Generating Membership Values And Fuzzy Association Rules From Numerical Data

Dr.R.Radha
S.D.N.B.Vaishnav College For Women,
Chromepet, Chennai, Tamil Nadu, India

Dr.S.P.Rajagopalan
Emirites, Professor, School of Computer Science and Engineering, M.G.R. University, Chennai.

Abstract*: The most important task in the design of fuzzy classification systems is to find a set of fuzzy rules from training data to deal with a specific classification problem. In this paper, a method to generate fuzzy rules from training data to deal with the data classification problem is presented. Partition method of interval is adopted in current classification based on associations (CBA). But this method cannot reflect the actual distribution of data and there exists the problem of sharp boundary. These type of problems can be approached with fuzzy representation of data. In this paper quantitative attributes are partitioned into several fuzzy sets by fuzzy C-Means algorithm and membership values are generated, and supervised association rule algorithm is used to discover interesting fuzzy association rules, which are used to build classification system. In this paper fuzzy classified association rules are generated and three classifiers namely C4.5, Naïvebayes , and ID3 are used for classification. Experiments are conducted on both primary and secondary data and accuracy of each of the classifiers are discussed with AUC-ROC curves. Quantitative values in databases generate very large number of rules. Using fuzzy linguistic values the generation of rules can be reduced and an objective measure is used further to filter the generated rules and present only the interesting rules*.

Keywords*: classification, quantitative attributes, Naivebayes, C4.5, ID3, fuzzy C-Means, fuzzy association rules, Supervised assoc rule*

I INTRODUCTION

A. Association Rule

Association rule mining was first proposed to find all rules in a basket data (also called transaction data) to analyze how items purchased by customers in a shop are related (one data record per customer transaction). The model is described in [1]. Given a set of transactions *D* (instances) , the problem of mining association rules is to such as multi-leveled equations, graphics, and tables are not prescribed, although the various discover all rules that have support and confidence greater than the user-specified minimum support (called *minsup*) and minimum confidence (called *minconf*). Association rule mining does not have a fixed target. That is, any item can appear on the right-hand-side or the left-hand-side of a rule. One of the efficient algorithms for mining association rules is the Apriori algorithm given in [1]. It generates rules in two steps:

*1. Find all the frequent item sets that satisfy m*insup

2.Generate all the association rules that satisfy *minconf* using the frequent item sets.

B. Classification

Classification is an important data-mining task. The data set used in a typical classification task consists of the descriptions of *M* data cases. Each data case is described by *I* distinct attributes. The *M* cases are also pre-classified into *C* known classes. The objective of the classification task is to find a set of characteristic descriptions (e.g. classification rules) for the *C* classes. This set of descriptions is often called a predictive model or classifier which is used to classify future (or test) cases into the *C* classes.

However building a classification does not give accurate result. Because the training data used is typically very noisy and has a highly imbalanced (or skewed) class distribution. Mostly the user is interested in data cases of a minority class, which is even harder to predict. This is called target selection problem. This problem is due to single *minconf* and *minsup* to all the classes.

Traditional association rule mining uses only a single *minsup* in rule generation, which is inadequate for unbalanced class distribution. In this paper this problem is solved by assigning different *minsup* and *minconf* to different classes rather than assigning the single *minsup* and *minconf* values to all the classes.

If the minsup is set too high, we may not find those rules that involve the minority class, which is the class that we are interested in. In order to find rules that involve the minority class, we have to set the minsup very low. This may cause combinatorial explosion because the majority class may have too many rules and most of them are over fitted with many conditions and covering very few data cases. These rules have little predictive value. They also cause increased execution time.

While a single *minsup* is inadequate for our application, a single *minconf* also causes problems. For example, in a database, it is known that only 5% of the people are buyers and 95% are non-buyers. If we set the *minconf* at 96%, we may not be able to find any rule of the buyer class because it is unlikely that the database contains reliable rules of the buyer class with such a high confidence. If we set a lower confidence, say 50%, we will find many rules that have the confidence between 50-95% for the non-buyer class and such rules are meaningless

In order to solve this problem we have adopted different *minconf* and *minsup* for rules of different classes. For minimum supports, the user has to give *minsup* called *u_minsup* which is then distributed to all the classes according to class distribution in the data as follows:

$$minsup(C_i) = u\_minsup \times \frac{f(C_i)}{|D|} \quad (1)$$

where $f(C_i)$ is the number of C_i class cases in the training data. $|D|$ is the total number of cases in the training data. The reason for using this formula is to give rules with the frequent (negative) class a higher *minsup* and rules with the infrequent (positive) class a lower *minsup*. This ensures that we will generate enough rules with the positive class and will not produce too many meaningless rules for the negative class. Here positive class means the less frequent rules, which will be more surprising. The negative class means the frequent rules occurring i.e. general case, which are not much interesting.

For minimum confidence, we use the following formula(2) to automatically assign minimum confidence to each class:

$$minconf(C_i) = \frac{f(C_i)}{|D|} \quad (2)$$

C. Class Association Rule

The use of association rules for classification is restricted to problems where the instances can only belong to a discrete number of classes. The reason is that association rule mining is only possible for nominal attributes. However, association rules in their general form cannot be used directly. The head Y of an arbitrary association rule $X \Rightarrow Y$ is a disjunction of items. Every item, which is not present in the rule body, may occur in the head of the rule.

When we want to use rules for classification, we are interested in rules that are capable of assigning a class membership. Therefore the head Y of a class association rule $X \Rightarrow Y$ is restricted to one item. The attribute of this attribute-value-pair has to be the class attribute. According to this, a class association rule is of the form $X \Rightarrow Y(a_i)$ where a_i is the class attribute and $x \subseteq \{a_1,...,a_{i-1},a_{i+1},...a_n\}$. In this paper the adapted version of Supervised association rule is used to generate classified association rules.

D. Classified Fuzzy Association Rule

When dealing with quantitative attributes, their domains are usually divided into equal-width or equal-frequency intervals called discretization[13]. But it introduces some problems. The first problem is that equiwidth partitioning cannot embody the actual distribution of the data. The second problem is caused by the sharp boundary. In most cases, the resulting intervals are not too meaningful and are hard to understand. Ref. [12] uses fuzzy set to soften partition boundary of the domains, and presents the concept of fuzzy association rules, but it does not present partition algorithm that can embody the actual distribution of the data and does not present the mining algorithm for fuzzy association rules which fits for large databases. Ref. [14] uses the relational fuzzy C-Means algorithm to partition the quantitative attributes into several linguistic terms, then the problem of mining association rules with linguistic terms is introduced by combining linguistic terms. The relational fuzzy C-Means algorithm can embody the actual distribution of the data. Furthermore, linguistic terms can soften partition boundary.

One of important applications of fuzzy set theory [31] is in the fuzzy classification systems. In this paper, a method to generate fuzzy rules from a set of training data for classification problem is presented. The proposed method is an extension of the fuzzy rule generation method presented in [42], but it generates less fuzzy rules and can get a higher average classification accuracy rate. Furthermore, a measure already discussed in paper [42] is used to assess the given fuzzy rules and are ranked. There are two issues that must be addressed in the classification system of fuzzy association rules. The first is that a huge number of rules could contain noisy information. The second is that a huge set of rules would extend the classification time. This could be an important problem in applications where fast responses are required. So fuzzy association rules should be pruned. For pruning hybrid objective measures can be used.

In order to mine fuzzy class association rules, a new database is build through original database T. First, the training data is converted into fuzzy attributes by finding membership values using fuzzy C-Means algorithm, and then these attributes are used to generate fuzzy rules. In this new database, the set of all fuzzy attributes are denoted I, the value of the j-th record in fuzzy attribute k $_y$ is still denoted $t_i(y_k)$. It is obvious that $t_i(y_k)$ falls in [0,1]. Let

$$X = \{y_1, y_2, ..., y_p\} \subset I \quad (3)$$

$$Y = \{y_{p+1}, y_{p+2}, ..., y_{p+q}\} \subset I \quad (4)$$

$$X \cap Y = \varphi \quad (5)$$

An association rule is an implication of the form $X \Rightarrow Y$. Because attributes in X and Y are fuzzy attributes, $X \Rightarrow Y$ is called fuzzy association rule. The support and confidence of fuzzy association rule are defined as follows [31].

Definition1. The support of X is defined as follows(6).

$$Sup(X) = \frac{\sum_{j=1}^{n}\prod_{m=1}^{p} t_j(y_m)}{n} \quad (6)$$

Fuzzy attribute sets with at least a minimum support are called frequent fuzzy attribute sets.

Definition 2. The support of $X \Rightarrow Y$ is defined as follows(7).

$$Sup(X) = \frac{\sum_{j=1}^{n}\prod_{m=1}^{p+q} t_j(y_m)}{n} \quad (7)$$

Definition 3. The confidence of $X \Rightarrow Y$. is defined as follows(8):

$$conf = \frac{sup}{sup(X)} \quad (8)$$

Because $t_i(y_k)$ falls in [0,1], all subsets of a frequent fuzzy attribute set must also be frequent according to definition 1. With the above finding, the input can be directly modified for supervised association rule algorithm to mine fuzzy association rules.

Supervised association rule algorithm is applied on this data to discover interesting fuzzy association rules, which are used to build classification system. A user interface is created to get the data in its original form. In the fuzzy inference process of the program the data are converted into fuzzy sets, which in turn are converted into domain, defined linguistic terms. This converted data base is given as input parameters to Naïve bays, C4.5 and ID3 classifiers which is then used to generate classified fuzzy association rules. Different measures like the confidence, support, PS, IS and SQSPR values are calculated for the discovered rules and the results are displayed.

In this paper, in section I, methods of association rules, class association rules, classification and fuzzy classified association rules are explored. In section II, some of the related work are discussed. In section III the concept of Fuzzy C-Means algorithm is explained. In section IV the algorithm used for deriving the fuzzy class association rule is given. In Section V, the problem under taken for experimentation is discussed. Section VI discuses the experimental results of the specific domains. Section VII makes a concise conclusion of the study and points out future work in this area.

II. RELATED WORK

A. Association Rule

Traditional association rule mining algorithms can only be applied to data mining problems with categorical features. For a data mining problem with quantitative features, it is necessary to transform each quantitative feature into discrete intervals. Many discretization algorithms have been proposed for this purpose. Kamber et al proposed one such algorithm to mine multidimensional association rules using statistically discretization of quantitative features and data cubes based on predefined concept hierarchies [32]. The ARCS [34] algorithm mines quantitative association rules by dynamically discretizing quantitative attributes based on binding, where "adjacent" association rules may be combined by clustering. Techniques for mining quantitative rules based on x-monotone and rectilinear regions were presented by Fukuda et al [23], and Yoda et al. [50]. A non-grid-based technique for mining quantitative association rules, which uses a measure of partial completeness, has been proposed by Srikant and Agrawal [44]. The distance-based association rule mining algorithm [39] can mine distance-based association rules to capture the semantics of interval data, where intervals are defined by clustering.

B. Classified Association Rule

The problem of classification is concerned with the mining of a set of production rules that can allow the values of an attribute in a database to be accurately predicted based on those of other attributes [38]. Association rules can be used for classification (CBA)[38].

Some research works have been carried out to utilize "crisp" association rules for classification. In 1997, Lent et al., proposed a method, Association Rule Clustering System, or ARCS, to mine association rules based on clustering and then employ the rules for classification [35]. The classification by aggregating emerging patterns, called CAEP, is proposed by Dong et al [19]. Association based decision tree [47], called ADT, is a different classification algorithm based on association rules, combined with decision tree pruning techniques. Baralis et al [5] proposed "Live and Let Live" (L3), for associative classification. Liu et al proposed a framework, named associative classification, to integrate association rule mining and classification [38]. Li et al proposed an algorithm "Classification based on Multiple Association Rules" (CMAR), which utilizes multiple class-association rules for accurate and efficient classification [36]. This method extends an efficient mining algorithm, FP-growth [25], constructs a class distribution- associated FP-trees, and predicts the unseen sample within multiple rules, using weighted χ^2. Liu and Li's [37] approaches generate the complete set of association rules as the first step, and then select a small set of high quality rules for prediction.

Yin et al proposed, "Classification based on Predictive Association Rules" (CPAR) [49], which combines the advantages of both associative classification and traditional rule-based classification. Using association rules for classification helps to solve the understandability problem [17,41] in classification rule mining.

C. Fuzzy Classified Association Rule

All the above methods have the disadvantage that they involve crisp cutoffs for quantitative features. Fuzzy logic can

be introduced into the system to allow "fuzzy" thresholds or boundaries to be defined. Fuzzy logic is demonstrated to be a superior mechanism to enhance interpretability of these discrete intervals. Many fuzzy association rule mining algorithms have been proposed in recent research works.

Lee et al., [34] uses a membership threshold to transform fuzzy transactions into crisp ones before looking for binary association rules in the set of crisp transactions. This algorithm can diminish the granularity of quantitative features. Chan et al introduced F-APACS to employ linguistic terms for representing the reveal regularities and exceptions for mining fuzzy association rules [8]. In [3, 4 and 9], Au et al also proposed a series of algorithms to employ a set of predefined linguistic labels using adjusted difference and weight of evidence to measure the importance and accuracy of fuzzy association rules. These two measures can avoid the need for a user to provide importance thresholds, but has the drawback of making symmetric the adjusted difference and thus, when a rule $C \Rightarrow A$ is found to be interesting, then $A \Rightarrow C$ will be too.

Kaya et al. [33] proposed a clustering method that employs multi-objective Genetic Algorithm for the automatic discovery of membership functions used in determining fuzzy quantitative association rules. Chen et al [16, 15] have considered the case in which there are certain fuzzy taxonomic structures reflecting partial belonging of one item to another in the hierarchy. To deal with these situations, association rules are required to be of the form $X \Rightarrow Y$ where either X or Y is a collection of fuzzy sets.

Delgado et al define "fuzzy transactions", which can be applied to quantitative features. They also propose an algorithm to mine "fuzzy association rules" based on these "fuzzy transactions" [18]. Frosini, G et al, [21] proposed a novel method for feature selection based on a modified fuzzy C-Means algorithm with supervision (MFCMS). Hanif D. Sherali et al.,[27] present a global optimization algorithm to solve the fuzzy clustering problem, where each data point is to be assigned to (possibly) several clusters, with a membership grade assigned to each data point that reflects the likelihood of the data point belonging to that cluster. Srinivasa K G et al., [45] use fuzzy C-Means to cluster the data points of the feature vectors for feature extraction. Xiao Ying Wang et al., in his paper [48] used two data clustering techniques, Hierarchical Cluster Analysis (HCA) and Fuzzy C-Means (FCM) clustering, to classify sets of oral cancer cell data without a pre-processing procedure. The performances of these two techniques are compared and their differences are discussed. The FCM method was found to perform significantly better. Ta-Wei Hung [46] proposes a new fuzzy clustering-based approach to fuzzy system identification based on the bi-objective fuzzy C-Means (BOFCM) cluster analysis. Akara Sopharak et al., [2] the authors investigated and proposed automatic methods of exudates detection on low-contrast images taken from non-dilated pupils. The process has two main segmentation steps, which are coarse segmentation using Fuzzy C-Means clustering and fine segmentation using morphological reconstruction.

In recent years, many research works have been conducted for fuzzy association rules mining. However, to our best knowledge, there are very few works focusing on fuzzy association rule mining on supervised classification problems. Hu et al proposed to extract "fuzzy associative classification rules" in "fuzzy grids" that are generated by fuzzy partitioning on each input feature [27]. Chatterjee et al propose a fuzzy pattern classifier named Influential Rule Search Scheme (IRSS) [10]. Nauck[40] has developed a learning algorithm that creates mixed fuzzy rules involving both categorical and numeric attributes.

III. FCM ALGORITHM (FUZZY C-MEANS)

The FCM algorithm, also known as Fuzzy ISODATA, is one of the most frequently used methods in pattern recognition. It is based on minimization of the objective function (9) to achieve good classifications.

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{ij})^{m}\left\|x_i - v_j\right\|^{2} \qquad (9)$$

J(U,V) is a squared error clustering criterion, and solutions of minimization of (9) are least-squared error stationary points of *J(U,V)*. The expression, $X=\{x_1, x_2, \ldots x_n\}$ is a collection of data, where n is the number of data points. $V=\{v_1,v_2,\ldots v_c\}$ is a set of corresponding cluster centers in the data set *X*, where *c* is the number of clusters. μ_{ij} is the membership degree of data x_i to the cluster center v_j . Meanwhile μ_{ij} has to satisfy the following conditions:

$$\mu_{ij} \in [0,1], \quad \forall_i = 1,..n, \forall_j = 1,..c \qquad (10)$$

$$\sum_{j=1}^{c}\mu_{ij}, \qquad \forall_i = 1,..n \qquad (11)$$

where $U = (\mu_{ij})_{n*c}$ is a fuzzy partition matrix, $\left\|x_i - v_j\right\|$ represents the Euclidean distance between x_i and v_j , parameter *m* is "fuzziness index" and is used to control the fuzziness of membership of each datum in the range $m \in [1,\alpha]$. In this experimentation the value of m=2.0 was chosen. Although there is no theoretical basis for the optimal selection of m, this has been chosen because the value has been commonly applied within the literature. The following steps can perform the FCM algorithm:

1) Initialize the cluster centers $V=\{v_1,v_2,\ldots v_c\}$, or initialize the membership matrix μ_{ij} with random value and make sure it satisfies conditions (10) and (11) and then calculate the centers.
2) Calculate the fuzzy membership μ_{ij} using(12)

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}} \quad (12)$$

where $d_{ij} = \left\| x_i - v_j \right\|, \forall_i = 1,...n, \forall_j = 1,...c$

3) Compute the fuzzy centers v_j using (13)

$$v_j = \frac{\sum_{i=1}^{n} (\mu_{ij})^m x_i}{\sum_{i=1}^{n} (\mu_{ij})^m}, \ \forall_i = 1,...c \quad (13)$$

4) Repeat steps 2) and 3) until the minimum J value is achieved.
5) Finally, defuzzification is necessary to assign each data point to a specific cluster (i.e. by setting a data point to a cluster for which the degree of the membership is maximal).

IV. ALGORITHM

In the previous work [42] we divide the domain of each input variable into several overlapping length intervals and each interval is associated with a fuzzy triangular membership function. Then, we assign each membership function a label to represent it. If a value x_j in the membership function associated with label L_j has the largest membership value, then the value x_j is converted into L_j. In this paper we have used fuzzy C-Means algorithm to generate the membership values instead of using triangular functions to generate membership values.

Let '*D*' be a database and '*d*' represents the records in it. Let $I = \{ I_1, I_2, ..I_n \}$ be the attributes in it. The numerical attributes under consideration, for example I_j is further divided into '*k*' number of classes such that *k*=*c*, where '*c'* is the number of clusters formed by using Fuzzy C-Means algorithm.

Fuzzy Classified Association algorithm process

1. Randomly initialize the number of clusters to '*c*' where $2 \le c \le \sqrt{n}$ where '*n*' is the number of records present in the database *D*.
2. Initialize fuzzy membership μ_{ij} , subject to conditions eq(10) and eq(11) being satisfied.
3. Calculate data centers using eq(13).
4. Find the degree of membership for the attribute I_j where *j* values ranges from $1 < j <= c$
5. Find the maximum membership value for the attribute I_j from step(4) and assign appropriate linguistic label to it. Now these attributes can be called as fuzzy attributes.
6. for *i* =1 to *n do*
   use α - cut value as the users threshold to decide the destination class
   end for
7. The above-generated conditional (class) attribute can now be called as target class or output class and all the other fuzzy attributes are called as input values.
8. These fuzzy attributes are given as input parameters to classifiers (Naïve Bays, ID3, C4.5)
9. The classified attributes by step (8) are given as input parameters to supervised association algorithm giving individual *minsup* and *minconf* values by equation (1) and (2) with user threshold *u_minsup*=0.02.
10. For all generated rules do
    a) calculate SQSPR values given in eq(14) in section V
    b) fix a threshold and filter the rules above or equal to that value
    c) add these rules to a list of interesting rules.

    End for

A system has been implemented to perform the post analysis of the generated association rules. Past research on inductive learning has mostly been focused on techniques for generating concepts or rules from datasets (e.g.,[11],[20],[30]). Limited research has been done on what happens after a set of rules has been induced. It is assumed that these rules will be used directly by an expert system or some human user to infer solutions for specific problems within a given domain. Having obtained a set of rules is not the end of the story, post-analysis of rules has to be done. The motivation for performing post-analysis of the rules comes from realizing the fact that using a learning technique on a dataset does not mean that the user knows nothing at all about the domain and the dataset. This is particularly true if the user is a human being. Typically, the human user does have some pre-conceived notions or knowledge about the learning domain.

V. PROBLEM DISCUSSION

Experiments are conducted on two datasets namely veterinary dataset (primary data) from TamilNadu Veterinary and Animal Sciences University (TVASU), Chennai, India and liver disorder dataset (secondary data) from UCI ML Repository database.

A. Liver Disorder Data

The liver disorder data has been taken from the UCI ML Repository Database. This dataset consists of 345 records with 6 variables in each record. The 6 variables are Mean Corpuscular Volume(MCV), ALkaline Phosphatase (ALP) ,ALanine Transaminase (ALT),ASpartate Transaminase(AST) , Gamma Glutamyl Transpeptidase (GGT).

1) Liver Function Tests

This include liver enzyme, which are groups of clinical biochemistry laboratory blood assays designed to give information about the state of a patient's liver. Most liver diseases cause only mild symptoms initially, while it is vital that these diseases be detected early. Hepatic involvement in some diseases can be of crucial importance. This testing is performed by a Medical technologist on a patient's serum or plasma which is collected by a phlebotomist. The data obtained are in numerical form. These data are converted into fuzzy linguistic labels by finding the membership values. The membership values are found using the fuzzy C-Means algorithm

2) Veterinary Data

This real data set consists of 9500 records and from these, 6572 records are selected as others are left because of null values occurrence. The data consists of 10 attributes namely identification number, owner name, address, species name, breed name, age, sex, examinationdate, diagnosis and special procedures. After collecting the data, the data was converted into usable format. The age attribute was given as input for fuzzy C-Means clustering algorithm and the data was converted into fuzzy values by forming 5 clusters and was categorized as *veryyoung*, *young*, *middle*, *old* and *veryold* animals. The categorized attributes are given as input to the classifiers, and then using the supervised association rule generator, fuzzy classified association rules are generated. In order to evaluate the performance of classifiers, 10-fold cross-validation is used. The error rates of the classifiers are given in the table. The minimum error rate shows that the classifiers behave optimistically in classifying the fuzzy data.

Finally the ROC curves are drawn with False Positive Rate(FPR) in X-axis and True Positive Rate(TPR) in Y-axis for each classification. The results are shown for individual classifiers. The AUC-ROC values of all the datasets are given in the tables 15, 16 and 17.

VI. RESULTS AND DISCUSSION

In a real-world application of supervised learning, we have a training set of examples with labels, and a test set of examples with unknown labels. The whole point is to make predictions for the test examples. However, in research or experimentation we want to measure the performance achieved by a learning algorithm. To do this a test set consisting of examples with known labels are used. The classifier is trained on the training set and is applied on the test set. The performance is measured by comparing the predicted labels with the true labels. A common rule of thumb is to use 70% of the database for training and 30% for testing. In this work also the same is followed. The performance of the classifier is measured through cross-validation

The performances of the three classifier are expressed with Area Under the Receiver Operator Characteristic (AUC-ROC) curves, which show how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples. The AUC value can indicate a model's generalization capability as a function of varying a classification threshold. An AUC value of 1 represents a perfect classification, while an AUC value of 0.5 represents a worthless model.The result of three classifier namely C4.5, Naïve Bayes and ID3 are compared

A. Metric Used For Rule Interestingness

The generated rules are filtered based on their interestingness using the objective measures SQSPR which is given in (14) below

$$\text{SQSPR}= \sqrt{\frac{P(A \cap B)}{P(A)} - \frac{P(A \cap B)}{P(B)}} \quad (14)$$

SQSPR= 0 , If $\frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{P(B)}$

This shows that both antecedent and consequent are having equal contribution in framing the rule and are independent of each other and rules of these types can be eliminated.

SQSPR increases if $\frac{P(A \cap B)}{P(A)} \rangle \frac{P(A \cap B)}{P(B)}$ so the antecedent plays a major role in deciding the consequent

Even this also contains uninteresting rules. From this interesting rules can be filtered using SQSPR [42]. Table 1 gives the comparative result of using other techniques and fuzzy classified rule generation.Among the filtered rules few are presented in table 2, 3 and 4 . The filtered sample result shown in the tables gives an idea about the factors to be considered for taking any decisions. These measures show the strength and the interestingness of the discovered rule.

We filtered the interesting rules as those, which are having high values for SQSPR. By having a look at these rules one can identify the combination of enzymes level that are responsible for the status of the patient

SQSPR decreases if $\frac{P(A \cap B)}{P(A)} \langle \frac{P(A \cap B)}{P(B)}$ This shows the negative correlation.

TABLE 1. COMPARATIVE RESULT IN LIVER DATA SET

Methods Used	Number Of Rules Generated
Apriori Mr	1170
Apriori Pt	2140
Fuzzy classified Assoc Rule using triangular membership generation	370
Fuzzy classified Assoc Rule using fuzzy C-Means	133

TABLE 2 CLASSIFIED ASSOC RULES FOR PATIENT CONDITION TO BE NORMAL

Rule no	Antecedent	conseq	conf	SQSPR
115	"fuzmcv=low" - "fuzast=low" - "fuzratio=high"	normal	1	0.98039
125	"fuzmcv=low" - "fuzalt=medium" - "fuzast=low"	normal	1	0.98039
92	"fuzmcv=medium"- "fuzggt=medium" - "fuzdr=medium"	normal	1	0.977911

TABLE 3. CLASSIFIED ASSOC RULES FOR PATIENT CONDITION TO BE CRITICAL

Rule no	Antecedent	conseq	conf	SQSPR
11	"fuzmcv=medium" - "fuzalk=low"	critical	1	0.964724
12	"fuzalk=low" - "fuzast=low"	critical	1	0.964724
82	"fuzalk=medium" - "fuzalt=high"- "fuzast=medium"	critical	1	0.964724

TABLE 4. CLASSIFIED ASSOC RULES FOR PATIENT CONDITION TO BE SERIOUS

Rule no	Antecedent	conseq	conf	SQSPR
15	"fuzdr=extreme" - "fuzratio=high"	serious	1	0.973329
16	"fuzast=high" - "fuzratio=extreme"	serious	1	0.973329
18	"fuzmcv=low" - "fuzast=extreme"	serious	1	0.973329
24	"fuzalk=low" - "fuzast=extreme"	serious	1	0.973329

B. Liver Data Set

Even if we use fuzzy sets around 2141 rules are generated using APRIORI PT(Christian Borgelt's APRIORI algorithm using prefix tree) algorithm. If they are classified using classifiers (C4.5, NaiveBayes and ID3) techniques and then using Supervised association rule algorithm the number of generated rules are narrowed down and generated based on specific target. Nearly 370 fuzzified association rules are generated in using fuzzy triangular membership values. Generating the fuzzy membership values using the fuzzy C-Means algorithm and then assigning the linguistic labels for them respectively and then using them for generating association rules generate only 133 rules only.

C. Veterinary Data

As described for the liver data set here also fuzzy classified association rules are generated. Table 5 and 6 shows the sample agewise and areawise analysis of the veterinary dataset. Table 5 gives an idea that *veryyoung* , *young and old* ones are the most frequent visitors of the hospital for treatment. The diagnosis part shows for what purpose they are brought to the hospital. Table 6 shows that most of the animals visiting the hospital are from northchennai. So if the diagnosis part is analyzed the doctors can see the cause and can arrange medical camps in these places and can take remedial measures for this area animals.

In table 7 the calculated individual *minsup* and *minconf* values, which are used for mining of liver data set, is given. In table 8 and 9 the calculated individual *minsup* and *minconf* values, which are used for mining of veterinary data set for agewise analysis and areawise analysis respectively are given. Table 10 shows the number of rows generated without classification, and with a common minimum support and minimum confidence.

Table 11 shows the number of classifier fuzzy association rules generated for liver disorder data with the filtering condition. Table 12 and table 13 shows the number of classified fuzzy association rules generated for veterinary data agewise and areawise respectively with the filtering condition. In this three tables (table 11, table 12, table 13)
Column1 – classifier used

Column2 – number of rules generated initially.

Column3 – represents if the SQSPR value is 0.90, then how many interesting rules can be filtered from column2

Column4 - represents if the SQSPR value is 0.98, then how many interesting rules can be filtered from column2

Column5 - represents if the SQSPR value is 0.99, then how many interesting rules can be filtered from column2

Table 14 a) and Table 14 b) discusses the performance of classifiers in terms of error rate and execution time respectively. One can find that among the three classifiers the Naïve Bayes performs very well in both execution time as well as have less error rate.

TABLE 5: FUZZY ASSOCIATION RULES FOR VETERINARY DATA – (AGE WISE)

Rule no	anteced	conseq	Supp	conf	SQSPR
19	"species=Canine" - "diagnosis= demodicosis"	Very young	0.01	1	0.9935
27	"sex=m" - "diagnosis= deworming"	Very young	0.01	1	0.9935
34	"species=Canine" - "diagnosis= limping"	Very young	0.01	1	0.9935
49	"sex=m" - "diagnosis= checkup"	Very young	0.01	1	0.9935
23	"breed=nd" - "diagnosis= wound"	Very young	0.01	1	0.9934

TABLE 6: FUZZY ASSOCIATION RULES FOR VETERINARY DATA – (AREA WISE)

Rule no	anteced	conseq	Supp	conf	SQSPR
119	"breed=spitz" - "fuzage= veryyoung" - diagnosis= dermatitis"	North chennai	0.0065	1	0.995
18	"breed=nd" - "diagnosis= deworming"	North chennai	0.0068	1	0.994
106	"breed=nd" – "fuzage= young" - "diagnosis =gastritis"	North chennai	0.007	1	0.994
116	"breed=spitz" - "fuzage= veryyoung" - "diagnosis= inappetance"	North chennai	0.007	1	0.994

TABLE 7 LIVER DISORDER DATASET – USER THRESHOLD = 0.02

Status of patient	No.of rows	*minsup*	*minconf*
Normal	165	0.0095	0.478
Critical	136	0.0078	0.39
Serious	44	0.0025	0.127

TABLE 8 VETERINARY DATASET-AGEWISE – USER THRESHOLD = 0.02

Age	No.of rows	*minsup*	*minconf*
Veryyoung	3178	0.00967	0.4835
Young	1612	0.00490566	0.245283
middle	899	0.002734	0.1367
Old	632	0.00192	0.096166
Veryold	251	0.0007638	0.038192

TABLE 9 VETERINARY DATASET- AREAWISE – USER THRESHOLD = 0.02

Area	No.of rows	*minsup*	*minconf*
Centralchennai	1943	0.0059	0.2957
Southchennai	908	0.0028	0.1382
Westchennai	1594	0.0049	0.2425
Northchennai	2127	0.0065	0.3236

TABLE 10 RULES WITHOUT CLASSIFICATION

	No.of rules	Common Sup	Common conf
Apriori(vetenary)	2224	0.003999	0.1999
Apriori (vetenary)	0	0.1	0.50
Apriori (liver data)	7064	0.025	0.12
Apriori (liver data)	1117	0.066	0.331

TABLE 11 CLASSIFIER FUZZY RULES GENERATED – LIVER DISORDER DATASET

Classifier	No.of Rules	SQSPR =0.90	(SQSPR =0.92)	(SQSPR =0.93)
C4.5	226	43	41	32
Naïve bayes	256	43	38	31
Id3	117	52	27	14

TABLE 12 CLASSIFIED FUZZY RULES GENERATED – VETERINARY AGEWISE

Classifier	No.of Rules	SQSPR =0.90	SQSPR =0.98	SQSPR = 0.99
C4.5	353	168	67	24
Naïve bays	356	215	72	25
Id3	346	239	109	38

TABLE 13 CLASSIFIED FUZZY RULES GENERATED – VETERINARY AREAWISE

Classifier	No.of Rules	SQSPR =0.90	SQSPR =0.98	SQSPR =0.99
C4.5	375	63	42	18
Naïve bayes	412	170	25	13
Id3	441	194	139	58

TABLE 14 a) COMPARISON OF CLASSIFIERS WITH ERROR RATE

Classifier	Liver disorder data (Error rate)	Veterinary agewise (Error rate)	Veterinary areawise (Error rate)
C4.5	0.31	0.21	0.35
Naïve Bayes	0.26	0.2	0.34
ID3	0.3145	0.2	0.35

TABLE 14 b) COMPARISON OF CLASSIFIERS WITH TIME OF EXECUTION

Classifier	Time (ms)	Time (ms)	Time (ms)
C4.5	42041	13092776	193067
Naïve Bayes	38986	8262	46216
ID3	35251	54608	47538

D. Result Analysis Through AUC-ROC Curves

The traditional academic point system to roughly guide the performance evaluation on the AUC (Area Under Curve) metric is given as follows:

Perfect prediction = 1.0
Excellent prediction = 0.9
Good prediction = 0.8
Mediocre prediction = 0.7
Poor prediction = 0.6
Random prediction = 0.5
Something wrong < 0.5

In this table 15, 16a),16b), 17 gives the AUC values of the classifiers ID3, C4.5 and Naïve Bayes for liver disorder dataset and veterinary data set agewise analysis and veterinary data set area wise analysis respectively. The high value of AUC shows that the classifier model generated predict very well as almost all the AUC values are between 0.98 to 1. This shows that fuzzy logic application to data mining applications enhances and gives more accurate results.

ROC curves for liver disorder dataset

FIGURE 1. ROC-CURVE FOR ALL CONDITIONS (ID3)

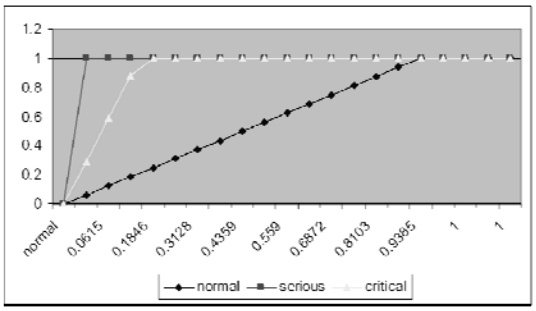

FIGURE 2. ROC- CURVE FOR ALL CONDITIONS (NAÏVE BAYS)

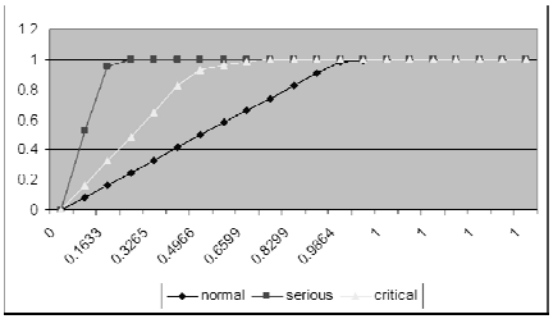

FIGURE 3. ROC-CURVE FOR ALL CONDITIONS (C4.5)

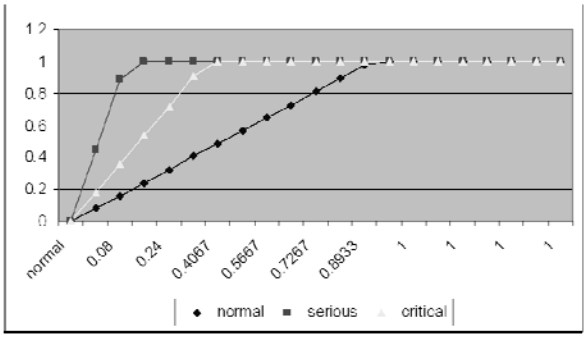

TABLE 15 AUC VALUES FOR LIVER DISORDER DATASET

Classifier	Normal	Serious	critical
Id3	1	0.9915	0.9979
C4.5	0.9989	0.9977	0.9985
Naïve bayes	0.9979	0.9966	0.9919

ROC-curves for veterinary dataset analyzed agewise

FIGURE 4. ROC-CURVE FOR ALL AGES (ID3)

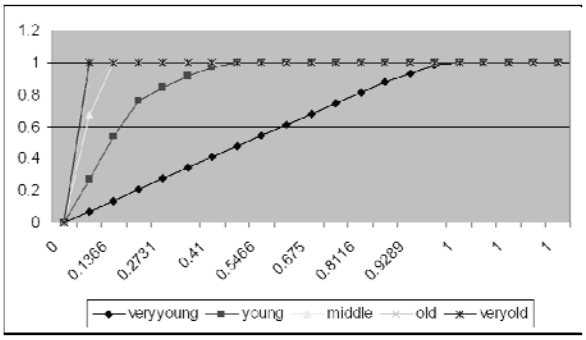

FIGURE 5. ROC-CURVE FOR ALL AGES (NAÏVE BAYS)

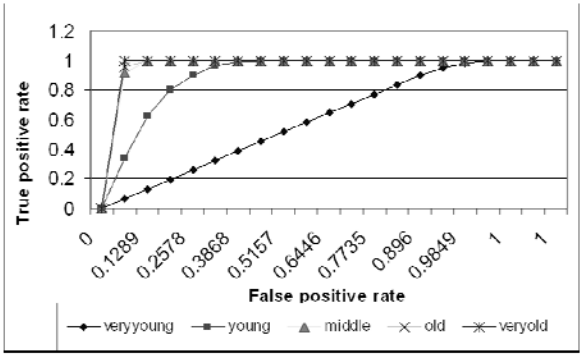

FIGURE 6. ROC-CURVE FOR ALL AGES (C4.5)

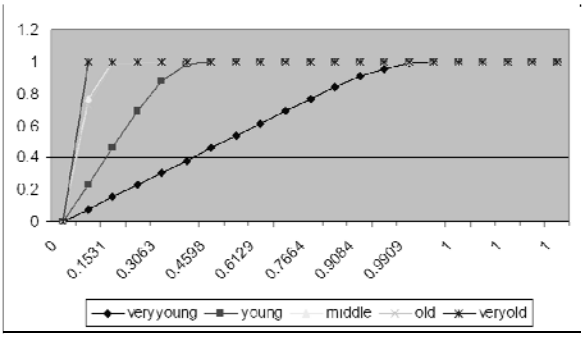

TABLE 16 a) AUC VALUES FOR VETERINARY DATASET ANALYZED AGEWISE

Classifier	veryyoung	Young
Id3	0.9834	0.9785
C4.5	0.9905	0.9946
Naïve bayes	0.9885	0.9751

TABLE 16 b) AUC VALUES FOR VETERINARY DATASET ANALYZED AGEWISE

Classifier	Middle	Old	veryold
Id3	0.9699	0.9947	0.9774
C4.5	0.9949	0.9934	0.9794
Naïve bayes	0.9881	0.9921	0.9769

ROC-curves for veterinary dataset analyzed areawise

FIGURE 7. ROC-CURVE FOR ALL AREAS (ID3)

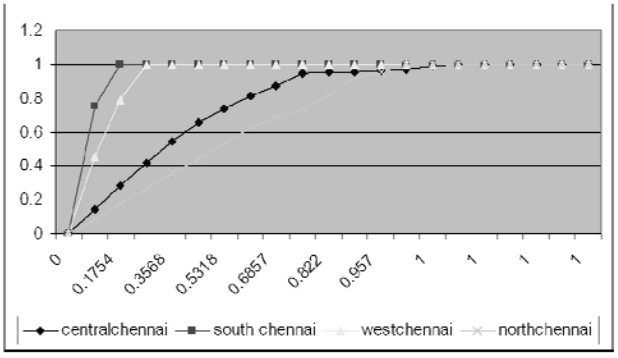

FIGURE 8. ROC-CURVE FOR ALL AREAS (NAÏVE BAYES)

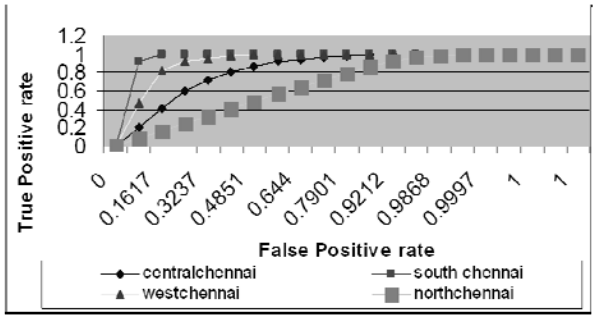

FIGURE 9. ROC-CURVE FOR ALL AREAS (C4.5)

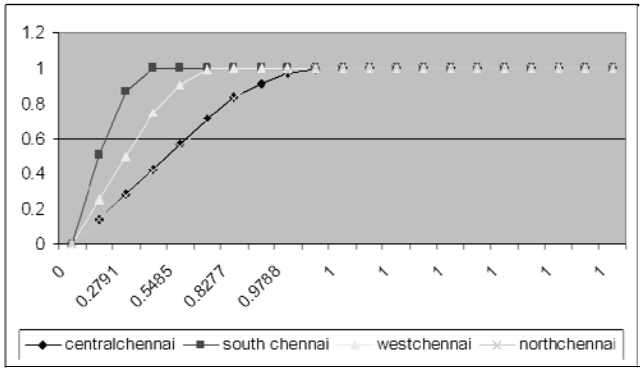

TABLE 17 a) AUC VALUES FOR VETERINARY DATASET ANALYZED AREAWISE

Classifier	Westchennai	northchennai
Id3	0.9803	0.9091
C4.5	0.9907	0.9932
Naïve bayes	0.9819	0.977

TABLE 17 B) AUC VALUES FOR VETERINARY DATASET ANALYZED AREAWISE

Classifier	Centralchennai	Southchennai
Id3	0.9442	0.9902
C4.5	0.9881	0.9904
Naïve bayes	0.9511	0.9916

References

[1] R.Agarwal, and R. Srikant. '*Fast Algorithms for Mining Association Rules*,' *In Proceedings of VLDB-94, 1994*

[2] Akara Sopharak, Bunyarit Uyyanonvara "*Automatic Exudates Detection From Diabetic Retinopathy Retinal Image Using Fuzzy C-Means And MorphologicalMethods"* Proceedings of the third IASTEDinternational conference advances in computer science and technology, april 2-4, 2007, Phuket, Thailand.

[3] W. H. Au and K. C. C. Chan. *'An effective algorithm for discovering fuzzy rules in relational databases,'* In Proc.7th IEEE Int. Conf. Fuzzy Systems, Anchorage, AK, 1998, pages 1314-1319.

[4] W. H. Au and K. C. C. Chan. '*FARM: a data mining system for discovering fuzzy association rules,'* In Proc. 8th IEEE Int. Conf. Fuzzy Systems, Seoul, Korea, 1999, pages 1217-1222.

[5] E.Baralis, P. Gazza, '*A lazy approach to pruning classification rules,*' Proc. IEEE Int. Conf. on Data Mining (ICDM'04), pages 35-42, 2002.

[6] P.Bosc, O. Dubois, O. Pivert, H. Prade, '*On fuzzy association rules based on fuzzy cardinalities,*' Proc. of IEEE Int. Conf. on fuzzy system, ages 461-464, 2001.

[7] P.Bosc, O. Pivert, '*On some fuzzy extensions of association rules,'* Proc. of joint 9th IFSA World Congress and 20th NAFIPS Int. Conf., pages 1104- 1109, 2001.

[8] K. C. C. Chan and W.-H. Au, '*Mining fuzzy association rules,*' Proc. Of 6th Int.Conf. Information Knowledge Management, pp. 209–215, Las Vegas, NV, 1997.

[9] K. C. C. Chan and W. H. Au. '*Mining fuzzy Association rules in database containing relational and transactional data*,'. In Data Mining and Computational Intelligence, A.Kandel, M. Last, and H. Bunke, Eds. NewYorks: Physica-Verlag, 2001, pages 95-114.

[10] A. Chatterjee and A. Rakshit, '*Influential Rule Search Scheme (IRSS)-A New Fuzzy Pattern Classifier*,' IEEE Transactions on Knowledge and Data Engineering, Volume 16, Issue. 8, pp. 881-893, August 2004

[11] Chen, Y. C. and Chen, S. M. 2000. '*A new method to generate fuzzy rules for fuzzy classification systems*,' Proceedings of the 2000 Eighth national Conference on Fuzzy Theory and Its applications, Taipei, Taiwan, Republic of China.

[12] Chen, S. M. and Chen, Y. C. 2002. '*Automatically constructing membership functions and generating fuzzy rules using genetic algorithms,'* Cybernetics and Systems: An International Journal, 33, 8: Yung-Chou Chen, Li-Hui Wang, and Shyi-Ming Chen 52 *Int. J. Appl. Sci. Eng., 2006.4, 1* 841-863.

[13] Chen, S. M., Kao, C. H., and Yu, C. H. 2002.'*Generating fuzzy uzzy rules from training data containing noise for handling classification*

problems,' Cybernetics and Systems: An International Journal, 33, 7: 723-749.

[14] Chen, S. M. and Yu, C. H. 2004. '*A new method for handling fuzzy classification problems using clustering techniques,*'. International Journal of applied Science and Engineering, 2, 1: 90-104.

[15] G.Chen and Q. Wei. '*Fuzzy association rules and the extended mining algorithms. Information Sciences,*' vol 147, pages 201-228, 2002.

[16] G.Chen, Q. Wei, and E. Kerre. '*Fuzzy data mining: Discovery of fuzzy generalized association rules*,' In Recent Issues on Fuzzy database, G. Bordogna and G. Pasi, Eds. Physica-Verlag, 2000. "Studies in Fuzziness and Soft Computing" Series.

[17] P. Clark and S. Matwin. '*Using qualitative models to guide induction learning,'* In Proc. 1993 Int. Conf. on Machine Learning (ICML'93), pages 49--56, Amherst, MA, 1993.

[18] M. Delgado, N. Marin, D. Sanchez and M. A. Vila, '*Fuzzy Association Rules: General Model and Applications,*' IEEE Transactions on Fuzzy Systems, Volume 11, Issue. 2, pp. 214-225, April 2003.

[19] G. Dong and J. Li, '*Efficient mining of emerging patterns:Discovering trends and differences,'* In Proc. 1999 Int.Conf. on Knowledge Discovery and Data Mining (KDD'99), pages 43-52, San Diego, CA, Aug. 1999.

[20] Fisher, R. 1936. '*The use of multiple measurements in taxonomic problems,'* Ann. Eugenics, 7: 179-188

[21] Frosini, G.; Lazzerini, B.; Marcelloni, F. '*A modified fuzzy C-Means algorithm for feature selection,'* Fuzzy Information Processing Society, 2000. NAFIPS. 19th International Conference of the North American Volume, Issue , 2000 Page(s):148 - 152.

[22] A.W.C. Fu, M.H. Wong, S.C. Sze, W.C. Wong, W.L. Wong, and W.K. Yu, '*Finding fuzzy sets for the mining of fuzzy association rules for numerical attributes*,' In Proc. Int. Symp. on Intelligent Data Engineering And Learning (Ideal'98), Hong Kong,1998, pages 263-268

[23] T. Fukuda, Y. Morimoto, S, Morishita, and T.Tokuyama. '*Data mining using twodimensional optimized association rules: Scheme, algorithms and visualization*,' In Proc. 1996 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'96), page 13-23, Montreal, Canada, June 1996

[24] T. Fukuda, Y. Morimoto, S. Morishita, T.Tokuyama. '*Mining optimized association rules for numeric attributes*,' In Proc. 1996 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'96), pages 182-191,Montreal, Canada, June 1996.

[25] J. Han, J. Pei and Y. Yin. '*Mining frequent patterns without candidate generation*,' In Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD' 00), pages 1-12, Dallas, TX, May 2000.

[26] Hanif D. Sherali , Jitamitra Desai, '*A Global Optimization RLT-based Approach for Solving the Fuzzy Clustering Problem,'* Journal of Global Optimization, volume 33, issue 4 (December 2005) pp: 597-615, 2005

[27] Y.-C. Hu, R.-S. Chen and G.-H. Tzeng, '*Mining fuzzy association rules for classification problem*,' Computers and Industrial Engineering, Volume 43, Issue 4, pp. 735-750, 2002.

[28] Yi-Chung Hu, Ruey-Shun Chen, Gwo-Hshiung Tzeng, '*Discovering fuzzy association rules using fuzzy partition methods*,' Knowledge Based Systems, vol. 16, pp. 137-147. 2003.

[29] Yi-Chuang Hu, Gwo-Hshiung Tzeng, '*Elicitation of classification rules by fuzzy data mining*,' Engineering Applications of Artificial Intelligence, Vol. 16,pp. 709-716,2003.

[30] Hayushi, Y. 1992. 'Fuzzy neural expert system with automated extraction of fuzzy if-then rules from a trained neural network,' In "Analysis and Management of Uncertainty: Theory and Applications" (edited by Ayyub, B. M.,Gupta, M. M., and Kanal, L. N.), North-Holland, Amsterdam: 171-181.

[31] Jianjiang Lu, Bbaowen Xu, Hongji Yang, 2003, '*A classification method of fuzzy association rules,'* IEEE International workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications ,8-10,September,Lviv, Ukraine.

[32] M. Kamber, J.Han, and J. Y. Chiang. Metarule-guided mining of multidimensional association rules using data cubes. In Proc. 1997 Int. Conf. on Knowledge Discovery and Data Mining (KDD'97),pp. 207-210, Newport Beach, CA, Aug. 1997.

[33] M.Kaya, R. Alhajj. '*Integrating mutlti-objective genetic algorithms into clustering for fuzzy association rules mining,'* In Proc. of the Fourth IEEE Int. Conf. on Data Mining (ICDM'04), pp. 431-434, Brighton, UK, Nov.2004.

[34] J. H. Lee and H. L. Kwang, '*An extension of association rules using fuzzy sets*,' Proc of IFSA'97, 1997.

[35] A. Lent, A. Swami, and J. Widom. '*Clustering association rules*,' Proc. of 1997 International Conference on Data Engineering (ICDE'97), pp. 220-231, Birmingham, England, Apr, 1997

[36] W. Li, J. Han and J. Pei, '*CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules,*' Proc. of ICDM 2001, pp369-376, 2001.

[37] J.Li and H. Liu, '*Kent ridge bio-medical data set repository*,' Available at http://sdmc.lit.org.sg/GEDatasets/Datasets.html.

[38] A. Liu, W. Hsu, and Y. Ma, '*Integrating classification and association rule mining*,' Proc. of 4th International conference on knowledge discovering and data mining KDD'98, pp 80-86, 1998.

[39] R. J. Miller and Y.Yang. '*Association rules over interval data,'* In Proc. 1997 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'97), pages 452-462, Tucson, AZ, May 1997.

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

[41] M. J. Pazzani, S. Mani, and W. R. Shankle. '*Beyond concise and colorful: Learning intelligible rules,*' In Knowledge Discovery and Data Mining, pages 235--238, 1997

[42] R.Radha, S.P.Rajagopalan, 2007, '*Assessing the interestingness of discovered knowledge using a hybrid approach based on fuzzy concepts*,' International Journal of Soft Computing 2(2): 243-248,2007 , Medwell Jounals , 2007.

[43] D. Singh, P. G. Febbo, K. Ross, et al, '*Gene expression correlates of clinical prostate cancer behavior*,' Cancer Cell 1:203-209, 2002

[44] R. Srikant and R. Agrawal. '*Mining quantitative association rules in large relational tables,'* In Proc. 1996 ACM-SIGMOD Int. Conf. on Manage of Data (SIGMOD'96), pages 1-12, Montreal, Canada, June 1996.

[45] Srinivasa K G *, Venugopal K R 1 and L M Patnaik '*Feature Extraction using Fuzzy C - Means Clustering forData Mining Systems,'* IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.3A,March 2006 Pp 230-236

[46] Ta-Wei Hung '*The bi-objective fuzzy C-Means cluster analysis for TSK fuzzy system identification*,' Springer Science, Business Media, LLC 2007

[47] K.Wang, S. Zhou, and Y. He, '*Growing decision trees on support-less association rules*,' In Proc. of the sixth ACM-SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'00), pages 265-269,Boston,MA,Aug. 2000.

[48] Xiao Ying Wang, Jon Garibaldi, Turhan Ozen , '*Application of the Fuzzy C-Means Clustering Method on the Analysis of non Preprocessed FTIR Data for Cancer Diagnosis*,' Fuzzy Optim Decis Making (2007) 6:51–61

[49] X. Yin and J. Han, '*CPAR: Classification based on Predictive Association Rules*,' Proc. Of SIAM Int. Conf. on Data Mining (SDM'03), pp. 331-335, San Francisco, CA, 2003.

[50] K. Yoda, T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. '*Computing optimized rectilinear regions for association rules*,' In Proc. 1997 Int. Conf. on Knowledge Discovery and Data Mining (KDD'97), pages 96-103, Newport Beach, CA, Aug. 1997