Fuzzy Sets in Machine Learning and Data Mining

Eyke Hüllermeier

Philipps-Universität Marburg
Department of Mathematics and Computer Science
Marburg, Germany
eyke@informatik.uni-marburg.de

Abstract

Machine learning, data mining, and several related research areas are concerned with methods for the automated induction of models and the extraction of interesting patterns from empirical data. Automated knowledge acquisition of that kind has been an essential aspect of artificial intelligence since a long time and has more recently also attracted considerable attention in the fuzzy sets community. This paper briefly reviews some typical applications and highlights potential contributions that fuzzy set theory can make to machine learning, data mining, and related fields. In this connection, some advantages of fuzzy methods for representing and mining vague patterns in data are especially emphasized.

1 Introduction

As the conception of intelligence is inseparably connected with the ability to learn from experience and to adapt to new situations, it is hardly astonishing that machine learning has always been considered as an integral part of the field of artificial intelligence (AI). In fact, the key role of learning and adaptation is almost self-evident for the connectionist approach to AI, since representing knowledge with (artificial) neural networks becomes possible only by "training" such networks. However, in view of the fact that the "knowledge acquisition bottleneck" turned out to be one of the key problems in the design of intelligent and knowledge-based systems, the importance of automated methods for knowledge acquisition has also been realized for more symbol-oriented approaches like rule-based systems (for which expert knowledge could in principle be injected by hand). In any case, machine learning is certainly one of the most important subfields of contemporary AI, regardless of the particular facet.

A research field closely related to machine learning is that of *knowledge discovery and data mining*. As a response to the progress in digital data acquisition and storage technology, along with the limited human capabilities in analyzing and exploiting large amounts of data, this field has recently emerged as a new research discipline, lying at the intersection of statistics, machine learning, data management, and other areas. According to a widely accepted definition, knowledge discovery refers to the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable structure in data [25]. The central step within this process is *data mining*, the application of computational methods and algorithms for extracting useful patterns from potentially very large data sets. Meanwhile, knowledge discovery has established itself as a new, independent research field, including its own journals and conferences.

As mentioned before, the fields of machine learning and data mining are closely related, and a strict separation between them is neither reasonable nor desirable. A rough distinction that we shall make throughout the paper relates to the distinction between the performance tasks of *pattern discovery* and *model induction*. While we consider the latter to be the core problem of machine learning, the former is more in the realm of data mining. A typical example of model induction is learning a *classifier* from a set of training examples, i.e., a function $C : \mathcal{X} \rightarrow \mathcal{Y}$ that (hypothetically) assigns a class $y = C(x) \in \mathcal{Y}$ to every potential input x from an underlying instance space $\mathcal{X}$. One of the most important criteria for this type of problem is generalization performance, that is, the predictive accuracy of an induced model. According to our view, data mining is of a more explorative nature, and patterns discovered in a data set are usually of a *local* and *descriptive* rather than of a *global* and *predictive* type. A typical example of such a pattern is an *association rule*, to be discussed in more detail in Section 4. For pattern discovery methods, evaluation criteria are more diverse and often more difficult to quantify; essentially, patterns should be “interesting” or “useful” in one way or the other. Our distinction between machine learning and data mining can roughly be seen as a “modern” or extended distinction between descriptive and inductive statistics. Anyway, we repeat that this distinction is very rough and represents a quite simplified view, which is not an *opinio communis* (for example, some people prefer having an even more general view of data mining that includes machine learning as a special case). Under the heading ML&DM we shall subsequently subsume both machine learning and data mining, as well as related fields like various forms of data analysis (distinguished by adjectives like multivariate, exploratory, Bayesian, intelligent, ...).

In fuzzy set theory (FST), one of the cornerstones of soft computing, aspects of knowledge representation and reasoning have dominated research for a long time, at least in that part of the theory which lends itself to intelligent systems design and applications in AI. Yet, problems of automated learning and knowledge acquisition have more and more come to the fore in recent years, and numerous contributions to ML&DM have been made in the meantime. The aim of this paper is to convey an impression of the current status and prospects of FST in ML&DM, especially highlighting potential features and advantages of fuzzy in comparison with non-fuzzy approaches.

The remainder of the paper is organized as follows:[1] Section 2 presents a collection of typical applications of FST in ML&DM; the examples given are representative though not complete, and the section is definitely not a comprehensive review of the literature. In Section 3, we try to highlight in a more systematic way the potential contributions that FST can make to machine learning and data mining. One of the core advantages of fuzzy methods for data mining, namely an increased expressiveness that contributes to representing and mining vague patterns in data, is discussed in more detail and illustrated in the context of association analysis in Section 4. Section 5 completes the paper with some concluding remarks.

2 Typical Applications of Fuzzy Set Theory

The tools and technologies that have been developed in FST have the potential to support all of the steps that comprise a process of model induction or knowledge discovery. In particular, FST can already be used in the data selection and preparation phase, e.g., for modeling vague data in terms of fuzzy sets [55], to "condense" several crisp observations into a single fuzzy one, or to create fuzzy summaries of the data [38]. As the data to be analyzed thus becomes fuzzy, one subsequently faces a problem of analyzing fuzzy data, i.e., of *fuzzy data* analysis [3].

The problem of analyzing fuzzy data can be approached in at least two principally different ways. First, standard methods of data analysis can be extended in a rather generic way by means of an extension principle, that is, by "fuzzifying" the mapping from data to models. A second, often more sophisticated approach is based on embedding the data into more complex mathematical spaces, such as fuzzy metric spaces [16], and to carry out data analysis in these spaces [17].

If fuzzy methods are not used in the data preparation phase, they can still be employed in a later stage in order to analyze the original data. Thus, it is not the data to be analyzed that is fuzzy, but rather the methods used for analyzing the data (in the sense of resorting to tools from FST). Subsequently, we shall focus on this type of fuzzy data analysis (where the adjective "fuzzy" refers to the term *analysis*, not to the term *data*), which is predominant in

[1] Sections 2 and 3 closely correspond to parts of the related survey paper [35].

ML&DM.[2] In the following, we focus on fuzzy extensions of some well-known machine learning and data mining methods without repeating the original methods themselves; thus, we assume basic familiarity with these methods.

2.1 Fuzzy Cluster Analysis

Many conventional clustering algorithms, such as the prominent k-means algorithm, produce a clustering structure in which every object is assigned to one cluster in an unequivocal way. Consequently, the individual clusters are separated by sharp boundaries. In practice, such boundaries are often not very natural or even counterintuitive. Rather, the boundary of single clusters and the transition between different clusters are usually "smooth". This is the main motivation underlying fuzzy extensions to clustering algorithms [28]. In fuzzy clustering, an object may belong to different clusters at the same time, at least to some extent, and the degree to which it belongs to a particular cluster is expressed in terms of a fuzzy membership. The membership functions of the different clusters (defined on the set of observed data points) is usually assumed to form a partition of unity. This version, often called probabilistic clustering, can be generalized further by weakening this constraint as, e.g., in possibilistic clustering [37]. Fuzzy clustering has proved to be extremely useful in practice and is now routinely applied also outside the fuzzy community (e.g., in recent bioinformatics applications [26]).

2.2 Learning Fuzzy Rule-Based Systems

The most frequent application of FST in machine learning is the induction or the adaptation of rule-based models. This is hardly astonishing, since rule-based models have always been a cornerstone of fuzzy systems and a central aspect of research in the field, not only in ML&DM but also in many other subfields, notably approximate reasoning and fuzzy control. (Often, the term *fuzzy system* implicitly refers to fuzzy *rule-based* system.)

Fuzzy rule-based systems can represent both classification and regression functions, and different types of fuzzy models have been used for these purposes. In order to realize a regression function, a fuzzy system is usually wrapped in a "fuzzifier" and a "defuzzifier": The former maps a crisp input to a fuzzy one, which is then processed by the fuzzy system, and the latter maps the (fuzzy)

[2] By now, the analysis of fuzzy data is still hampered by the non-availability of such data. It is likely to become more important in the future, once the enabling data engineering technology, allowing to acquire, store, and process fuzzy data on a large scale, has been established.

output of the system back to a crisp value. For so-called Takagi-Sugeno models, which are quite popular for modeling regression functions, the defuzzification step is unnecessary, since these models output crisp values directly.

In the case of classification learning, the consequent of single rules is usually a class assignment (i.e. a singleton fuzzy set). [3] Evaluating a rule base (*à la* Mamdani-Assilian) thus becomes trivial and simply amounts to "maximum matching", that is, searching the maximally supporting rule for each class. Thus, much of the appealing interpolation and approximation properties of fuzzy inference gets lost, and fuzziness only means that rules can be activated to a certain degree. There are, however, alternative methods which combine the predictions of several rules into a classification of the query [12]. In methods of that kind, the degree of activation of a rule provides important information. Besides, activation degrees can be very useful, e.g., for characterizing the uncertainty involved in a classification decision.

A plethora of strategies has been developed for inducing a fuzzy rule-based system from the data given, and we refrain from a detailed exposition here. Especially important in the field of fuzzy rule learning are hybrid methods that combine FST with other (soft computing) methodologies, notably evolutionary algorithms and neural networks. For example, evolutionary algorithms are often used to optimize ("tune") a fuzzy rule base or for searching the space of potential rule bases in a (more or less) systematic way [13]. Quite interesting are also *neuro-fuzzy* methods [43]. For example, one idea is to encode a fuzzy system as a neural network and to apply standard methods (like backpropagation) in order to train such a network. This way, neuro-fuzzy systems combine the representational advantages of fuzzy systems with the flexibility and adaptivity of neural networks.

2.3 Fuzzy Decision Tree Induction

Fuzzy variants of decision tree induction have been developed for quite a while (e.g. [56,36]) and seem to remain a topic of interest even today [47–49] (see [44] for a recent approach and a comprehensive overview of research in this field). In fact, these approaches provide a typical example for the "fuzzification" of standard machine learning methods. In the case of decision trees, it is primarily the "crisp" thresholds used for defining splitting predicates (constraints), such as `size` ≤ 181, at inner nodes that have been criticized: Such thresholds lead to hard decision boundaries in the input space, which means that a slight variation of an attribute (e.g. `size` $= 182$ instead of `size` $= 181$) can entail a completely different classification of an object (e.g., of a person characterized

[3] More generally, a rule consequent can suggest different classes with different degrees of certainty.

by size, weight, gender, ...) Moreover, the learning process becomes unstable in the sense that a slight variation of the training examples can change the induced decision tree drastically.

In order to make the decision boundaries "soft", an obvious idea is to apply fuzzy predicates at the inner nodes of a decision tree, such as `size` $\in$ `TALL`, where `TALL` is a fuzzy set (rather than an interval). In other words, a fuzzy partition instead of a crisp one is used for the splitting attribute (here `size`) at an inner node. Since an example can satisfy a fuzzy predicate to a certain degree, the examples are partitioned in a fuzzy manner as well. That is, an object is not assigned to exactly one successor node in a unique way, but perhaps to several successors with a certain degree. For example, a person whose size is 181 cm could be an element of the `TALL`-group to the degree, say, 0.7 and of the complementary group to the degree 0.3.

The above idea of "soft recursive partitioning" has been realized in different ways. Moreover, the problems entailed by corresponding fuzzy extensions have been investigated. For example, how can splitting measures like information gain (expressed in terms of entropy), originally defined for ordinary sets of examples, be extended to fuzzy sets of examples [14]? Or, how can a new object be classified by a fuzzy decision tree?

2.4 Fuzzy Association Analysis

The use of fuzzy sets in connection with association analysis, to be discussed in more detail in Section 4.2.1, has been proposed by numerous authors (see [10,15] for recent overviews), with motivations closely resembling those in the case of rule learning and decision tree induction. Again, by allowing for "soft" rather than crisp boundaries of intervals, fuzzy sets can avoid certain undesirable threshold effects [54], this time concerning the quality measures of association rules (like support and confidence) rather than the classification of objects. Moreover, identifying fuzzy sets with linguistic terms allows for a comprehensible and user-friendly presentation of rules discovered in a database.

Many standard techniques for association rule mining have been transferred to the fuzzy case, sometimes in a rather ad-hoc manner. Indeed, publications on this topic are often more concerned with issues of data preprocessing, e.g., the problem of finding good fuzzy partitions for the quantitative attributes, rather than the rule mining process itself. Still, more theoretically-oriented research has recently been started [22]. For example, the existence of different types of fuzzy rules [24] suggests that fuzzy associations can be interpreted in different ways and, hence, that the evaluation of an association cannot be independent of its interpretation. In particular, one can raise the question which generalized logical operators can reasonably be applied in order to evaluate fuzzy associations, e.g., whether the antecedent part and the consequent part should be combined in a conjunctive way (à la Mamdani rules) or by means of a generalized implication (as in implication-based fuzzy rules) [29]. Moreover, since standard evaluation measures for association rules can be generalized in many ways, it is interesting to investigate properties of particular generalizations and to look for an axiomatic basis that supports the choice of specific measures [22].

2.5 Fuzzy Methods in Case-Based Learning

The major assumption underlying case-based learning (CBL) is a commonsense principle suggesting that "similar problems have similar solutions". This "similarity hypothesis" serves as a basic inference paradigm in various domains of application. For example, in a classification context, it translates into the assertion that "similar objects have similar class labels". Similarity-based inference has also been a topic of interest in FST, which is hardly astonishing since similarity is one of the main semantics of fuzzy membership degrees [51,53]. Along these lines, a close connection between case-based learning and fuzzy rule-based reasoning has been established in [19,21]. Here, the aforementioned "similarity hypothesis" has been formalized within the framework of fuzzy rules. As a result, case-based inference can be realized as a special type of fuzzy set-based approximate reasoning.

A possibilistic variant of the well-known k-nearest neighbor classifier, which constitutes the core of the family of CBL algorithms, has been presented in [32]. Among other things, this paper emphasizes the ability of possibility theory to represent partial ignorance as a special advantage in comparison to probabilistic approaches. In fact, this point seems to be of critical importance in case-based learning, where the reliability of a classification strongly depends on the existence of cases that are similar to the query.

The use of OWA-operators as generalized aggregation operators in case-based learning has been proposed in [57]. In fact, there are several types of aggregation problems that arise in CBL. One of these problems concerns the derivation of a global degree of similarity between cases by aggregating *local* similarity degrees pertaining to individual (one-dimensional) attributes. (This problem is indeed a fundamental one that appears in various guises, not only in CBL. In fuzzy association analysis, for example, the problem of deriving the degree of occurrence of an itemset in a transaction from the degrees of occurrence of individual items is very similar.) Usually, this is done by means of a simple linear combination, and this is where OWA-operators provide an interesting, more flexible alternative. A second aggregation problem in CBL concerns the combination of the evidences in favor of different class labels that come from the neighbors of the query case. In [33], it is argued that cases retrieved from a case library must not be considered as independent information sources, as implicitly done by most case-based learning methods. To take interdependencies between the neighbored cases into account, a new inference principle is developed that combines potentially interacting pieces of evidence by means of the (discrete) Choquet-integral. This method can be seen as a generalization of weighted nearest neighbor estimation.

2.6 Possibilistic Networks

So-called graphical models, including Bayesian networks [45] and Markov networks [39], have been studied intensively in recent years. The very idea of such models is to represent a high-dimensional probability distribution (defined on the Cartesian product of the domains of all attributes under consideration) in an efficient way, namely by factorizing it into several low-dimensional conditional or marginal distributions.

By their very nature, graphical models of the above kind provide a suitable means for representing *probabilistic* uncertainty. However, they cannot easily deal with other types of uncertainty such as imprecision or incompleteness. This has motivated the development of *possibilistic networks* as a possibilistic counterpart to probabilistic networks [5,6,8]. This approach relies upon possibility theory as an underlying uncertainty calculus, which makes it particularly suitable for dealing with imprecise data (in the form of set-valued specifications of attribute values). For example, the interpretation of possibility distributions in [8] is based on the so-called context model [27], hence possibility degrees are considered as a kind of upper probability.

3 Potential Contributions of Fuzzy Set Theory

In the following, we highlight and critically comment some potential contributions that FST can make to machine learning and data mining.

3.1 Graduality

The ability to represent gradual concepts and fuzzy properties in a thorough way is one of the key features of fuzzy sets. This aspect is also of primary importance in the context of ML&DM. In machine learning, for example, the formal problem of *concept learning* has received a great deal of attention. A concept is usually identified with its extension, that is a subset C of an underlying set (universe) U of objects. For example, C might be the concept "dog" whose extension is the set of dogs presently alive, a subset of all creatures on earth. The goal of (machine) learning is to induce an *intensional* description of a concept from a set of (positive and negative) examples, that is a characterization of a concept in terms of its properties (a dog has four legs and a tail, it can bark, ...). Now, it is widely recognized that most natural concepts have non-sharp boundaries. To illustrate, consider concepts like woods, river, lake, hill, street, house, or chair. Obviously, these concepts are vague or fuzzy, in that one cannot unequivocally say whether or not a certain collection of trees should be called a wood, whether a certain building is really a house, and so on. Rather, one will usually agree only to a certain extent that an object belongs to a concept. Thus, an obvious idea is to induce *fuzzy concepts*, that are formally identified by a fuzzy rather than a crisp subset of U. Fuzzy concepts can be characterized in terms of fuzzy predicates (properties) which are combined by means of generalized logical connectives. In fact, one should recognize that graduality is not only advantageous for expressing the concept itself, but also for modeling the qualifying properties. For example, a "firm ground" is a characteristic property of a street, and this property is obviously of a fuzzy nature (hence it should be formalized accordingly).

Likewise, in data mining, the patterns of interest are often vague and have boundaries that are non-sharp in the sense of FST. To illustrate, consider the concept of a "peak": It is usually not possible to decide in an unequivocal way whether a timely ordered sequence of measurements has a "peak" (a particular kind of pattern) or not. Rather, there is a gradual transition between having a peak and not having a peak. Taking graduality into account is also important if one must decide whether a certain property is frequent among a set of objects, e.g., whether a pattern occurs frequently in a data set. In fact, if the pattern is specified in an overly restrictive manner, it might easily happen that none of the objects matches the specification, even though many of them can be seen as approximate matches. In such cases, the pattern might still be considered as "well-supported" by the data (see also Section 4).

Unfortunately, the representation of graduality is often foiled in machine learning applications, especially in connection with the learning of predictive models. In such applications, a fuzzy prediction is usually not desired, rather one is forced to come up with a definite final decision. Classification is an obvious example: Eventually, a decision in favor of one particular class label has to be made, even if the object under consideration seems to have partial membership in several classes simultaneously. This is the case both in theory and practice: In practice, the bottom line is the course of action one takes on the basis of a prediction, not the prediction itself. In theory, a problem concerns the performance evaluation of a fuzzy classifier: The standard benchmark data sets have crisp rather than fuzzy labels. Moreover, a fuzzy classifier cannot be compared with a standard (non-fuzzy) classifier unless it eventually outputs crisp predictions.

Needless to say, if a fuzzy predictor is supplemented with a "defuzzification" mechanism (like a winner-takes-all strategy in classification), many of its merits are lost. In the classification setting, for instance, a defuzzified fuzzy classifier does again produce hard decision boundaries in the input space. Thereby, it is actually reduced to a standard classifier. Moreover, if a classifier is solely evaluated on the basis of its predictive accuracy, then all that matters is the decision boundaries it produces in the input space. Since a defuzzified fuzzy classifier does not produce a decision boundary that is principally different from the boundaries produced by alternative classifiers (such as decision trees or neural networks), fuzzy machine learning methods do not have much to offer with regard to generalization performance. And indeed, fuzzy approaches to classification do usually *not* improve predictive accuracy.

Let us finally note that "graduality" is of course not reserved to fuzzy methods. Rather, it is inherently present also in many standard learning methods. Consider, for example, a concept learner (binary classifier) $c : \mathcal{X} \rightarrow [0, 1]$ the output of which is a number in the unit interval, expressing a kind of "propensity" of an input x to the concept under consideration. Classifiers of such kind abound, a typical example is a multilayer perceptron. In order to extend such classifiers to multi-class problems (involving more than two classes), one common approach is to apply a one-against-all strategy: For each class y, a separate classifier $c_y(\cdot)$ is trained which considers that class as the concept to be learned and, hence, instances of all other classes as negative examples. The prediction for a new input x is then given by the class that maximizes $c_y(x)$. Now, it is of course tempting to consider the $c_y(x)$ as (estimated) membership degrees and, consequently, the collection $\{c_y(x) \,|\, y \in \mathcal{Y}\}$ of these estimations as a fuzzy classification.

3.2 Granularity

Granular computing, including FST as one its main constituents, is an emerging paradigm of information processing in which "information granules" are considered as key components of knowledge representation [4]. A central idea is that information can be processed on different levels of abstraction, and that the choice of the most suitable level depends on the problem at hand.

As a means to trade off accuracy against efficiency and interpretability, granular computing is also relevant for ML&DM, not only for the model induction or pattern discovery process itself, but also for data pre- and post-processing, such as data compression and dimensionality reduction [41]. For example, one of the most important data analysis methods, cluster analysis, can be seen as a process of information granulation, in which data objects are combined into meaningful groups so as to convey a useful idea of the main structure of a dataset.

3.3 Interpretability

A primary motivation for the development of fuzzy sets was to provide an interface between a numerical scale and a symbolic scale which is usually composed of linguistic terms. Thus, fuzzy sets have the capability to interface quantitative patterns with qualitative knowledge structures expressed in terms of natural language. This makes the application of fuzzy technology very appealing from a knowledge representational point of view. For example, it allows association rules discovered in a database to be presented in a linguistic and hence comprehensible way. In fact, the user-friendly representation of models and patterns is often emphasized as one of the key features of fuzzy methods.

The use of linguistic modeling techniques does also produce some disadvantages, however. A first problem concerns the ambiguity of fuzzy models: Linguistic terms and, hence, models are highly subjective and context-dependent. It is true that the imprecision of natural language is not necessarily harmful and can even be advantageous. [4] A fuzzy controller, for example, can be quite insensitive to the concrete mathematical translation of a linguistic model. One should realize, however, that in fuzzy control the information flows in a reverse direction: The linguistic model is not the end product, as in ML&DM, it rather stands at the beginning.

It is of course possible to disambiguate a model by complementing it with the semantics of the fuzzy concepts it involves (including the specification of membership functions). Then, however, the complete model, consisting of a qualitative (linguistic) and a quantitative part, becomes cumbersome and will not be easily understandable. This can be contrasted with interval-based models, the most obvious alternative to fuzzy models: Even though such models do certainly have their shortcomings, they are at least objective and not prone to context-dependency.

Another possibility to guarantee transparency of a fuzzy model is to let a user of a data mining system specify all fuzzy concepts by hand, including the fuzzy partitions for all of the variables involved in the study under consideration. This is rarely done, however, mainly for two reasons. Firstly, the job is of course

[4] See Zadeh's principle of incompatibility between precision and meaning [58].

tedious and cumbersome if the number of variables is large. Secondly, much flexibility for model adaptation is lost, because it is by no means guaranteed that accurate predictive models or interesting patterns can be found on the basis of the fuzzy partitions as pre-specified by the user. In fact, in most methods the fuzzy partitions are rather *adapted* to the data in an optimal way, so as to maximize the model accuracy or the interestingness of patterns.

A second problem with regard to transparency concerns the complexity of models. A rule-based classifier consisting of, say, 40 rules each of which has a condition part with 5-7 antecedents, will hardly be comprehensible as a whole, even if the various ingredients might be well understandable. Now, since models that are simple, e.g., in the sense of including only a few attributes or a few rules, will often not be accurate at the same time, there is obviously a conflict between accuracy and understandability and, hence, the need to find a trade-off between these criteria [9].

In fact, this trade-off concerns not only the size of models, but also other measures that are commonly employed in order to improve model accuracy. In connection with rule-based models, for example, the *weighing* of individual rules can often help to increase the predictive accuracy. On the other hand, the interpretation of a set of weighted rules becomes more difficult.

3.4 Robustness

It is often claimed that fuzzy methods are more robust than non-fuzzy methods. Of course, the term "robustness" can refer to many things, e.g., to the sensitivity of an induction method toward violations of the model assumptions.[5] In connection with fuzzy methods, the most relevant type of robustness concerns sensitivity toward variations of the data. Generally, a learning or data mining method is considered robust if a small variation of the observed data does hardly alter the induced model or the evaluation of a pattern.

A common argument supporting the claim that fuzzy models are in this sense more robust than non-fuzzy models refers to a "boundary effect" which occurs in various variants and is arguably an obvious drawback of interval-based methods. This effect refers to the fact that a variation of the boundary points of an interval can have a strong influence on a model or a pattern. In fact, it is not difficult to construct convincing demonstrations of this effect: In association analysis (cf. Section 2.4), for example, a small shift of the boundary of an interval can have a drastic effect on the support of an association rule if many data points are located near the boundary. This effect is alleviated when using fuzzy sets instead of intervals.

[5] This type of sensitivity is of special interest in robust statistics.

Despite the intuitive persuasiveness of such examples, there is still no clear conception of the concrete meaning of *robustness*. Needless to say, without a formal definition of robustness, i.e., certain types of robustness measures, one cannot argue convincingly that one data mining method is more robust than another one. For example, it makes a great difference whether robustness is understood as a kind of *expected* or a kind of *worst-case* sensitivity: It is true that a shifting of data points can have a stronger effect on, say, the support of an interval-based association rule than on the support of a fuzzy association. However, if the data points are not located in the boundary region of the intervals, it can also happen that the former is not affected at all, whereas a fuzzy rule is almost always affected at least to some extent (since the "boundary" of a fuzzy interval is much wider than that of a standard interval). Consequently, if robustness is defined in terms of the *average* rather than the *maximal* change, the fuzzy approach might not be more robust than the non-fuzzy one.

3.5 Representation of Uncertainty

Machine learning is inseparably connected with uncertainty. To begin with, the data presented to learning algorithms is imprecise, incomplete or noisy most of the time, a problem that can badly mislead a learning procedure. But even if observations are perfect, the generalization beyond that data, the process of induction, is still afflicted with uncertainty. For example, observed data can generally be explained by more than one candidate theory, which means that one can never be sure of the truth of a particular model.

Fuzzy sets and possibility theory have made important contributions to the representation and processing of uncertainty. In ML&DM, like in other fields, related uncertainty formalisms can complement probability theory in a reasonable way, because not all types of uncertainty relevant to machine learning are probabilistic and because other formalisms are more expressive than probability.

To illustrate the first point, consider the problem of inductive reasoning as indicated above: In machine learning, a model is often induced from a set of data on the basis of a *heuristic* principle of inductive inference, such as the well-known Occams's razor. As one can never be sure of the truth of the particular model suggested by the heuristic principle, it seems reasonable to specify a kind of *likelihood* for all potential candidate models. This is done, e.g., in Bayesian approaches, where the likelihood of models is characterized in terms of a posterior probability distribution (probability of models given the data). One can argue, however, that the uncertainty produced by heuristic inference principles such as Occam's razor is not necessarily of a probabilistic nature and, for example, that the derivation of a *possibility distribution* over the model space is a viable alternative. This idea has been suggested in [31] in connection with decision tree induction: Instead of learning a single decision tree, a possibility distribution over the class of all potential trees is derived on the basis of a possibilistic variant of Occam's razor.

The second point, concerning the limited expressivity of probability distributions, was already indicated in Section 2.5, where we mentioned that possibility distributions are more suitable for representing partial ignorance in case-based learning. Similarly, possibility theory is used for modeling incomplete and missing data in possibilistic networks (cf. Section 2.6) as well as other data analysis methods, such as formal concept analysis [18].

3.6 Incorporation of Background Knowledge

Roughly speaking, inductive (machine) learning can be seen as searching the space of candidate hypotheses for a most suitable model. The corresponding search process, regardless whether it is carried out in an explicit or implicit way, is usually "biased" in various ways, and each bias usually originates from a sort of background knowledge. For example, the *representation bias* restricts the hypothesis space to certain types of input-output relations, such as linear or polynomial relationships. Incorporating background knowledge is extremely important, because the data by itself would be totally meaningless if considered from an "unbiased" point of view [42].

As demonstrated by other application fields such as fuzzy control, fuzzy set-based modeling techniques provide a convenient tool for making expert knowledge accessible to computational methods and, hence, to incorporate background knowledge in the learning process. This can be done in various ways and on different levels.

One very obvious approach is to combine modeling and learning in rule-based systems. For example, an expert can describe an input-output relation in terms of a fuzzy rule base (as in fuzzy control). Afterward, the membership functions specifying the linguistic terms that have been employed by the expert can be adapted to the data in an optimal way.[6] In other words, the expert specifies the rough structure of the rule-based model, while the fine-tuning ("model calibration") is done in a data-driven way. Let us note that specifying the structure of a model first and adapting that structure to the data afterward is a general strategy for combining knowledge-based and data-driven modeling,

[6] Here, the expert implements a kind of *search bias*, as it determines the starting point of the search process and, hence, the first local optimum to be found.

which is not reserved to rule-based models; it is used, for example, in graphical models (cf. Section 2.6) as well.

An alternative approach, called constraint-regularized learning, aims at exploiting fuzzy set-based modeling techniques within the context of the regularization (penalization) framework of inductive learning [34]. Here, the idea is to express vague, partial knowledge about an input-output relation in terms of fuzzy constraints and to let such constraints play the role of a penalty term within the regularization approach. Thus, an optimal model is one that achieves an optimal trade-off between fitting the data and satisfying the constraints.

Expert knowledge can also be exploited in the form of user feedback in interactive data analysis. Here, the idea is to let the user, to some extent, guide the learning or data mining process, which can be especially beneficial in the exploration of high-dimensional data spaces. A nice example for an approach of that kind is the recent paradigm of knowledge-based clustering [46].

3.7 Aggregation, Combination, and Information Fusion

The problem to aggregate or combine partial or intermediary results frequently occurs in inference processes of various kind, and is also relevant to ML&DM. A simple example is the aggregation of the individual predictions of classifiers in ensemble methods. Essentially, this can be considered as a problem of information fusion [2]. Likewise, in nearest neighbor classification, each neighbor provides a certain amount of evidence in favor of the class it belongs to. To make a final decision, this evidence must be aggregated either way. Problems of this kind call for suitable aggregation operators.

Aggregation operators, both logical and arithmetical, are also used by ML&DM methods for representing relationships between attributes in models and patterns. In decision tree induction, for example, each inner node represents an equality or an inequality predicate, and these predicates are combined in a conjunctive way along a path of a tree.

A large repertoire of generalized logical (e.g., t-norms and t-conorms) and arithmetical (e.g., Choquet- and Sugeno-integral) operators have been developed in FST and related fields. These operators can be usefully applied to the aforementioned problems (e.g., [11,2,57]).

Moreover, conventional learning methods can be extended in a straightforward way by replacing standard operators by their generalized versions. In fact, several examples of this idea have been presented in previous sections. The general effect of such generalizations is to make models more flexible.

For example, while a standard decision tree can only produce axis-parallel decision boundaries, these boundaries can become non-axis-parallel for fuzzy decision trees where predicates are combined by means of a t-norm. Now, it is well-known that learning from empirical data will be most successful if the underlying model class has just the right flexibility, since both over- and underfitting of a model can best be avoided in that case. Therefore, the question whether or not a fuzzy generalization will pay off cannot be answered in general: If the original (non-fuzzy) hypothesis space is not flexible enough, the fuzzy version will probably be superior. On the other hand, if the former is already flexible enough, a fuzzification might come along with a danger of overfitting.

4 Mining Fuzzy Patterns in Data

This section will discuss in more detail one of the advantages of fuzzy methods which is, in the author's opinion, one of the key contribution of FST to data mining. More specifically, it will be argued that the increased expressiveness of fuzzy methods, which is mainly due to the ability to represent *graded* properties in an adequate way, is useful for both feature extraction and subsequent dependency analysis. Here, we proceed from the standard representation of data entities in terms of *feature vectors*, i.e., a fixed number of features or attributes, each of which represents a certain property of an entity. For example, if the data entities are employees, possible features might be gender, age, and income. A common goal of feature-based methods in then to analyze relationships and *dependencies* between the attributes.

4.1 Fuzzy Feature Extraction and Pattern Representation

Many features of interest, and therefore the patterns expressed in terms of these features, are inherently fuzzy. As an example, consider the so-called "candlestick patterns" which refer to certain characteristics of financial time series. These patterns are believed to reflect the psychology of the market and are used to support investment decisions. Needless to say, a candlestick pattern is fuzzy in the sense that the transition between the presence and absence of the pattern is gradual rather than abrupt; see [40] for an interesting fuzzy approach to modeling and discovering such patterns.

To give an even simpler example, consider a discrete time series of the form

$$x = \Big(x(1), x(2) \ldots x(n)\Big) ,$$

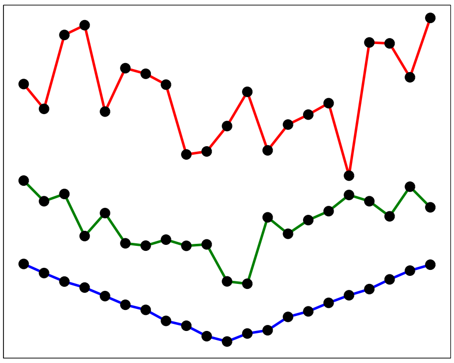

Fig. 1. Three exemplary time series that are more or less “decreasing at the beginning”.

i.e., a timely ordered sequence of measurements. To bring one of the topical application areas of fuzzy data mining into play, one may think of x as the expression profile of a gene in a microarray experiment, i.e., a timely ordered sequence of expression levels. For such profiles, the property (feature) “decreasing at the beginning” might be of interest, e.g., in order to express patterns like

$$P: \begin{array}{l}\text{“A profile which is decreasing at the beginning} \\ \quad \text{is typically increasing at the end.”}\end{array} \tag{1}$$

Again, the aforementioned pattern is inherently fuzzy, in the sense that a time series can be more or less decreasing at the beginning. To begin with, it is unclear which time points belong to the “beginning” of a time series, and defining it in a non-fuzzy (crisp) way by a subset $B = \{1, 2 \ldots k\}$, for a fixed $k \in \{1 \ldots n\}$, comes along with a certain arbitrariness and does not appear fully convincing. Moreover, the human perception of “decreasing” will usually be tolerant toward small violations of the standard mathematical definition, which requires

$$\forall\, t \in B \,:\, x(t) \geq x(t+1), \tag{2}$$

especially if such violations may be caused by noise in the data.

Fig. 1 shows three exemplary profiles. While the first one at the bottom is undoubtedly decreasing at the beginning, the second one in the middle is clearly not decreasing in the sense of (2). According to human perception, however, this series is still approximately or, say, almost decreasing at the beginning. In other words, it does have the corresponding (fuzzy) feature to some extent.

By modeling features like “decreasing at the beginning” in a non-fuzzy way, that is, as a Boolean predicate which is either true or false, it will usually become impossible to discover patterns such as (1), even if these patterns are to some degree present in a data set.

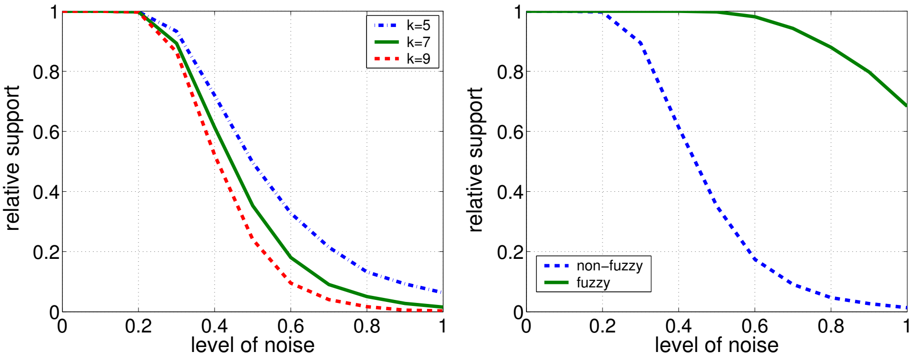

Fig. 2. Left: Relative support of pattern (1) as a function of the level of noise (σ) and various values of k. Right: Comparison with the relative support for the fuzzy case.

To illustrate this point, consider a simple experiment in which 1,000 copies of an (ideal) profile defined by $x(t) = |\, t - 11 \,|$, $t = 1 \ldots 21$, are corrupted with a certain level of noise. This is done by adding an error term to each value of every profile; these error terms are independent and normally distributed with mean 0 and standard deviation σ. Then, the relative support of the pattern (1) is determined, i.e., the fraction of profiles that still satisfy this pattern in a strict mathematical sense:

$$(\forall\, t \in \{1 \ldots k\} \,:\, x(t) \geq x(t+1)) \wedge (\forall\, t \in \{n-k \ldots n\} \,:\, x(t-1) \geq x(t))$$

Fig. 2 (left) shows the relative support as a function of the level of noise (σ) and various values of k. As can be seen, the support drops off quite quickly. Consequently, the pattern will be discovered only in the more or less noise-free scenario but quickly disappears for noisy data.

Fuzzy set-based modeling techniques offer a large repertoire for generalizing the formal (logical) description of a property, including generalized logical connectives such as t-norms and t-conorms, fuzzy quantifiers such as `FOR-MOST`, and fuzzy relations such as `MUCH-SMALLER-THAN`. Making use of these tools, it becomes possible to formalize descriptions like "for all points t at the beginning, $x(t)$ is not much smaller than $x(t+1)$, and for most points it is even strictly greater" in an adequate way:

$$F_1(x) \overset{\mathrm{df}}{=} \Big(\widetilde{\forall}\, t \in B \,:\, x(t+1) > x(t)\Big) \otimes \Big(\forall\, t \in B \,:\, \neg\, \mathtt{MS}(x(t+1), x(t))\Big)\,, \quad (3)$$

where B is now a *fuzzy* set characterizing the beginning of the time series, $\widetilde{\forall}$ is an exception-tolerant relaxation of the universal quantifier, $\otimes$ is a t-norm, and `MS` is a fuzzy `MUCH-SMALLER-THAN` relation; we refrain from a more detailed description of these concepts at a technical level.

In any case, (3) is an example for a fuzzy definition of the feature "decreasing at the beginning" (we do by no means claim that it is the best characterization) and offers an alternative to the non-fuzzy definition (2). According to (3), every time series can have the feature to some extent. Analogously, the fuzzy feature "increasing at the end" (F_2) can be defined. Fig. 2 (right) shows the relative support

$$\mathrm{supp}(P) = \frac{1}{1000} \sum_{x_i} \mathrm{supp}_{x_i}(P) = \frac{1}{1000} \sum_{x_i} F_1(x_i) \otimes F_2(x_i) \tag{4}$$

of the pattern P for the fuzzy case, again as a function of the noise level. As can be seen, the relative support also drops off after a while, which is an expected and even desirable property (for a high enough noise level, the pattern will indeed disappear). The support function decreases much slower, however, so the pattern will be discovered in a much more robust way.

The above example shows that a fuzzy set-based modeling can be very useful for extracting certain types of features. Besides, it gives an example of increased robustness in a relatively specific sense, namely robustness of pattern discovery toward noise in the data. In this connection, let us mention that we do not claim that the fuzzy approach is the only way to make feature extraction more adequate and pattern discovery more robust. For example, in the particular setting considered in our example, one may think of a probabilistic alternative, in which the individual support $\mathrm{supp}_{x_i}(P)$ in (4) is replaced by the probability that the underlying noise-free profile does satisfy the pattern P in the sense of (2). Apart from pointing to the increased computational complexity of this alternative, however, we like to repeat our argument that patterns like (1) are inherently fuzzy in our opinion: Even in a completely noise-free scenario, where information is exact and nothing is random, human perception may consider a given profile as somewhat decreasing at the beginning, even if it does not have this property in a strict mathematical sense.

Finally, one may argue that our formalization, which involves the choice of a fuzzy quantifier, a t-norm, and so on, is to some extent arbitrary. This is indeed a valid argument. In general, however, this is unavoidable and a typical property of model building. A model, whether fuzzy, probabilistic, or deterministic, is always a simplified and approximate image of reality, and a fuzzy model of the human conception of "decreasing at the beginning" is perhaps not more arbitrary than a probabilistic model of, say, a certain ecological or economic system. Besides, there are of course also means to calibrate a fuzzy model. In our example, this could be done on the basis of a number of exemplary judgments, that is, a number of times series x with associated values $F_1(x)$. The parameters of the model could then be tuned so as to reproduce these values as good as possible.

4.2 Mining Gradual Dependencies

4.2.1 Association Analysis

As already mentioned in Section 2.4, association analysis [1,52] is a widely applied data mining technique that has been studied intensively in recent years. The goal in association analysis is to find “interesting” associations in a data set, that is, dependencies between so-called itemsets $\mathcal{A}$ and $\mathcal{B}$ expressed in terms of rules of the form $\mathcal{A} \rightharpoonup \mathcal{B}$. To illustrate, consider the well-known example where items are products and a data record (transaction) $\mathcal{I}$ is a shopping basket such as $\{\texttt{butter}, \texttt{milk}, \texttt{bread}\}$. The intended meaning of an association $\mathcal{A} \rightharpoonup \mathcal{B}$ is that, if $\mathcal{A}$ is present in a transaction, then $\mathcal{B}$ is likely to be present as well. A standard problem in association analysis is to find all rules $\mathcal{A} \rightharpoonup \mathcal{B}$ the *support* (relative frequency of transactions $\mathcal{I}$ with $\mathcal{A} \cup \mathcal{B} \subseteq \mathcal{I}$) and *confidence* (relative frequency of transactions $\mathcal{I}$ with $\mathcal{B} \subseteq \mathcal{I}$ among those with $\mathcal{A} \subseteq \mathcal{I}$) of which reach user-defined thresholds `minsupp` and `minconf`, respectively.

In the above setting, a single item can be represented in terms of a binary (0/1-valued) attribute reflecting the presence or absence of the item. To make association analysis applicable to data sets involving numerical variables, such attributes are typically discretized into intervals, and each interval is considered as a new binary attribute. For example, the attribute `temperature` might be replaced by two binary attributes `cold` and `warm`, where $\texttt{cold} = 1$ ($\texttt{warm} = 0$) if the temperature is below 10 degrees and $\texttt{warm} = 1$ ($\texttt{cold} = 0$) otherwise.

A further extension is to use fuzzy sets (fuzzy partitions) instead of intervals (interval partitions), and corresponding approaches to fuzzy association analysis have been proposed by several authors (see e.g. [10,15] for recent overviews). In the fuzzy case, the presence of a feature subset $\mathcal{A} = \{A_1 \ldots A_m\}$, that is, a *compound feature* considered as a conjunction of primitive features $A_1 \ldots A_m$, is specified as

$$\mathcal{A}(x) = A_1(x) \otimes A_2(x) \otimes \ldots \otimes A_m(x),$$

where $A_i(x) \in [0,1]$ is the degree to which x has feature A_i, and $\otimes$ is a t-norm serving as a generalized conjunction.

There are different motivations for a fuzzy approach to association rule mining. For example, again pointing to the aspect of *robustness*, several authors have emphasized that, by allowing for “soft” rather than crisp boundaries of intervals, fuzzy sets can avoid undesirable boundary effects (see e.g. [54]). In this context, a boundary effect occurs if a slight variation of an interval boundary causes a considerable change of the evaluation of an association rule, and therefore strongly influences the data mining result.

In the following, we shall emphasize another potential advantage of fuzzy association analysis, namely the fact that association rules can be represented in a more *distinctive* way. In particular, working with fuzzy instead of binary features allows for discovering *gradual* dependencies between variables.

4.2.2 Gradual Dependencies between Fuzzy Features

On a logical level, the meaning of a standard (association) rule $\mathcal{A} \rightharpoonup \mathcal{B}$ is captured by the material conditional, i.e., the rule applies unless the consequent $\mathcal{B}$ is true and the antecedent $\mathcal{A}$ is false. On a natural language level, a rule of that kind is typically understood as an IF–THEN construct: If the antecedent $\mathcal{A}$ holds true, so does the consequent $\mathcal{B}$.

In the fuzzy case, the Boolean predicates $\mathcal{A}$ and $\mathcal{B}$ are replaced by corresponding fuzzy predicates which assume truth values in the unit interval $[0, 1]$. Consequently, the material implication operator has to be replaced by a generalized connective, that is, a suitable $[0, 1] \times [0, 1] \rightarrow [0, 1]$ mapping. In this regard, two things are worth mentioning. Firstly, the choice of this connective is not unique, instead there are various options. Secondly, depending on the type of operator employed, fuzzy rules can have quite different semantical interpretations [24].

A special type of fuzzy rule, referred to as *gradual rules*, combines the antecedent $\mathcal{A}$ and the consequent $\mathcal{B}$ by means of a *residuated* implication operator $\rightsquigarrow$. The latter is a special type of implication operator which is derived from a t-norm $\otimes$ through residuation:

$$\alpha \rightsquigarrow \beta \overset{\mathrm{df}}{=} \sup\{\, \gamma \,|\, \alpha \otimes \gamma \leq \beta \,\}. \tag{5}$$

As a particular case, so-called *pure* gradual rules are obtained when using the following implication operator:[7]

$$\alpha \rightsquigarrow \beta = \begin{cases} 1 & \text{if } \alpha \leq \beta \\ 0 & \text{if } \alpha > \beta \end{cases} \tag{6}$$

The above approach to modeling a fuzzy rule is in agreement with the following interpretation of a gradual rule: “THE MORE the antecedent $\mathcal{A}$ is true, THE MORE the consequent $\mathcal{B}$ is true” [50,23], for example “The larger a truck, the slower it is”. More specifically, in order to satisfy the rule, the consequent must be *at least* as true as the antecedent according to (6), and the same principle applies for other residuated implications, albeit in a somewhat relaxed form.

[7] This operator is the core of all residuated implications (5).

The above type of *implication-based* fuzzy rule can be contrasted with so-called *conjunction-based* rules, where the antecedent and consequent are combined in terms of a t-norm such as minimum or product. Thus, in order to satisfy a conjunction-based rule, both the antecedent and the consequent must be true (to some degree). As an important difference, note that the antecedent and the consequent play a symmetric role in the case of conjunction-based rules but are handled in an asymmetric way by implication-based rules.

The distinction between different semantics of a fuzzy rule as outlined above can of course also be made for association rules. Formally, this leads to using different types of support and confidence measures for evaluating the quality (interestingness) of an association [29,20]. Consequently, it may happen that a data set supports a fuzzy association $\mathcal{A} \rightharpoonup \mathcal{B}$ quite well in one sense, i.e., *according to a particular semantics*, but not according to another one.

The important point to notice is that these distinctions cannot be made for non-fuzzy (association) rules. Formally, the reason is that fuzzy extensions of logical operators all coincide on the extreme truth values 0 and 1. Or, stated the other way round, a differentiation can only be made on intermediary truth degrees. In particular, the consideration of gradual dependencies does not make any sense if the only truth degrees are 0 and 1.

In fact, in the non-fuzzy case, the point of departure for analyzing and evaluating a relationship between features or feature subsets $\mathcal{A}$ and $\mathcal{B}$ is a contingency table:

	$\mathcal{B}(y) = 0$	$\mathcal{B}(y) = 1$	
$\mathcal{A}(x) = 0$	n_{00}	n_{01}	$n_{0\bullet}$
$\mathcal{A}(x) = 1$	n_{10}	n_{11}	$n_{1\bullet}$
	$n_{\bullet 0}$	$n_{\bullet 1}$	n

In this table, n_{00} denotes the number of examples x for which $\mathcal{A}(x) = 0$ and $\mathcal{B}(x) = 0$, and the remaining entries are defined analogously. All common evaluation measures for association rules, such as support (n_{11}/n) and confidence ($n_{11}/n_{1\bullet}$) can be expressed in terms of these numbers.

In the fuzzy case, a contingency table can be replaced by a *contingency diagram*, an idea that has been presented in [30]. A contingency diagram is a two-dimensional diagram in which every example x defines a point $(\alpha, \beta) = (\mathcal{A}(x), \mathcal{B}(x)) \in [0,1] \times [0,1]$. A diagram of that type is able to convey much more information about the dependency between two (compound) features $\mathcal{A}$ and $\mathcal{B}$ than a contingency table. Consider, for example, the two diagrams depicted in Fig. 3. Obviously, the dependency between $\mathcal{A}$ and $\mathcal{B}$ as suggested by the left diagram is quite different from the one shown on the right. Now, consider the non-fuzzy case in which the fuzzy sets $\mathcal{A}$ and $\mathcal{B}$ are replaced by

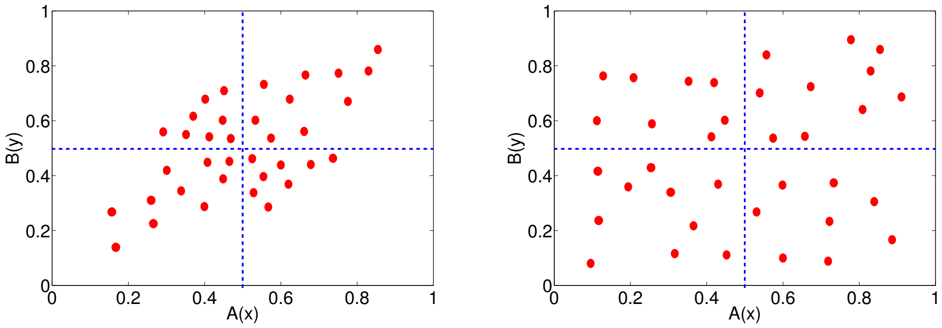

Fig. 3. Two contingency diagrams reflecting different types of dependencies between features $\mathcal{A}$ and $\mathcal{B}$.

crisp sets $\mathcal{A}_{bin}$ and $\mathcal{B}_{bin}$, respectively, for example by using a $[0,1] \rightarrow \{0,1\}$ mapping like $\alpha \mapsto (\alpha > 0.5)$. Then, identical contingency tables are obtained for the left and the right scenario (in the left diagram, the four quadrants contain the same number of points as the corresponding quadrants in the right diagram). In other words, the two scenarios cannot be distinguished in the non-fuzzy case.

In [30], it was furthermore suggested to analyze contingency diagrams by means of techniques from statistical regression analysis. Amongst other things, this offers an alternative approach to discovering gradual dependencies. For example, the fact that a linear regression line with a significantly positive slope (and high quality indexes like a coefficient of determination, R^2) can be fit to the data suggests that indeed a higher $\mathcal{A}(x)$ tends to result in a higher $\mathcal{B}(x)$, i.e., the more x has feature $\mathcal{A}$ the more it has feature $\mathcal{B}$. This is the case, for example, in the left diagram in Fig. 3. In fact, the data in this diagram supports an association $\mathcal{A} \rightharpoonup \mathcal{B}$ quite well in the sense of the THE MORE–THE MORE semantics, whereas it does not support the non-fuzzy rule $\mathcal{A}_{bin} \rightharpoonup \mathcal{B}_{bin}$.

Note that a contingency diagram can be derived (and remains 2-dimensional) not only for simple but also for compound features, that is, feature subsets representing conjunctions of simple features. The problem, then, is to derive regression-related quality indexes for all potential association rules in a systematic way, and to filter out those gradual dependencies which are well-supported by the data in terms of these indexes. For corresponding mining methods, including algorithmic aspects and complexity issues, we refer to [30]; see also [7] for an alternative, non-parametric approach to mining fuzzy gradual dependencies.

Before concluding this section, let us note that the two approaches for modeling gradual dependencies that we have presented, the one based on fuzzy gradual rules and the other one using statistical regression analysis, share similarities but also show differences. In particular, the logical modeling of gradual dependencies via suitable implication operators does not assume a relationship between $\mathcal{A}(x)$ and $\mathcal{B}(x)$ which is, say, indeed "strictly increasing". For example, if $\mathcal{B}(x) \equiv 1$, then the rule $\mathcal{A} \rightharpoonup \mathcal{B}$ will be perfectly satisfied, even though $\mathcal{B}(x)$ is constant and does not increase with $\mathcal{A}(x)$. In fact, more specifically, the semantical interpretation of a gradual rule should be expressed in terms of a *bound* on the degree $\mathcal{B}(x)$ rather than the degree itself: The more x is in $\mathcal{A}$, the higher is the guaranteed *lower bound* of the membership of x in $\mathcal{B}$. Seen from this point of view, the statistical approach is perhaps even more in line with the intuitive understanding of a THE MORE–THE MORE relationship.

5 Conclusions

The previous sections have shown that FST can contribute to machine learning and data mining in various ways. Needless to say, for most of the issues that were addressed, a fuzzy approach will not be the only solution. Still, FST provides a relatively flexible framework in which different aspects of machine learning and data mining systems can be handled in a coherent way. In this regard, let us again highlight the following points:

1. FST has the potential to produce models that are more comprehensible, less complex, and more robust; fuzzy information granulation appears to be an ideal tool for trading off accuracy against complexity and understandability.
2. In data mining, fuzzy methods appear to be especially useful for representing "vague" patterns, a point of critical importance in many fields of application.
3. FST, in conjunction with possibility theory, can contribute considerably to the modeling and processing of various forms of uncertain and incomplete information.
4. Fuzzy methods appear to be particularly useful for data pre- and postprocessing.

Despite the fact that substantial contributions have already been made to all of the aforementioned points, there is still space for improvement and a high potential for further developments. For example, concerning the first point, we already mentioned that notions like "comprehensibility", "simplicity", or "robustness" still lack an underlying formal theory including a quantification of their intuitive meaning in terms of universally accepted measures. Likewise, the fourth point has not received enough attention so far. In fact, even though FST seems to be especially qualified for data pre- and postprocessing, e.g., for feature generation (cf. Section 4), data summarization and reduction, approximation of complex and accurate models, or the (linguistic) presentation of data mining results, previous research has still more focused on the inductive reasoning or data mining process itself. Therefore, we see a high potential for future work in this area, especially against the background of the current trend to analyze complex and heterogeneous information sources that are less structured than standard relational data tables.

Acknowledgments: The author thanks two anonymous reviewers for their constructive comments and useful suggestions.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th Conference on* VLDB, pages 487–499, Santiago, Chile, 1994.

[2] A. Al-Ani and M. Deriche. A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence. *Journal of Artificial Intelligence Research*, 17:333–361, 2001.

[3] H. Bandemer and W. Näther. *Fuzzy Data Analysis.* Kluwer Academic Publishers, Dordrecht, 1992.

[4] A. Bargiela and W. Pedrycz. *Granular Computing: An Introduction.* Kluwer Academic Publishers, Boston, Dordrecht, London, 2005.

[5] S. Benferhat, D. Dubois, L. Garcia, and H. Prade. On the transformation between possibilistic logic bases and possibilistic causal networks. *International Journal of Approximate Reasoning*, 29(21):35–173, 2002.

[6] S. Benferhat and S. Smaoui. Hybrid possibilistic networks. *International Journal of Approximate Reasoning*, 44(3):224–243, 2007.

[7] F. Berzal, JC. Cubero, D Sanchez, JM. Serrano, and MA. Vila. An alternative approach to discover gradual dependencies. *Int. Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 15(5):559–570, 2007.

[8] C. Borgelt and R. Kruse. *Graphical Models – Methods for Data Analysis and Mining.* Wiley, Chichester, 2002.

[9] J. Casillas, O. Cordon, F. Herrera, and L. Magdalena, editors. *Interpretability Issues in Fuzzy Modeling.* Studies in Fuzziness and Soft Computing. Springer-Verlag, Berlin, 2003.

[10] G. Chen, Q. Wei, E. Kerre, and G. Wets. Overview of fuzzy associations mining. In *Proc. ISIS–2003, 4th International Symposium on Advanced Intelligent Systems.* Jeju, Korea, September 2003.

[11] S. Cho and J. Kim. Combining multiple neural network by fuzzy integral for robust classification *IEEE Transactions on Systems, Man, and Cybernetics*, 25:380–384, 1995.

[12] O. Cordon, MJ. del Jesus, and F. Herrera. Analyzing the reasoning mechanisms in fuzzy rule based classification systems. *Mathware & Soft Computing*, 5:321–332, 1998.

[13] O. Cordon, F. Gomide, F. Herrera, and F. Hoffmann anf L. Magdalena. Ten years of genetic fuzzy systems: current framework and new trends. *Fuzzy Sets and Systems*, 141(1):5–31, 2004.

[14] TH. Dang, B. Bouchon-Meunier, and C. Marsala. Measures of information for inductive learning. In *Proc. IPMU-2004*, Perugia, Italy, 2004.

[15] M. Delgado, N. Marin, D. Sanchez, and MA. Vila. Fuzzy association rules: general model and applications. *IEEE Transactions on Fuzzy Systems*, 11(2):214–225, 2003.

[16] P. Diamond and P. Kloeden. *Metric Spaces of Fuzzy Sets: Theory and Applications*. World Scientific, Singapur, 1994.

[17] P. Diamond and H. Tanaka. Fuzzy regression analysis. In R. Slowinski, editor, *Fuzzy Sets in Decision Analysis, Operations Research and Statistics*, pages 349–387. Kluwer, 1998.

[18] D. Dubois, F. Dupin de Saint-Cyr, and H. Prade. A possibility-theoretic view of fomal concept analysis. *Fundamenta Informaticae*, 76:1–19, 2006.

[19] D. Dubois, F. Esteva, P. Garcia, L. Godo, R. Lopez de Mantaras, and H. Prade. Fuzzy set modelling in case-based reasoning. *International Journal of Intelligent Systems*, 13:345–373, 1998.

[20] D. Dubois, E. Hüllermeier, and H. Prade. A systematic approach to the assessment of fuzzy association rules. *Data Mining and Knowledge Discovery*, 13(2):167.

[21] D. Dubois, E. Hüllermeier, and H. Prade. Fuzzy set-based methods in instance-based reasoning. IEEE *Transactions on Fuzzy Systems*, 10(3):322–332, 2002.

[22] D. Dubois, E. Hüllermeier, and H. Prade. A note on quality measures for fuzzy association rules. In *Proc.* IFSA–03*, 10th International Fuzzy Systems Association World Congress*, number 2715 in LNAI, pages 677–648, Istambul, 2003. Springer-Verlag.

[23] D. Dubois and H. Prade. Gradual inference rules in approximate reasoning. *Information Sciences*, 61(1,2):103–122, 1992.

[24] D. Dubois and H. Prade. What are fuzzy rules and how to use them. 84:169–185, 1996.

[25] UM. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996.

[26] AP. Gasch and MB. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11):1–22, 2002.

[27] J. Gebhardt and R. Kruse. A possibilistic interpretation of fuzzy sets by the context model. In *IEEE International Conference on Fuzzy Systems*, pages 1089–1096, 1992.

[28] F. Höppner, F. Klawonn, F. Kruse, and T. Runkler. *Fuzzy Cluster Analysis.* Wiley, Chichester, 1999.

[29] E. Hüllermeier. Implication-based fuzzy association rules. In *Proc.* PKDD–01*, 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 241–252, Freiburg, Germany, 2001.

[30] E. Hüllermeier. Association rules for expressing gradual dependencies. In *Proceedings* PKDD–02*, 6th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 200–211, Helsinki, Finland, 2002.

[31] E. Hüllermeier. Possibilistic induction in decision tree learning. In *Proc.* ECML–02*, 13th European Conference on Machine Learning*, pages 173–184, Helsinki, Finland, 2002.

[32] E. Hüllermeier. Possibilistic instance-based learning. *Artificial Intelligence*, 148(1–2):335–383, 2003.

[33] E. Hüllermeier. Cho-k-NN: A method for combining interacting pieces of evidence in case-based learning. In *Proceedings* IJCAI–05*, 19th International Joint Conference on Artificial Intelligence*, pages 3–8, Edinburgh, Scotland, 2005.

[34] E. Hüllermeier, I. Renners, and A. Grauel. An evolutionary approach to constraint-regularized learning. *Mathware and Soft Computing*, 11(2–3):109–124, 2004.

[35] E. Hüllermeier. Fuzzy sets in machine learning and data mining: Status and prospects. *Fuzzy Sets and Systems*, 156(3):387–406, 2005.

[36] CZ. Janikow. Fuzzy decision trees: Issues and methods. *IEEE Transactions on Systems, Man, and Cybernetics*, 28(1):1–14, 1998.

[37] R. Krishnapuram and J.M. Keller. A possibilistic approach to clustering. IEEE *Transactions on Fuzzy Systems*, 1(2):98–110, 1993.

[38] A. Laurent. Generating fuzzy summaries: a new approach based on fuzzy multidimensional databases. *Intelligent Data Analysis Journal*, 7(2):155–177, 2003.

[39] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50(2):157–224, 1988.

[40] CHL. Lee, A. Liu, and WS. Chen. Pattern discovery of fuzzy time series for financial prediction. *IEEE Transactions on Knowledge and Data Engineering*, 18(5):613–625, 2006.

[41] T. Y. Lin, Y.Y. Yao, and L.A. Zadeh, editors. *Data Mining, Rough Sets and Granular Computing.* Physica-Verlag, Heidelberg, 2002.

[42] T.M. Mitchell. The need for biases in learning generalizations. Technical Report TR CBM–TR–117, Rutgers University, 1980.

[43] D. Nauck, F. Klawonn, and R. Kruse. *Foundations of Neuro-Fuzzy Systems.* Wiley and Sons, Chichester, UK, 1997.

[44] C. Olaru and L. Wehenkel. A complete fuzzy decision tree technique. *Fuzzy Sets and Systems*, 138(2), 2003.

[45] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[46] W. Pedrycz. *Knowledge-Based Clustering.* John Wiley & Sons, 2005.

[47] W. Pedrycz and Z.A. Sosnowski. Designing decision trees with the use of fuzzy granulation. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 30(2):151–159, 2000.

[48] W. Pedrycz and Z.A. Sosnowski. The design of decision trees in the framework of granular data and their application to software quality models. *Fuzzy Sets and Systems*, 123(3):271–290, 2001.

[49] W. Pedrycz and Z.A. Sosnowski. C-fuzzy decision trees. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 35(4):498–511, 2005.

[50] H. Prade. Raisonner avec des règles d'inférence graduelle - Une approche basée sur les ensembles flous. *Revue d'Intelligence Artificielle*, 2(2):29–44, 1988.

[51] E.H. Ruspini. Possibility as similarity: The semantics of fuzzy logic. In P.P. Bonissone, H. Henrion, L.N. Kanal, and J.F. Lemmer, editors, *Uncertainty In Artificial Intelligence 6.* Elsevier Science Publisher, 1991.

[52] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In VLDB–95, *Proceedings of 21th International Conference on Very Large Data Bases*, pages 11–15, Zurich, September 1995.

[53] T. Sudkamp. Similarity as a foundation for possibility. In *Proc. 9th IEEE Int. Conference on Fuzzy Systems*, pages 735–740, San Antonio, 2000.

[54] T. Sudkamp. Examples, counterexamples, and measuring fuzzy associations. *Fuzzy Sets and Systems*, 149(1), 2005.

[55] R. Viertl. *Statistical Methods for Non-Precise Data.* CRC Press, Boca Raton, Florida, 1996.

[56] R. Weber. Fuzzy-ID3: a class of methods for automatic knowledge acquisition. In *IIZUKA-92, Proc. of the 2nd Intl. Conf. on Fuzzy Logic*, volume 1, pages 265–268. 1992.

[57] R.R. Yager. Soft aggregation methods in case based reasoning. *Applied Intelligence*, 21:277–288, 2004.

[58] L.A. Zadeh. New approach to the analysis of complex systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(1), 1973.