion of positives.

3.2 What works

Subgroup *description* is to identify the very same subgroups in a way as comprehensive and informative as possible. In other words, even if subgroup *discovery* results are presented in a human-readable form, the users are left alone to map these results to synonyms that can be more meaningful in the context of the application. In a domain with thousands of influence variables, however, the users cannot be expected to bear all the (possibly even multivariate) interactions in their minds. Subgroup description is thus required to provide any reasonable explanation as long as there is no evidence that some explanation is void or unjustified.

4 A Tool that Suits the Experts

4.1 What we tried

We had a look at several commercially available data mining suites and tools. Unfortunately, any of these fell short of the requirements outlined in Section 2.2.

As an overall observation, they were rather inaccessible. Even if they allowed for interaction at the model building level, they could not present information like measures in a way suiting the experts. On the other hand, tools of this kind offer their methods in a very generic fashion so that the typical domain expert does not know where to start. In short, we believe that the goal conflict between flexibility and guidance can hardly be solved by any general-purpose application, where the greatest simplification potential, namely domain adaption, remains unexploited.

4.2 What works

We ended up in programming a tool of our own. Figure 1 shows a simplified view of our tool's process model. It emerged as the union of our experts' workflows und thus offers guidance even for users not overly literate in data mining. At the same time, it does not constrain the user to a single process but allows going deeper and gain flexibility whereever the user is able and willing to.

For example, the users start with extracting data for further analysis. We tried to keep this step simple and hide the complexities as much as possible. The

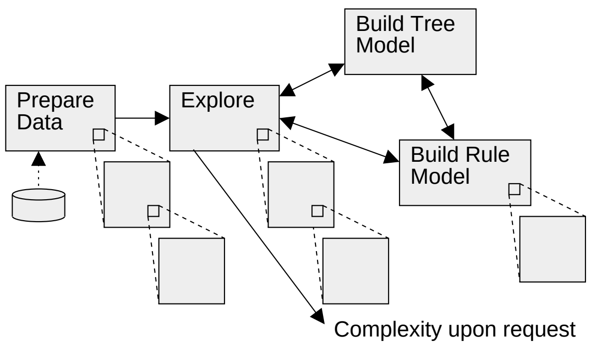


Fig. 1. Coarse usage model of our tool. There is a fixed process skeleton corresponding to the legacy workflow. The user can just go through, or gain more flexibility (and complexity) upon request.

user just selects the vehicle subset and the influence variables he likes to work with. A meta data based system cares about joins, aggregations, discretizations or other data transformation steps. Of course this kind of preprocessing is domain specific, but still flexible enough to adapt to changes and extensions.

In the course of their analyses, the experts often want to derive variables of their own. This is an important point where they introduce case-specific background knowledge. The system allows them to do so, up to the full expressiveness of mathematical formulas.

In a similar fashion, the system offers both standard reports, suiting the experts' needs in most of the cases, up to individually configurable diagrams. For the sake of model induction, our tool offers currently two branches that interact and complement each other: decision trees and rule sets.

5 Interactive Decision Trees

Subgroup discovery (and description) can be mapped to partitioning the instance set into multiple decision tree leaves. At least one tree path should represent a description of an interesting subgroup. In fact, decision tree induction roughly corresponds to what our experts had been doing even before getting in touch with data mining. Hence, decision trees were our first method we chose.

5.1 What we tried

To quickly provide the users with explanation models, it was proximate to build decision trees automatically as is typically done when inducing tree-based classifiers. However, the experts deemed the results unusable most of the time, because the split attributes that had been selected by any of the common top-down tree induction algorithm were often uninformative or meaningless to them: The top-ranked variable was seldom the factually most relevant one.

For some time, we experimented with different measures. Literature suggests measures such as information gain, information gain ratio, χ^2 p-value, or gini index, to mention the most important ones.

However, in an exemplary analysis case, the variable that gave the essential hint to the expert was ranked 27th by information gain, 41st by gain ratio, 36th by p-value and 33rd by gini index. We conclude that an automatic induction process hardly could have found a helpful tree.

5.2 What works

This is where interactivity comes into action. Building trees interactively relieves the measure of choice from the burden of selecting the single "best" split attribute. The idea is almost trivial: Present the attributes in an ordered list and let the expert make tentative choices until he finds one he considers plausible.

What remains is the problem of how to rank the attributes in a reasonable way. But even for ranking, the aforementioned statistical measures proved little helpful. We explain this by the fact that they are measures designed for classificator induction, trying to separate the classes in the best possible way. But as illustrated in Section 3, this is not the primary goal in our application.

Most of the time, we deal with two-class problems anyway: the positive class versus the contrasting rest. Hence, we can use the measure *lift* (the factor by which the positive class rate in a given node is higher than the positive class rate in the root node). To complement the lift value of a tree node, we use the *recall* of the positive class. Both lift and recall are readily understandable for the users as they have immediate analogies in their domain. Now, focusing on high-lift paths, the users can successively split tree nodes to reach a lift as high as possible while maintaining nodes with substantial instance counts.

In order to condense this into a suitable attribute ranking, we must group attribute values (along with the current node's children). We require the resulting split to create at most k children, where typically $k = 2$ so as to force binary splits. This ensures both that the split is "handy" and easily understood by the user, and that the subsequent attribute ranking can be based consistently on the child node with the highest lift.

To group the children in a reasonable way, we simply sort them by lift. Then, keeping their linear order, we cluster them using several heuristics: merge smalles nodes first, merge adjacent nodes with lowest lift difference. Lift and recall of the resulting highest-lift node are finally combined to a one-dimensional measure (weighted lift, or "explanational power") in order to create the ranking.

Grouping is automatically performed during attribute assessment. Still, the users can interactively undo and redo the grouping or even arrange the attribute values into any form that they desire. This is important to further incorporate background knowledge, e.g. with respect to ordered domains, geographical regions, or, in particular, components that are used in certain subsets of vehicles and should, thus, be considered together.

As an alternative to a ranked list, the user can still get the more natural two-dimensional presentation of the split attributes (Figure 2). Similar to within a

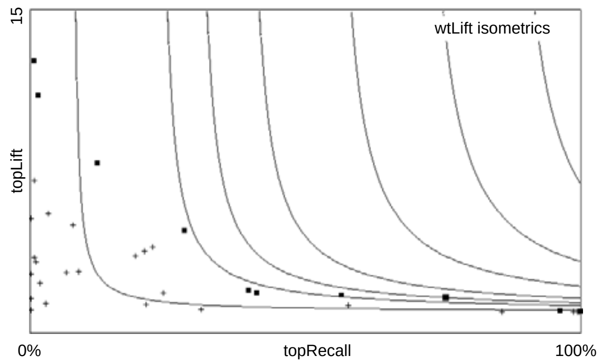


Fig. 2. Quality space for the assessment of split attributes. Each dot represents an attribute, plotted over recall (x axis) and lift (y axis) of the best (possibly clustered) child that would result. Dots are plotted bold if there is no other dot that is better in both dimensions. The curves are isometrics according to the coverage-weighted lift.

ROC space, every such attribute is plotted as a point. We use recall and lift as the two dimensions.

6 Interactive Rule Sets

As an important data property we mentioned that influences interact in a way that some quality issues do not occur until several influences coincide. While decision tree building is intuitive, its search is greedy and thus may miss interesting combinations. So the experts asked for an automatic, more comprehensive search. This led us to rule sets.

6.1 What we tried

A well-known subgroup discovery algorithm is CN2-SD [2]. It induces rules by sequential covering: By heuristic search, find a rule that is best according to some statistical measure. Reduce the weights of the covered examples, and re-iterate until no reasonable rule can be found any more.

The first handicap of this procedure is the same as with decision trees: There is no measure that could guarantee to select the best influence, here: rule.

But even the hope that a good rule will be among the subsequently mined ones need not hold: Imagine there are two rules describing exactly the same example set. CN2-SD will never find both, because by modifying the examples' weights, the two rules' ranks will change simultaneously. This however runs counter the idea of subgroup *description*, in other words, comprehensiveness at the textual level rather than mere subset identification.

6.2 What works

We thus came up with an exhaustive search (within constraints). It is realized by an association rule miner with fixed consequence. This is far from being new, and like us, many research groups think about how to handle redundancy within the results.

What we like to point out here is that once again, the idea of interactivity produced a simple but effective solution. The expert is enabled to control a CN2-SD like sequential covering. He picks a rule he recognizes as "interesting" or "already known". This is comparable to selecting a decision tree split attribute. Several measures, fitting into his mindset, support him with his choice. The instance set is then modified so as to remove the marked influence, and the expert can re-iterate to find the next interesting rule.

7 Module Interaction

The key property that makes a tool more than the sum of its components, however, is the facility of module interaction. This is still only partly implemented, but our users strongly request for it. Indeed it is the feature that allows them to flexibly apply the methods offered and to take out the respective best of them.

Such sometimes trivial but practically important features include:

- Extracting instance subsets as covered by a rule or tree path and exchanging them within the modules for deeper analyses or visualization.
- Building a tree with a path as described by a rule in order to take a closer look at the respective contrast sets.
- Deriving new variables from tree paths or rule antecedents.

8 Conclusion

We reported on our experiences of applying data mining methods in a domain where data is difficult, analysis tasks change structurally case by case, and thus a great amount of background knowledge is indispensable. Many approaches suggested in the literature turned out either too constrained or too complex to be offered without major adaption. In such a setting, we consider it best to stick to simple methods, provide these in a both flexible and understandable way, and settle on interactivity.

Still, there is a wide field to explore. At many points of the process, methods are needed that support the experts and reduce their routine work load as much as possible.

References

1. M. Connelly. Chrysler's LaSorda on quality: Fix it now. *Automotive News*, May 9th 2005.
2. N. Lavrač, P. A. Flach, B. Kavšek, and L. Todorovski. Rule induction for subgroup discovery with CN2-SD. In *ECML/PKDD'02 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, 2002.

Computer Vision Controlled Autonomous Mobile Working Machines

Alexander Stolpmann

University of Applied Sciences Braunschweig/Wolfenbüttel
Institute for Applied Computer Science
stolpmann@fh-wolfenbuettel.de
hochschule@artificial-vision.de

Abstract. This article describes the use of camera based systems to control autonomous ground conveyors and mobile working machines. Next to the modern approach towards intelligent automated guided vehicles the use of computational intelligence such as neural networks, fuzzy systems, and evolutionary computation is highlighted. An outlook will be given what can be expected in the near and far future and how it will influence society.

1 Introduction

In the course of ongoing automation in all industrial areas and everyday life the demand for complex control systems is increasing. A lot of routine work still has to be done by individuals as the requirements for interactivity with the surroundings are high.

In many areas it would be beneficial, either in terms of safety or monetary aspects, to automate machine and material movement. Therefore systems are needed that can take over the tasks done by humans. To reach highest possible flexibility such systems shall need no different surroundings than what is available for a human operator. In most applications a mix between manual and automated systems will occur.

2 Ground Conveyors and Mobile Working Machines

Ground conveyors (Fig.1), also called industrial trucks, are the group of vehicles used for the non-continuous transport of goods in a working environment. They are not bound to railtracks, run on wheels[1] and are steerable. Best known is the fork-lift[2] and the airfield tug, but vehicles come as small as low-level order pickers and as large as container stackers.

Mobile working machines (Fig.2) are comparable to the ground conveyors. They are not used for the transportation of goods, but are being driven to the

[1] Wheels are most common, but crawler-tracks, hover cushion, or even legs are possible, too.
[2] Counterbalanced truck

(a) Fork-Lift (b) Airfield Tug (c) Reacher Stacker

Fig. 1. Ground Conveyors

place of work. Best known are road-rollers, street-sweepers, or excavators, but here, too, sizes and usage vary considerably. Among the rare specimen is the guardrail post setting machine as mentioned below or the huge bucket-wheel excavators used for open-pit mining.

(a) Road Roller (b) Street Sweeper (c) Excavator

Fig. 2. Mobile Working Machines

2.1 Reason for Automation

Most of the above mentioned vehicles are being used in an enclosed environment. This has the benefit that the boundary conditions are known and that special behaviour rules can be imposed on the work force. In many cases it is technically, logistically or financially not viable to adapt the constraints to the mobile machines, often the only givens are the availability of an even track and space to manoeuvre. All other conditions, like traffic, objects lying in the way, pedestrians, or changing light conditions, have to be handled by the operator.

As humans are very versatile, but never perfect, they are less well suited for repetitive tasks. Especially distraction or fatigue are reasons for mistakes, which in the case of moving machines can lead to accidents. Since the start of the industrialisation machines took over perseverative tasks, at first the automatic loom, later more complex machines like welding robots.

The very high demands for personal safety paired with the need for low costs asks for the automation of vehicles. A single person can overlook a number of autonomous vehicles from a safe and worker friendly place.

Additional areas that are well suited are those that are prohibitive for humans due to heat, contamination, noise or just size[3] or distance[4].

2.2 Surroundings

What are the surroundings where the automation of the control of vehicles can take place? In short: anywhere.

Saying this it has to be said that the available computational power is far from adequate. Also has the software not been developed to the extend needed for the use in any imaginable setting. But this again is true for a human driver, too.

For a start areas are chosen with well defined conditions. As long as all the possible occurrences have been classified with an appropriate action assigned, the demands for the control system have been met.

2.3 Limitations

The above said is correct and will work well in theory. The real life is the problem. In most cases the environmental complexity is by far too large to handle it with a straight forward control concept. Intelligent solutions are needed.

The main problem is the handling of changed conditions. As independent of concept any proposed system needs sensory data it is a question of which information can be gathered through these sensors and can the gained data be classified.

Two examples to clarify the problems. A very simple driverless vehicle uses a mechanism that steers to a new direction after a given distance. The distance is being measured by the rotation of the wheels. The first and obvious problem pose objects lying in the path. A second problem is wheel-slip. As this simple system is blind against the surrounding it will be doomed to fail through the simplest problems. But even a system with highly developed sensory data, i. e. one that is using cameras, can be fooled. Cameras, just like the human eye and brain, can detect objects by measuring contrast. A simple sheet of paper can, if only a single image is used for classification, be identified as a hindering object. The same can happen through sunlight. Shadows cast by the sun have a much higher contrast than a white line on tarmac. A camera alone will not be enough if the following software cannot handle the complexity of the received data.

[3] Sewers

[4] Planet Mars

2.4 Developments

The field of automated guided vehicles (AGV), a term used for mainly in-house driverless ground conveyors, is only emerging. Available AGVs use either wired or wireless sensors for navigation.[5]

Fig. 3. Wireless AGV

Wired systems need a radio frequency transmitting wire in the ground which the vehicle will follow. It is not a free moving vehicle and circumnavigating an obstacle is not within the realms of its possibilities. Avoiding a collision has to be handled by other means, usually a soft emergency stop bumper[6].

The wireless navigation is done by applying reflective targets along the path which are detected by a laser scanner. From the received data the position can be calculated. Such a machine is to a certain extent free where it will drive within a path. Additionally the targets are cheaper and easily repositioned. Still it is a very limited approach.

2.5 New Approach

The new approach by the author is the very old one used by every human: vision.

There are two good reasons for using vision:

- Images are very detailed and hold all the needed information.
- Humans, and this includes the scientists that develop visual systems, use their eyes for at least 70% of sensory data and therefore know vision best. Only the transfer of this knowledge to a technical system is complex and has to be solved.

[5] For easily accessed further information start here: [1, 6].
[6] Well known from lift doors.

In the long term all vehicles, including public and privat transport, will use visual systems for assistance or complete control.

3 Concept

At the time of writing two systems designed and made by the author exist. The first is a camera based steering control system for a guardrail post setting machine (Fig. 4). This machine positions and rams posts into the hard shoulder of streets and motorways. Crash barriers are afterwards attached to the posts. The machine works fully automatically and only needs a chalked out line for guidance.[7]

(a) Machine (b) Camera Control System

Fig. 4. Guardrail Post Setting Machine with Camera Based Control System

The second system, a camera guided autonomous forklift (CamGAF), is used for research purposes.[12, 15, 16] A model of a forklift, 3/4" scale, has been chosen to allow testing of new algorithms in the laboratory (Fig. 5). Such a model does not have enough space for an on-board computer. Therefore the sensor and actuator signals are wirelessly transmitted to and from a PC. The sensors consist of three cameras, two with wide angle lenses for monitoring the surroundings, one pan and tilt camera for detailed images. Image processing routines on the PC continuously interpret the image content and decide on how to steer the forklift to

[7] Details are classified.

keep it inside the track, avoid an obstacle or pick up a pallet. The automatically generated control sequences are transmitted to the vehicle through a wireless remote control. The research project provides evidence of how computer vision can be used for the control of autonomous vehicles and in driver assistance systems.

Fig. 5. Camera Guided Autonomous Forklift (CamGAF)

Through the modular layout of the system it is possible to include and verify new approaches, especially in image processing, classification, interpretation, and control. The self-optimisation of those modules and their parameters is also possible.

3.1 Description of Modules

The system consists of control circles. There is the outer, more visible circle which contains all the necessary hardware, and the inner software circle.

Starting with the light that is being reflected or radiated from the objects that span the 3-dimensional space for the vehicle. This light, containing all the information needed to control the vehicle, is gathered by the main sensor, the camera. After passing the optics the light is transformed into electrical signals by a CCD or CMOS chip. These signals are being transferred, wired or wirelessly, to a digitiser and then on to a computer. What kind of computer, embedded or PC-based, is being used is facultative and depends on the constraints given by the environment in which the machine is use and on the needs of the software. After the software, more about that below, has reached a decision, the action, mainly steering and speed, is transmitted to the actuators of the machine. This again can be done wired or wirelessly and might need additional hardware for the process.

As the CamGAF is rather small the image data processing is performed off-board on a standard computer. The video signal is being transmitted wirelessly (5.8 GHz) to a video receiver which is coupled to the PC by means of a digitiser. The control signals send out from the PC are pass through an embedded computer to a radio remote control unit. From there the signals are radioed (40 MHz) to an on-board receiver which controls the servos and thus the movement of the forklift.

The software running on the computer has to handle a number of tasks. An image processing module is used for image preparation and object detection. Depending on the properties of the image high-level statistical methods could be applied for image segmentation.[14, 21] The classification, too, depends on the complexity of the image contents. If the information within the image is unpretentious a simple blob analysis or template match will provide the desired details. The following classification and interpretation is just as straight forward. With growing image entropy the complexity of the analysis has to be matched. The classification modules will use methods like fuzzy clustering [7, 8, 19] or neural networks [3]. It has to be said that such methods are work intensive, especially at the setup stage. The long term desired goal is a self-learning system.

With growing complexity of the software system and growing number of modules involved the manual handling, e. g. setting parameters, it no longer viable. Automatic optimisation using genetic algorithms is a way to handle such problems.[13, 14, 18, 20] This is used to feed back information among the modules to gain better performance (cf. Fig. 6).

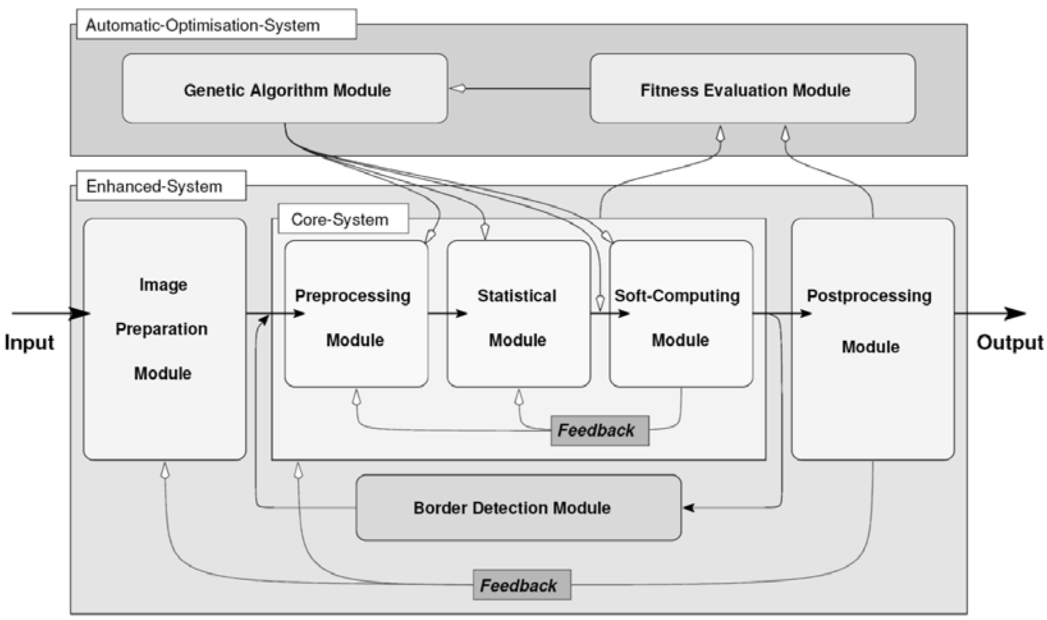


Fig. 6. Software Modules with Automatic Optimisation

3.2 On-Board vs. Off-Board Computing

The CamGAF utilises off-board computing while the post setting machine uses embedded controllers. It is not generally possible to define which layout is better or worse. The Tab. 1 lists the pros and cons for both systems.

Table 1. The Pros and Cons

On-Board Computing	Off-Board Computing
completely independent	machine has to be within reach
relatively small size	very small on-board, large off-board (PC)
higher costs	PC is standard equipment
no extra costs for data transfer	extra transmitting equipment needed
higher energy consumption	low on-board energy consumption
software development less user-friendly	easy software development
limited choice of hardware	large hardware variety
good for mass production	good for laboratory and scientific work

For future systems it can be useful to mix both approaches, i. e. having an on-board control system which switches to off-board if higher processing power is needed. Off-board systems can also double as manual remote control systems. A single operator can thus control a large number of vehicles, as a direct intervention is only needed for unforeseen occasions.

4 The Use of Soft-Computing for Autonomous Vehicles

In most cases there is a demand for high computational intelligence. At the same time systems need to be easily programmable or customisable, either by an operator, or by a self-learning system.

As the system replaces a human driver the driver's knowledge and capabilities have to be utilised. This can be archived using a soft-computing approach.

In many areas of image processing uncertainty can be handled through the use of fuzzy systems. There is the greyness ambiguity in the early stages, the geometrical fuzziness, and, if used, statistical fuzziness, and the uncertain knowledge during the analysing and interpretation of the data. Using crisp data means loosing information which can be important at later stages.[14, 22]

As the number of possible approaches is large and the quality of results depend on the properties of the initial data, i. e. the image captured by the camera, it is still highly case dependent which modules are used for a specific task. In the future, when computational power is no longer a mayor restraint, the self optimising within the system will automatically choose the appropriate software modules. Such software systems will be dynamic with specialised solutions for individual tasks.

In the following some fuzzy approaches towards vehicle and traffic control shall be elucidated.

4.1 Fuzzy Speed and Fuzzy Steering

The well known fuzzy control mechanism can be applied for the speed control and for steering. Even though it would be possible to calculate an exact speed and steering angle, this is not necessarily beneficial. The image contents is vague in itself. This is due to deficiencies in optical properties, sensor resolution and dynamics, and image noise to name but a few. Additionally it is not necessary to measure the position of a border line or an obstacle, but to know where about it is. This reduces the calculatory load and eases the definition and handling.

The classification modules supply fuzzified data which then uses a fuzzy rulebase to obtain a result. This approach makes it very simple to add or change rules. This can be used to include information about the load carried, track conditions, light conditions or even if people are likely to cross the path.

Additionally this will help to handle mechanical hysteresis. Thus it can be avoided that through constant readjustment of speed or steering angle the mechanical and electrical parts are unnecessary stressed.

4.2 Fuzzy Moving Object Positioning

Next to the control of the vehicle itself the surrounding has to be addressed. As by definition the track is not known before it is being traversed, but only to a certain extend its properties, it is not a simple task to identify other moving objects. If the surroundings would be known in detail the contained static objects could be subtracted and all what is left is a possible obstacle. In simple cases possible objects may be defined through shape or colour. But as life is not that simple and the proposed systems shall be used next to human driven machines approaches with higher computational intelligence have to be applied.

Possible solutions which are being tested are the use of image series [5] and the use of stereo camera systems [4].

Here, too, the exact identification of objects, i. e. position, size, shape, etc., is not necessary. It is enough to derive a classification into moving speed and moving direction classes. These can be used as fuzzified input data for the speed and steering control routines mentioned above.

5 Other Uses and Additional Needs

Next to autonomous mobile working machines a vision based control system can also be used for driver assistance systems. In the beginning such systems will help the driver to avoid dangerous situations. This may be as simple as an ultrasonic based parking assistance, which is available for every modern car, or lane crossing detectors available for some trucks and few cars. Such approaches are still rather simple compared to the possibilities opened by the use of camera based systems.

For the development of integrated system, i. e. on-board intelligent cameras that handle all vehicle control tasks, better and more flexible development tools are needed. A project has been started to utilise the possibilities of PC-based graphical programming tools for computer vision. During the experimental stages the PC system is used and later the software structure is being transferred onto the intelligent camera, using the cameras own optimised function libraries.[9]

5.1 Driver Assistance Systems

Almost every car producer is working on driver assistance systems.[4, 17] At the moment camera based systems still play a minor role. This is due to the involved complexity and the missing knowledge in this field. Too few scientists have addressed this field so far, which will change industrial and community life considerably. Not only the recent demand for intelligent camera surveillance systems underlines this.

With ongoing development of the driver assistance systems more and more tasks will be taken over by the system. This is of course always a question of acceptance, but other electromechanical assistance systems, like anti-lock braking systems (ABS) or electronic stability control (ESC), have long been accepted despite first scepticism[8].

5.2 Lane Detection

To keep a vehicle within a lane the borders have to be detected. This is the easiest was to identify a lane.[2, 12, 15, 16] While the vehicle is moving with low speed the detection can be done with high precision close to the vehicle. The above mentioned post setting machine uses such a system, as the maximal speed is about 3 kph. The CamGAF has a higher speed and uses a forward looking camera position, as a considerable distance is being travelled between frames. This layout allows realistic speeds, i. e. like those used by human drivers, even in confined areas.

Cars again are able to drive at even higher speeds. Depending on which frame rate per second is possible a car may has covered a distance of its own length when driving fast. As shown in [11] an number of frames are needed to calculate lane borders with high reliability. This means that the system needs to obtain information from the road far ahead of the vehicle. To handle the variabilities of road situations, e. g. rain, snow, darkness, sunshine with shadows, worn out markings, etc., a complex image processing and classification system is needed. The above mentioned project uses a special clustering algorithm for classification and is able to detect lane borders in all conditions without change of parameter settings.

5.3 Distance Control

The distance or interval control is important to keep enough space for breaking in an emergency. In the long term communication among cars will avoid such

[8] Mainly among the writing guild.

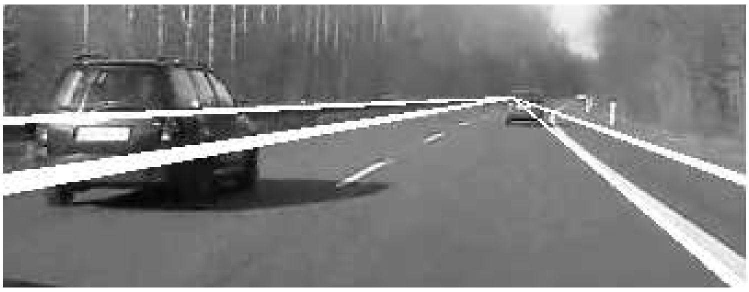


Fig. 7. Lane Border Detection

dangers, but for the time being assistance systems have to measure the distance just like a human driver does. Some executive class cars use radar for this task, but being an active system this has its draw backs. A camera system works passive, like the human eyes. It was shown in [5] that it is possible to estimate the distance and the relative movement of other cars on a motorway by just using a monocular camera.

Fig. 8. Vehicle Distance Control

The equipment needed was a standard video camera to capture the images. The calculation was performed on a PC in real time.

These two together with other solutions will form future driver assistance and self-driving systems.

6 Summary

This article gives an insight into the emerging technology of intelligent vision controlled autonomous vehicles. A number of subjects, that are still part of ongoing research, have been addressed and ideas for solutions were presented. Soft-computing is among the important techniques for classification and control tasks, as well as for self-optimisation of the system.

Self-driving vehicles, either in a working environment or among normal street traffic, will take an important role in the near future. It is estimated that within less than two decades all vehicles will have the capability of driverless movement.

Before that fully automated ground conveyers and material handling will be in widespread use in the industry. Additional needs will arise, like for example the theoretically well known truck-and-trailer problem [10, 23, 24], and be solved, more often than not, with the aid of soft-computing and camera based systems.

References

1. Automated Guided Vehicle. URL of the English Wikipedia entry on Automated Guided Vehicle (AGV). `http://en.wikipedia.org/wiki/Automated_Guided_Vehicle`.
2. Claus Bammann. Entwicklung eines Selbstfahrsystems basierend auf Microcontrollern. Master's thesis, Fachhochschule Braunschweig/Wolfenbüttel, Wolfenbüttel, to be published in 2006.
3. Gail A. Carpenter, Stephen Grossberg, Natalya Markuzon, John Reynolds, and David B. Rosen. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, 3(4):698–713, September 1992.
4. Sabine Eisentraut. Stand der Forschung und Entwicklung zu Fahrerassistenzsystemen. Master's thesis, Fachhochschule Braunschweig/Wolfenbüttel, Wolfenbüttel, 2005.
5. Niko Felekidis. Erkennung und Verfolgung von Kraftfahrzeugen in Bewegtbildern durch Bildverarbeitung sowie die Klassifizierung der Bewegungsrichtung. Master's thesis, Fachhochschule Braunschweig/Wolfenbüttel, Wolfenbüttel, 2004.
6. Fahrerloses Transportfahrzeug. URL of the German Wikipedia entry on Fahrerloses Transportfahrzeug (FTF). `http://de.wikipedia.org/wiki/Fahrerloses_TransportfahrzeugFTF`.
7. Frank Höppner, Frank Klawonn, and Rudolf Kruse. *Fuzzy-Clusteranalyse: Verfahren für die Bilderkennung, Klassifikation und Datenanalyse*. Vieweg Verlag, Braunschweig, Wiesbaden, 1997.
8. Leonard Kaufmann and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York, 1990.
9. Michael Klein. Entwicklung einer grafischen Programmierumgebung zur automatischen Generation von Quelltext für eine Intelligente Kamera. Master's thesis, Fachhochschule Braunschweig/Wolfenbüttel, Wolfenbüttel, to be published in 2006.
10. Seong-Gon Kong and Bart Kosko. Adaptive fuzzy systems for backing up a truck-and-trailer. *IEEE Transactions on Neural Networks*, 3(2):211–223, March 1992.
11. Dario Matos. Erkennung von Fahrbahnrändern in Bewegtbildern durch Bildverarbeitung. Master's thesis, Fachhochschule Braunschweig/Wolfenbüttel, Wolfenbüttel, 2004.
12. Nils Rückemann. Ein bildgesteuertes Selbstfahrsystem für ein Flurförderzeug. Master's thesis, Fachhochschule Braunschweig/Wolfenbüttel, Wolfenbüttel, to be published in 2006.
13. Eberhard Schöneburg, Frank Heinzmann, and Sven Feddersen. *Genetische Algorithmen und Evolutionsstrategien*. Addison-Wesley (Deutschland), Bonn, 1. edition, 1994.
14. Alexander Stolpmann. *An Intelligent Soft-Computing Texture Classification System*. PhD thesis, University of Glamorgan, Pontypridd, UK, 2005.
15. Alexander Stolpmann. Durch Bildverarbeitung gesteuertes selbstfahrendes Flurförderzeug. In *Wissen schaffen. Zukunft gestalten. Forschungsland Niedersachsen*. 2006.

16. Alexander Stolpmann. Selbstfahrender Gabelstapler – gesteuert durch Bildverarbeitung. Technical report, ti – Technologie-Informationen niedersächsischer Hochschulen, 2006.
17. Alexander Stolpmann. Kamerabasierte Fahrerassistenzsysteme. In *Optische Technologien in Niedersachsen.* to be published in 2006.
18. Alexander Stolpmann and Laurence S. Dooley. Genetic algorithms for automised feature selection in a texture classification system. In *Proceedings of the 4^{th} International Conference on Signal Processing*, Beijing, 1998.
19. Alexander Stolpmann and Laurence S. Dooley. About the use of fuzzy clustering for texture classification. In *Proceedings of the 18^{th} International Conference of the North American Fuzzy and Information Processing Society, NAFIPS'99*, New York, 1999.
20. Alexander Stolpmann and Laurence S. Dooley. A texture classification system with automatic feature vector optimization using genetic algorithms. In *Proceedings of the 7^{th} European Congress on Intelligent Techniques and Soft Computing, EUFIT'99*, Aachen, 1999.
21. Alexander Stolpmann and Laurence S. Dooley. A texture classification system using statistical and soft-computing methods. In *Texture Analysis Book*, Machine Perception and Artificial Intelligence (MPAI) series. World Scientific, 2000.
22. Hamid R. Tizhoosh. *Fuzzy-Bildverarbeitung: Einführung in Theorie und Praxis.* Springer, Berlin, Heidelberg, New York, 1998.
23. Ralphe Wiggins. Docking a truck: A genetic fuzzy approach. *AI Expert*, pages 28–35, May 1992.
24. Dieter Zöbel. Trajectory segmentation for the autonomous control of backward motion for truck and trailer. *IEEE Transactions on Intelligent Transportation Systems*, 4(2):59–66, 6 2003.

Neuro-Fuzzy-Modul for the Prognosis and Control of a Power Plant Using Renewable Sources of Energy

E. Förster

Delta Automation GmbH
Hallesche Straße 66, 06258 Schkopau, Germany
email: e.forester@t-online.de

Extended Abstract

With intelligent local electricity supply systems it will be possible to optimize the electricity supply integral “from below” over closed-loop control circuits. With the scenario of the participation of renewable sources of energy in this system, there has to be found a stable management for such power producer connected systems. An essential part of such a power producer data management system is an efficient prognosis tool. Apart from hydraulic energy, photovoltaic and wind power belong also anaerobe gas forming reactions to the renewable sources of energy (deposit gas, sludge gas, rotten gas, biological gas). By all those power plant types the planning with classical balance equations or rather balance equation systems is hardly practicable. Because of this there was chosen a Neuro-Fuzzy-approach for the GPMS (Green Power Management System). The processing was motivated by the following considerations: Prognosis information can be won from meteorological predictions and the measurement-technological acquisition of state variables with the help of experience based policy. There are already historical data records for the variation of service providence in combination with the variation of state variables for already existing installations.

In the future it must be possible to integrate economic distinctive numbers (energy price) in the expectation values. The optimization of the power plant inset can then be formed on economical, ecological and storage optimized criteria. At present the prototype of GPMS is being tested. The first results will be introduced here.

Modelling Corporate Strategy with the Fuzzy Balanced Scorecard

Prof. Dr. Volker Nissen

TU Ilmenau, FG Wirtschaftsinformatik für Dienstleistungen,
PF 10 05 65, D-98684 Ilmenau,
volker.nissen@tu-ilmenau.de

Abstract. This article discusses the importance of fuzzy aspects within the context of the balanced scorecard. This is followed by demonstrating possible approaches towards fuzziness as based on the fuzzy set theory within the framework of an advanced fuzzy balanced scorecard. The result is a realistic differentiated model of a corporate strategy. One implementation among others is in strategic simulations.[1]
Key words: balanced scorecard, fuzziness, fuzzy set theory, strategic planning

1 Brief Overview of the Balanced Scorecard

Widespread agreement that key figure systems solely aimed at the profitability of a company alone do not suffice in supporting corporate management has existed for a considerable time [4, p. 350]. In a survey of German companies Schonmann comes to the conclusion that "the traditional balance and accounting oriented planning and management concepts are not suitable to master (...) the dynamic and turbulent corporate environment." [12, p. 106, transl.]

Almost one third of the companies surveyed has already reacted to the deficit and introduced *Performance Measurement* systems. Since the mid-1980s performance measurement subsumes the new conception of key-figure based instruments used in corporate planning and management and can be defined, according to Gleich, as the assessment of the effectiveness and efficiency of the performance and the achievement potential of the most varied objects within the company (for instance, employees, organisational units, processes). A prerequisite here is the setting up of a performance measurement system which contains quantifiable key figures of differing dimensions (for instance, related to costs, time, quality, innovative ability).

The development of the balanced scorecard (BSC) by Kaplan and Norton represents a milestone in modern performance measurement. It supplements the traditional accounting oriented view of a company with further non-monetary dimensions which are seen as driving factors for future performance. Typically in a BSC distinctions are made between the following perspectives of a company, although variations may occur:

[1] These descriptions are a further development of previous work in [9, 10].

- Finance perspective: objectives and relevant key figures which stand in direct relation to the financial results of a company
- Business process perspective (also internal perspective): objectives and the relevant key figures for the achievement potential of the business processes
- Customer perspective: objectives and relevant key figures for the achievement potential of the company in terms of the market
- Learning and development perspective: objectives and key figures which reflect the company's potential concerning future market requirements.

The BSC forms the framework for the application of a strategy for a corporate business unit. Using the BSC the vision of the company for the future should be communicated throughout the whole organisation. A strategy is regarded by Kaplan and Norton as a catalogue of hypotheses on the cause-and-effect relationships of strategic objectives. Objectives are represented by key figures in the BSC. Key figures reflect critical success factors against competition. The aim is to create a balance between the internal- and external-oriented measurements.

Management objectives are substantiated by giving target values for each of the key figures. The key figure system of the BSC should clarify the hypotheses on the causal relationships between the targets (and their key figures). These interrelationships form causal chains connecting the 'subjective' performance drives (early indicators) and the 'objective' critical result key figures (late indicators). Here the causal chains are finally linked to the financial aims.

The high-priority strategic objectives and their interconnections as expressed in the corporate strategy form the basis of the BSC key figure system. These are then broken down to key figure level. The connections may arise within the same scorecard perspective as well as between one perspective and another [2]. The focus of this procedure lies on a consistent key figure system of objectives and their related key figures, with which plans for the next three to five years can be made.

A BSC conceptualised in this way expresses basic assumptions of business. It helps in evaluating the available potential and provides indications of how to optimise finances [5, p.143-145, p.293-296], [1, p.79-81], significantly different from the historical and purely financial key figure system of the past.

The BSC not only forms a key figure system but should also be understood as a widespread tool in strategic management. It serves the following purposes:

- Harmonising and prioritising strategic objectives needed within the enterprise.
- Revealing the effect relationships between the strategic targets within the various business perspectives
- Communicating strategic objectives and measures required to achieve them
- Breaking down a previously-defined management strategy for the individual business units

[2] A detailed example of a cause-effect relationship between early and late indicators is given by Kaplan and Norton for the National Insurance company cf. [5, p.154]

– Strategic feedback-learning which, if necessary, can lead to the adjustment of the strategy or individual strategic targets.

As a management tool every BSC must be individually tailor-made for the strategy of each business unit. At its highest level, that is a BSC for a company or a company business sector, the objectives and the measures required to achieve these objectives are comparatively global and abstract. Therefore, the BSC approach allows for hierarchically derived BSCs for subordinate organisational units, such as departments and teams. In this way the superior objectives are broken down to the level of local measures to which every employee can contribute.

2 Related Work

The author is aware of only one similar work, done by Pochert [11] who independently developes the concept of a fuzzy Balanced Scorecard. In her thesis, she starts from the assumption that each strategic target should initially be viewed as one fuzzy set, modelled with a trapezoidal membership function. This fuzzy set represents an estimation of what planning experts of the company expect as a realistic value for this target in the future[3].

Before the different scorecard perspectives are connected through cause-and-effect relationships of strategic objectives, each fuzzy target is converted with the help of company-specific expert knowledge to a set of linguistic terms. These terms represent various levels of possible achievement for the fuzzy target. In her model, no intersections between neighbouring fuzzy sets occur for any key figure [11, p.228]. An example of Pochert's modelling approach for fuzzy targets is given in figure 1, concerning the staff motivation.

Explicitly referring to the principle of fuzzy control [11, p.212], cause-and-effect relationships between targets in the different scorecard perspectives are modelled with fuzzy rule bases that contain knowledge about the interconnections between the targets. In her model, Pochert assumes that input for performance drivers within the BSC comes in the form of fuzzy linguistic terms, such as "unmotivated" for staff motivation, and "unqualified" for staff qualification [11, p. 217 - 218], and the result key figure value, "complaint response time" in this example, is determined through the application of rules. The fuzzy BSC is then used as a scenario tool where different inputs can be used to forecast result key figures on a coarse level: "With the help of this fuzzyfied instrument the planners are in a position to make predictions about future developments of the company (...)" [11, p. 234, transl.].

While it is acknowledged that Pochert, on a general level, correctly identifies the value of fuzzy set theory for the BSC, details of her approach appear problematic. There seems to be a fundamental misconception of the BSC approach in her work. The BSC is a tool to help implement a predefined company strategy using a top-down approach. Pochert, however, starts from rough estimations (fuzzy

[3] For examples concerning financial targets see, for instance, [11, p. 117 - 118].

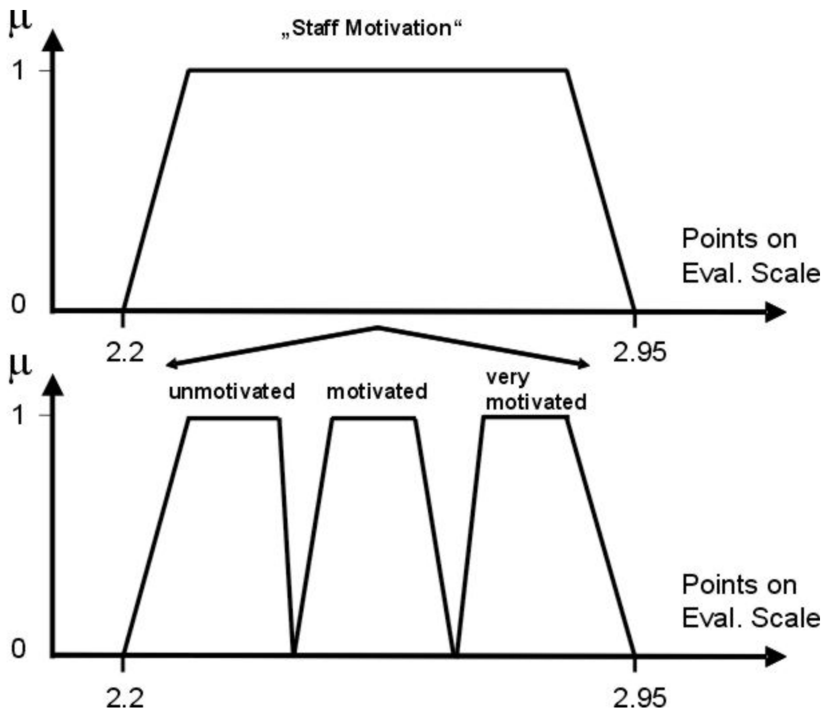


Fig. 1. Fuzzy modelling of the strategic target "staff motivation"and convertion to linguistic terms by [11, p. 214]

intervals) of input key figures and bottom-up derives fuzzy intervals for result key figures, such as the ROI, using the BSC as a forecast tool for management.

Consequently, in her own example, possible ROI-results span the interval from - 6% to +32% [11, p. 228]. This is unrealistic for a top financial key figure that is intended to help steer the company at a strategic level. Management must set financial targets in a much more precise way, even though the planned ROI may be stated as a (narrow) interval, and then answer the question "What results are required for related strategic objectives in other BSC perspectives to finally arrive at the ROI target level we aim for?". Simulations with a fuzzy-BSC can help to answer this question and identify the most effective performance drivers, as will become clear in our own approach described below.

It also a point for discussion to model fuzzy sets without intersections when you claim to imitate the principle of fuzzy control. Of course it simplifies the application of fuzzy expert rules when the input comes in the form of linguistic terms instead of sharp values. Due to this assumption by Pochert, only one rule will apply in each rule-base at a time for a given combination of inputs. However, this only very remotely resembles the concept of a technical fuzzy controller. In fact, a fuzzy controller would use sharp inputs and fuzzify them based on intersecting fuzzy sets. This in turn means that more than one rule can be activated in a rule base, and one gets the desired smooth output behaviour of the fuzzy system. Without intersections of fuzzy sets, the behaviour of the fuzzy rule-based system will be partly erratic.

Finally, Pochert did not implement her approach. Her work is strictly at the conceptional level without delivering a proof of principle prototype.

3 Fuzzy Aspects of the Balanced Scorecard

In a company-specific BSC concept and implementation certain problems occur again and again in practice:

– There are no links between the objectives within the various BSC dimensions and the interconnections between the objectives are not fully understood. Thus, the BSC simply becomes a collection of key figures instead of forming a consistent key figure system.
– The calculation of qualitative key figures (e.g. customer satisfaction) is questionable.
– There is a lack of decision-making aids when selecting and giving priority to suitable strategic measures for achieving the objectives strived for.

These difficulties may be alleviated if the existing fuzziness within the BSC can be exposed and explicitly modelled. Subsequently, each area where fuzziness is of significance will be discerned. The possibilities of examining the fuzzy aspects as based on the fuzzy set theory within the framework of an advanced Fuzzy Balanced Scorecard (fuzzy BSC) are then demonstrated.

Fuzziness is still often regarded as negative, as do Kaplan and Norton who pejoratively use the term "fuzzy key figure" in connection with the employee perspective: "If investment is to be made .... in the knowledge and qualifications of a member of staff, more than simply a fuzzy key figure will be required .... Tangible results should materialise (...)" [5, p. 246, transl.]. This point of view does not go far enough as the clear, non-fuzzy representation in a model often provides a distorted view of reality. In decision-making models this can result in poor decisions.

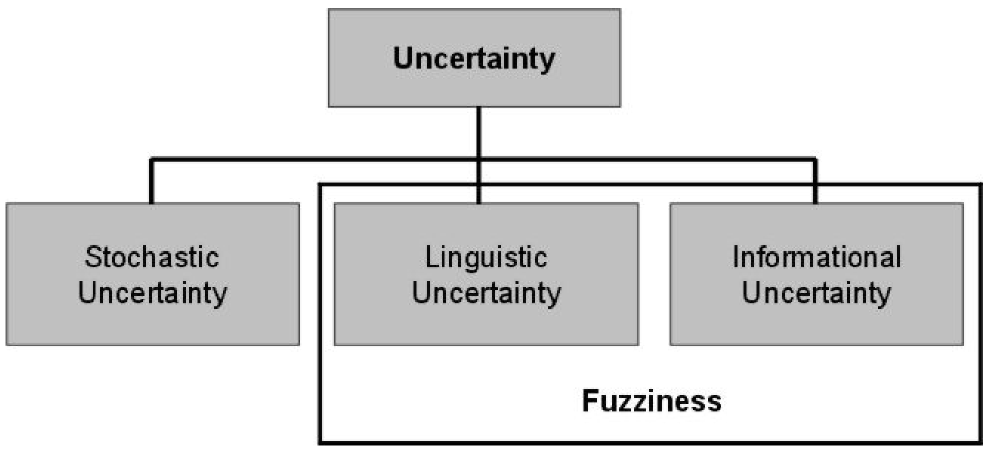


Fig. 2. Forms of uncertainty (according to Zimmermann et al.[13])

Fuzziness is a form of uncertainty. Zimmermann differentiates between three kinds of uncertainty (figure 2) [13, p. 3-7]:

– Stochastic uncertainty
– Linguistic uncertainty
– Informational uncertainty

Stochastic uncertainty can be modelled on the basis of the theory of probability. Events can only be described in a clear bivalent way (occurred or not occurred). Probability seems to evoke an appearance of exactness. (Example: "the probability of contracting X disease is 0.2"). Fuzzy events, for example one that occurred in part, cannot be represented.

In linguistic uncertainty the cause lies in a lack of precision and the indefinite nature of human language. (Example: "high" increase of sales). A form of linguistic uncertainty are fuzzy relations. Fuzzy relations exist when several objects are placed in a fuzzy relation with each other. (Example: A is "much bigger" than B).

Informational uncertainty exists when a high quantity of descriptors is necessary to describe a term clearly. (Example: "creditworthy" person).

In the following, linguistic and informational uncertainty are to be considered as fuzziness.

1. Fuzzy area: cause-and-effect relationships (fuzzy relations)
   Besides strategic result key figures (late indicators), it is important within the BSC key-figure system to identify the relevant performance drivers (early indicators) and, by doing so, to connect them with the result key figures in such a way that company strategy is best expressed. One of the main tasks of management is to be aware of the presumed cause-effect (or end- means) relationships.
   If possible, effect relationships between the key figures should be quantified against each other [5, p. 17]. Often, however, a precise quantification proves to be complicated and imprecise. Kaplan and Norton also recognise that "initially the influences of effect must be evaluated subjectively and qualitatively" [5, p. 245, transl.]. Then, an explicit modelling on a level above that of the graphic model for the key figure interconnections occurs only seldom. This limits the number of possible applications of the BSC in simulations.
   The effect relationships to be modelled in the BSC represent fuzzy relations between the objectives and/or sub-objectives (as represented by the key figures and performance objectives). The effect relationships between the interconnected targets are indeed co-rotating in nature - a selection criteria for the objectives to be included in the scorecard perspectives - but the relationships in detail may be far more complex, particularly when several performance drivers all flow into the same result key figure. Non-linear interconnections and, above all, varying intensive compensatory relationships become possible here.
2. Fuzzy area: qualitative measurements, consolidation between components and qualitative key figures

Strategic feedback and organisational learning at the management level are described by Kaplan and Norton as the most important aspects of the BSC approach [5, p. 15]. In regular intervals it is, therefore, necessary to carry out checks as to whether the monetary as well as the non-monetary objectives have been fulfilled. However, in all cases objectives are to be quantified in terms of target values for key figures. Correspondingly, qualitative measurements must be depending on their character converted into quantifiable terms.
In quantitative key figure system there may be limits to the significance of each key figure or even in some cases to the significance of the whole key figure system [2, p. 346]. One reason is that too great a loss of information occurs when qualitative aspects (and other forms of fuzziness) become important [6, p. 209]. In a long-term area highly dependent on human assessment like strategic management it is reasonable to suppose that such a situation will occur.
Practical examples show that qualitative sets such as image or service quality are often formed as a weighted or not weighted average (index) of a number of components[4]. Technically this situation poses many questions for the model. Implicit assumptions on the independency and mutual compensation between the objectives (or rather between their respective key figures) are made. Above all, compensatory relationships require attention here. A low value for one or more components can either be compensated in full or partially by the high value of other components. Likewise one component's role in the total result depends on the result measured for another component. In such cases it is preferable to aim at a model which is as realistic as possible with a low loss of information. How this is to be achieved with the assistance of the fuzzy set theory is demonstrated at a later stage in this article.

3. Fuzzy area: identification and focussing of required action
   Suitable measures should guarantee that the defined performance objectives should be reached. Management action is generally called for when the gap between the actual value and the target value of the key figure is "sufficiently wide". The fuzziness can be defined to the effect that varying degrees of inconsistency between the actual value and a target value is either satisfactory or disconcerting or anything between. The conventional procedure for establishing a clear boarder between tolerable and intolerable inconsistencies (traffic light logic) cannot be applied appropriately here.
   In the following, the fuzzy areas as mentioned above will be discussed within the framework of the fuzzy balanced scorecard. The following possibilities arise as a result:
   – Objective relationships can be modelled explicitly
   – A realistic model is created for consolidating subordinate components with a key figure
   – Fuzzy key figures provide indications of to what degree an objective has been fulfilled as well as the reliability of the assessment

[4] For instance the Metro Bank calculated service qualty as an index of various components cf. [5, p. 115]

– The selection of measures for the fulfilment of strategic objectives may be supported by simulations with the fuzzy BSC.

4 Fuzzy Modelling in a Fuzzy Balanced Scorecard

4.1 The fuzzy set concept

Since the mid 1960s the fuzzy set theory has been used for developing a theoretical basis in order to model fuzziness. Fuzzy systems and fuzzy methods have a solid mathematical basis. In classical set theory, an element x out of a basic set $X(x \in X)$ either definitely belongs to a set A or it definitely does not belong to A. However, for many real circumstances such a sharp distinction does not render an appropriate representation. In fact, gradual membership prevails in reality. Thus a fuzzy set $\tilde{A}$ is characterized by the fact that the membership of an element x to $\tilde{A}$ can be indicated by a real number which is usually standardized on the range of values [0.1] (uniform interval), thus describing formally a fuzzy set $\tilde{A}$ by a real value function $\mu_{\tilde{A}}$ is:$\mu_{\tilde{A}} : X \to [0, 1]$. Such a function is called membership function. Herein, a value $\mu_{\tilde{A}}(x) = 0$ means that x does not belong to the fuzzy set $\tilde{A}$, while a value $\mu_{\tilde{A}}(x) = 1$ indicates full membership. Values within the interval $0 \leq \mu_{\tilde{A}} \leq 1$ indicate a partial membership of x in the set $\tilde{A}$.

The classical, non-fuzzy set A can be interpreted as a special fuzzy set, for which only two alternatives, no membership or full membership, exist.

Fuzzy sets are very useful for representing vague concepts, wherein the basic set can be continuous as well as discrete. If the basic set is discrete, the result is the representation of a fuzzy set $\tilde{A}$ as a list of value-pairs. Each pair entails an element of the basic set as well as its membership value with respect to $\tilde{A}$:

$$\tilde{A} = \{(x_1, \mu_{\tilde{A}}(x_1)); \cdots; (x_n, \mu_{\tilde{A}}(x_n))\}, \qquad \forall x \in \mathrm{X}$$

However, only elements with strictly positive membership values are incorporated. Figure 3 shows some characteristic examples of membership functions with continuous basic sets.

The basic operators of classical set theory such as intersection and union have been enhanced for fuzzy sets. The intersection of two fuzzy sets $\tilde{A}$ and $\tilde{B}$ is defined by the following membership function:

$$\mu_{\tilde{A}\cap\tilde{B}}(x) = min\,\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x) \mid x \in \mathrm{X}\} \qquad \text{(Minimum-Operator)}$$

The union of two fuzzy sets and is defined by the following membership function:

$$\mu_{\tilde{A}\cup\tilde{B}}(x) = max\,\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x) \mid x \in \mathrm{X}\} \qquad \text{(Maximum-Operator)}$$

Special significance for the simulation of human decision-making behaviour can be attributed to the so-called compensatory fuzzy operators, which will be addressed at a later stage by the example of the γ operator. For a more detailed description of the fundamentals of the fuzzy set theory, see the relevant literature [13, 7, 14].

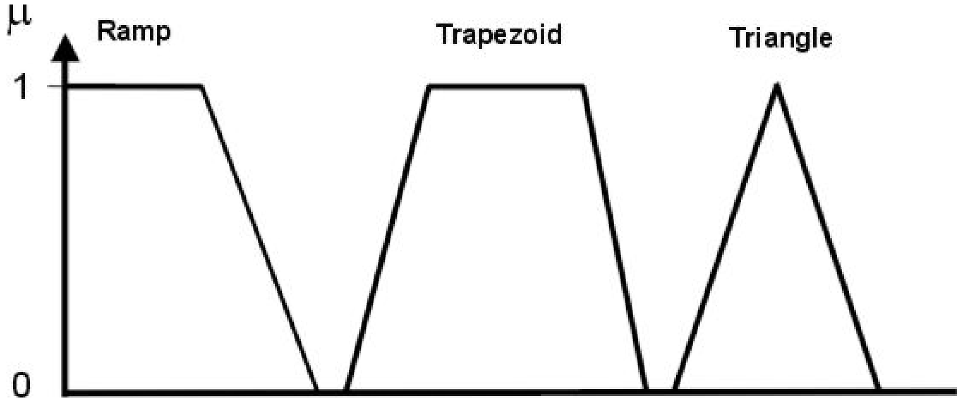


Fig. 3. Examples of membership functions

4.2 The Fuzzy Balanced Scorecard

The consideration of the fuzzy aspects as mentioned above in connection with a fuzzy BSC involves minor changes to the process of creating a BSC:

1. In modelling the strategic target values, for each key figure of the fuzzy BSC a relation is established between the key figure value and the degree of satisfaction on the decision-maker's side. With the help of a basic set of reasonable key figure values, these degrees of satisfaction can be represented as a fuzzy set, in which the extreme value 0 signifies total dissatisfaction and the opposite extreme value 1 signifies total satisfaction. Values between 0 and 1 signify partial satisfaction.

In this respect, figure 4 shows two examples of performance drivers following the process- related measurements for a European insurance company. The first key figure concerns the availability of IT systems, the second the average backlog in the processing of orders for each member of staff. The actual values of 97% availability and the average backlog of 3.5 days in order processing result in degrees of satisfaction regarding the achievement of the strategic targets with the fuzzy set membership values of 0.5 and 0.7 respectively.

The achievement of targets for result key figures is modelled in basically the same way. Attention must be paid to the fact that the membership function runs monotonously, that the value 0 is acquired at exactly one position in the basic set and the value 1 at exactly another, and that for all other values the membership value is strictly positive. Without these additional conditions, it is not possible to derive a sharp value for the corresponding key figure from the fuzzy target realisation. Figure 5 shows an example of the result key figure 'revenue increase'[5].

[5] The restriction applies to all key figures which result from others within a causal chain. If this restriction should prove to be problematic in individual cases, the targets and cause-effect relationships can be modelled in correspondence to the procedure for component aggregation to qualitative key figures as described later in this article. In most cases however, a revision of the underlying key figure will be sufficient.

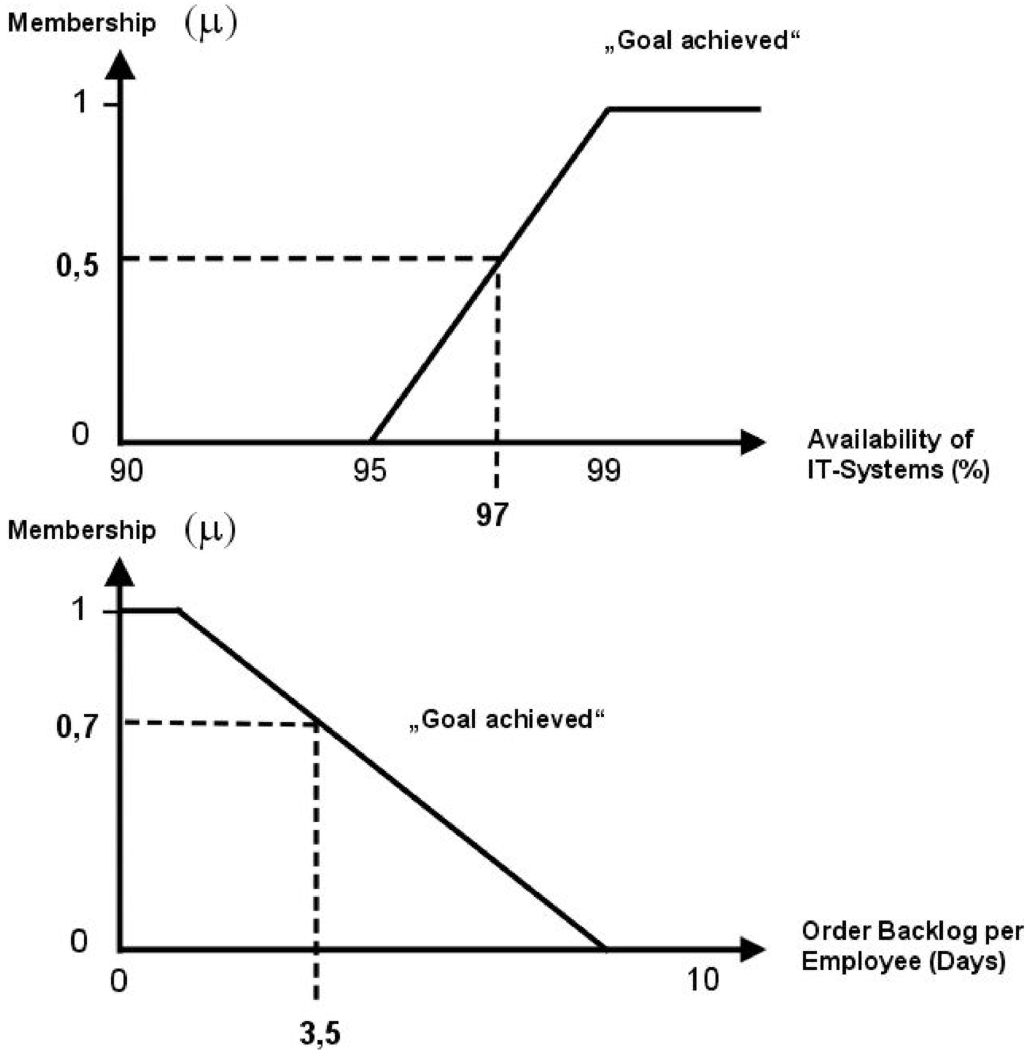


Fig. 4. Fuzzy sets for target realisation (performance drivers without preliminary causal chain)

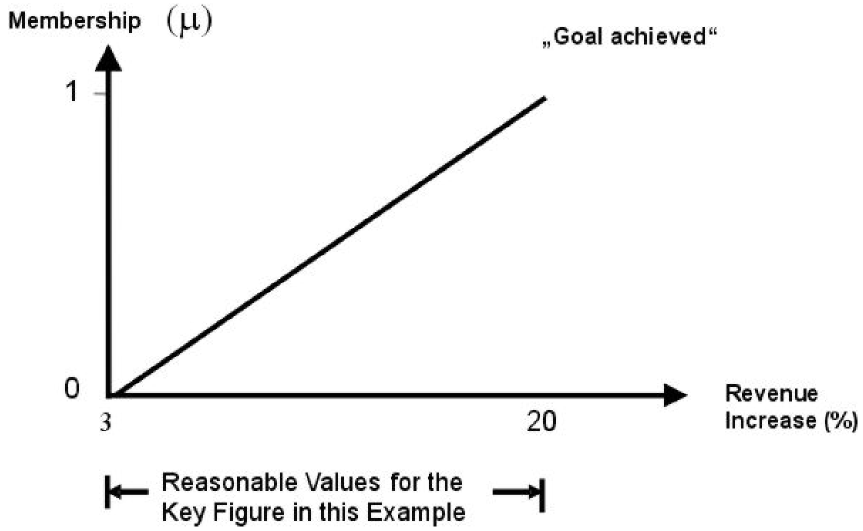


Fig. 5. Fuzzy set for the target realisation (result key figure with preliminary causal chain)

2. The cause-and-effect relationships between performance drivers and result key figures, as assumed in the corporate strategy, must be determined and recorded more accurately than in the past. The case that a result key figure solely depends on one performance driver, is relatively trivial. It is based on the hypothesis that these values develop in unison, albeit with a certain time lag. Much more interesting are those cases where several key figures (drivers) meet in one key figure (result). Here the assumed interrelations between the drivers and the resulting value need to be represented more accurately. In many cases it is inadequate to assume that the effect of several performance drivers on the result is independent. Instead, there are often more or less definite compensatory relations between the performance drivers.

For example, the operator, which can be traced back to Zimmermann and Zysno, can be used to model these fuzzy correlations and may be defined for two fuzzy sets $\tilde{A}$ and $\tilde{B}$ as follows:

$$\mu_{comp}(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x), \gamma) =$$
$$(\mu_{\tilde{A}}(x)\mu_{\tilde{B}}(x))^{1-\gamma}(\mu_{\tilde{A}}(x) + \mu_{\tilde{B}}(x) - \mu_{\tilde{A}}(x)\mu_{\tilde{B}}(x))^{\gamma} \mid x \in \mathrm{X}, \gamma \in [0,1]$$

The parameter γ represents the so called degree of compensation with the extreme values $\gamma = 0$ for "no compensation" (meaning no willingness to compromise) and $\gamma = 1$ for "full compensation" (meaning full willingness to compromise). In the key value system of a fuzzy BSC, an assumed degree of compensation must be defined for each target relationship. For practical reasons it is advisable to also quantify the expected *time lag* between cause and effect.

Figure 6 shows a larger section of a fuzzy BSC in its overall context. It follows the key value system of a large insurance company. Performance drivers and result key figures show fuzziness in degrees of satisfaction with regard to

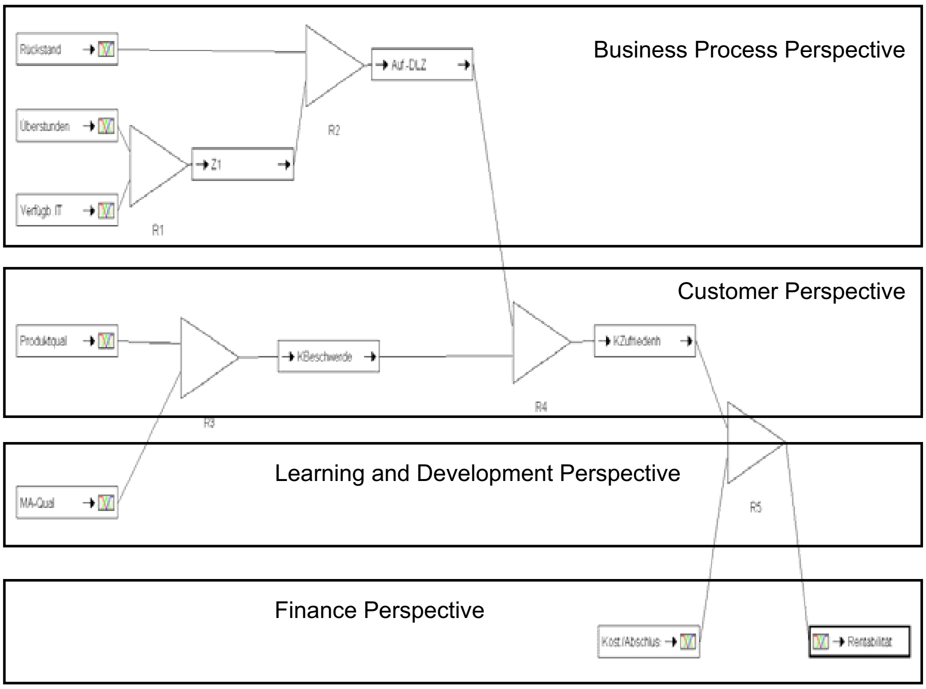


Fig. 6. Example for a fuzzy BSC (extract) for an insurance company

the achievement of objectives. The connections within and between the perspectives are made by parameterised, compensatory fuzzy operators (represented as triangles in Figure 6). The financial result figures as well as all other results can be shown as fuzzy in form (degree of satisfaction) or as sharp key figures (e.g. profitability equals 12%).

3. When aggregating several subordinate key figures (components), which should not be shown in the scorecard, into one qualitative BSC key figure (e.g. 'service quality', 'image' or 'staff qualification'), weighted averages should not be generated without some reflection. Instead it makes sense to establish those components which are independent and those which must be seen in context of their effect. The next step is to clarify assumed correlations. A possible method for doing so, in correspondence to the procedure explained above for interconnections of strategic targets, is first to establish the degrees of compensation between the components. The combination is then formed by compensatory fuzzy operators such as the γ operator as described above.

Another more differentiated but also more laborious modelling requires the mutual dependencies to be expressed as a number of if-then-rules. The following is an example of such a rule:
IF the average number of years of service for the staff is 'HIGH'
AND the number of training days per member of staff and per year is 'MEDIUM',
THEN staff qualification is 'HIGH'.

These rules include terms such as 'medium' and 'high', which as so called *linguistic terms* and can be modelled by fuzzy sets, for example triangle and trapezoid functions (Figure 7). In this case the actual sharp key figure of 55% university graduates results in memberships of approximately 0.8 for the 'medium' linguistic term and 0.2 for the 'high' linguistic term. This procedure is called 'fuzzifying' the actual sharp key figure value. Figure 8 shows an example of how, according

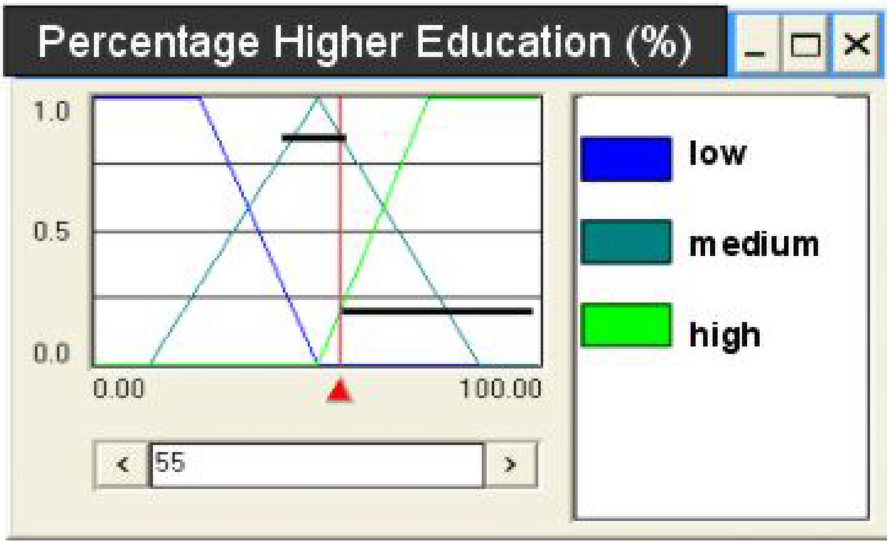


Fig. 7. Three linguistic terms ('low', 'medium' and 'high') as fuzzy sets of the basic set 'proportion of university graduates'

to these rules, the key figure of 'employee qualification' as a qualitative figure may be made up of the secondary quantitative components 'number of training days per member of staff', 'relevant years of service per member of staff' and 'proportion of university graduates'.

Table 1 shows an extract from the rule set. Here the rule assumptions are combined in an exemplary way by the minimum operator (AND). The appropriate selection of the fuzzy operator is however problem-specific, and in the context of the fuzzy BSC it should be chosen in a way that reflects the view of the management.

The rule sets of fuzzy rule-based systems incorporate several special features. They need to be neither complete nor consistent. Moreover, rules can be weighted in order to express confidence in their correctness. In our example, the entrance values simultaneously activate three rules. The use of a fuzzy inference mechanism which shall not be explained here[6] , initially reaches a fuzzy result for the key figure of 'employee qualification' (Figure 9). The membership to the fuzzy sets of 'medium' and 'high' is about 0.7 and 0.5 respectively. This result

[6] An explanation of the inference mechanism can be found in [8].

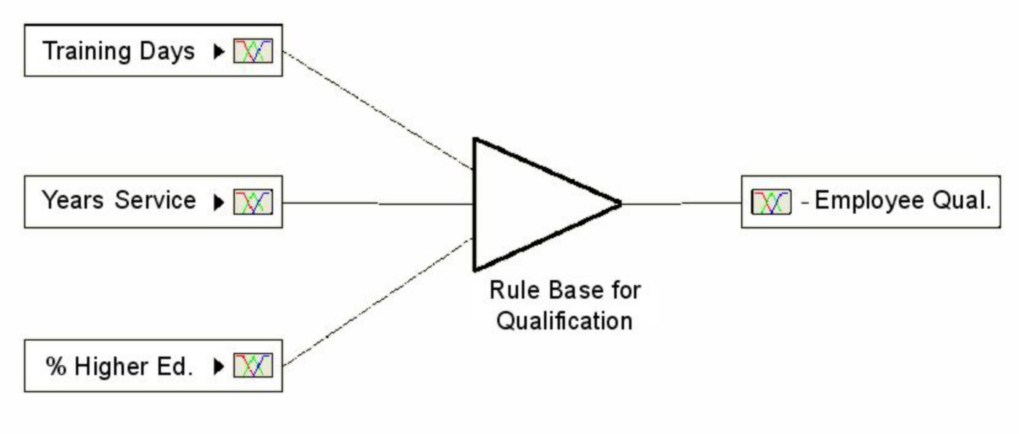


Fig. 8. Example of fuzzy rule-based aggregation

Table 1. Excerpt of the rule set (first three rules are active)

Rule No.	Training Days	Years of Service	% Higher Education	Fuzzy Operator	Rule Weight	Employee Qualification	Degree of Member-ship
1	low	high	high	**AND**	0,9	high	**0,18**
2	low	high	medium	**AND**		medium	**0,70**
3	low	high	medium	**AND**	0,7	high	**0,49**
4	low	low	high	**AND**	0,5	medium	0,00
5	high	low	high	**AND**		medium	0,00

indicates a considerable uncertainty regarding the assessment of employee qualification. In an optional step, the fuzzy result can now be converted to a sharp value by means of an appropriate defuzzifying procedure. During this step, the information on the amount of uncertainty is however lost. The defuzzified value for this example lies around 0.59. It may also represent the management's degree of satisfaction with regard to the achievement of the objective for 'employee qualification' in the fuzzy BSC.

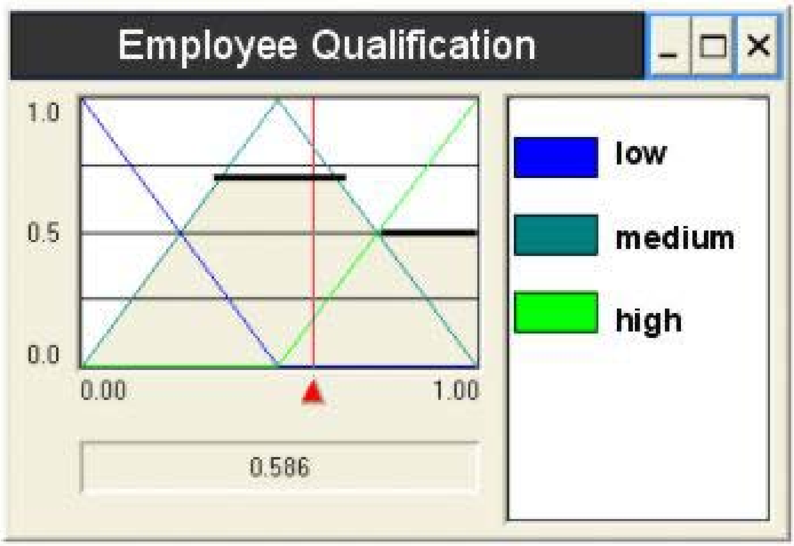


Fig. 9. Membership degrees to the fuzzy sets 'medium' and 'high' employee qualification (0.7 and 0.5 respectively) and defuzzified target realisation (0.586)

The aggregating procedure for qualitative key figures proposed here can also be used for modelling the achievement of objectives and cause-and-effect relationships of strategic targets in the overall fuzzy BSC. However in comparison to the aforementioned procedure aggregation which makes do with a single fuzzy set per key figure, it is a more laborious procedure.

4.3 Simulating with the Fuzzy Balanced Scorecard

The fuzzy BSC represents a corporate strategy model. Due to the explicit modelling of cause-effect-relationships, it is suitable as a simulation tool for several purposes[7]. Firstly, it can be used to establish whether the assumed target correlations within the scorecard reflect the reality. Once the key figure values of the performance drivers are specified, the modelled target interconnections allow the

[7] The differentiated form of modelling allows simulations to be more precise and more flexible than those that Kaplan and Norton carried out with the traditional BSC. cf. [5, p. 246-249].

simulation of the values that are to be expected for the result key figures after a certain amount of time (*time lag*). Should the results measured in practice for these key figures differ greatly from the predictions, the model should be thoroughly revised. Several causes for the variations are possible. Apart from errors within the model (missing performance drivers, incorrectly modelled fuzzy sets or target relationships etc.), changes in the external conditions may call for a new strategy[8].

In an extension of this line of thought, it appears possible to optimise the parameters within the model of a fuzzy BSC automatically in such a way that the corporate strategy is mirrored in the best possible way. The parameters of compensatory fuzzy operators in particular, such as the degree of compensation of the y operator, allow for a simple and flexible possibility of changing the effect relationships within the model. Such a parameter optimisation approach would take as inputs the actual measurements of key figure values for the performance drivers. The objective function to be minimised then can be constructed from the sum of squared differences between actually measured and fuzzy BSC-predicted values for the result key figures.

A second type of simulation can help in making the right decisions for reaching strategic objectives. Measures for improving the result key figures apply at the performance drivers. Limited corporate resources necessitate the use of measures that promise the best possible effect. Since the effect relationships in the fuzzy BSC are modelled explicitly, alternative scenarios can be run through in which the actual values at the performance drivers are changed individually or simultaneously, and the effects on the result key figures are determined. The measures to be chosen should then improve the actual situation at the most effective point of departure.

Thirdly, a systematic simulation with maximum values at the performance drivers can help determine whether the target result values can be reached in the scope of the model or not. If not, then the rule bases or membership functions must be revised, or additional performance drivers must be integrated within the model.

5 Conclusion

In this paper, the relevance of vagueness and qualitative information in strategic management was taken as a starting point to discuss aspects of fuzziness within the context of the well-known Balanced Scorecard concept. Fuzzy set theory offers adequate modelling options to capture these aspects in a fuzzy BSC and, thus, arrive at a more realistic model of a company strategy than in the classical BSC approach. Moreover, the fuzzy BSC is an interesting simulation device. Itt

[8] It should be noted that the fuzzy BSC can of course supply only such values which lie within the modelled range of the corresponding key figure. With result key figures, degrees of satisfaction with the value 0 therefore supply a sharp maximum value, and degrees of satisfaction with the value 1 a sharp minimum value for the corresponding key figure.

can help management to identifiy where limited ressources are best employed to most effectively improve the actual situation. Simulations can also help to identify and remove flaws in the BSC-model of the company strategy.

The concept of a fuzzy balanced scorecard as it was presented and implemented prototypically on the basis of a generic fuzzy tool, can be developed into several further directions. Firstly, it would be possible for the time lag with which the changes in the early indicators affect the subsequent late indicators, to be modelled explicitly in order to further increase the realism of the corporate strategy model.

Secondly, an explanation component could be developed to explain the model results verbally and would, thus, make further improvements in its comprehensibility and also in the benefits for the management. Such an explanation component has already been developed in prototype by Kuhl [6, p. 218-221] in the course of a fuzzy key figure system for an inventory problem. This approach that evaluates the degrees of membership for explanation purposes can be adapted for the fuzzy BSC.

References

1. Beyer, R.: Ist die Balanced Scorecard ein innovativer Ansatz oder ein herkömmliches Kennzahlensystem? In: Scherer, A.G.; Alt, J.M. (eds.): Balanced Scorecard in Verwaltung und Non-Profit-Organisationen. Schäffer-Poeschel, Stuttgart, 2002, S. 73 - 89.
2. Bramsemann, R.: Handbuch Controlling - Methoden und Techniken, 2. Aufl., München and Wien, 1990.
3. Gleich, R.: Performance Measurement. In: Die Betriebswirtschaft 57 (1997) 1, S. 114 - 117.
4. Horváth, P.: Der Einsatz von Kennzahlen im Rahmen des Controlling. In: Wirtschaftswissenschaftliches Studium, (1983) 7, S. 349 - 356.
5. Kaplan, R.S.; Norton, D.P.: Balanced Scorecard. Strategien erfolgreich umsetzen. Schäffer-Poeschel, Stuttgart, 1997 (translated by Horváth, P.).
6. Kuhl, J.: Angepaßte Fuzzy-Regelungssysteme. Unitext, Göttingen, 1996.
7. Nauck, D. und Kruse, R.: Fuzzy-Systeme und Neuro-Fuzzy-Systeme. In: Biethahn, J.; Hönerloh, A.; Kuhl, J.; Leisewitz, M.-C.; Nissen, V.; Tietze, M. (eds.): Betriebswirtschaftliche Anwendungen des Soft Computing. Vieweg, Braunschweig and Wiesbaden, 1998, S. 35-54.
8. Nissen, V.: Wissenskonservierung mittels eines fuzzy-regelbasierten Systems für die Aufwandsplanung von Beratungsprojekten. In: Zeitschrift für Planung & Unternehmenssteuerung 14 (2003) 3, S. 243 - 258.
9. Nissen, V.: Potenziale der Fuzzy Set Theorie für das Kompetenzmanagement in Beratungsunternehmen. In: Biethahn, J.; Lackner, A.; Nissen, V. (eds.): Information Mining und Wissensmanagement in Wissenschaft und Wirtschaft. AFN: Göttingen 2004, S. 49 - 72.
10. Nissen, V.: Die Fuzzy Balanced Scorecard als Instrument der Unternehmensführung. In: Mönch, L.; Beyer, J. (eds.): Aspekte der Wirtschaftsinformatik: Methoden, Werkzeuge und Anwendungen, SCS Publishing: San Diego and Erlangen, 2005, 71 - 90.

11. Pochert, B.: Konzeption einer unscharfen Balanced Scorecard. Cuvillier: Göttingen, 2005.
12. Schomann, M.: Wissensorientiertes Performance Measurement. DUV, Wiesbaden, 2001.
13. Zimmermann, H.-J.; Angstenberger, J.; Lieven, K.; Weber, R. (eds.): Fuzzy Technologien: Prinzipien, Werkzeuge, Potentiale. VDI, Düsseldorf, 1993.
14. Zimmermann, H.-J.: Fuzzy Set Theory - and Its Applications, 4. ed., Boston: Kluwer, 2001.

Machine Learning Methods for Safety-Related Domains: Status and Perspectives

Clemens Otte, Sebastian Nusser, and Werner Hauptmann

Siemens AG, Corporate Technology, Otto-Hahn-Ring 6, 81730 Munich, Germany, {clemens.otte, sebastian.nusser.ext, werner.hauptmann}@siemens.com

Abstract. Machine learning provides significant advantages due to its data-driven modelling capabilities. However, these strengths are finding only slow acceptance in safety-related applications. This paper gives a survey of machine learning in safety-related domains and discusses methods to validate the safety of the learned model. Hybrid approaches play a dominant role, Neuro-Fuzzy techniques in particular. We briefly review the pros and cons of Neuro-Fuzzy and discuss a special form of a mixture-of-experts approach to meet safety requirements. We conclude with an outline of further research perspectives.

1 Introduction

Machine learning methods are successfully employed in a wide range of applications. But in the community of safety research and Verification & Validation, machine learning methods and especially "black box" methods like artificial neural networks (ANNs) are regarded with suspiciousness (e.g. international safety standard IEC-61508 currently discourages the use of ANNs and more generally, knowledge-based computing and AI).

However, the increasing complexity of safety-related systems has raised a strong interest [1, 2] in utilizing the advantages of machine learning like non-linear modeling capabilities and learning from examples. The driving factors for applying machine learning in safety-related systems typically are:

- *Reduction of development time and cost*: in contrast to performing a manual model design and calibration, the model is either automatically learned from training data or parameters of an existing model are adapted based on training data.
- *Improved performance*: many machine learning methods like ANNs, support vector machines (SVMs) or Gaussian processes (GPs) are known for a good performance in modeling non-linear functions. They usually outperform conventional simpler methods (an example follows in Sect. 3.2); in particular, this is true for high-dimensional input spaces where human experts are often unable to provide an adequate analytical model.
- *Increased efficiency*: (multi-dimensional) look-up tables, which are commonly used for control tasks, may have prohibitive memory requirements. They can be replaced by verified ANNs to gain a much more compact representation, e.g. [3] suggests this for aircrafts.

Safety-related systems are systems whose malfunction or failure may lead to death or serious injury of people, loss or severe damage of equipment, or environmental harm. They are deployed, for instance, in aviation, automotive industry, medical systems and process control.

Required bounds for failure rates vary depending on the application: a control application performing a direct intervention in the controlled system (e.g. plant or car) has stronger safety requirements than a monitoring system or an advisory system (e.g. cytological screening). Different industries also use different requirements. For example, [1] reports the accepted failure rate to be less than 10^{-9} failures per hour in aircraft systems and less than 10^{-4} failures per hour in monitoring systems for nuclear power plants. For typical statistical testing procedures, such values become unrealistic since the amount of test data is limited. Thus, formal verification approaches should be preferred instead of purely statistical testing methods. To allow (formal) verification by an expert, the model must provide some kind of interpretability.

Machine learning methods returning symbolic outputs, like decision trees or (fuzzy) rule sets, can be interpreted by experts; however, their performance in terms of accuracy is often insufficient. In contrast, ANNs, SVMs and further methods (e.g. GPs) have a higher accuracy but their interpretability by experts is limited. The tradeoff between model performance and model interpretability thus becomes a crucial aspect in applying machine learning methods in the field of safety-related domains.

The paper is organized as follows: Section 2 discusses prerequisites for machine learning in safety-related applications and reviews the most common validation methods, which try to prove the correctness of a learned model. Learning methods in safety-related systems have to incorporate expert knowledge about the application; this ensures that safety constraints (e.g. bounds for parameters) are met and it also compensates for limited amounts of data. Therefore, methods deployed in practice are typically hybrid approaches, which are described in Sect. 3. They combine data-driven learning algorithms with symbolic forms of knowledge representation. Neuro-Fuzzy methods are well-known members of this family; we briefly discuss their pros and cons. Less-known in the safety community are mixture-of-experts approaches; we argue that under certain conditions they can complement or replace Neuro-Fuzzy in safety-related systems. We draw conclusions and outline perspectives for further research in Sect. 4.

2 Learning & Validation

For the reasons mentioned before, the application of machine learning in safety-related systems becomes increasingly important. Thus, there is the need to transfer the main requirements of safety analysis defined in safety standards like IEC-61508 or DEF STAN 00-55 into the field of machine learning. There are several approaches to establish a standard or guideline for certifying the use of machine learning methods (especially ANNs) in safety-related applications, e.g. [4, 2]. Taylor gives in [1] a general review. Kurd et al. [5, 6] suggest a safety lifecycle and safety criteria based on Neuro-Fuzzy models.

In the following, we will focus on the methodical aspects of learning and validation, instead of focusing on standardization issues.

Specialized approaches like error correction neural networks[1] (ECNN) [7] or Zakrzewski's approach[2] [3] are not further discussed in this paper.

2.1 Robust Learning

A common problem in the field of machine learning is that the amount of training data is limited and, thus, the training data can represent only a few situations of a real-world problem. Hence, interpolation between and extrapolation around training data is needed. Furthermore, training data may include noisy, irrelevant, corrupted or missing values. Thus, methods like data cleaning, feature selection or feature extraction, active learning, and, in particular, regularization are needed to provide a robust solution. We do not go into details here as most of these topics are well-known and covered by standard textbooks like [8]; for neural networks, [9] discusses some learning tricks and appropriate network architectures.

2.2 Validation Methods

Robust learning is necessary but not sufficient to provide solutions for safety-related problems. A model that shows a good generalization performance on available data might still fail on completely new data. It must be ensured that valid outputs are provided for the whole input domain. Therefore, the results must be interpretable by experts, e.g., by rule extraction methods, visualization techniques or by mathematical formalisms.

Rule extraction and knowledge insertion: Some safety-related standards, e.g. the UK defense standard [10], require:

> 36.5.2 Proof obligations shall be: Constructed to verify that the code is a correct refinement of the Software Design and does nothing that is not specified.

Interpretation in symbolic form is the easiest way to allow such formal verification. Thus, rule extraction methods are commonly preferred for validation.

Key issues are fidelity, accuracy and comprehensibility of the rules. *Fidelity* means that the symbolic representation accurately models the network from which it was extracted; and *accuracy* terms the ability of the extracted representation to make accurate predictions on previously unseen samples. The *comprehensibility* refers to traceability and explanation capability of the symbolic representation by human domain experts.

According to [11], there are three basic types of rule extraction techniques: decompositional, pedagogical, and eclectic. The *decompositional* approaches are

[1] ECNN is a model-design-based approach. It estimates the internal prediction error by an additional layer in the network structure to provide a robust solution. ECNNs are designed for modeling dynamical systems (e.g. time series prediction).

[2] Zakrzewski uses an already validated reference implementation (e.g. look-up-table) as deterministic data generator for validation. This approach is only applicable for low-dimensional problems.

focused on extracting rules at the level of individual units (i.e. the hidden and output neurons) of an ANN. By regarding the level of activation of the single neurons it is possible to extract rules as subsets of positively weighted incoming links. Typically, special types of network architectures are necessary. The *pedagogical* approaches treat the neural network as a black box. Rules are extracted from changes in the levels of input and output units. An advantage of this kind is that one can examine already existing networks. Unfortunately, pedagogical approaches become computational expensive due to the fact, that they usually perform a grid search on input-output-combinations. An *eclectic* approach combines elements of both former ones, i.e. knowledge about the internal structure of an ANN (e.g. weights, number of layers) are used to complement symbolic learning.

For the knowledge insertion task, the initial knowledge is typically represented by fuzzy rules; these rules are used to initialize the network before training. After the training procedure refined fuzzy rules (i.e. modified membership functions) can be extracted from the network.

A framework based on rule extraction and knowledge insertion is the so-called *safety lifecycle* given in [5]. In [12] several approaches to extract comprehensible descriptions (rules and decision trees) from learning systems (e.g. ANNs) are developed and discussed. Kolman [13] uses a fuzzy model that can be transformed into a usual Multilayer Perceptron (MLP) and vice versa — the disadvantage of this approach is that the number of extracted rules rises exponentially with the number of inputs. Instead of extracting rule systems, there are also approaches to extract decision trees, e.g. [14]. Furthermore, there are approaches to extract finite state machines and/or rules from Recurrent Neural Networks (RNN) [15]. For an introduction into RNNs, see e.g. [7].

Visualization techniques: Visualization techniques can assist in improving the readability of an ANN. They provide an intuitive mapping between inputs and outputs and can assist in detecting errors or anomalies within the model. The following two methods serve as samples of common visualization techniques.

Hinton diagrams provide a compact visual display of the weights and biases related to particular units in a network; the topological information of the neural network is represented as data matrices. Each weight is represented by a rectangle whose color (black or white) indicates the sign, and the size is associated with the magnitude. The Hinton diagram allows getting a picture of the relationships between the weights, but it is difficult to show a large network clearly. Thus, Hinton diagrams are more useful during network design than for explaining the learned input-output relationship.

For multi-dimensional data sets, *2-D slices of input-output space* can be used to get an impression of the behavior of the model – the problem is that such slices do not provide any information about the complete input-output-space and it is hard to find interesting and relevant slices for a large number of dimensions.

Visualization techniques can help to improve the understanding of the functioning of an ANN. The most difficult aspect of these techniques is that for high-dimensional problems or advanced network architectures the visualization becomes too complex. Thus, visualization techniques should be combined with other methods to provide sufficient information about the suitability of a model. A survey of a number of visualization techniques of ANNs can be found in [16].

Statistical estimation of reliability: The basic idea is to use error bars to estimate the uncertainty of the model's output. There are two methods discussed in this subsection, which provide confidence estimation: MLP with evidence propagation (MLPev) [17] and Gaussian processes (GPs) [18]). GP and MLPev are originally designed for regression tasks and not for classification tasks but they show also good performance for classification tasks.

MLP with evidence propagation: MLPev extends the MLP to overcome the need of large test sets for cross-validation to assess the network's performance on unseen data. The evidence procedure provides an objective criterion for comparing alternative neural network solutions and for setting free parameters. MLPev is based on a Bayesian framework [19] to evaluate the evidence, i.e. the probability of the data D given the model H_i, $P(D|H_i)$. The evidence procedure allows adding error bars to the predicted output of an ANN. The size of the error bars varies approximately with the inverse data density, so that the error bars are broader in regions where the training data density is low. By setting two hyperparameters it is possible to define the expected smoothness of the model as well; more details and an implementation can be found in [17].

Gaussian Process: The GP[3] can be seen as generalization of a Gaussian distribution to a space of functions. A GP is specified by its mean function and its covariance function, like a Gaussian distribution is specified by a mean value and a covariance matrix. The function underlying the observed data is assumed to be a single sample from this Gaussian distribution over the function space. Good introductions can be found in [18] and [20].

The advantage of using a GP is that one can determine confidence bands for Gaussian processes quite as simple as confidence intervals for Gaussian distributions. Unfortunately, the GP method has a poor performance w.r.t. space and time on data sets with more than 1000 samples since it stores all training data - thus, it is not suitable for embedded controller solutions (e.g. airbag deployment). In order to avoid performance problems, [21] introduced Bayesian committee machines as finite dimensional approximation of GPs, which show slightly the same predictive performance as the original GP.

3 Hybrid Approaches

In hybrid approaches, the learning capabilities of neural networks are combined with the possibility of including expert knowledge and/or a physical model of the underlying process. Typical examples are Neuro-Fuzzy methods, which are briefly reviewed in the following. In the tradeoff between interpretability and accuracy, Neuro-Fuzzy systems typically put more weight on the former. Therefore, in the second part of this section we discuss mixture-of-experts methods as an alternative or even complementary kind of hybrid approach.

[3] The Gaussian process is also known as Kriging in the field of spatial statistics.

3.1 Neuro-Fuzzy

The motivation for using Neuro-Fuzzy is that the resulting system can be interpreted in terms of fuzzy rules, which makes it attractive for safety-related applications. For example, the classification task of assigning an input $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ to a set C of (non-fuzzy) class labels can be solved by a model consisting of rules R_k with

$$R_k : \text{if } x_1 \text{ is } \mu_{k1} \text{ and } \ldots \text{ and } x_n \text{ is } \mu_{kn} \text{ then (class is } c) \ , \quad c \in C \tag{1}$$

where $\mu_{k1}, \ldots, \mu_{kn}$ are fuzzy sets learned from data [22–24]. Some approaches use a weight $w_k \in [0,1]$ to control the impact of the rule R_k on the classification decision. The effects of rule weights on the decision boundary are discussed in [25] and it is shown that different weights often lead to non-axis-parallel decision boundaries. However, a semantic justification of the weights may be problematic; a probabilistic interpretation is given in [26].

A general and comprehensive survey of rule generation methods can be found in [27]. Kurd et al. [28] use a constrained Fuzzy Self-Organizing Map to extract fuzzy rules for controlling safety-critical systems; parameter constraints ensure that semantic safety bounds are not violated. Note that parameter constraints from safety requirements are a general concept that can also be introduced in many of the other rule generation methods.

Finally, there are approaches that combine (neuro-) fuzzy rules with support vector machines (SVMs) [29, 30]. SVMs use a regularized form of learning, which often results in a good generalization performance of the classifier even in high-dimensional input spaces. The objective is to combine the comprehensibility of fuzzy rules with the good SVM classification performance. The number of generated fuzzy rules, however, is often still quite high, which makes the result difficult to interpret.

PROS AND CONS: Research on Neuro-Fuzzy systems started in the early 1990s and many techniques have now reached a maturity, which makes it feasible to deploy them in selected safety-related applications. For example, Siemens has used Neuro-Fuzzy to implement the decision logic in fire-detectors [2], for the identification of driving situations in cars [31], and for the calibration of airbag systems in automotive safety-electronics. The latter has particularly severe safety requirements because a wrong decision (e.g. airbag triggering in a non-crash situation) cannot be rectified. Additionally, training data is scarce as it usually results from expensive crash tests. Therefore, expert knowledge is used to constrain the complete learning process and to carefully validate the final fuzzy rules.

The main challenge for Neuro-Fuzzy methods in safety-related systems is that those approaches providing easily interpretable rules often do not have a sufficient accuracy (and vice versa). Thus, in practice, several methods for feature selection, model initialization, model reduction and model tuning must be combined [24, 32]. In particular, the following issues should be considered carefully:

Input Independence Assumption: The area of the input space that is mapped to a particular class in (1) is described by its projections on the coordinate axes only. If we can interpret each fuzzy set $\mu_{ki}(x_i)$ as a probability density function $p(x_i \,|\, c)$, which is often straight-forward if some normalization issues are ignored, and choose the multiplication as $\top$-norm for *and*, which is commonly done, then the area covered by the rule is described by the joint probability density function $p(x_1, \ldots, x_n \,|\, c) = p(x_1 \,|\, c) \cdots p(x_n \,|\, c)$. It is known that this equation only holds if the inputs x_i are mutually independent. Decorrelating the inputs during preprocessing may help to avoid accuracy losses, but may also affect their interpretability. Note that [33, 25, 24] show that class boundaries of fuzzy rules need not to be axes-parallel (unlike those of common decision trees), which may be favorable in case of correlated inputs.

Membership Functions: The membership functions of the fuzzy sets should be able to model monotonic concepts (e.g. sigmoids) as well as vague prototypes (e.g. Gaussians, triangles). If a human expert is able to state a monotonic relationship between an input and the class then this should be reflected in the choice of a monotonic membership function; otherwise, the classifier often shows poor extrapolating capabilities.[4] Unfortunately, approaches based on RBF-networks or SVMs typically cannot model monotonic relationships well.

Accuracy Loss in Change of Representation: Approaches that have to extract the final fuzzy sets from an internal representation of the network (e.g. projections onto coordinate axes and approximation by piecewise linear functions) suffer from a loss of accuracy. In contrast, methods that directly optimize the fuzzy sets in their final representation form (e.g. NEFCLASS [23]) avoid this problem. However, the drawback of the latter is that usually gradient-based optimization cannot be used due to non-differentiable membership functions and heuristics must be employed instead.

3.2 Mixture-of-experts

A fundamental problem in applying non-linear data-driven models in safety-related applications is that basically their correctness can only be guaranteed in regions of the input space covered by available training and test data. Even if cross-validation methods indicate a good generalization performance, the model may still fail if it is confronting completely new data. In the following, we argue that mixture-of-experts methods can solve this problem under certain conditions.

Initially, the mixture-of-experts (ME) approach was developed to design a system in which different neural networks are responsible for modeling different regions in input space (see e.g. [34]). The ME calculates a function $h : \mathbb{R}^n \to \mathbb{R}$ as a weighted average of M "experts", $f_i(\boldsymbol{x})$, $i = 1, \ldots, M$, using input-dependent weights $g_i(\boldsymbol{x})$.

$$h(\boldsymbol{x}) = \sum_{i=1}^{M} g_i(\boldsymbol{x}) f_i(\boldsymbol{x}), \qquad g_i(\boldsymbol{x}) \geq 0, \quad \sum_{i=1}^{M} g_i(\boldsymbol{x}) = 1 \ .$$

[4] Monotonic membership functions prohibit the interpretation as probability densities.

To be used in a safety-related application we assume the following conditions:

1. There must be at least one *global* expert, i.e. an expert providing a correct (but possibly suboptimal) output in the *complete* input space. This may be e.g. a physical model or a carefully validated fuzzy rule base.
2. *Local* experts may be used to optimize or overrule the global one. However, their respective weight g_i must be zero outside the regions in input space where the local expert has been validated. Additionally, in regions with a low data density the weight should be small in order to rely more on the conservative global expert.

In a variation of this scheme we may alternatively combine the experts in a multiplicative way: $h(\boldsymbol{x}) = \prod_{i=1}^{M} g_i(\boldsymbol{x}) f_i(\boldsymbol{x})$ with $g_i(\boldsymbol{x}) > 0$. Then, the product $g_i(\boldsymbol{x}) f_i(\boldsymbol{x})$ must be 1 outside the validated regions of the local expert. The latter is implemented in the following application.

Example: Industrial Process Control. In this non-linear regression problem, a mixture of an analytical (physical) model of the controlled process and ANNs is used to predict the rolling force in a steel mill. The analytical model (AM) acts as the global expert and ensures a baseline performance, which may be suboptimal. RBF networks multiplicatively correct the output of the AM; for unknown inputs they have been designed to produce a correction factor close to 1, which means the system then relies on the AM. However, for regions of input space where the networks have been trained, they significantly improve the accuracy. The application is deployed at about 40 industrial customers worldwide. Further details are described in [35–37].

4 Conclusions

This paper briefly surveys relevant methods for learning and validation in safety-related domains. Extensive testing, based on e.g. cross-validation or resampling [8], is necessary but not sufficient for validation. Formal proofs of the correctness of the learned model, however, are in most cases infeasible[5]. The remedy is to include safety and semantic constraints in the learning algorithm for ensuring a certain level of interpretability of the learned model by experts, who have to validate it. Neuro-Fuzzy methods are typical examples for this approach (e.g. [28]); they are proven in real applications but finding the right balance between interpretability and accuracy is often a challenge.

In a special form of the mixture of experts approach, a combination of several models improves the accuracy while safety requirements, e.g. reliable extrapolation in regions of the input space not covered by training data, can be met; an example is described in Sect. 3.2.

The recent years in machine learning have seen a particular interest in kernel-based methods, e.g. SVMs for classification, which have some advantages in comparison to conventional ANNs (e.g. convex error function not affected by local

[5] One exception is [3], where deterministic bounds for the approximation error of a neural network are given. However, the correct outputs must be known for the complete input space.

minima) [38]. Using these advantages in safety-related systems would be appealing. However, to achieve this, further research is necessary with the following objectives: (1) increased interpretability of the model and the possibility to include domain knowledge (e.g. via kernel function), (2) reliable confidence levels for the (classification) output, and (3) for classification problems handling of different misclassification costs, which often occur in safety-related systems. The first two are particularly important if training data is scarce and may have a different sample distribution than the data in the deployment phase. For example, an airbag system may encounter quite different crashes in reality than those used for training, i.e. the training samples of the classes are not drawn randomly but biased due to standardized crash tests. Therefore, a model should indicate when its output is less reliable. A combination with a Bayesian framework, as described in [39], could be helpful.

References

1. Taylor, B.J.: Methods and Procedures for the Verification and Validation of Artificial Neural Networks. Springer (2005)
2. Lisboa, P.J.G.: Industrial use of safety-related artificial neural networks. Contract research report 327/2001, Liverpool John Moores University (2001)
3. Zakrzewski, R.: Verification of a trained neural network accuracy. In: Proc. of Int. Joint Conf. on Neural Networks (IJCNN'01). (2001) 1657–1662
4. Bedford, D.F., Austin, J., Morgan, G.: Requirements for a standard certifying the use of artificial neural networks in safety critical applications (1996)
5. Kurd, Z., Kelly, T.P.: Safety lifecycle for developing safety-critical artificial neural networks. In: Proc. of SAFECOMP'03. (2003)
6. Kurd, Z., Kelly, T.P., Austin, J.: Safety criteria and safety lifecycle for artificial neural networks. In: Proc. of Eunite 2003. (2003)
7. Zimmermann, H.G., Grothmann, R., Schäfer, A.M., Tietz, C.: Modeling Large Dynamical Systems with Dynamical Consistent Neural Networks. (2006, to appear)
8. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer (2001)
9. Neuneier, R., Zimmermann, H.G.: How to train neural networks. In: Neural Networks: Tricks of the Trade, Springer (1998) 373–423
10. Ministry of Defence, UK: Defence standard 00-55(part 1)/issue 2 (1997) Requirements for safety related software in defence equipment.
11. Andrews, R., Diederich, J., Tickle, A.B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowl.-Based Syst. **8**(6) (1995) 373–389
12. Craven, M.W.: Extracting comprehensible models from trained neural networks. PhD thesis, University of Wisconsin, Madison (1996)
13. Kolman, E., Margaliot, M.: Are artificial neural networks white boxes? IEEE Transact. on Neural Networks **16**(4) (2005) 844–852
14. Boz, O.: Extracting decision trees from trained neural networks. In: KDD, ACM (2002) 456–461
15. Jacobsson, H.: Rule extraction from recurrent neural networks: A taxonomy and review. Neural Comput. **17**(6) (2005) 1223–1263
16. Craven, M., Shavlik, J.: Visualizing learning and computation an artificial neural networks. Technical report, University of Wisconsin, Madison (1991)

17. Nabney, I.: NETLAB: algorithms for pattern recognitions. Advances in pattern recognition. Springer (2002)
18. MacKay, D.J.C.: Introduction to gaussian processes. In Bishop, C.M., ed.: Neural Networks and Machine Learning. NATO ASI Series. Kluwer (1998) 133–166
19. Mackay, D.J.C.: Bayesian methods for adaptive models. PhD thesis, California Institute of Technology, Pasadena (1992)
20. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
21. Schwaighofer, A., Tresp, V.: Transductive and inductive methods for approximate gaussian process regression. In Becker, S., Thrun, S., Obermayer, K., eds.: NIPS, MIT Press (2002) 953–960
22. Nauck, D., Kruse, R.: Obtaining interpretable fuzzy classification rules from medical data. Artificial Intelligence in Medicine **16**(2) (1999) 149–169
23. Nauck, D.: Fuzzy data analysis with NEFCLASS. Int. J. Approx. Reasoning **32**(2-3) (2003) 103–130
24. Abonyi, J., Roubos, J.A., Szeifert, F.: Data-driven generation of compact, accurate, and linguistically-sound fuzzy classifiers based on a decision-tree initialization. Int. J. Approx. Reasoning **32** (2003) 1–21
25. Ishibuchi, H., Nakashima, T.: Effect of rule weights in fuzzy rule-based classification systems. IEEE Transact. on Fuzzy Systems **9** (2001) 506–515
26. Nürnberger, A., Borgelt, C., Klose, A.: Improving naive bayes classifiers using neuro-fuzzy learning (1999)
27. Mitra, S., Hayashi, Y.: Neuro-fuzzy rule generation: Survey in soft computing framework. IEEE Transact. on Neural Networks **2**(3) (2000) 748–768
28. Kurd, Z., Kelly, T.: Using fuzzy self-organising maps for safety critical systems. In: Proc. of SAFECOMP'04, Springer (2004) 17–30
29. Chen, Y., Wang, J.Z.: Support vector learning for fuzzy rule-based classification systems. IEEE Transact. on Neural Networks **11**(6) (2003) 716–728
30. Lin, C., Yeh, C., Liang, S., Chung, J., Kumar, N.: Support-vector-based fuzzy neural network for pattern classification. IEEE Transact. on Neural Networks **14**(1) (2006) 31–41
31. Hauptmann, W., Graf, F., Heesche, K.: A neuro-fuzzy approach for the sensor-based driving situation identification. In: Proc. of ICONIP. (1997)
32. Casillas, J., Cordon, O., Herrera, F., Magdalena, L., eds.: Accuracy Improvements in Linguistic Fuzzy Modeling. Springer (2003)
33. Klose, A.A.: Partially Supervised Learning of Fuzzy Classification Rules. PhD thesis, University of Magdeburg (2004)
34. Tresp, V.: Committee machines. In Hu, Y.H., Hwang, J.N., eds.: Handbook of Neural Network Signal Processing. CRC Press (2002)
35. Schlang, M., Feldkeller, B., Lang, B., Poppe, T., Runkler, T.: Neural computation in steel industry. In: Proc. European Control Conf. ECC 99, VDI-Verlag (1999)
36. Jansen, M., Broese, E., Feldkeller, B., Poppe, T.: How neural networks are proving themselves in rolling mill process control. In Weinberger, Kratzer, V., ed.: Metals, Mining and More, Siemens AG (ATD MP) (1999)
37. Broese, E., Löffler, H.: Artificial neural networks. In Raabe, D., Roters, F., Barlat, F., Chen, L.Q., eds.: Continuum Scale Simulation of Engineering Materials, Fundamentals - Microstructures - Process Applications, Wiley (2004)
38. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2001)
39. Chu, W.: Bayesian approach to support vector machines. PhD thesis, National University of Singapor (2003)