

Information Mining

Rudolf Kruse

Dept. of Knowledge Processing and Language Engineering
Otto-von-Guericke-University of Magdeburg
D-39106 Magdeburg, Germany
E-mail: rudolf.kruse@cs.uni-magdeburg.de

Abstract

In response to the explosion of collected, stored, and transferred data, *Data Mining* has emerged as a new research area. However, up to now research has mainly been oriented at highly structured data and the goal to obtain understandable results has often been neglected. *Information Mining* tries to combine the analysis of heterogeneous information sources with the prominent aim of producing comprehensible results. Since the objective 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.

Keywords: Data mining, information mining, heterogeneous data.

1 Introduction: Mining in Heterogeneous Data

As the computer industry produces more powerful processors and storage devices every year, it becomes easier and cheaper to collect, store and process large digital archives of all kinds of data, including documents, images, sounds or video. However, exploiting the information contained in these archives in an intelligent way often turns out to be fairly difficult. In reply to this challenge a new area of research has emerged, known as “Knowledge Discovery in Databases” or “Data Mining”. Although the standard definition of knowledge discovery and data mining [5] only speaks of discovery in *data*, thus not restricting the type

and the organization of the data to work on, research mostly concentrated on highly structured 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. Therefore we suggest to concentrate on *information mining*, which we see as an extension of traditional data mining to identify understandable patterns in *heterogeneous information sources*.

There have been some publications on mining in heterogeneous data like texts, images or sounds. However, most of the reported work deals with *information retrieval*, i.e. indexing and the recovery of pieces of information matching a query. *Information mining*, on the other hand, aims at discovering *knowledge*, i.e. more general patterns within objects or collections of objects. Examples of knowledge that can be extracted from textual archives are document categorization, term or phrase associations, information extraction from text, or text summarization [10]. In image databases typical mining tasks comprise supervised or unsupervised classification of images or objects in images (cf. [9, 13, 4]). The extraction of patterns from moving images is even more challenging. An interesting approach is given in [1], where basketball games are analyzed. Spatial data like geometric models or geographic data are another source of information that can be mined for dependencies and relationships.

Of course, we cannot expect to find data mining algorithms that are generally applicable to all mentioned kinds of information sources. The approaches will always strongly depend on pre-processing to extract characterizing features from the specific type of media. For this purpose, techniques from ma-

chine vision, signal processing or text analysis will be used. However, to enable data mining in these feature spaces we suppose that it is crucial to have algorithms that can easily incorporate expert background knowledge. For research this results in the challenge to develop theories and scalable techniques that can extract knowledge from large, multi-relational, and high-dimensional information sources, and that close the semantic gap between structured data and human notions and concepts. That is, they should 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.

2 Strengths of Fuzzy Set Models

It is undisputed that language is a humans most effective tool to structure his experience and to model his 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* [12].

A fundamental property of linguistic terms is their inherent vagueness, i.e., they have “fuzzy” boundaries: Well-known examples include the terms *pile of sand* and *bald*. In both cases no precise number of hairs or grains of sand, respectively, can be given which separates the situations in which the terms are applicable from those in which they are not.

Fuzzy set theory provides excellent means to model the “fuzzy” boundaries of linguistic terms by introducing gradual memberships. 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. All of these interpretations are needed in applications and have proven useful for solving practical problems. In addition, fuzzy sets also turned out to be worth considering when non-linguistic, but imprecise (i.e. set-valued) information has to be modeled.

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

3 Neuro-Fuzzy-Systems

One way to use fuzzy systems in data analysis is 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* [7]. 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 comprehensibility. Important aspects of the interpretability of a fuzzy system are the number of rules in the rule base, the number of variables used in each rule, and the partitioning of the variables with fuzzy sets associated with meaningful linguistic labels.

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 start with initial fuzzy partitions to create a rule base [8]. By providing initial fuzzy sets, it is easy to find meaningful linguistic labels. The rule learning, which can be done in a single pass through the training data [11], has been implemented in the neuro-fuzzy classification system NEFCLASS [7]. The most recent implementation of the NEFCLASS approach has been extended by appealing features like treatment of missing values, the ability to use data with numeric and symbolic attributes and automatic pruning strategies to get more compact and readable rule bases.

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, and merging of information originating from multiple sources to provide a better insight and understanding of the

phenomena under consideration. Some aspects of information fusion can be implemented by NEFCLASS.

If a fuzzy classifier is created from data, its training is usually supervised, i.e. each pattern is labeled. 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. 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 classifications according to the context model [6]. These can be used by the NEFCLASS learning algorithm.

Another aspect of information fusion that is implemented by NEFCLASS is the possibility to integrate prior expert knowledge in form of fuzzy rules and information obtained from data. If prior knowledge about the classification problem is available, then the rule base of the fuzzy classifier can be initialized with suitable fuzzy rules. The learning algorithm performs fusion of expert opinions and observations by completing and/or correcting the rule base.

4 Dependency Analysis with Possibilistic Graphical Models

Since reasoning in high-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 modeling, *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. W.r.t. this quantitative component, the conditional independence graph describes a *decomposition* 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 modeled. Therefore recent research has concentrated on learning graphical models from databases of sample cases.

Due to the origin of graphical modeling 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 [2, 3], 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 analyze is highly imprecise and thus offer interesting perspectives for future research.

5 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. 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 interfaces. In order to achieve this user-friendliness, often certain (limited) reductions in performance and solution quality are accepted.

In the field of information mining, fuzzy methods can be a valuable means to reach a trade-off between correctness, completeness, and efficiency on the one hand and manageable solutions for more and more complex systems on the hand. However, a formal theory of utility in which the simplicity of a system is taken into account, is a lasting challenge for the fuzzy community to meet.

Acknowledgements

I am grateful to C. Borgelt and A. Klose for their support in the preparation of this manuscript.

References

- [1] I. Bhandari, E. Colet, J. Parker, Z. Pines, and R. Pratap. Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. *Data Mining and Knowledge Discovery* 1:121–125, 1997
- [2] C. Borgelt, J. Gebhardt, and R. Kruse. Chapter F1.2: Inference Methods. In: E. Ruspini, P. Bonissone, and W. Pedrycz, eds. *Handbook of Fuzzy Computation*. Institute of Physics Publishing Ltd., Bristol, United Kingdom 1998
- [3] C. Borgelt and R. Kruse. *Graphical Models — Methods for Data Analysis and Mining*. Wiley, Chichester, United Kingdom 2001 (to appear)
- [4] U. Fayyad und P. Smyth, Ed., *Image Database Exploration: Progress and Challenges*. AAAI Press, Menlo Park, CA, 1993
- [5] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds. *Advances in Knowledge Discovery and Data Mining*. AAAI Press / MIT Press, Cambridge, MA 1996
- [6] R. Kruse, J. Gebhardt, and F. Klawonn. *Foundations of Fuzzy Systems*. Wiley, Chichester, United Kingdom 1994
- [7] D. Nauck, F. Klawonn, and R. Kruse. *Foundations of Neuro-Fuzzy Systems*. Wiley, Chichester, United Kingdom 1997
- [8] D. Nauck and R. Kruse. Chapter D.2: Neuro-fuzzy Systems. In: E. Ruspini, P. Bonissone, and W. Pedrycz, eds. *Handbook of Fuzzy Computation*. Institute of Physics Publishing Ltd., Bristol, United Kingdom 1998
- [9] C. Ordonez and E. Omiecinski. Discovering association rules based on image content. *Proc. IEEE Forum on Research and Technology: Advances in Digital Libraries*. Baltimore, Maryland, 1999
- [10] M. Rajman and R. Besançon. Text mining: natural language techniques and text mining applications. *Proc. 7th IFIP 2.6 Working Conference on Database Semantics (DS-7)*. Chapman & Hall IFIP proceedings series, 1997
- [11] L.-X. Wang and J.M. Mendel. Generating fuzzy rules by learning from examples. *IEEE Trans. Syst., Man, Cybern.* 22:1414–1227. IEEE Press, Piscataway, NJ, 1992
- [12] L.A. Zadeh. Fuzzy Logic = Computing With Words. *IEEE Transactions on Fuzzy Systems* 4:103–111. IEEE Press, Piscataway, NJ, 1996
- [13] O.R. Zaiane, J. Han, Z.N. Li, J.Y. Chiang and S. Chee. MultiMedia-Miner: a system prototype for multimedia data mining. *Proc. 1998 ACM-SIGMOD Conf. on Management of Data*. Seattle, Washington, 1998