Problems and Prospects in Fuzzy Data Analysis

Rudolf Kruse[1], Christian Borgelt[1], and Detlef Nauck[2]

[1] Otto-von-Guericke University of Magdeburg
Faculty of Computer Science (FIN-IWS)
Universitaetsplatz 2, D-39106 Magdeburg, Germany
{rudolf.kruse,christian.borgelt}@cs.uni-magdeburg.de

[2] Intelligent Systems Research Group
BT Labs, Adastral Park, Martlesham Heath
Ipswich IP5 3RE, United Kingdom
detlef.nauck@bt.com

Abstract. In meeting the challenges that resulted from the explosion of collected, stored, and transferred data, *Knowledge Discovery in Databases* or *Data Mining* has emerged as a new research area. However, the approaches studied in this area have mainly been oriented towards highly structured and precise data. In addition, the goal to obtain understandable results is often neglected. Therefore we suggest concentrating on *Information Mining*, i.e. the analysis of heterogeneous information sources with the main aim of producing comprehensible results. Since the aim of fuzzy technology has always been to model linguistic information and to achieve understandable solutions, we expect it to play an important role in information mining.

1 Introduction: A View of Information Mining

Due to modern information technology, which produces ever more powerful computers every year, it is possible today to collect, store, transfer, and combine huge amounts of data at very low costs. Thus an ever-increasing number of companies and scientific and governmental institutions can afford to build up large archives of documents and other data like numbers, tables, images, and sounds. However, exploiting the information contained in these archives in an intelligent way turns out to be fairly difficult. In contrast to the abundance of data there is a lack of tools that can transform this data into useful information and knowledge. Although a user often has a vague understanding of the data and its meaning—he can usually formulate hypotheses and guess dependencies— he rarely knows:

- where to find the "interesting" or "relevant" pieces of information;
- whether these pieces of information support his hypotheses and models;
- whether (other) interesting phenomena are hidden in the data,
- which methods are best suited to find the needed pieces of information in a fast and reliable way;

– how the data can be translated into human notions that are appropriate for the context in which it is needed.

In reply to these challenges a new area of research has emerged, which has been named "Knowledge Discovery in Databases" or "Data Mining". Although the standard definition of knowledge discovery and data mining [1] only speaks of discovery in *data*, thus not restricting the type and the organization of the data to work on, it has to be admitted that research up to now concentrated on highly structured data. Usually a minimal requirement is relational data. Most methods (e.g. classical methods like decision trees and neural networks) even demand as input a single uniform table, i.e. a set of tuples of attribute values. It is obvious, however, that this paradigm is hardly adequate for mining image or sound data or even textual descriptions, since it is inappropriate to see such data as, say, tuples of picture elements. Although such data can often be treated successfully by transforming it into structured tables using feature extraction, it is not hard to see that methods are needed which yield, for example, descriptions of what an image depicts, and other methods which can make use of such descriptions, e.g. for retrieval purposes.

Another important point to be made is the following: the fact that pure neural networks are often seen as data mining methods, although their learning result (matrices of numbers) is hardly interpretable, shows that in contrast to the standard definition the goal of *understandable* patterns is often neglected. Of course, there are applications where comprehensible results are not needed and, for example, the prediction accuracy of a classifier is the only criterion of success. Therefore interpretable results should not be seen as a *conditio sine qua non*. However, our own experience—gathered as part of several cooperative ventures with industry—is that modern technologies are accepted more readily, if the methods applied are easy to understand and the results can be checked against human intuition. In addition, if we want to gain insight into a domain, training, for instance, a neural network is not of much help.

Therefore we suggest concentrating on *information mining*, which we see as an extension of data mining and which can be defined in analogy to the KDD definition given in Fayyad et al [1] as follows:

> Information mining is the non-trivial process of identifying valid, novel, potentially useful, and *understandable* patterns in *heterogeneous information sources*.

The term *information* is thus meant to indicate two things: in the first place, it points out that the heterogeneous sources to mine can already provide *information*, understood as expert background knowledge, textual descriptions, images and sounds, etc, and not only raw data. Secondly, it emphasizes that the results must be *comprehensible* ("must provide a user with information"), so that a user can check their plausibility and can get insight into the domain from which the data comes.

For research this results in the challenges:

- to develop theories and scalable techniques that can extract knowledge from large, dynamic, multi-relational, and multi-medial information sources,
- to close the semantic gap between structured data and human notions and concepts, i.e. to be able to translate computer representations into human notions and concepts and vice versa.

The goal of fuzzy systems has always been to model human expert knowledge and to produce systems that are easy to understand. Therefore we expect fuzzy systems technology to play a prominent role in the quest to meet these challenges. In the following we try to point out how fuzzy techniques can help with information mining.

2 Strengths of Fuzzy Set Models

Although there is still some philosophical discussion going on as to whether a (symbolic) language is necessary for consciousness and thinking abilities, it is undisputed that language is humans' most effective tool to structure their experience and to model their environment. Therefore, in order to represent the background knowledge of human experts and to arrive at understandable data mining results, it is absolutely necessary to model linguistic terms and do what Zadeh so pointedly called *computing with words* [2].

A fundamental property of linguistic terms is their inherent vagueness, i.e. they have "fuzzy" boundaries: for each linguistic term there usually are some phenomena to which it can clearly be applied and some others, which cannot be described using this term. But in between these phenomena there lies a "penumbra" of phenomena for which it is not definite whether the term is applicable or not. Well-known examples include the terms *pile of sand* (which is the basis of the classic *sorites* paradox) and *bald*. In both cases no precise number of grains of sand or hairs, respectively, can be given which separates the situations in which the terms are applicable from those in which they are not.

The reason for this inherent vagueness is that for practical purposes full precision is not necessary and may even be a waste of resources. To quote an example by Wittgenstein [3], the request "Please stay around here!" is, of course, inexact. It would be more precise to draw a line on the ground, or, because the line has a certain width and thus would still not be fully exact, to use a colour boundary. But this precision would be entirely pointless, since the inexact request can be expected to work very well.

Fuzzy set theory provides excellent means to model the "fuzzy" boundaries of linguistic terms by introducing gradual memberships. In contrast to classical set theory, in which an object or a case either is a member of a given set (defined, e.g. by some property) or not, fuzzy set theory makes it possible that an object or a case belongs to a set only to a certain degree, thus modelling the penumbra of the linguistic term describing the property that defines the set.

Interpretations of membership degrees include *similarity*, *preference*, and *uncertainty*: they can state how similar an object or case is to a prototypical one, they can indicate preferences between suboptimal solutions to a problem, or they can model uncertainty about the true situation, if this situation is described in imprecise terms. Drawing on Wittgenstein's example as an illustration, we may say that the locations "around here" are (for example, with respect to the person being in sight or calling distance) sufficiently similar to "here", so that the request works well. Or we may say that it would be preferred, if the person stayed exactly "here", but some deviation from "here" would still be acceptable. Finally, if we tell someone to stay "around here" and then go away, we are uncertain about the exact location this person is in at a given moment. It is obvious that all of these interpretations are needed in applications and thus it is not surprising that they have all proven useful for solving practical problems. They also turned out to be worth considering when non-linguistic, but imprecise, i.e. set-valued information has to be modelled.

In general, due to their closeness to human reasoning, solutions obtained using fuzzy approaches are easy to understand and to apply. Due to these strengths, fuzzy systems are the method of choice, if linguistic, vague, or imprecise information has to be modelled.

3 Fuzzy Set Methods in Data Mining

The research in knowledge discovery in databases and data mining has led to a large number of suggestions for a general model of the knowledge discovery process. A recent suggestion for such a model, which can be expected to have considerable impact, since it is backed by several large companies like NCR and DaimlerChrysler, is the CRISP-DM model (CRoss Industry Standard Process for Data Mining) [4].

The basic structure of this process model is depicted in Fig. 1. The circle indicates that data mining is essentially a circular process, in which the evaluation of the results can trigger a re-execution of the data preparation and model generation steps. In this process, fuzzy set methods can profitably be applied in several phases.

The *business understanding* and *data understanding* phases are usually strongly human centred and only little automation can be achieved here. These phases serve mainly to define the goals of the knowledge-discovery project, to estimate its potential benefit, and to identify and collect the necessary data. In addition, background domain knowledge and meta knowledge about the data is gathered. In these phases, fuzzy set methods can be used to formulate, for instance, the background domain knowledge in vague terms, but still in a form that can be used in a subsequent modelling phase. Furthermore, fuzzy database queries are useful to find the data needed and to check whether it may be useful to take additional, related data into account.

In the *data preparation* step, the gathered data is cleaned, transformed, and maybe properly scaled, to produce the input for the modelling techniques. In

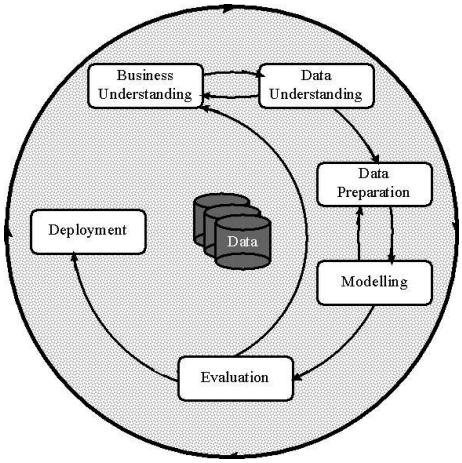

Fig. 1. The CRISP-DM Model.

this step fuzzy methods may, for example, be used to detect outliers, e.g. by *fuzzy clustering* the data [5, 6] and then finding those data points that are far away from the cluster prototypes.

The *modelling* phase, in which models are constructed from the data in order, for instance, to predict future developments or to build classifiers, can, of course, benefit most from fuzzy data analysis approaches. These approaches can be divided into two classes. The first class, *fuzzy data* analysis [7], consists of approaches that analyse fuzzy data—data derived from imprecise measurement instruments or from the descriptions of human domain experts. An example from our own research is the induction of *possibilistic graphical models* [8] from data which complements the induction of the well-known probabilistic graphical models. The second class, fuzzy *data analysis* [9], consists of methods that use fuzzy techniques to structure and analyze crisp data, for instance, *fuzzy clustering* for data segmentation and rule generation and *neuro-fuzzy systems* for rule generation.

In the *evaluation* phase, in which the results are tested and their quality is assessed, the usefulness of fuzzy modelling methods becomes most obvious. Since they yield interpretable systems, they can easily be checked for plausibility against the intuition and expectations of human experts. In addition, the results can provide new insights into the domain under consideration, in contrast to, e.g. pure neural networks, which are black boxes.

To illustrate the usefulness of fuzzy data analysis approaches, in the following sections we discuss the topics generating fuzzy rules from data and learning possibilistic graphical models in a little more detail.

4 Rule Generation with Neuro-Fuzzy-Systems

In order to use fuzzy systems in data analysis, it must be possible to induce fuzzy rules from data. To describe a fuzzy system completely we need to determine a rule base (structure) and fuzzy partitions (parameters) for all variables. The data driven induction of fuzzy systems by simple heuristics based on local computations is usually called *neuro-fuzzy* [10]. If we apply such techniques, we must be aware of the trade-off between precision and interpretability. A fuzzy solution is not only judged for its accuracy, but also—if not especially—for its simplicity and readability. The user of the fuzzy system must be able to comprehend the rule base.

Important points for the interpretability of a fuzzy system are that:

- there are only few fuzzy rules in the rule base;
- there are only few variables used in each rule;
- the variables are partitioned by few meaningful fuzzy sets;
- no linguistic label is represented by more than one fuzzy set.

There are several ways to induce the structure of a fuzzy system. Cluster-oriented and hyperbox-oriented approaches to fuzzy rule learning create rules and fuzzy sets at the same time. Structure-oriented approaches need initial fuzzy partitions to create a rule base [11].

Cluster-oriented rule learning approaches are based on fuzzy cluster analysis [5, 6], i.e. the learning process is unsupervised. Hyperbox-oriented approaches use a supervised learning algorithm that tries to cover the training data by overlapping hyperboxes [12]. Fuzzy rules are created in both approaches by projection of clusters or hyperboxes. The main problem of both approaches is that each generated fuzzy rule uses individual membership functions and thus the rule base is hard to interpret. Cluster-oriented approaches additionally suffer from a loss of information and can only determine an appropriate number of rules, if they are iterated with different fixed rule base sizes.

Structure-oriented approaches avoid all these drawbacks, because they do not search for (hyperellipsoidal or hyperrectangular) clusters in the data space. By providing (initial) fuzzy sets before fuzzy rules are created the data space is structured by a multidimensional fuzzy grid. A rule base is created by selecting those grid cells that contain data. This can be done in a single pass through the training data. This way of learning fuzzy rules was suggested in Wang and Mendel [13]. Extended versions were used in the neuro-fuzzy classification system NEFCLASS [10]. NEFCLASS uses a performance measure for the detected fuzzy rules. Thus the size of the rule base can be determined automatically by adding rules ordered by their performance until all training data is covered. The performance measure is also used to compute the best consequent for each rule.

The number of fuzzy rules can also be restricted by including only the best rules in the rule base. It is also possible to use pruning methods to reduce the number of rules and the number of variables used by the rules. In order to obtain meaningful fuzzy partitions, it is better to create rule bases by structure-oriented learning than by cluster-oriented or by hyperbox-oriented rule learning.

The latter two approaches create individual fuzzy sets for each rule and thus provide less interpretable solutions. Structure-oriented methods allow the user to provide appropriate fuzzy partitions in advance such that all rules share the same fuzzy sets. Thus the induced rule base can be interpreted well.

After the rule base of a fuzzy system has been generated, we must usually train the membership function in order to improve the performance. In NEFCLASS, for example, the fuzzy sets are tuned by a simple backpropagation-like procedure. The algorithm does not use gradient-descent, because the degree of fulfilment of a fuzzy rule is determined by the minimum, and non-continuous membership function may be used. Instead a simple heuristics is used that results in shifting the fuzzy sets and in enlarging or reducing their support.

The main idea of NEFCLASS is to create comprehensible fuzzy classifiers, by ensuring that fuzzy sets cannot be modified arbitrarily during learning. Constraints can be applied in order to make sure that the fuzzy sets still fit their linguistic labels after learning. For the sake of interpretability we do not want adjacent fuzzy sets to exchange positions, we want the fuzzy sets to overlap appropriately, etc.

We will not describe more details of learning fuzzy rules here, but refer to the paper on "NEFCLASS-J – A Java based Soft Computing Tool" in this volume. In the next section we discuss some aspects of information fusion that can be implemented by neuro-fuzzy systems.

5 Information Fusion with Neuro-Fuzzy Models

If neuro-fuzzy methods are used in information mining, it is useful to consider their capabilities in fusing information from different sources. Information fusion refers to the acquisition, processing, exploitation, and merging of information originating from multiple sources to provide a better insight and understanding of the phenomena under consideration. There are several levels of information fusion. Fusion may take place at the level of data acquisition, data pre-processing, data or knowledge representation, or at the model or decision making level. On lower levels where raw data is involved, the term (sensor) *data fusion* is preferred. Some aspects of information fusion can be implemented by NEFCLASS. For a conceptual and comparative study of fusion strategies in various calculi of uncertainty see Gebhardt and Kruse [14] and Dubois et al [15].

If a fuzzy classifier is created based on a supervised learning problem, then the most common way is to provide a data set, where each pattern is labelled—ideally with its correct class. That means we assume that each pattern belongs to one class only. Sometimes it is not possible to determine this class correctly due to a lack of information. Instead of a crisp classification it would also be possible to label each pattern with a vector of membership degrees. This requires that a vague classification is obtained in some way for the training patterns, e.g. by partially contradicting expert opinions.

Training patterns with fuzzy classifications are one way to implement information fusion with neuro-fuzzy systems. If we assume that a group of n experts

provide partially contradicting classifications for a set of training data we can fuse the expert opinions into fuzzy sets that describe the classification for each training pattern. According to the context model, we can view the experts as different observation contexts [16]. The training then reflects fusion of expert opinions on data set level.

Due to the capabilities of its learning algorithms NEFCLASS can handle such training data in the process of creating a fuzzy classifier. However, it does not implement fusion on data set level itself. For information fusion in neuro-fuzzy environments like NEFCLASS we usually consider three operator schemes:

$$\begin{array}{lcl} \text{fuse}(R, R') & : & \text{fuse two rule sets } R \text{ and } R', \\ \text{induce}(D) & : & \text{induce a rule set from a given data set } D, \\ \text{revise}(R, D) & : & \text{revise a rule set in the light of a data set } D. \end{array}$$

An aspect of information fusion that is implemented by NEFCLASS is to integrate expert knowledge in form of a set of fuzzy rules R and knowledge induced from a data set D:

$$\text{fuse}(R, \text{induce}(D)).$$

If expert knowledge about the classification problem is available, then the rule base of the fuzzy classifier can be initialized with suitable fuzzy rules before rule learning is invoked to complete the rule base. If the algorithm creates a rule from data that contradicts with an expert rule then we can, for example:

- always prefer expert rule;
- always prefer the learned rule; or
- select the rule with the higher performance value.

In NEFCLASS we determine the performance of all rules over the training data and in case of contradiction the better rule prevails. This reflects fusion of expert opinions and knowledge obtained from observations. Note that providing a rule base and tuning it, e.g. by modifying membership functions, is not information fusion but knowledge revision or update:

$$\text{revise}(R, D).$$

In this case the rule base is seen as prior knowledge and the tuned rule base is posterior knowledge. This approach is also known in Bayesian statistics, where a given prior probability distribution is revised by additional evidence to a posterior distribution [17, 18]. Since NEFCLASS is mainly used to train a fuzzy rule base it usually performs

$$\text{revise}(\text{fuse}(R, \text{induce}(D)))$$

if an expert's rule base is given in advance.

Because NEFCLASS is able to resolve conflicts between rules based on rule performance, it is also able to fuse expert opinions on fuzzy rule level:

$$\text{fuse}(R, R').$$

Rule bases R and R' from different experts can be entered into the system. They will then be fused into one rule base and contradictions are resolved automatically by deleting from each pair of contradicting rules the rule with lower performance.

After all contradictions between expert rules and rules learned from data were resolved, usually not all rules can be included into the rule base, because its size is limited by some criterion. In this case we must decide whether:

- to include expert rules in any case; or
- to include rules by descending performances values.

The decision depends on the trust we have in the expert's knowledge and in the training data. A mixed approach can be used, e.g. include the best expert rules and then use the best learned rules to complete the rule base.

A similar decision must be made, when the rule base is pruned after training, i.e. is it acceptable to remove an expert rule during pruning, or must such rules remain in the rule base. In NEFCLASS expert rules and rules induced from data are not treated differently.

An example of information fusion in neuro-fuzzy system with an application to stock index prediction can be found in Siekmann et al [19].

6 Dependency Analysis with Graphical Models

Since reasoning in multi-dimensional domains tends to be infeasible in the domains as a whole—and the more so, if uncertainty and imprecision are involved—decomposition techniques, that reduce the reasoning process to computations in lower-dimensional subspaces, have become very popular. In the field of graphical modelling, *decomposition* is based on dependence and independence relations between the attributes or variables that are used to describe the domain under consideration. The structure of these dependence and independence relations are represented as a graph (hence the name graphical models), in which each node stands for an attribute and each edge for a direct dependence between two attributes. The precise set of dependence and (conditional) independence statements that hold in the modeled domain can be read from the graph using simple graph theoretic criteria, for instance, d-separation, if the graph is a directed one, or simple separation, if the graph is undirected.

The conditional independence graph (as it is also called) is, however, only the *qualitative* or *structural component* of a graphical model. To do reasoning, it has to be enhanced by a *quantitative component* that provides confidence information about the different points of the underlying domain. This information can often be represented as a distribution function on the underlying domain, for example, a probability distribution, a possibility distribution, a mass distribution, etc. With respect to this quantitative component, the conditional independence graph describes a *factorization* of the distribution function on the domain as a whole into conditional or marginal distribution functions on lower-dimensional subspaces.

Graphical models make reasoning much more efficient, because propagating the evidential information about the values of some attributes to the unobserved ones and computing the marginal distributions for the unobserved attributes can be implemented by locally communicating node and edge processors in the conditional independence graph.

For some time the standard approach to construct a graphical model has been to let a human domain expert specify the dependency structure of the considered domain. This provided the conditional independence graph. Then the human domain expert had to estimate the necessary conditional or marginal distribution functions, which then formed the quantitative component of the graphical model. This approach, however, can be tedious and time consuming, especially, if the domain under consideration is large. In addition, it may be impossible to carry it out, if no or only vague knowledge is available about the dependence and independence relations that hold in the domain to be modelled. Therefore recent research has concentrated on learning graphical models from databases of sample cases.

Due to the origin of graphical modelling research in probabilistic reasoning, the most widely known methods are, of course, learning algorithms for Bayesian or Markov networks. However, these approaches—as probabilistic approaches do in general—suffer from certain deficiencies, if imprecise information, understood as set-valued data, has to be taken into account. For this reason recently possibilistic graphical models also gained some attention [8], for which learning algorithms have been developed in analogy to the probabilistic case. These methods can be used to do dependency analysis, even if the data to analyse is highly imprecise, and can thus offer interesting perspectives for future research.

We have implemented these methods as a plug-in for the well-known data mining tool *Clementine* (ISL/SPSS). Its probabilistic version is currently used at DaimlerChrysler for fault analysis.

7 Possibilistic Graphical Models

A *possibility distribution* π on a universe of discourse Ω is a mapping from Ω into the unit interval, i.e. $\pi : \Omega \to [0, 1]$, see Zadeh [20] and Dubois and Prade [21]. From an intuitive point of view, $\pi(\omega)$ quantifies the degree of possibility that $\omega = \omega_0$ is true, where ω_0 is the actual state of the world: $\pi(\omega) = 0$ means that $\omega = \omega_0$ is impossible, $\pi(\omega) = 1$ means that $\omega = \omega_0$ is possible without any restrictions, and $\pi(\omega) \in (0, 1)$ means that $\omega = \omega_0$ is possible only with restrictions, i.e. that there is evidence that supports $\omega = \omega_0$ as well as evidence that contradicts $\omega = \omega_0$.

Several suggestions have been made for semantics of a *theory of possibility* as a framework for reasoning with uncertain and imprecise data. The interpretation of a degree of possibility we prefer is based on the context model [22, 16]. In this model possibility distributions are seen as *information-compressed* representations of (not necessarily nested) random sets and a degree of possibility as the one-point coverage of a random set [23].

To be more precise: Let ω_0 be the actual, but unknown state of a domain of interest, which is contained in a set Ω of possible states. Let $(C, 2^C, P)$, $C = \{c_1, c_2, \ldots, c_m\}$, be a finite probability space and $\gamma : C \to 2^\Omega$ a set-valued mapping. C is seen as a set of contexts that have to be distinguished for a set-valued specification of ω_0. The contexts are supposed to describe different physical and observation-related frame conditions. $P(\{c\})$ is the (subjective) probability of the (occurrence or selection of the) context c.

A set $\gamma(c)$ is assumed to be the *most specific correct set-valued specification* of ω_0, which is implied by the frame conditions that characterize the context c. By "most specific set-valued specification" we mean that $\omega_0 \in \gamma(c)$ is guaranteed to be true for $\gamma(c)$, but is not guaranteed for any proper subset of $\gamma(c)$. The resulting *random set* $\Gamma = (\gamma, P)$ is an imperfect (i.e. imprecise *and* uncertain) specification of ω_0. Let π_Γ denote the *one-point coverage of* Γ (the *possibility distribution induced by* Γ*)*, which is defined as

$$\pi_\Gamma : \Omega \to [0, 1], \quad \pi_\Gamma(\omega) = P\left(\{c \in C \mid \omega \in \gamma(c)\}\right).$$

In a complete model the contexts in C must be specified in detail to make the relationships between all contexts c_j and their corresponding specifications $\gamma(c_j)$ explicit. But if the contexts are unknown or ignored, then $\pi_\Gamma(\omega)$ is the total mass of all contexts c that provide a specification $\gamma(c)$ in which ω_0 is contained, and this quantifies the *possibility of truth* of the statement "$\omega = \omega_0$" [22, 24].

That in this interpretation a possibility distribution represents uncertain *and* imprecise knowledge can be understood best by comparing it to a probability distribution and to a relation. A probability distribution covers *uncertain*, but *precise* knowledge. This becomes obvious, if one notices that a possibility distribution in the interpretation described above reduces to a probability distribution, if $\forall c_j \in C : |\gamma(c_j)| = 1$, i.e. if for all contexts the specification of ω_0 is precise. On the other hand, a relation represents *imprecise*, but *certain* knowledge about dependencies between attributes. Thus, not surprisingly, a relation can also be seen as a special case of a possibility distribution, namely if there is only one context. Hence the context-dependent specifications are responsible for the imprecision, the contexts for the uncertainty in the imperfect knowledge expressed by a possibility distribution.

Although well-known for a couple of years [25], a unique concept of possibilistic independence has not been fixed yet. In our opinion, the problem is that possibility theory is a calculus for uncertain *and* imprecise reasoning, the first of which is related to probability theory, the latter to relational theory (see above). But these two theories employ different notions of independence, namely stochastic independence and lossless join decomposability. Stochastic independence is an *uncertainty-based* type of independence, whereas lossless join decomposability is an *imprecision-based* type of independence. Since possibility theory addresses both kinds of imperfect knowledge, notions of possibilistic independence can be uncertainty-based or imprecision-based.

With respect to this consideration two definitions of possibilistic independence have been justified [26], namely uncertainty-based possibilistic independence, which is derived from *Dempster's rule of conditioning* [27] adapted to

possibility measures, and imprecision-based possibilistic independence, which coincides with the well-known concept of *possibilistic non-interactivity* [21]. The latter can be seen as a generalization of lossless join decomposability to the possibilistic setting, since it treats each α-cut of a possibility distribution like a relation.

Because of its consistency with the *extension principle* [28], we confine ourselves to possibilistic non-interactivity. As a concept of possibilistic independence it can be defined as follows: let X, Y, and Z be three disjoint subsets of variables in V. Then X is called *conditionally independent* of Y given Z with respect to π, abbreviated $X \perp\!\!\!\perp_\pi Y \mid Z$, iff

$$\forall \omega \in \Omega: \quad \pi(\omega_{X \cup Y} \mid \omega_Z) = \min\{\pi(\omega_X \mid \omega_Z), \pi(\omega_Y \mid \omega_Z)\}$$

whenever $\pi(\omega_Z) > 0$, where $\pi(\cdot \mid \cdot)$ is a non-normalized conditional possibility distribution, i.e.

$$\pi(\omega_X \mid \omega_Z) = \max\{\pi(\omega') \mid \omega' \in \Omega \wedge \mathrm{proj}_X^V(\omega) = \omega_X \wedge \mathrm{proj}_Z^V(\omega) = \omega_Z\}.$$

Both mentioned types of possibilistic independence satisfy the *semi-graphoid axioms* [29, 30]. Possibilistic independence based on Dempster's rule in addition satisfies the intersection axiom and thus can be used within the framework of the valuation-based systems already mentioned above [31]. However, the intersection axiom is related to uncertainty-based independence. Relational independence does not satisfy this axiom, and therefore it cannot be satisfied by possibilistic non-interactivity as a more general type of imprecision-based independence.

Similar to probabilistic networks, a possibilistic network can be seen as a decomposition of a multi-variate possibility distribution. The factorization formulae can be derived from the corresponding probabilistic factorization formulae (for Markov networks) by replacing the product by the minimum.

Just as for probabilistic networks, it is possible in principle to estimate the quality of a given possibilistic network by exploiting its factorization property. For each $\omega \in \Omega$ the degree of possibility computed from the network is compared to the degree of possibility derived from the database to learn from. But again this approach can be costly.

Contrary to probabilistic networks, the induction of possibilistic networks from data has been studied much less extensively. A first result, which consists in an algorithm that is closely related to the $K2$ algorithm for the induction of Bayesian networks, was presented in Gebhardt and Kruse [32]. Instead of the Bayesian evaluation measure used in $K2$, it relies on a measure derived from the *nonspecificity* of a possibility distribution. Roughly speaking, the notion of nonspecificity plays the same role in possibility theory that the notion of *entropy* plays in probability theory. Based on the connection of the imprecision part of a possibility distribution to relations, the nonspecificity of a possibility distribution can also be seen as a generalization of *Hartley information* [33] to the possibilistic setting.

In Gebhardt and Kruse [34] a rigid foundation of a learning algorithm for possibilistic networks is given. It starts from a comparison of the nonspecificity

of a given multi-variate possibility distribution to the distribution represented by a possibilistic network, thus measuring the loss of specificity, if the multi-variate possibility distribution is represented by the network. In order to arrive at an efficient algorithm, an approximation for this loss of specificity is derived, which can be computed locally on the hyperedges of the network. As the search method a generalization of the optimum weight spanning tree algorithm to hypergraphs is used. Several other heuristic local evaluation measures, which can be used with different search methods, are presented in Borgelt and Kruse [35, 36].

It should be emphasized, that, as already discussed above, an essential advantage of possibilistic networks over probabilistic ones is their ability to deal with imprecision, i.e. multi-valued, information. When learning possibilistic networks from data, this leads to the convenient situation that missing values in an observation or a set of values for an attribute, all of which have to be considered possible, do not pose any problems.

8 Concluding Remarks

In knowledge discovery and data mining as it is, there is a tendency to focus on purely data-driven approaches in a first step. More model-based approaches are only used in the refinement phases (which in industry are often not necessary, because the first successful approach wins—and the winner takes all). However, to arrive at truly useful results, we must take background knowledge and, in general, non-numeric information into account and we must concentrate on comprehensible models.

The complexity of the learning task, obviously, leads to a problem: when learning from information, one must choose between (often quantitative) methods that achieve good performance and (often qualitative) models that explain what is going on to a user. This is another good example of Zadeh's principle of the incompatibility between precision and meaning. Of course, precision and high performance are important goals. However, in the most successful fuzzy applications in industry such as intelligent control and pattern classification, the introduction of fuzzy sets was motivated by the need for more human-friendly computerized devices that help a user to formulate his knowledge and to clarify, to process, to retrieve and to exploit the available information in a most simple way. In order to achieve this user-friendliness, often certain (limited) reductions in performance and solution quality are accepted.

So the question is: what is a good solution from the point of view of a user in the field of information mining? Of course, correctness, completeness, and efficiency are important, but in order to manage systems that are more and more complex, there is a constantly growing demand to keep the solutions conceptually simple and understandable. This calls for a formal theory of utility in which the simplicity of a system is taken into account. Unfortunately such a theory is extremely hard to come by, because for complex domains it is difficult to measure the degree of simplicity and it is even more difficult to assess the gain

achieved by making a system simpler. Nevertheless, this is a lasting challenge for the fuzzy community to meet.

References

1. Fayyad U., Piatetsky-Shapiro G., Smyth P. and Uthurusamy R., Eds.: 'Advances in Knowledge Discovery and Data Mining'. Cambridge, MA, USA, AAAI Press / MIT Press (1996).
2. Zadeh L.: 'Fuzzy logic = computing with words'. IEEE Transactions on Fuzzy Systems, Vol. 4, pp. 103–111 (1996).
3. Wittgenstein L.: 'Philosophical Investigations'. Englewood Cliffs, NJ, USA, Prentice Hall (1973 (first published 1952)).
4. Chapman P., Clinton J., Khabaza T., Reinartz T. and Wirth R. 'The crisp-dm process model', (1999). Available from http://www.ncr.dk/CRISP/.
5. Bezdek J., Keller J., Krishnapuram R. and Pal N.: 'Fuzzy Models and Algorithms for Pattern Recognition and Image Processing'. The Handbooks on Fuzzy Sets. Norwell MA, USA, Kluwer (1998).
6. Höppner F., Klawonn F., Kruse R. and Runkler T.: 'Fuzzy Cluster Analysis'. Chichester, England, J. Wiley & Sons (1999).
7. Kruse R. and Meyer K.: 'Statistics with Vague Data'. Dordrecht, Netherlands, Reidel (1987).
8. Borgelt C., Gebhardt J. and Kruse R.: 'Chapter f1.2: Inference methods'. In 'Handbook of Fuzzy Computation', E. Ruspini, P. Bonissone, and W. Pedrycz, Eds. Institute of Physics Publishing Ltd., Bristol, United Kingdom (1998).
9. Bandemer H. and Näther W.: 'Fuzzy Data Analysis'. Dordrecht, Netherlands, Kluwer (1992).
10. Nauck D., Klawonn F. and Kruse R.: 'Foundations of Neuro-Fuzzy Systems'. Chichester, England, J. Wiley & Sons (1997).
11. Nauck D. and Kruse R.: 'Chapter d.2: Neuro-fuzzy systems'. In 'Handbook of Fuzzy Computation', P. B. E. Ruspini and W. Pedrycz, Eds. Institute of Physics Publishing Ltd., Bristol, UK (1998).
12. Berthold M. and Huber K.: 'Constructing fuzzy graphs from examples'. Int. J. Intelligent Data Analysis, Vol. 3(1), pp. 37–51 (1999).
13. Wang L.-X. and Mendel J.: 'Generating fuzzy rules by learning from examples'. IEEE Trans. Syst., Man, Cybern., Vol. 22, pp. 1414–1227 (1992).
14. Gebhardt J. and Kruse R.: 'Parallel combination of information sources'. In 'Handbook of Defeasible Reasoning and Uncertainty Management Systems. Vol. 3: Belief Change', D. Gabbay and P. Smets, Eds. Kluwer, Dordrecht, Netherlands, pp. 329–375 (1998).
15. Dubois D., Prade H. and Yager R.: 'Merging fuzzy information'. In 'Approximate Reasoning and Fuzzy Information Systems', D. D. J.C. Bezdek and H. Prade, Eds. Kluwer, Dordrecht, Netherlands, pp. 335–402 (1999).
16. Kruse R., Gebhardt J. and Klawonn F.: 'Foundations of Fuzzy Systems'. Chichester, England, J. Wiley & Sons (1994).
17. Cooke R.: 'Experts in Uncertainty: Opinion and Subjectivity Probability in Science'. New York, NY, Oxford University Press (1991).
18. Fisher D. and Lenz H.: 'Learning from Data.'. No. 112 in Lecture Notes in Statistics. New York, NY, Springer (1996).

19. Siekmann S., Gebhardt J. and Kruse R.: 'Information fusion in the context of stock index prediction'. In 'Proc. ECSQARU'99' (London, UK) (1999).
20. Zadeh L.: 'Fuzzy sets as a basis for a theory of possibility'. Fuzzy Sets and Systems, Vol. 1, pp. 3–28 (1978).
21. Dubois D. and Prade H.: 'Possibility Theory'. New York, NY, Plenum Press (1988).
22. Gebhardt J. and Kruse R.: 'The context model — an integrating view of vagueness and uncertainty'. Int. Journal of Approximate Reasoning, Vol. 9, pp. 283–314 (1993).
23. Nguyen H.: 'Using random sets'. Information Science, Vol. 34, pp. 265–274 (1984).
24. Gebhardt J. and Kruse R.: 'Possinfer — a software tool for possibilistic inference'. In 'Fuzzy Set Methods in Information Engineering: A Guided Tour of Applications', H. P. D. Dubois and R. Yager, Eds. Wiley, New York, NY, pp. 407–418 (1996).
25. Hisdal E.: 'Conditional possibilities, independence, and noninteraction'. Fuzzy Sets and Systems, Vol. 1, pp. 283–297 (1978).
26. de Campos L., Gebhardt J. and Kruse R.: 'Syntactic and semantic approaches to possibilistic independence'. Technical report, University of Granada Spain, and University of Braunschweig, Germany (1995).
27. Shafer G.: 'A Mathematical Theory of Evidence'. Princeton, NJ, Princeton University Press (1976).
28. Zadeh L.: 'The concept of a linguistic variable and its application to approximate reasoning'. Information Sciences, Vol. 9, pp. 43–80 (1975).
29. Dawid A.: 'Conditional independence in statistical theory'. SIAM Journal on Computing, Vol. 41, pp. 1–31 (1979).
30. Pearl J. and Paz A.: 'Graphoids: a graph based logic for reasoning about relevance relations'. In 'Advances in Artificial Intelligence 2', B. B. et al, Ed. North Holland, Amsterdam, Netherlands, pp. 357–363 (1987).
31. Shafer G. and Shenoy P.: 'Local computations in hypertrees'. Working paper 201, School of Business, University of Kansas, Lawrence, KS (1988).
32. Gebhardt J. and Kruse R.: 'Learning possibilistic networks from data'. In 'Proc. 5th Int. Workshop on Artificial Intelligence and Statistics' (Fort Lauderdale, FL), pp. 233–244 (1995).
33. Hartley R.: 'Transmission of information'. The Bell Systems Technical Journal, Vol. 7, pp. 535–563 (1928).
34. Gebhardt J. and Kruse R.: 'Tightest hypertree decompositions of multivariate possibility distributions'. In 'Proc. Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU'96)' (Granada, Spain), pp. 923–927 (1996).
35. Borgelt C. and Kruse R.: 'Evaluation measures for learning probabilistic and possibilistic networks'. In 'Proc. 6th IEEE Int. Conf. on Fuzzy Systems' (Barcelona, Spain), pp. 669–676 (1997).
36. Borgelt C. and Kruse R.: 'Some experimental results on learning probabilistic and possibilistic networks with different evaluation measures'. In 'Proc. 1st Int. J. Conf. on Qualitative and Quantitative Practical Reasoning, ECSQARU-FAPR'97' (Bad Honnef, Germany), pp. 71–85 (1997).