Fuzzy Model based classification of Non-Inter compatible Data: Improvisation on Comprehending Heterogeneous Information

Deepak V P
Software Engineer, Larsen & Toubro Infotech, Chennai, India
Deepak.VP@lntinfotech.com

Ananth Kumar V
Software Engineer, Oracle, Bangalore, India
ananthkumar.venkateshswaran@oracle.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 at highly structured and precise data pertaining to a single dimension mostly. In addition, the goal to obtain understandable results is often neglected. 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 heterogeneous information mining. The objective of the paper is to analyze heterogeneous information sources with the prominent aim of producing comprehensible results.

Keywords: Data Mining, Heterogeneous data, Fuzzy models, Information mining, Decision making, Hypothesis.

1 Introduction

Data is also being pulled in from many more sources. With the explosion in Internet use, data pours into data warehouses very quickly. Fuzzy Logic and Fuzzy Expert Systems define fuzzy logic as a "Superset of conventional (Boolean) logic that has been extended to handle the concept of partial truths - truth values between completely true and completely false". It is the process of identifying interesting patterns and describing them in a concise and meaningful manner. Most of the techniques used in feature extraction are highly mathematical and are quantitative in nature. To fully exploit all the attributes of an object present in the data set, one must use the qualitative attributes. These can then also be used in describing the result such that the result can be easily understood. Fuzzy logic bridges this gap and allows the use of qualitative/linguistic terms in formation of extraction rules.

The amount of information being collected by businesses, companies and agencies is large. A recent advance in technologies to automate and improve data collection has only increased the volumes of data. The need for collecting data is to extract useful information. Data mining is primarily the process of knowledge discovery in databases.

The data of interest is the previously unknown and potentially useful information contained in the database

In contrast to the abundance of data there is a lack of tools that can transform these data into useful information and knowledge. Although a user often has a vague understanding of his data and their meaning — the user can usually formulate hypotheses and guess dependencies—, the user rarely knows

a. Where to find "Relevant" pieces of information
b. Whether the information extracted is supportive of the hypothesis considered.
c. Whether more "relevant" data can be extracted from the data.
d. How can the information be extracted in more reliable and faster way?

2 Pertinence of Fuzzy Set Models for Mining Heterogeneous Information

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.

Fuzzy sets are known for their pertinence to predictability rather than precision. They help in cutting out the unwanted computation of high voluminous data from data warehouses and data marts by inducing near to precision fuzzy rules that dissect the data repositories to bring out relevance required by the user. The process of setting up a model for this purpose involves linguistic computation cutting down the process of incorporating complex mathematical models.

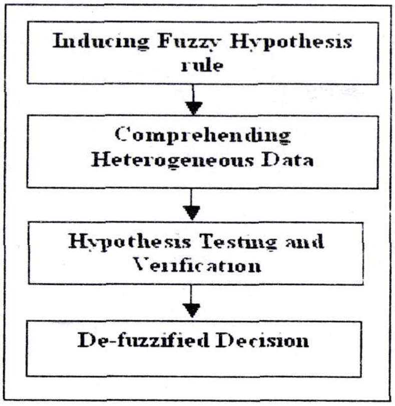
Inducing Fuzzy Hypothesis rule
Comprehending Heterogeneous Data
Hypothesis Testing and Verification
De-fuzzified Decision

Figure 1: Simplified process diagram

3 Fuzzy Model Based Classification of Non-Inter Compatible Data

3.1 Inducing Fuzzy Hypothesis Rule

Fuzzy sets are a generalization of crisp sets providing increased expressive power and comprehensibility. There have been many attempts to induce fuzzy concept representations from data. These include fuzzy neural networks, fuzzy decision trees, genetic algorithms, grid methods, and clustering.

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

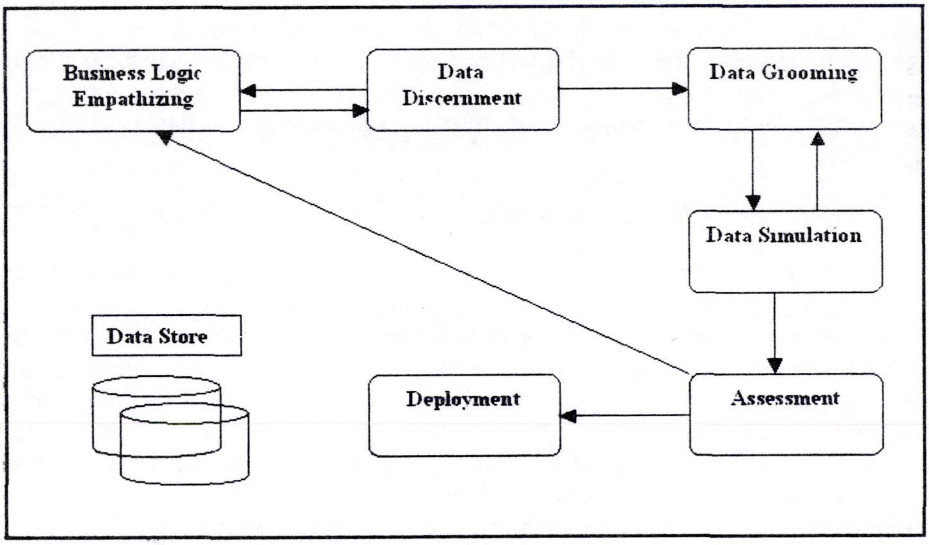

Figure 2: Process Methodology defining Business Intelligence

The “Business Logic Empathizing” and “Data Discernment” phases are usually strongly human centered 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 modeling 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 modeling techniques. In this step fuzzy methods may, for example, be used to detect outliers, e.g., by fuzzy clustering the data and then finding those data points that are far away from the cluster prototypes. The modeling 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, consists of approaches that analyze 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 possibility graphical models from data, which complements the induction of the well-known probabilistic graphical models. The second class, fuzzy data analysis, 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 modeling 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.

3.2 Simplifying the process of comprehending Heterogeneous Data

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. If we apply such techniques, we must be aware of the trade-of 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

a. There are only few fuzzy rules in the rule base,
b. There are only few variables used in each rule,
c. The variables are partitioned by few meaningful Fuzzy set,
d. 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. Cluster-oriented rule learning approaches are based on fuzzy cluster analysis 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. 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.

3.3 Hypothesis Testing and Verification (heterogeneous data mining)

Structure-oriented approaches avoid all these drawbacks, because they do not search for (hyper ellipsoidal or hyper rectangular) 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. The rule base is created by selecting the grid cells containing data. This can be done in a single pass through the training data. 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 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. For example, the fuzzy sets are tuned by a simple back propagation-like procedure. The algorithm does not use gradient-descent, because the degree of fulfillment of a fuzzy rule is determined by the minimum and noncontiguous membership functions 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 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. The approach to generate fuzzy classifiers from data has the following features:

a. Structure-oriented fuzzy rule learning,
b. Automatic determination of the number of rules,
c. Treatment of missing values (without imputation), the ability to use data with both numeric and symbolic attributes
d. Constrained fuzzy set learning, and

e. Automatic pruning strategies

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. 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.

If a fuzzy classifier is created based on a supervised learning problem ˜ L, then the most common way is to provide a data set, where each pattern is labeled—ideally with its correct class. That is, 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. The training then reflects fusion of expert opinions on the data set level. Another aspect of information fusion is to integrate 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 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

a. Always prefer the expert rule,
b. Always prefer the learned rule, or Select the rule with the higher performance value.

It is necessary that 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 observations. Because we are able to resolve conflicts between rules based on rule performance, it is also able to fuse expert opinions on the fuzzy rule level. Rule bases from different experts can be entered as prior knowledge. They will 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.

3.4 Improvisation of Hypothesis results: De-fuzzified Decision

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 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. 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 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.

4 Illustrative Example

Land evaluation is the assessment of land performance when used for specific purposes. As such it provides a rational basis for taking land use decisions based on analysis of relations between land use and land, giving estimations of required inputs and projected outputs, and they must took into account the environmental effects of agricultural practices. Land evaluation deals with two majors aspects of land: physical resources such as soil, topography and climate, and socioeconomic resources like farm size or management level. The need for optimal use of land has never been greater than now at the present. The classic models of land evaluation (based on the limiting factor) and land evaluation rules (from which models are established) assess these factors separately and they do not consider this cultivation as a global system. On the other hand, user knowledge has a vital importance to evaluate the crop, since user (farmer) is the main agent who works directly on the system and who receives any decision.

4.1 Land Evaluation Description

This experiment aims to provide a robust and reliable environment which shall equip the user, with all the necessary details pertaining to Productivity of any particular desired Crop, based on the Geo-spatial data such as Land Humidity, Rainfall, Soil Density and a host of other details, which shall be used to Predict whether the particular User-specified Region is suitable for growth of the concerned Crop.

The experiment must have the following features:

a. It must have the complete set of Satellite Imagery, for Land Humidity, soil Density, etc. from which the input Data Set can be mined from.

b. The User must specify the region of Choice, for which the particular Data shall be extracted and the Report shall be generated.

The development of the project is as follows:

a. Different Maps of the same region and of the same size are stored in the database.
b. The user is then asked to click on one particular area and that area is selected from all the maps in the database. This is done by retrieving 50 pixels long and wide around the clicked region.
c. The clipped images are then subject to the analyzer. Intensity values in the region are first noted and then, noting all the values of a particular map, the most common value is taken out as input to the next level.

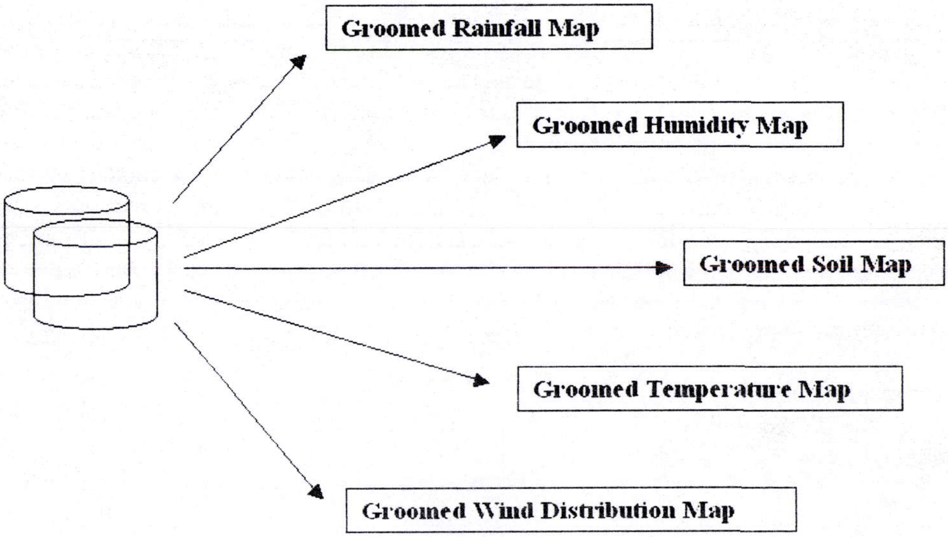

Figure 3: Clipping & Grooming of user-selected data

d. The inputs, when fed into the fuzzy inference engine, produce output corresponding to the values from all the maps.
e. With the output received from the fuzzy engine, the productivity of each crop that is specified is found and given to the user who can then easily find out which crop will provide maximum output.

5 Conclusion

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 is 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.

Acknowledgements

We would like to thank Dr. Daniel M. Dubois, President of CASYS'05 for his continued support.

References

J.C. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Series: The Handbooks on Fuzzy Sets. Kluwer, Norwell, MA, USA 1998

H. Bandemer and W. N¨ather. Fuzzy Data Analysis. Kluwer, Dordrecht, Netherlands 1992

J.C. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Series: The Handbooks on Fuzzy Sets. Kluwer, Norwell, MA, USA 1998

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

D. Dubois, H. Prade, and R.R. Yager. Merging Fuzzy Information. In: J.C. Bezdek, D. Dubois, and H. Prade, eds. Approximate Reasoning and Fuzzy Information Systems, (Series: Handbook of Fuzzy Sets), 335–402. Kluwer, Dordrecht, Netherlands 1999

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, USA 1996

F. H¨oppner, F. Klawonn, R. Kruse, and T. Runkler. Fuzzy Cluster Analysis. J. Wiley & Sons, Chichester, England 1999

R. Kruse, J. Gebhardt, and F. Klawonn. Foundations of Fuzzy Systems. J. Wiley & Sons, Chichester, England 1994

D. Nauck, F. Klawonn, and R. Kruse. Foundations of Neuro-Fuzzy Systems. J. Wiley & Sons, Chichester, England 1997

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, USA 1992

L.A. Zadeh. Fuzzy Logic "Computing With Words. IEEE Transactions on Fuzzy Systems 4:103–111". IEEE Press, Piscataway, NJ, USA 1996

Chen, M-S., Han., J., and Yu, P.S, 1996: Data Mining: An Overview from a Database Perspective, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6

Hambaba, M. L., 1996: Intelligent Hybrid System for Data Mining, Proceedings of IEEE/IAFE 1996 Conference on Computer Intelligence for Financial Engineering.

Clark, P., Niblett, T.: The CN2 rule induction algorithm. Machine Learning 3 (1989) 261–2842. Quinlan, J.: c4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA (1993)

Cohen, W.: Fast effective rule induction. In: Proceedings 12th International Conference on Machine Learning, Morgan Kaufmann (1995) 115–123

Domingos, P.: Unifying instance-based and rule-based induction. Machine Learning 24 (1996) 141–168

Sebag, M., Schoenauer, M.: A rule-based similarity measure. In Wess, S., Althoff, K.D., Richter, M.M., eds.: Topics in case-based reasoning. Springer Verlag (1994) 119–130

T. Lindgren and H. Bostr¨om. Resolving rule conflicts with double induction. In Proc. of the 5th International Symposium on Intelligent Data Analysis

P. Bartlett R. Schapire, Y. Freund and W. Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics.

Kandel, A., Pacheco, R., Martins, A., and Khator, S.: The Foundations of Rule-Based Computations in Fuzzy Models. In: Pedrycz W. (ed.): Fuzzy Modelling, Paradigms and Practice. Kluwer, Boston (1996) 231-263

Last, M., and Kandel, A.: Automated Perceptions in Data Mining, invited paper, to appear in the proceedings of the 8th International Conference on Fuzzy Systems

Mendenhall, W., Reinmuth, J.E., Beaver, R.J.: Statistics for Management and Economics. Duxbury Press, Belmont, CA (1993)