A SURVEY: FUZZY SET THEORY IN DATA MINING

Komal Sahedani*

Abstract: *The present article provides a Review of the available literature on data mining & fuzzy set. Usually, data mining is considered as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. In our data-driven data mining model, knowledge is originally existed in data, but just not understandable for human. Data mining is taken as a process of transforming knowledge from data format into some other human understandable format like rule, formula, theorem, etc.. Generally fuzzy sets are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster.*

Keywords: *Data mining, Fuzzy Set, Knowledge discovery from data (KDD).*

*C.E. Department, R.K. University, Rajkot

I. INTRODUCTION

The digital revolution has made digitized information easy to capture and fairly inexpensive to store [1], [2].With the development of computer hardware and software and the rapid computerization of business, huge amount of data have been collected and stored in databases. The rate at which such data is stored is growing at a phenomenal rate. As a result, traditional *ad hoc* mixtures of statistical techniques and data management tools are no longer adequate for analyzing this vast collection of data. Several domains where large volumes of data are stored in centralized or distributed databases include the following.

- Financial Investment: Stock indexes and prices, interest rates, credit card data, fraud detection [3].
- Health Care: Several diagnostic information stored by hospital management systems[4].
- Manufacturing and Production: Process optimization and trouble shooting [5].
- Telecommunication network: Calling patterns and fault management systems.
- Scientific Domain: Astronomical observations [6], genomic data, biological data.
- The World Wide Web [7].

The data mining refers to *extracting or "mining" knowledge from large amounts of data*. Mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both "data" and "mining" became a popular choice. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

Fuzzy logic has rapidly become one of the most successful of today's technologies for developing sophisticated control systems. The reason for which is very simple. Fuzzy logic addresses such applications perfectly as it resembles human decision making with an ability to generate precise solutions from certain or approximate information. While other approaches require accurate equations to model real-world behaviors, fuzzy design can accommodate the ambiguities of real-world in human language and logic. Although genetic algorithms and neural networks can perform just as well as fuzzy logic in many cases, fuzzy logic has the advantage that the solution to the problem can be cast in terms that human operators can understand, so that their experience can be used in th design of the controller. This makes it easier to mechanize tasks that are already successfully performed by humans.

The present article provides an overview of the available literature on data mining, that is scarce, in the fuzzy set. Section II describes the basic notions of knowledge discovery in databases, and data mining. Some challenges are highlighted.

This is followed in Section III by a survey explaining the role of the fuzzy set. Section IV concludes the article.

II. KDD & DATA MINING

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD.

The subject of KDD has evolved, and continues to evolve, from the intersection of research from such fields as databases, machine learning, pattern recognition, statistics, artificial intelligence, reasoning with uncertainties, knowledge acquisition for expert systems, data visualization, machine discovery, and high performance computing. KDD systems incorporate theories, algorithms, and methods from all these fields. Many successful applications have been reported from varied sectors such as marketing, finance, banking, manufacturing, and telecommunications.

Database theories and tools provide the necessary infrastructure to store, access and manipulate data. *Data warehousing* [2], a recently popularized term, refers to the current business trends in collecting and cleaning transactional data, and making them available for analysis and decision support.

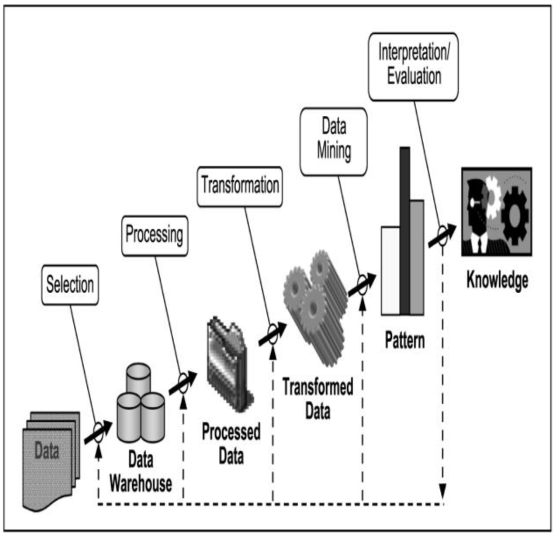


Figure 1 KDD Process

Knowledge discovery as a process is depicted in Figure 1 and consists of an iterative sequence of the following steps:

1. **Data cleaning** (to remove noise and inconsistent data)
2. **Data integration** (where multiple data sources may be combined)1
3. **Data selection** (where data relevant to the analysis task are retrieved fromthe database)
4. **Data transformation** (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
5. **Data mining** (an essential process where intelligent methods are applied in order to extract data patterns)
6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

Data mining involves fitting models to or determining patterns from observed data. The fitted models play the role of inferred knowledge. Deciding whether the model reflects useful knowledge or not is a part of the overall KDD process for which subjective human judgment is usually required. Typically, a data mining algorithm constitutes some combination of the following three components.

- The model: The function of the model (e.g., classification, clustering) and its representational form (e.g., linear discriminants, neural networks). A model contains parameters that are to be determined from the data.
- The preference criterion: A basis for preference of one model or set of parameters over another, depending on the given data. The criterion is usually some form of goodness-of-fit function of the model to the data, perhaps tempered by a smoothing term to avoid over fitting, or generating a model with too many degrees of freedom to be constrained by the given data.
- The search algorithm: The specification of an algorithm for finding particular models and parameters, given the data, model(s), and a preference criterion. A particular data mining algorithm is usually an instantiation of the model/preference/search

components. The more common model functions in current data mining practice include the following.

1) **Classification** : classifies a data item into one of several predefined categorical classes.
2) **Regression** : maps a data item to a realvalued prediction variable.
3) **Clustering** : maps a data item into one of several clusters, where clusters are natural groupings of data items based on similarity metrics or probability density models.
4) **Rule generation** : extracts classification rules from the data.
5) **Discovering association rules** : describes association relationship among different attributes.
6) **Summarization** : provides a compact description for a subset of data.
7) **Dependency modeling** : describes significant dependencies among variables.

Top 10 Challenging Problems in Data Mining Research:

1) *Developing a Unifying Theory of Data Mining*: The current state of the art of data-mining research is too "ad-hoc" so needs unifying Research.

2) *Scaling Up for High Dimensional Data and High Speed Streams*: ultra-high dimensional classification problems (millions or billions of features, e.g., bio data) & ultra High speed data streams.

3) *Sequential and Time Series Data*: How to efficiently and accurately cluster, classify and predict the trends?

4) *Mining Complex Knowledge from Complex Data*: many objects are not independent of each other, and are not of a single type. Mine the rich structure of relations among objects. E.g.: interlinked Web pages, social networks, metabolic networks in the cell.

5) *Data Mining in a Network Setting*: Linked data between emails, Web pages, blogs, citations, sequences and people.

6) *Distributed Data Mining and Mining Multi-agent Data*: Need to correlate the data seen at the various probes.

7) *Data Mining for Biological and Environmental Problems*: Biological data mining, such as HIV vaccine design, DNA, chemical properties,3D structures, and functional properties -> need to be Fused, Environmental data mining, Mining for solving the energy crisis.

8) *Data-mining-Process Related Problems*: The composition of data mining operations,Data cleaning, with logging capabilities, Visualization and mining automation.

9) *Security, Privacy and Data Integrity*: How to ensure the users privacy while their data are being mined?

10) *Dealing with Non-static,Unbalanced and Cost-sensitive Data*: Real world data are large (10^5 features) but only < 1% of the useful classes. [8]

III. FUZZY SET

The modeling of imprecise and qualitative knowledge, as well as the transmission and handling of uncertainty at various stages are possible through the use of fuzzy sets. Fuzzy logic is capable of supporting, to a reasonable extent, human type reasoning in *natural* form. It is the earliest and most widely reported constituent of soft computing.

Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth- truth values between "completely true" and "completely false". As its name suggests, it is the logic underlying modes of reasoning which are approximate rather than exact. The importance of fuzzy logic derives from the fact that most modes of human reasoning and especially common sense reasoning are approximate in nature.

The essential characteristics of fuzzy logic as founded by Zader Lotfi are as follows.

- In fuzzy logic, exact reasoning is viewed as a limiting case of approximate reasoning.
- In fuzzy logic everything is a matter of degree.
- Any logical system can be fuzzified.
- In fuzzy logic, knowledge is interpreted as a collection of elastic or, equivalently , fuzzy constraint on a collection of variables .
- Inference is viewed as a process of propagation of elastic constraints.

Fuzzy sets are generalized sets which allow for a graded membership of their elements. Usually the real unit interval [0; 1] is chosen as the member-ship degree structure.

Let X be a space of points, with a generic element of X denoted by x. Thus X = {x}.

A fuzzy set A in X is characterized by a membership function fA(*x*)which associates with each point in X a real number in the interval [0,1], with the values of fA(*x*) at x representing the "grade of membership" of x in A. Thus, the nearer the value of fA(*x*) to unity, the higher the grade of membership of *x* in A.

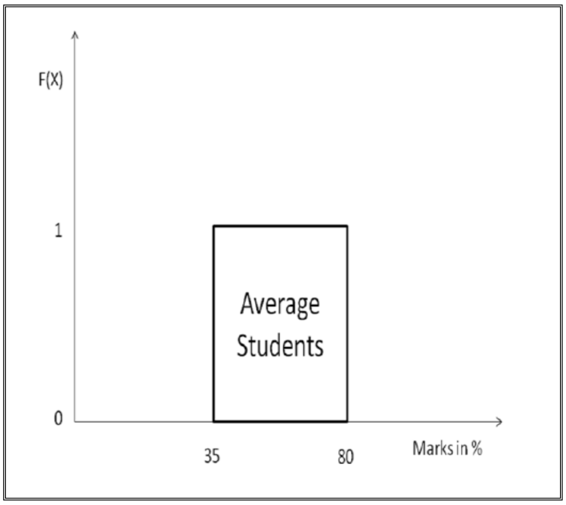
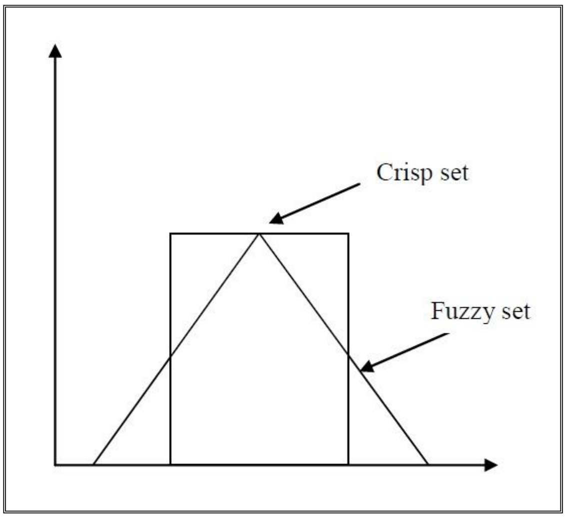


Figure 2 Example of sharp boundary problem



Suppose we have three range of marks of any examination.

F(x) is a function such that

0<X<=35 then f(X) = fail students

35<X<=80 then f(X) = average students

X>80 then f(X) = intelligent students

From the above scenario now suppose a student got 79.8% of marks then he is a average student. But if he got 80.1 marks then the tag will intelligent. But the student who got 79.8 % of marks is also an intelligent student, which is not the huge difference between both the conditions. Figure 2 show the scenario for boundary shape problem, which happened above.

There are some basic approaches to solve the sharp boundary problem.

1. Quantitative approach
2. Fuzzy Taxonomic Structures
3. Approximate Item set Approach

To resolve the sharp boundary problem by using Quantitative approach divide the variable marks into three fuzzy sets. The fuzzy sets and their membership functions will have to be defined by a domain expert. For easy demonstration, we will just define the borders of the sets and split the overlapping part equally between the so generated fuzzy sets. For an example, we will use the following borders for the fuzzy sets of the variable marks: Fail={0−35}, Average students={33−70}, intelligent students={70−∞}. The generated fuzzy sets is shown in Figure 2 . For all areas having no overlap of the sets,the support will simply be 1 for the actual itemset. If there is an overlap, the membership can be computed by using the borders of the overlapping fuzzy sets. The added support will here always sum up to 1.

Fuzzy set theory has been used more and more habitually in intellectual systems because of its simplicity and similarity to human reasoning.

There is a growing indisputable role of fuzzy set technology in the realm of data mining [9]. Various data browsers have been implemented using fuzzy set theory [10]. Analysis of real-world data in data mining often necessitates simultaneous dealing with different types of variables, *viz.*, categorical/symbolic data and numerical data. Nauck [11] has developed a learning algorithm that creates *mixed* fuzzy rules involving both categorical and numeric attributes. Pedrycz [12] discusses some constructive and fuzzy set-driven computational vehicles of knowledge discovery, and establishes the relationship between data mining and fuzzy modeling. The role of fuzzy sets is categorized below based on the different functions of data mining that are modeled.

1)Clustering: Data mining aims at sifting through large volumes of data in order to reveal useful information in the form of new relationships, patterns, or clusters, for decision-making by a user [13].

2) *Association Rules:* An important area of data mining research deals with the discovery of *association rules* [14]. An association rule describes an interesting association relationship among different attributes.

3) *Functional Dependencies:* Fuzzy logic has been used foranalyzing inference based on functional dependencies (FDs), between variables, in database relations. Fuzzy inference generalizes both imprecise (set-valued) and precise inference. Similarly, fuzzy relational databases generalize their classical and imprecise counterparts by supporting fuzzy information storage and retrieval [15].

4) *Data Summarization:* Summary discovery is one of the major components of knowledge discovery in databases. This provides the user with comprehensive information for grasping the essence from a large amount of information in a database. Fuzzy set theory is also used for data summarization [16].

5) *Web Application:* Mining typical user profiles and URL associations from the vast amount of access logs is an important component of Web personalization, that deals with tailoring a user's interaction with the Web information space based on information about him/her. Nasraoui *et al.* [17] have defined a *user session* as a temporally compact sequence of Web accesses by a user and used a dissimilarity measure between two Web sessions to capture the organization of a Web site. Their goal is to categorize these sessions using Web mining.

6) *Image Retrieval:* Recent increase in the size of *multimedia* information repositories, consisting of mixed media data, has made content-based image retrieval (CBIR) an active research area [18].

The fuzzy set used in Chemical,Medical ,Educational and also many other fields.

IV. CONCLUSION

Current research in data mining mainly focuses on the discovery algorithm and visualization techniques. There is a growing awareness that, in practice, it is easy to discover a huge number of patterns in a database where most of these patterns are actually obvious, redundant, and useless or uninteresting to the user. To prevent the user from being overwhelmed by a large number of uninteresting patterns, techniques are needed to identify only the useful/interesting patterns and present them to the user. Fuzzy sets, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster. They have been mainly used in clustering, discovering association rules and functional dependencies, summarization, time series analysis, web applications, and image retrieval.

REFERENCES

[1] U. Fayyad and R. Uthurusamy, “Data mining and knowledge discovery in databases,” *Commun. ACM*, vol. 39, pp. 24–27,1996.

[2] W. H. Inmon, “The data warehouse and data mining,” *Commun. ACM*, vol. 39, pp. 49–50, 1996

[3] J. A. Major and D. R. Riedinger, “EFD—A hybrid knowledge statisticalbased system for the detection of fraud,” *Int. J. Intell. Syst.*, vol. 7, pp. 687–703, 1992.

[4] R. L. Blum, Discovery and Representation of Causal Relationships From a Large Time-Oriented Clinical Database: The RX Project. New York: Spinger-Verlag, 1982, vol. 19. of Lecture Notes in Medical Informatics.

[5] R. Heider, Troubleshooting CFM 56-3 Engines for the Boeing 737—Using CBR and Data-Mining, Spinger-Verlag, New York, vol. 1168, pp. 512–523, 1996. *Lecture Notes in Computer Science*.

[6] U. Fayyad, D. Haussler, and P. Stolorz, “Mining scientific data,” *Commun. ACM*, vol. 39, pp. 51–57, 1996.

[7] O. Etzioni, “The world-wide web: Quagmire or goldmine?,” *Commun. ACM*, vol. 39, pp. 65–68, 1996.

[8] 10 challenging problems in data mining research by Qiang Yang, Hong Kong Univ. of Sci. & Tech. and Xindong Wu, University of Vermont,2005.

[9] R. R. Yager, “Database discovery using fuzzy sets,” *Int. J. Intell. Syst.*, vol. 11, pp. 691–712, 1996.

[10] J. F. Baldwin, “Knowledge from data using fuzzy methods,” *Pattern Recognition Lett.*, vol. 17, pp. 593–600, 1996.

[11] D. Nauck, “Using symbolic data in neuro-fuzzy classification,” in *Proc. NAFIPS 99*, New York, June 1999, pp. 536–540.

[12] W. Pedrycz, “Fuzzy set technology in knowledge discovery,” *Fuzzy Sets Syst.*, vol. 98, pp. 279–290, 1998.

[13] P. Piatetsky-Shapiro and W. J. Frawley, Eds., *Knowledge Discovery in Databases*. Menlo Park, CA: AAAI/MIT Press, 1991.

[14] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” in *Proc. 1993 ACMSIGMOD Int. Conf. Management Data*, Washington, DC, May 1993, pp. 207–216.

[15] J. Hale and S. Shenoi, “Analyzing FD inference in relational databases,” *Data Knowledge Eng.*, vol. 18, pp. 167–183, 1996.

[16] D. H. Lee and M. H. Kim, “Database summarization using fuzzy ISA hierarchies,” *IEEE Trans. Syst., Man, Cybern. B*, vol. 27, pp. 68–78, 1997.

[17] O. Nasraoui, R. Krishnapuram, and A. Joshi, “Relational clustering based on a new robust estimator with application to web mining,” in *Proc. NAFIPS 99*, New York, June 1999, pp. 705–709.

[18] S. K. Pal, A. Ghosh, and M. K. Kundu, Eds., *Soft Computing for Image Processing*. Heidelberg, Germany: Physica-Verlag, 2000.

[19] Sushmita Mitra*,* and Pabitra Mitra, “Data Mining in Soft Computing Framework:: A Survey”, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 1 ,pp. 3-14, JANUARY 2002.

[20] jiawei Han and Micheline Kamber , “Data Mining Concepts and Techniques”,2nd Edition.