ISSN 2229-5518

A Study on Applications of Fuzzy Set Theory in Datamining

Santhosh Kottam, Dr. Varghese Paul

Abstract— Data Mining Researchers are working for scalable and efficient data analysis algorithms apt for data mining functionalities. Data mining is an interdisciplinary area and it does include database systems, statistics, machine learning, visualization and information science. Data uncertainty and incomplete information raise challenges in Data mining. New mathematical approaches- Fuzzy Set, Rough Set and Soft Set-have wide applications in dealing uncertain data. This paper is a study on applications of the Fuzzy set theories in the domain of Data mining. We first discuss an introduction to Crisp set, Fuzzy set and Uncertain Data mining. In addition to these topics our paper is covering applications of Fuzzy sets in Association Analysis and Clustering. Finally this paper gives a brief on advantages of Fuzzy set theory.

Index Terms— Crisp set, Fuzzy set, Uncertain, Clustering, Apriori, Compatibility, Graduality and itemsets

1 INTRODUCTION

INFORMATION Technology is not a single discipline field. It consists of a number multidisciplinary subjects and technologies. Objectives of all the technologies are convert data into meaningful and valuable information. Among the subjects Data Mining has a great attention in the Information Industry.

Data Mining refers to develop or filter knowledge from large amount of unstructured data. Increasing popularity of Data Mining is really amazing and obvious. One of the main reasons for wide acceptance of Data Mining in the knowledge discovery domain is present world filled with huge amount of data and the imminent need for turning such data into useful information and knowledge. What kind of Data can be mined? Data mining researchers are solving this question and they are working for scalable and efficient data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Mining users must have idea regarding what kind of data have to retrieve from data repositories. But this not practices. Data mining experts proposes a set of models and patterns for mining data. These patterns are known as Data Mining functionalities Table 1. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Some of the functionalities are Concept/class description, mining frequent patterns, Association and Correlation, Classification and prediction, Cluster analysis and Outlier analysis [6].

TABLE 1

DATA MINING FUNCTIONALITIES AND TECHNIQUES

Functionality	Techniques	Application
Association	• Set Theory • Statistics • Bayesian Classification	Cross sell
Estimation	• Neural Network • Statistics • Time Series	Exchange Rate - Estimation Stock Price - Estimation
Classification	• Decision Tree • Fuzzy • Neural Network • Genetic Algorithm	Credit Embezzle Market - Segmentation
Prediction	• Regression • Neural Network • Decision Tree	Churn Prediction Fraudster- Prediction
Segmentation	• Neural Network • Statistics • Genetic Algorithm • Decision Tree	Market- segmentation

Most Challenging dilemma in data mining domain is invention of efficient algorithms for finding hidden patterns from incomplete and inconsistent Information Systems. Study on efficiency of such algorithms for implementing Data mining functionalities is interesting and knowledgeable. Many methods and algorithms are developed and tested in the mining world. Bayesian classification, decision tree classification, rules based classification, rough set and Fuzzy set are popular classification models nowadays used in commercial data mining systems. Clustering methods are classified in to following

- *Santhosh Kottam, Asst.Professor (Spl.Grade) & HOD, Department of MCA, Federal Institute of Science and Technology, Angamaly is currently pursuing PhD program in Computer Science in Bharathiyar University, India, PH-919446437774. E-mail: sankottam@gmail.com*
- *Dr. Varghese Paul, Professor & HOD, Department of Information Technology, Cochin University of Science & Technology, Cochin, India , PH-919446209996. E-mail: vp.itcusat@gmail.com*

ISSN 2229-5518

categories Partitioning, Hierarchal and Density based methods. Similarly data pre-processing forms - Data cleaning, Data Integration, Data selection, Data transformation and Data reduction-require effective methods to prepare data apt for Data Mining. Our paper discusses applications of fuzzy set in data mining functionalities. This theory not stick on a particular functionality instead we can apply in different forms. Many international workshops, conferences and seminars discussed strength of Fuzzy theory in Data Mining. Its supports both Discrete and Continuous valued attributes. Fuzzy set theory show strength in certain applications like evaluate fitness of other algorithms, market research, finance and health care etc. At present, certain research works in the Data Mining focus on applicability of Fuzzy set theory in following Data Mining functionalities- Rule Induction, Dimensionality Reduction, Association Rules, Relevance Analysis and Clustering.

2 Crisp Sets

In classical sets or Crisp sets are Binary. An element either belongs to the set or doesn't. Consider a domain X and a set A with objects $x_i \in X$. The membership function μA(x) is defined as

$$\mu_A(x) = \begin{cases} 1, & \text{if and only if } x \in A \\ 0, & \text{if and only if } x \notin A \end{cases}$$

Consider an example

P : the set of all people.
Y : the set of all young people.

Young={y | y=age(x) ≤ 25,x ∈ P}

Representation of the Crisp Set shown in Fig. 1

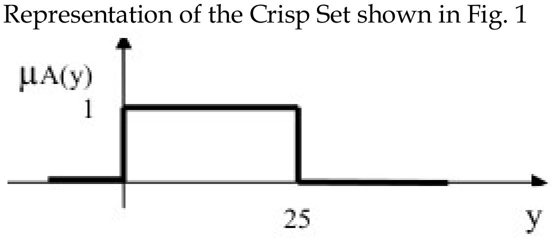


Fig. 1. Representation of the Crisp set

3 Fuzzy Sets

Fuzzy logic is differing from Classical set theory. An object x can be partially in A. μA(x) can take values between 0 and 1. Such sets are known as a Fuzzy sets. The value μA(x) is called the membership degree or membership grade [7]. An example for Fuzzy Set is shown in Fig. 2. There are certain properties for Fuzzy sets. The height of a fuzzy set hgt (A) is the Supremum (Maximum) of the membership grades of A.

$$hgt(A) = \sup_{x \in A} \mu A(x)$$

A Fuzzy set normal if hgt (A) =1. Any set that is not normal is called Subnormal. The support of a set A is the Crisp subset of A with nonzero membership grades. The core of a set A is the Crisp subset of A with membership grade equal to one.

Supp (A) $= \{x / \mu A(x) > 0\}$ and

Core (A) $= \{x / \mu A(x) = 1\}$

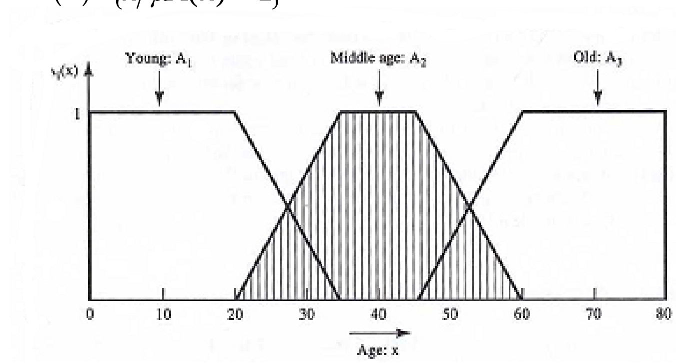


Fig. 2. Membership functions representing the concepts of a young, middle-aged, and old person

As demonstrated in this document, the numbering for sections upper case Arabic numerals, then upper case Arabic numerals, separated by periods. Initial paragraphs after the section title are not indented. Only the initial, introductory paragraph has a drop cap.

4 Uncertain Datamining

Data is often associated with uncertainty because of measurement inaccuracy, sampling discrepancy, outdated data sources, or other errors. These various sources of uncertainty have to be considered in order to produce accurate query and mining results. In recent years, there has been much research on the management of uncertain data in databases, such as the representation of uncertainty in databases and querying data with uncertainty. However, little research work has addressed the issue of mining uncertain data. We note that with uncertainty, data values are no longer atomic. To apply traditional data mining techniques, uncertain data has to be summarized into atomic values. Unfortunately, discrepancy in the summarized recorded values and the actual values could seriously affect the quality of the mining results [2] ,[5]. In recent years, there is significant research interest in data uncertainty management. There are a number of common data mining techniques, e.g., association rule mining, data classification, data clustering that need to be modified in order to handle uncertain data.

5 Association Analysis

5.1 Introduction

Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [Agrawal et al. 1993].It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [3].

Let S be a set of items ie S=$\{s1, s2,....sn\}$, and T be a collection of subsets of S. For X⊂S, we say that t∈T contains X if and only if X ⊆ t. An association rule has the form X→Y, where Y ⊂ S (|y|=1) and X∩Y=∅. The set X is called the antecedent of the rule while the set Y is called the consequent. Generally, a rule X→Y means that if a transaction contains X it very likely contains Y as well [9].

There are two parameters associated with a rule: Support and confidence. To define these parameters, we use Ts and Tc to denote the subset of T that contains both X and Y, and the subset of T that contains X respectively. It is obvious
Ts ⊆ Tc ⊆T.

Definition: The support of the rule X→Y obtained from dataset T is the ratio of the cardinality of Ts to the cardinality of T. Hence, the support of the rule is

$$S=\frac{|Ts|}{|T|}$$

Definition: The confidence of the rule X→Y obtained from dataset T is the ratio of the cardinality of Ts to the cardinality of Tc. Therefore, the confidence of the rule is

$$C=\frac{|Ts|}{|Tc|}$$

In this paper, our main concerns are support and confidence.

5.2 Association Analysis using Fuzzy Set

In solving real world problems, accurate data may not exist, or accurate computation may be impossible or not necessary. People need to make unambiguous high-quality decisions on inaccurate data. Fuzzy logic plays a vital role in ambiguous process of decision making. In traditional set theory, transactions either contain a specific item or not. Fuzzy set theory defines degree of membership of each item. First example shows how to generate association rules on a fuzzy set using Apriori algorithm. After that, same example defines associations by the use of one function: Compatibility by extending its traditional definition to present testing [1] [4].

5.3 A Sample Application

This sample application [1], assume that a dataset consists of symptoms of different. The symptoms could happen in different diseases. Each of the diseases is represented as a fuzzy set of symptoms. We perform the association analysis to determine what symptoms and their combination may occur most frequently which can give us some hints of what diseases are prevailing for the group of patients.

Data. We denote the symptom set as S = {A, B, C, D} and a disease as d. Here is a sample symptom set for four patients. The number attached with each symptom reflects the degree of presence for that symptom Table 2.

TABLE 2
PATIENT ID AND SYMPTOMS

ID	Symptoms
1	A(0.8), C(0.7), D(0.9)
2	B(0.6), C(0.5), E(0.4)
3	A(0.7), B(0.4), C(0.8), E(0.7)
4	B(0.9), E(0.8)

Finding Large Itemsets. From data in Table 2, we try to derive the large itemsets as we described in the previous section. We assume that large support should be no less that 0.38 i.e τ =0.38.

Let t_k be a fuzzy data item and tk=$\sum_j \beta jk / sj$ for k=1- - -N

We define support (sj) =$\sum_K \beta jk / N$

where 0<=βjk<=1 and N is the size of dataset. We also define fuzzy large itemsets of cardinality one by

$$A\tau\ (1) = \{sj \mid support\ (sj) \geq \tau\}$$

where τ is some rectified threshold for "large".

Here supports for all 1-itemsets are: supportA= 0.38, supportB=0.48, supportC=0.50, supportD=0.23 and supportE=0.48.
Hence, the large 1-itemset is: $A\tau^{(1)}$ = {A, B, C, E}.

We define the support of 2-itemset as:

$$Support\ (sj, sl) = \sum_k \frac{\beta jk \wedge \beta lk}{N}$$

The fuzzy large itemsets of cardinality two can be defined as

$$A\tau\ (2) =\{(sj, sl) \mid support\ (sj) \geq \tau\}$$

Since D is excluded, we only consider 2-itemset {AB, AC, AE, BC, BE, CE}. The supports for these 2-itemsets are: supportAB = 0.10, supportAC=0.35, supportAE=0.18, support BC=0.23, supportBE=0.40, and supportCE=0.28.

Hence, the large 2-itemset is: $A\tau^{(2)}$ = {BE}.

Similarly Support (sj, sl, St) =$\sum_k \frac{\beta jk \wedge \beta lk \wedge \beta tk}{N}$

$$A\tau\ (3) =\{(sj, sl, St) \mid support\ (sj, sl, St) \geq \tau\}$$

In general, we can define the support and large n-itemset $A\tau$ (n) as well for $n \leq N$.

Since $A\tau^{(2)}$ has only one member $A\tau^{(3)}$, $A\tau^{(4)}$ and $A\tau^{(5)}$ are empty.

Compatibility. Let d be a possible fuzzy consequent i.e $d=\sum \alpha i/Si$, where S1, S2, ……..Sp are items in S as defined in the previous section, and $0 \leq \alpha i \leq 1$ indicates to what extent the set of Si implies the presence of d. Our objective is to decide any possible relationships between d and a given fuzzy set of items $P=\sum \beta i/Si$. In our sample application we are interested in "large itemsets", S1, S2,……Sp.

We now define compatibility of d with P as
Comp(d,P)(u)=sup d(x)

Where the sup is taken over x such that P(x) =u. Thus, Comp (d,P) is a fuzzy subset of [0,1]. For the traditional setting of compatibility, readers may refer [8].

Provided that the symptoms for a patient is P=0.38/B+ 0.38/C + 0.32/BE. We try to find the compatibility he may suffer the disease d= 0.4/A +0.8/C+ 0.5/E. Based on the given data, the possible u's in Comp(d,P)(u)=sup d(x) over which Comp (d, P) $\neq$ 0 are 0.38 and 0.32. Thus, Comp (d, P) (0.38) =supd(x) where $x \in \{B, C\}$ so Comp (d, P) (0.38) = 0.8. Comp (d, P) (0.32) = 0 because d=0 on BE.

6 Clustering

6.1 Introduction

Clustering is an unsupervised learning process that can be used to organize a set of physical or abstract objects into classes of similar objects. It's an iterative process of finding better and better cluster centres. Objects from a cluster show high similarities in comparison to one another but are very dissimilar to objects in other clusters. Presently several clustering techniques are existing like partitioning methods, hierarchical methods, density-based methods etc. Among the techniques partitioning methods play an imperative role in clustering. Consider a data collection of n objects and a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$. All clusters together satisfy the following conditions: (1) each group must contain at least one object, and (2) each object must belong to exactly one group.

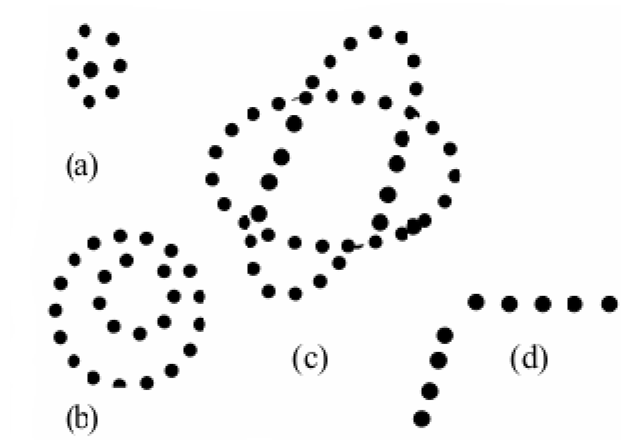


Fig. 3. Clusters of different shapes and Dimensions

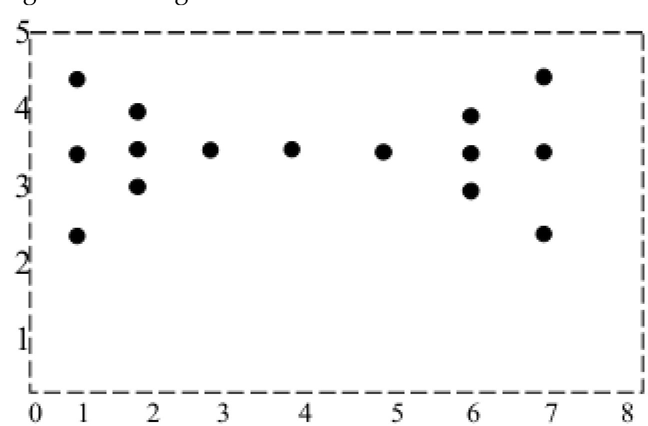
Crisp Clustering

In crisp clustering method certain data's has an ambiguous status concerning its cluster position. Consider a dataset A15[13] , be given in the Table 3.

TABLE 3
INPUT DATA

Dataset 1	Dataset 2
1.0	2.5
1.0	3.5
1.0	4.5
2.0	3.0
2.0	3.5
2.0	4.0
3.0	3.5
4.0	3.5
5.0	3.5
6.0	3.0
6.0	3.5
6.0	4.0
7.0	2.5
7.0	3.5
7.0	4.5

Apply classical set theory principles on A for clustering it into two separate clusters. A visual representation of this clustering shown in Fig 4.



ISSN 2229-5518

Fig. 4. Representation of dataset A15

Fig.5 visualizes two well separated clusters, all members of a cluster is assigned with a membership degree 1 and non-members are assigned a value zero. Consider, the point (4, 3.5), this point has an equal chance for belong to both clusters. Crisp method assigns this point to only one cluster, with membership degree 1.

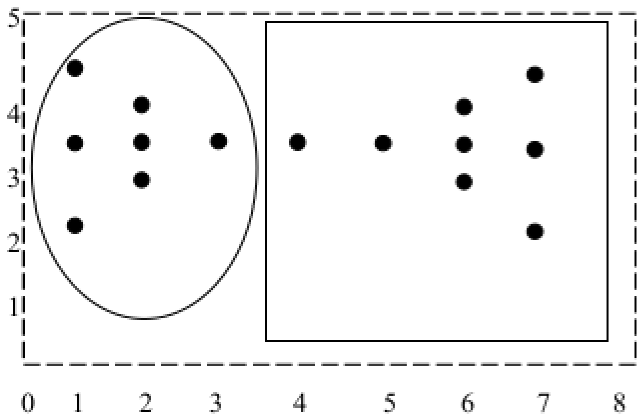


Fig. 5. Separation of data into two clusters

Consider a matrix U which defines all cluster subsets of dataset A. Each row of the matrix represents a cluster. Row elements are either 1 (stands for membership of a data item to a cluster) or 0 (stands for non membership). First row of U represents first subset of A, A1 and second row defines second cluster subset of A, A2. Data elements (4, 3.5) have been assigned to A1.

$$U=\begin{pmatrix} 0\,0\,0\,0\,0\,0\,0\,1 & 1 & 1 & 1 & 1 & 1 & 1\,1 \\ 1\,1\,1\,1\,1\,1\,1 & 0 & 0 & 0 & 0 & 0 & 0\,0\,0 \end{pmatrix}$$

This example discloses that crisp clustering may not give realistic picture of underlying data. Actually data items a5 and a6 represents patterns with mixture properties of clusters A1 and A2. Therefore these members cannot be fully assigned to either of these clusters. This shortcoming can be eliminated by using Fuzzy clustering methods.

7 Fuzzy Clustering

Fuzzy clustering is an approach operating towards fuzzy logic and it provides a flexible method to assigns the data points to the clusters. In Fuzzy clustering methods, objects belong to several clusters with different degrees of membership and assigned a membership degree between 1 and 0. Objects on the boundaries between several clusters are not forced to fully belong to one of the cluster. There are several Fuzzy clustering algorithms are existing like Hierarchical clustering methods, Graph-theoretic and Objective function methods. Objective function methods give the most precise formulation of the clustering. These models are formed by an objective function whose parameter's extreme values are defined as optimal clustering. Next we discuss one of the objective function methods applicable for point-prototype clustering known as Fuzzy C means clustering.

7.1 Fuzzy C Means Clustering

Fuzzy c-means clustering algorithms are iterative and its output is optimal for 'c' partitions, which obtained by minimizing the objective function J_{FCM}[14]. The basic FCM model is formulated as follows:

$$J(U,V) = \Sigma_{(i=1,c)} \Sigma_{(k=1,n)} (u_{ik})^m (D_{ik})^2;$$

$$(D_{ik})^2 = || x_k - vi ||^2$$

Subject to

$$\Sigma_{(i=1,c)} u_{ik}=1, \text{ for } k=1,2,....n$$

$$0<= u_{jk}<=1, \text{ for all } i=1,2,\ldots c \text{ and } k=1,2,\ldots n$$
$$0<= \Sigma_{(k=1,n)} u_{ik} <n, \text{ for } i=1,2,\ldots c$$

Given that

$$v_i = \Sigma_{(k=1,n)} (u_{ik})^m x_{ik} \,/\, \Sigma_{(k=1,n)} (u_{ik})^m,$$
for i=1,2,...c

$$u_{ik} = \forall j \begin{cases} \left(\Sigma_{(j=1,c)} \left(\frac{|| x_k - v_i ||}{|| x_k - vj ||} \right)^{2/m-1} \right)^{-1} & \text{if } ||xk - vj||>0, \\ 1 & \text{if } || xk- vj||=0 \\ m & \text{if } \exists j \neq || x_k - vj||=0 \end{cases}$$

Where

$x=\{x1, x2\ldots xn\} \subset R^p$ – are the feature data vectors

$v=\{v1,v2,\ldots\ldots..vc\} \subset R^p$ – are the cluster centres

$u= (u_{ik})_{cxn}$ – is the fuzzy partition matrix that contains the membership values of each feature vector x_k in each cluster i.

n-total number of data vectors in a given data set.
c-number of cluster
m-fuzzification parameter-fuzzifier, m>1.

Consider previous example, we can observe that Fuzzy C means clustering gives much better result on the point (4,3.5), since it assigns an equal membership degree to each of the clusters. This result given in the table 4.

TABLE 4

THE MEMBERSHIP FUNCTIONS FOR FUZZY CLUSTERING

Dataset 1	Dataset 2	Fuzzy C means clustering for A1	Fuzzy C means clustering for A2
1.0	2.5	0.0549	0.9451
1.0	3.5	0.0228	0.9772
1.0	4.5	0.0549	0.9451
2.0	3.0	0.0160	0.9840
2.0	3.5	0.0024	0.9976
2.0	4.0	0.0160	0.9840
3.0	3.5	0.1240	0.8760
4.0	3.5	0.5004	0.4996
5.0	3.5	0.8765	0.1235
6.0	3.0	0.9840	0.0160
6.0	3.5	0.9977	0.0023
6.0	4.0	0.9840	0.0160
7.0	2.5	0.9450	0.0550
7.0	3.5	0.9771	0.0229
7.0	4.5	0.9450	0.0550

8 FINDINGS

8.1 Graduality

According to Fuzzy set theory, patterns of data interesting to data mining are often vague and its boundaries are nonsharp. Fuzzy set theory shows a gradual transition between sharp and nonsharp boundaries. It represents gradual concepts and fuzzy attributes in an explicit mode. This feature is useful in the context of Data Mining.

Consider second sample application and data point (4, 3.5). Firstly Crisp clustering method applied on the data set. Observe results obtained after clustering process from Fig. 5. Since both clusters have equal chance to include the point (4, 3.5), by Crisp method it assigned the point to only one cluster with membership value 1. Regarding membership value certain uncertainties exist on the point (4, 3.5). The result obtained is not perfect. Reason for imperfect result is Crisp clustering method provides only two membership values either 1 or 0. This method classify data item to only one cluster.

Next come to Fuzzy C means clustering and verify result obtained from table 4. It assigns an equal membership value (0.5) to both clusters on point (4, 3.5). The point lies in both clusters and this gives a better result than Crisp clustering method.

8.2 Fuzzy Feature Extraction and Pattern Representation

Many data mining methods proceed from a representation of the entities with fixed number of features and attributes. These attributes represents properties of an entity. Objective of Feature based method is analyze relationships and dependencies between entity's attributes. Fuzzy methods represent graded properties in an adequate way, is useful for both feature extraction and dependency analysis.

In first sample application we started association analysis task by assigning a membership value lies between 0 and 1 to all symptoms of each patient. This convert data set to a Fuzzy data set. Next phase is to apply apriori algorithm and find fuzzy large itemsets for cadinality one, two, three, etc. Before starting the process we fixed large support for itemset as 0.38. Iteration process stoped in large 2-itemsets and one itemset selected {BE}. Application also defined a compatibility function for determining disease of the patient.

This application eliminates uncertainties from the symptoms membership value by assigning a value between 0 and 1. We can confirm that final result is better than association analysis using conventional method.

9 CONCLUSION AND FUTURE WORK

In this paper, we discussed role of fuzzy set in association analysis and Clustering using sample applications. Association Analysis sample application signify Compatibility is one of the parameter which determining presence of disease in patients by looking at symptoms. We generated suitable function similar to Compatibility and these functions are central in studying uncertainty in data and in making inference when information is vague and/or accurate.

Clustering example proposed a fuzzy C means mining algorithm for processing transaction data with quantitative values and discovering interesting patterns among them. Compared to conventional crisp-set-mining methods for quantitative data, this approach gets smoother mining results due to its fuzzy membership characteristics. And, when compared to fuzzy-mining methods, which take all fuzzy regions into consideration, this method achieves better time complexity since only the most important fuzzy term is used for each item.

In future we would like to study limitations of fuzzy sets in data processing and also wish to propose new mathematical models which produce better results in the domain of data mining.

REFERENCES

[1] Ping Chen, A. de Korvin, Chenyi Hu, "Association Analysis with Interval Valued Fuzzy sets and Body of Evidence", *IEEE International conference on Fuzzy Systems Proceedings*; 1; 518-523, 2002

[2] Shivangi Sengar, Sakshipriya, Urvashi Khatuja ,"Fuzzy Set Approaches To Data Mining of Association Rule", *International Journal of Computer Science and Informatics* (IJCSI) ISSN (PRINT): 2231 –5292, Vol-II,Iss-4,2012

[3] Qiankun Zhao, Sourav S. Bhowmick, "Association Rule Mining: A Survey", Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003

[4] Sandeep Kumar Singh, Ganesh Wayal, Niresh sharma, "A Review: Data Mining with Fuzzy Association Rule Mining", *International Journal of Engineering Research & Technology (IJERT)*, ISSN: 2278-0181 Vol. 1 Issue 5, July – 2012

[5] Chau, M., Cheng, R., and Kao, B., "Uncertain Data Mining: A New Research Direction," *in Proceedings of the Workshop on the Sciences of the Artificial*, Hualien, Taiwan, December 7-8, 2005

[6] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers

ISSN 2229-5518

[7] Guanrong Chen, Trung Tat Pham,"Introduction to Fuzzy Sets, Fuzzy Logic, and Fuzzy Control Systems", CRC Press LLC

[8] W. Pedrycz and F. Gomide, "An Introduction to Fuzzy Sets Analysis and Design", MIT Press, 1998

[9] D. Barbara, P. Chen, "Self-similar Mining of Association Rules over Interval Data", Technical report, ISE Dept., George Manson University, 2001

[10] T.P. Hong, C.S. Kuo, S.C. Chi, "A Data Mining Algorithm for TransactionDdata with Quantitative Values", Intell. Data Anal. 3 (5) (1999) 363-376.

[11] R. Srikant, R. Agrawal, "Mining Generalized Association Rules", *The Internat. Conf. on Very Large Databases*, 1995.

[12] Tzung-pei Hong, Kuei-Ying Lin, Shyue-Liang Wang, "Fuzzy Data Mining for Interesting Generalizes Association Rules", Elsevier 2002.

[13] Sadina Gagula-Palalic, "Fuzzy Clustering Models and Algorithms for Pattern Recognition", thesis for the degree of Master of Science in mechatronics and systems engineering 2008.

[14] H. Guldemır, A. Sengur, "Comparison of Clustering Algorithms for Analog Modulation Classification", Expert Systems with Applications, 30, 2006, 642–649