International Journal of Computational Intelligence and Informatics, Vol. 5: No. 4, March 2016

Interpretability of Fuzzy Clusters by Fuzzy Association Rules Using Cluster Based Fuzzy Partitioning

Swati Ramdasi
Computer Science Department
Savitribai Phule Pune University
Pune, India
jsp15@rediffmail.com

Shailaja Shirwaikar
Computer Science Department
Savitribai Phule Pune University
Pune, India
scshirwaikar@gmail.com

***Abstract*-Data mining is widely accepted and used tool for extracting interesting information from data. Associative rule mining and Clustering are descriptive techniques. Fuzzy approach has enhanced the power of both these techniques. Clustering is used in data processing for discretization and data reduction. However, Clustering suffers from interpretability problem. This paper presents a multi-step combination of the above two techniques which gives a better insight on the dataset and also identifies irrelevant attributes. It extends fuzzy association rule mining algorithm by using user defined support confidence framework. Several Clustering based methods are proposed and compared for fuzzy partitioning of individual attributes. Our proposed algorithm addresses the problem of interpretability of cluster by using expressive power of fuzzy rules as well as helps in improving quality of cluster by finding prime attributes contributing in cluster formation. The paper presents expected and interesting results obtained when the algorithm is applied on some known datasets.**

Keywords -Data Mining, Clustering, Fuzzy Association Rule mining, Fuzzy partitioning.

I. INTRODUCTION

Data mining is defined as the process of extraction of nontrivial, potentially useful knowledge from large datasets. It has very powerful tools and techniques which has widespread use in several application domains. Association Rule mining started with market basket analysis and has eventually established itself as a powerful mining technique and its use in various application domains is increasing rapidly. Amongst the several association rule mining algorithms, Apriori algorithm is the most simple to implement and extend to new situations. Most association rule mining algorithms assume Boolean attributes but in real life, attributes are quantitative or categorical in nature. To handle quantitative values, they are partitioned to give attribute range pairs.

Fuzzy partitioning removes the crisp boundaries and adds linguistic labels to attributes. The fuzzy association rules are closer to human interpretation such as 'young age → low income'. Traditionally, fuzzy partitioning is supervised, where experts decide boundaries of attribute intervals. Several partitioning method based on simple k - means algorithm are proposed for preprocessing of numerical data into fuzzy data.

An extension of Apriori Association rules mining algorithm is essential to mine fuzzy association rules. The algorithm is heavily based on support and confidence framework and requires membership functions to be extended to attribute sets. There are several clustering algorithms amongst which k-means algorithm has been extensively applied [60], [58] and [51].

Some of the issues related to clustering algorithm have been:

i) Interpretability of clusters

ii) Dependence of cluster quality on number of clusters specified

iii) Effect of noise or outliers and irrelevant dimensions on cluster quality.

In classical cluster analysis each datum must be assigned to exactly one cluster. Fuzzy cluster analysis relaxes this requirement by allowing membership degrees to deal with data that belong to more than one cluster at the same time. Fuzzy C means algorithm generates clusters which are probabilistic.

Each tuple belongs to every cluster with defined membership values. However interpretability of the generated clusters still remains a problem. This paper presents a combination of Fuzzy Apriori with clustering to provide an insight into the generated clusters in terms of their relationship with attribute values.

The paper is organized as follows. Section II presents related work containing some preliminaries Several Clustering based Fuzzy Partitioning alternatives and their comparative analysis is presented in section III. The main algorithm for Cluster Interpretability is presented in section IV. Section V is about the experimental setup and observations which is followed by conclusion in section VI.

II. RELATED WORK

Data mining is powerful technique with huge potential to contribute in proactive knowledge driven decisions by extraction of hidden patterns from large data sets. It helps business organizations by prediction of future trends and behavior. It is applied in various fields of human life due to availability of large amount of data in the form of web contents, records, documents, images, sound recordings, videos, scientific data by extracting knowledge that can be utilized for knowledge based decisions. Scope of data mining is varying from various fields such as Market analysis to Medical field including Insurance, Finance, Banking, Pharmacy, security and many more [14], [32], [43]. Various techniques are used for data mining such as Association rule mining, Classification, Clustering etc. all having their own strengths and applicability.

A. Association Rule Mining

Association rule mining is one of the important, widely used and highly researched techniques. This was first introduced by Agarwal [46] for Market basket analysis. Many algorithms were proposed for mining association rules such as Apriori, Eclet, Frequent pattern Growth [24], AIS algorithm etc. Apriori algorithm proposed by Agrawal, R. and Srikant, R. [46] became very popular. Rakesh Agrawal et al [2] discussed fast Algorithms for Mining Association Rules for discovering association rules between items in a large database of sales transactions. They introduced Apriori Hybrid approach to solve problem that fundamentally differs from the known algorithms by exploring the best features of the proposed algorithms. Various algorithms for mining association rules are, taxonomy based [44], RARM [13], Constrain based Apriori [48], PRICES [53], Matrix Algorithm [59], IDTE [50]. Association rule mining algorithms generate large number of rules and are reduced by filtering using various interest measures and this problem is addressed by many researchers [9], [7], [12], [31], [4].

An association rule is an implication of the form $A \rightarrow B$, where A and B are subsets of an attribute set and they are disjoint such that is $A \rightarrow B = \Phi$. The Apriori algorithm is the basic algorithm used for mining association rules. In Market basket analysis, every transaction in supermarket database D is represented as a Boolean for an item set $I=\{i_1,i_2,..,i_m\}$ containing m types of items. For each transaction t in database t[ik] is either 1 or 0 indicating the item is bought or not. The occurrence frequency of an itemset A, i.e. the number of transactions in the dataset D containing the itemset A is known as Support_Count or Absolute Support of the itemset A. The interestingness of the association rule is measured using support and confidence.

For a rule $A \rightarrow B$,

$$\text{Support } (A \rightarrow B) = \frac{\text{Support\_Count } (AUB)}{|D|} \quad (1)$$

$$\text{Confidence } (A \rightarrow B) = \frac{\text{Support\_Count}(A\ U\ B)}{\text{Support\_Count } (A)} \quad (2)$$

Discovering of association rules is two step processes

Discovering all frequent or large item sets having support greater than a minimum support threshold. Generating association rules from the frequent item sets having confidence greater than minimum confidence threshold.

Step1 uses the downward-closure property which guarantees that for a frequent item set, all its subsets are frequent and thus for an infrequent item set, all its supersets are infrequent. In step two, using minimum confidence criteria interesting association rules are generated.

B. Fuzzy association rule mining

Transaction data in real-world applications is not always binary but consist of quantitative values. Fuzzy set theory has been introduced in the process of mining quantitative association rules, which results in a new category of association rules called fuzzy association rules. In 1965, Zadeh first proposed Fuzzy set theory which concerns with quantifying and reasoning using linguistic variables.

Han [24] transformed quantitative data attributes into linguistic terms and discovered interesting associations among attributes using statistical analysis to remove need for user-specified thresholds. Main advantage of the approach is, it discovers both positive and negative association rules. Hong proposed similar fuzzy mining algorithm to mine fuzzy rules from quantitative transaction data [29]. It calculates the scalar cardinality of each linguistic term on all the transaction data. Toshihiko Watanabe [49] proposes a fast algorithm based on Apriori for extracting fuzzy association rules from database which improves the computational time of mining for real applications. Different variations of Apriori Algorithm for fuzzy association rule mining were proposed such as FCABTAR [15], FARME-D [39] and some more are found in [38].

A fuzzy association rule is then understood as a rule of the form (X, t(x)) → (Y, t(y)) where X and Y are attributes and t(x), t(y) are linguistic variables. For example (BP, "Low") → (Heart Attack , "High") can be well interpreted in natural language as 'if Blood Pressure is low chances for Heart Attack are High'.

Generation of fuzzy association rules using an appropriate Fuzzy Association Rule Mining (ARM) algorithm is three step processes.

i) Preprocessing of crisp dataset.

ii) Generation of Frequent item sets satisfying usefulness measures.

iii) Formation of Fuzzy Association rules satisfying interestingness measure thresholds.

i) *Preprocessing of crisp data:*

The very first step of data preprocessing is conversion of crisp quantitative or numerical attributes set into a fuzzy dataset. For a dataset containing *n* attributes and *m* linguistic terms, there will be $n \times m$ attribute linguistic pairs. Existing and proposed fuzzy partitioning methods are discussed in section III. As a result of processing, fuzzy dataset with $n \times m$ columns gets generated.

ii) *Use a fuzzy Apriori algorithm to generate frequent itemsets:*

Most of the algorithms used for generating fuzzy association rules are one or other kind of variations of basic Apriori algorithm that uses support as a measure for deciding usefulness of generated frequent itemsets. It employs an iterative approach, where *k* itemsets are used to explore (*k+1*) iemsets. First, the set of frequent 1-itemsets is found, denoted by L_1. L_1 is used to find L_2 by generating first candidate sets C_2. The set of frequent 2 item set L_2 is used to find L_3, and so on, until no more frequent *k* itemsets can be found. The algorithm reduces effort by making use of Apriori property that all non-empty subsets of frequent item sets must be frequent. An (attribute, term) pair will be considered frequent if its support count is more than the threshold defined by minimum support.

The support count of Attribute-Term set (x, T (k)) is defined as

$$\text{Support\_Count (x,T(k))} = \sum_{i=1}^{n} \mu_{T(k)}(x_i) \tag{3}$$

where $\mu_{T(k)}$ (x_i) denotes the membership value of x for linguistic variable T(k) in ith transaction.

The Support_Count can be extended to more than one attribute term sets as

$$\text{Support\_Count} (a_1\_t_1, a_2\_t_2,\ldots, a_k\_t_k) = \{ \mu_{t1} (a_1) \otimes \mu t_{t2} (a_2) \otimes\ldots.., \otimes \mu_{tk} (a_k)\} , \tag{4}$$

where there are several possibilities for the operator ⊗ which are discussed in next section.

iii) The association rules are generated by considering every subset *S* of frequent Item set *I* and generating the association rule *S*→ *(I-S)* and validating it using some interestingness measure. Confidence is popular measure used for interestingness but there are several other measures which need to be appropriately extended to use with fuzzy sets.

C. Selection of appropriate T-norm (x ⊗ y) for Fuzzy Association Rules

There are various ways to choose T-norm operator [43], [39] for defining support as described below.

- Goedal t-norm: $x \otimes y = \min(x, y)$ (5)
- Goguen t-norm : $x \otimes y = x.y$ (6)
- Lukasiewicz t-norm: $x \otimes y= \max (0, x + y – 1)$ (7)

- Drastic multiplication: x ⊗ y = x (if y=1), = y (if x=1), = 0 (if x, y < 1) (8)

It is obvious that using Goedal t-norm, number of frequent item sets generated are largest in comparison with others, hence it is used in further analysis.

D. Defining Support for Fuzzy Data

Relative Support is an important measure used in association rule generation and also in defining other interestingness measures. In Boolean transactions item can be either present or absent, hence relative support is defined as

$$\text{Support } (x \rightarrow y) \quad = \quad \frac{\text{Support\_ Count}(x \otimes y)}{N} \tag{9}$$

where *N* is total number of transactions

Here maximum possible support is *N* when item is present in all the transactions.

In fuzzy transactions maximum possible support need not be *N*. Hence for fuzzy data, *N* needs to be replaced by maximum possible value of membership function *MaxM* which is

$$\text{MaxM} \quad = \sum\nolimits_{i=1}^{n} \underset{j\in J,k\in L}{Max} \{ \mu_{lk}(a_{ij}) \} \tag{10}$$

where *i* vary over transactional dataset, *j* varies over columns representing attributes and *k* varies over set of linguistic variables *L*. Fuzzy support can be redefined as

$$\text{Fuzzy\_ Support } (x \rightarrow y) \quad = \quad \frac{\text{Support}(x \otimes y)}{\text{MaxM}} \tag{11}$$

Instead of dividing support by number of transactions, we divide by summarizing the maximum membership value among every item of each transaction. Use of this fuzzy support is more logical and appropriate in consideration of interestingness measures where support plays important role. We used computed values for some interestingness measures using support and fuzzy support.

Several interestingness measures are defined using support such as

$$\text{Fuzzy\_Conviction}(X \rightarrow Y) = \frac{\text{Fuzzy\_Support}(x) * \text{Fuzzy\_Support}(\sim y)}{\text{Fuzzy\_Support}(\text{x}\cup\sim\text{y})} \tag{12}$$

$$\text{Fuzzy\_Lift}(X \rightarrow Y) \quad = \quad \frac{\text{Confidence}(x \rightarrow y)}{\text{Fuzzy\_Support}(y)} \tag{13}$$

$$\text{Laplace } (X \rightarrow Y) = \frac{\text{Fuzzy\_Support}(x \rightarrow y)+1}{\text{Fuzzy\_Support}(x)+2} \tag{14}$$

$$\text{Jaccard} = \frac{\text{Fuzzy\_Support}(x \rightarrow y)}{\text{Fuzzy\_support}(x)+\text{FuzzySupport}(y)-\text{Fuzzy\_Support}(x \rightarrow y)} \tag{15}$$

III. FUZZY PARTITIONING

For datasets containing categorical or quantitative attributes, algorithms proposed for binary attributes are not convenient for association rule mining. The problem was addressed by Rakesh Agarwal & Shrikant [2]. Partitioning quantitative attributes into discrete intervals is major task in quantitative association rule mining. Various methods are proposed in literature for discretization of quantitative attributes so as to retain the original distribution of the attribute.

Methods include simple approach to replace quantitative attributes by Boolean values & use conventional association rule mining. Another approach is to divide quantitative attributes into number of partitions. Equi-depth and Equi-width partitioning methods [46] divide attributes into intervals accordingly. All these partitioning methods are used for crisp partitioning and all suffer from sharp boundary problem. Fuzzy partitioning is natural generalization for partitioning and the fuzzy sets and the membership functions determine partitions with more realistic intervals which may lead to correct, strong partitioning of attributes of data set. Many researchers have proposed various methods for fuzzy partitioning of quantitative data for generating fuzzy association rules [22], [11], [33].

The fuzzy set provides a smooth change near the boundaries of partitions and can express more sophisticated belongingness of data value to interval and this smooth transition of membership functions helps to eliminate the “sharp boundary problem”.

In fuzzy partitions data value can belong to more than one partition with some membership. This membership can be derived by using various membership methods. The fuzzy sets in most real-life datasets are Triangular, Trapezoidal or Gaussian-like depending on varied and heterogeneous nature of the datasets. Numerical data present in most real-life datasets can be converted into fuzzy sets using anyone of these membership functions , wherein a particular data point can belong to two or more fuzzy sets simultaneously.

A. *Fuzzy Membership function Alternatives*

The functions that identify a fuzzy set in their domain is termed membership function μ(m) having values between 0 to 1. A membership value $\mu(m) = 0$ indicates non-membership of the point m in the fuzzy set identified by the function μ, whereas value $\mu(m) = 1$ indicates the full membership of m in the fuzzy set.

Triangular membership function uses three variables *a*, *b* and *c* for deriving membership value for data. It covers all values when defined as overlapping function. Only single value has full that is 1 membership and membership value goes on increasing from *a to b* & decreases from *b* to *c* tending towards 0.

A trapezoidal MF is described by four parameters *a, b, c* and *d*. This can be reduced into triangular using b = c.

Trapezoidal Membership function is overlapping function that covers all values till infinity and last fuzzy trapezoidal interval is limited by infinity. A fuzzy interval defined in this fashion has full membership in points *b* to *c* and membership tends towards zero from *b* to *a* and from *c* to *d*. Gaussian Membership function is represented by two parameters *c* and σ where *c* represents its centre and σ represents its width Figure 1.

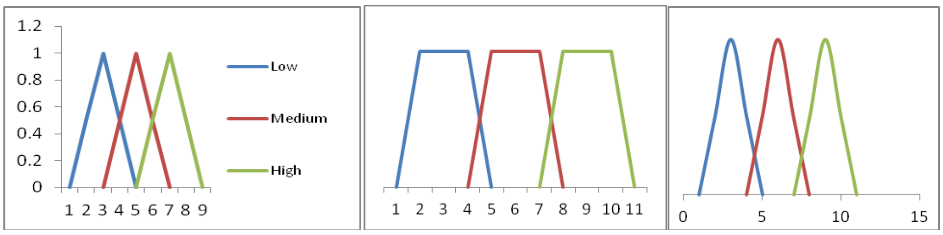


Figure 1. Triangular, Trapezoidal and Gaussian membership functions for overlapping fuzzy sets

B. *Supervised fuzzy partitioning method – Experts knowledge based*

This is one of the important partitioning techniques. Membership function is determined by Experts' knowledge or known perception by expert. This is used for partitioning attributes by forming fuzzy sets as intervals. The intervals are then labeled with a linguistic variable name. The set of consecutive labels are formed such that the order of the intervals is preserved. (eg. low, medium, high). Each interval is considered as an individual attribute. Membership degree of each data item is derived by using defined membership function. (E.g Use of trapezoidal membership function for Blood Pressure) The values for interval range are decided by expert in that domain (as *a, b, c*, and *d*) .The membership functions can be defined for both triangular and trapezoidal membership functions as

$$\text{Triangular } (x\text{: } a, b.\ c) = \max\left(\min\left(\frac{(x-a)}{(b-a)}, \frac{(c-x)}{(c-b)} \right), 0\right) \tag{16}$$

$$\text{Trapezoidal } (x\text{: } a, b, c, d) = \max\left(\min\left(\frac{(x-a)}{(b-a)} \quad \frac{(d-x)}{(d-c)} \right), 0\right) \tag{17}$$

where the values of *a, b, c, d* are provided by the Expert.

The values for cholesterol given by expert for *a, b, c* and *d* are 100, 120, 180 and 200 and the graphical representation of the trapezoidal membership function is as shown in Figure 2.

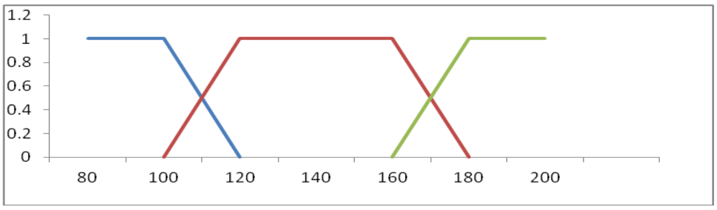


Figure 2. Experts Knowledge based Trapezoidal Membership function.

C. Unsupervised Cluster Based Approach to define Membership Functions

The fuzzy partitioning of quantitative attributes can be carried out in unsupervised manner using cluster centers obtained by *k* means clustering algorithm. The values of four variables *a, b, c, d* required for three linguistic variables for trapezoidal membership function can be computed using three ordered cluster centers C_1, C_2, C_3. In general for k linguistic variables the 2*(k-1)* values of a_i, b_i *(1<=i<=k-1)* can be computed using *k* cluster centers.

Method 1.

Equi-Spaced method: In this approach the interval between two cluster centers is divided into equal parts. Thus values of a_i and b_i are computed as follows.

$$a_i \;=\; \frac{(3C_i + C_{i+1})}{4} \qquad b_i \;=\; \frac{(C_i + 3C_{i+1})}{4} \tag{18}$$

A given data can be partitioned into five fuzzy partitions using five Cluster centers C_1, C_2, C_3, C_4 and C_5. The values of a_i and b_i can be computed and trapezoidal membership function can be obtained as shown in Figure 3 for the cluster centers $(C_1, C_2, C_3, C_4, C_5)$= {1, 2, 4, 5, 7}

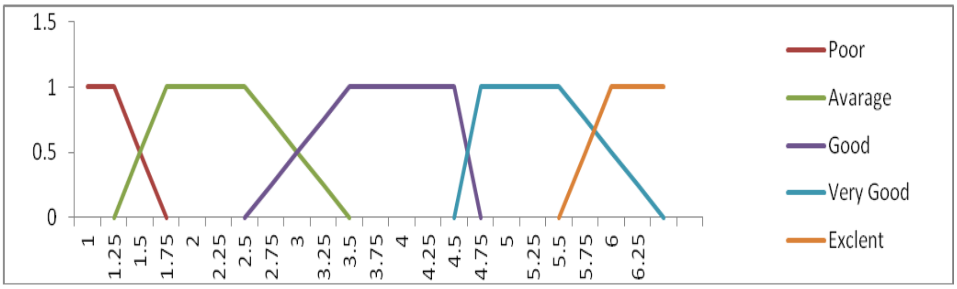


Figure 3. Five clusters showing cluster centers and values of a_i, b_i.

The trapezoidal membership values can be directly computed from cluster centers using following algorithm.

D. Algorithm for Fuzzy Partitioning using Method 1

INPUT:

I. Dataset *D* with attributes A = *{A₁, A2 …An}* where each A_j $(1 \leq j \leq n)$ is a quantitative attribute.

II. Number of linguistic variables *L = {l_1, l_2, l_3 ,….l_m}* where each attribute is associated with at least two (or more) linguistic variables and linguistic variables are associated with fuzzy sets

III. Ordered Cluster Centers *CC* {c_1,c_2,c_3…..*cm*}

OUTPUT:

Fuzzified Dataset with linguistic attributes {attribute, linguistic variable} pair

$$D_F = \{(A_1\_L_1, A_1\_L_2, \ldots A_1\_L_m,),(A_2\_L_1, A_2\_L_2, \ldots\ldots A_2\_L_m,), \ldots\ldots(A_n\_L_1, A_n\_L_2, \ldots A_n\_L_m,)\}$$

```
Steps:
For each Attribute $A_j$ (Column of Dataset)
    {
      Compute membership value for $A_i$, $L_1$ using equation 1.
         $\mu_1 = \max(min\left(1, \frac{C_1+3C_2-4x}{2(C_2-C_1)}\right), 0)$                [1]
         for each linguistic variable $L_i$,
        {
           Compute membership value for $A_i$,$L_i$ ( $1< i< n$) using equation 2.
              $\mu_i = \max(\min\left(\frac{4X-3C_{i-1}-C_i}{2(C_i-C_{i-1})}), 1, \frac{C_i+3C_{i+1}-4X}{2(C_{i+1}-C_i)}\right), 0)$      [2]
        }
              Compute membership value for $A_i$, $L_m$ using equation 3.
              $\mu_m = \max(min\left(1, \frac{X-C_{m-1}+3C_m}{2(C_m-C_{m-1})}\right), 0)$       [3]
    }
```

Using the corresponding membership functions defined with each fuzzy set using equation 1, 2 & 3, the original dataset D is changed into a fuzzy dataset D_F.

Method 2. Variance Spacing Proportionate: The k means clustering algorithm also provides apart from cluster centers the variance values for each cluster. The variance value indicates the spread of values around the cluster center. It can be used for proportionately spacing a_i and b_i between two cluster centers.

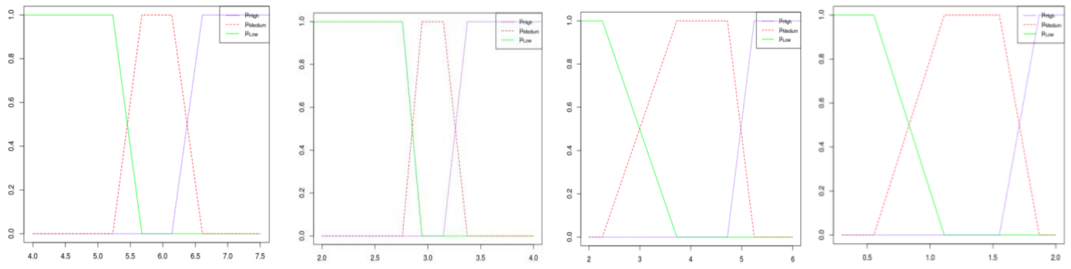
Thus a_i and b_i can be calculated using formula

$$a_i = C_i *(1-V_i) + V_i * C_{i+1} \quad (19)$$

$$b_i = C_{i+1}*(1-V_{i+1}) - C_i*V_{i+1} \quad (20)$$

where C_i is cluster centers and V_i are corresponding variance values.

Figure 4. Shows membership functions for different attributes for iris dataset using above method.



a) Sepal Length b) Sepal Width c) Petal Length d) Petal Width

Figure 4. Membership functions using Equi spaced method

Method 3: Variance spacing absolute: In this approach the spread around the cluster center is specified using absolute values of the variance. Thus a_i and b_i can be calculated using formula

$$a_i = C_i + V_i \quad (21)$$

$$b_i = C_{i+1} - V_{i+1} \quad (22)$$

E. Comparison of various fuzzy partitioning Approaches

We compared different methods of fuzzy partitioning by using a simple measure of quality of fuzzy partitions. Fuzzy partitioning assigns membership value to each data object with which it belongs to each partition. Membership value 1 indicates that data object fully belongs to that partition & 0 indicates that it does not belong to that partition. Membership values between 0.4 to 0.6 doses not clearly indicate the belongingness of data into any one of the partition and can be considered as poor partitioning.

We analyzed the iris data for same and carried out fuzzy partitions of this data for following methods.

1. Experts knowledge based fuzzy partitioning
2. Equi -spaced using cluster centers fuzzy partitioning
3. Variance based proportionate fuzzy partitioning
4. Variance based absolute fuzzy partitioning
5. Fuzzy C Means

Variance 2 method gives best results and hence used for further analysis. Figure 5 shows graphical representation of above mentioned methods. Table I Shows results various methods discussed above.

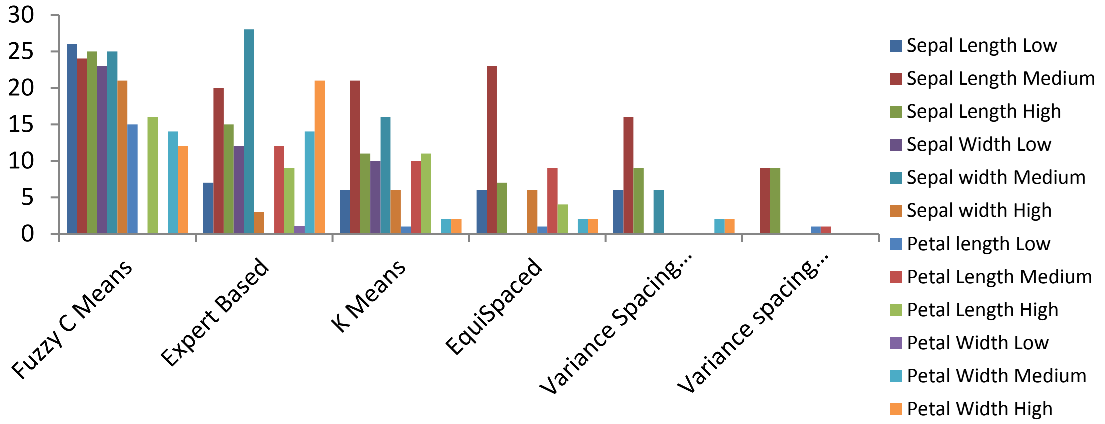
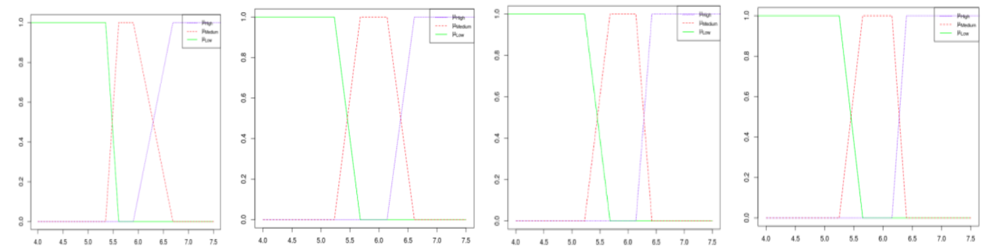


Figure 5. Graphical Representation of Table I

TABLE I. SHOWS RESULTS VARIOUS METHODS

Attributes Partitions Methods	No. of Records having μ as 0.4 < μ < 0.6												Total
	Sepal Length			Sepal Width			Petal Length			Petal Width			
	Low	*Medium*	*High*	*Low*	*Medium*	*High*	*Low*	*Medium*	*High*	*Low*	*Medium*	*High*	
Fuzzy C Means	26	24	25	23	25	21	15	0	16	0	14	12	201
Expert Based	7	20	15	12	28	3	0	12	9	1	14	21	142
K Means	6	21	11	10	16	6	1	10	11	0	2	2	96
Equi-Spaced method	6	23	7	0	0	6	1	9	4	0	2	2	60
Variance Spacing Proportionate	6	16	9	0	6	0	0	0	0	0	2	2	41
Variance spacing absolute	0	9	9	0	0	0	1	1	0	0	0	0	20



a) Expert Based b) EquiSpaced c) Variance Absolute d) Variance Proportionate

Figure 5. Shows membership functions for sepal length using our different approaches

IV. CLUSTER INTERPRETATION

Most of the research in clustering domain is related to number of clusters and initial cluster centers. Relatively less attention is given for cluster interpretation. However, it is important to highlight the prime attributes contributing to cluster formation. Fuzzy association rules uncover the dependencies among items in datasets as they have rich applicability over wide variety of problems. We have proposed a method that combines fuzzy association rules for interpretability of clusters.

It is commonly known that not all variables are important and not all the values of attributes are important for a particular cluster. Clusters can be properly interpreted if the attributes and specific values of attributes that play an important role in forming a cluster are known. The multistep algorithm proposed by us takes care of above problem. The method involves combination of Fuzzy clustering and Fuzzy association rule mining. This step by step algorithm is described below.

Algorithm for Cluster Interpretation

Input

1. Fuzzy Dataset generated by fuzzy Partitioning of each attribute of data set *D* into *k* linguistic variables {D_F}
2. Fuzzy clusters of data set *D* using Fuzzy *C* means Algorithm having *j* clusters. This will be a data set with *j* column containing membership values of each record of data set D in each cluster {D_C}

Output

Fuzzy association rules representing consequent as cluster variable and antecedent as item set with linguistic variables.

Steps.

Stage 1. Generate Extended Fuzzy data set D_F

Append each column of Cluster data D_C to fuzzy data set D_F {D_E}

Stage 2. Apply Extended Fuzzy Aprioi algorithm on extended fuzzy dataset D_E to get frequent Itemsets with support greater than minsupport, user defined threshold.

Stage 3. Filter frequent itemset containing last item of frequent itemset representing cluster variable.

Stage 4. Generate fuzzy association rules with consequent as cluster variable and antecedent as group of items along with linguistic variable using interest measures as confidence greater than user defined confidence threshold.

Generated fuzzy rules are of the form $C_j \rightarrow B_{lk}$, where $B_{lk} \sqsubseteq D_F$, and B_{lk} do not contain any two items that are associated with the same attribute (for instance, will not contain “sepal length low and sepal length high”).

To evaluate the performance of the proposed approach, tests were realized on various known datasets as well as on real data set also.

In next section, the experimental setup and experiments executed on various data sets are presented.

V. EXPERIMENTAL SETUP

A. *Experiment 1: IRIS data set*

The IRIS data set contains 150 records with measurements on three classes of Iris flower for Sepal Length, Sepal Width, Petal Length and Petal Width. The class label is one of the distinguished classes that are Iris Setosa, Iris Versicolor and Iris Virginica. The original dataset contains continuous values for each attribute and discrete values for class labels. To generate fuzzy association rules on this raw data, data must be transformed into fuzzy transactional dataset. In this work simple k means clustering algorithm with our proposed method is used to partition the four attributes. For designing parameterized membership function, trapezoidal membership function is used as it covers maximum area. On every attribute, three fuzzy sets are defined and linguistic variables are associated with each fuzzy set such as Low, Medium and High. We clustered crisp iris dataset using Fuzzy C Means clustering algorithm to generate three fuzzy clusters which may represent three classes of dataset that is class label in original dataset. We integrated the fuzzy transactional dataset by adding these three clusters as additional attributes to formulate new dataset which consists of 4*3 +3 attributes related to three fuzzy sets defined on four features (attributes) of dataset i.e. Sepal Length, Sepal width, Petal Length and Petal Width and three clusters representing classes.

The frequent item sets and the rules were generated by using extended fuzzy Apriori algorithm. The minimum support and confidence were set to =10% and 60%. With these parameters 160 frequent item sets were generated and 98 rules were obtained. The filtered set of rules satisfying the support threshold and best values for different measures for each cluster are shown in Table II.

TABLE II. SHOWS SUPPORT THRESHOLD AND BEST VALUES

Rule	Confidence	Conviction	Lift	Laplace	Jaccard
Cluster 3->Petal Width Medium	0.631941642	36.0164994	0.012768	0.627181161	0.488238
Cluster 3->Petal Width Medium & Petal Length Medium	0.676363343	32.1923104	0.013666	0.669124258	0.489103
Cluster 2->Petal Width Low	0.8125187	30.3350575	0.015596	0.800815244	0.676584
Cluster 2->Petal Width Low & Petal Length Low	0.811821526	30.0476916	0.015583	0.799981529	0.667461
Cluster 1->Petal Width High	0.667238348	47	0.014074	0.65999259	0.474456
Cluster 1->Petal Width High & Petal Length High	0.722984258	31.3952309	0.01525	0.711691053	0.468974

From obtained rules it is very clear that petal length and petal width are prime attributes in cluster formation. All three clusters can be clearly interpreted as one cluster is formulated with those having petal length and petal width as low another cluster is with medium petal length & medium petal width and third cluster is formed with high petal length and high petal width. Use of additional interest measures such as Conviction, Lift, Laplace, and Jaccard strongly provides conformation to the rule "Prime attributes in defining boundaries of three clusters are petal length and Petal width".

B. *Experiment 2: Glass data set*

This experiment is conducted on other well-known Glass data set. Data set comprises of nine attributes shown in Table III and results are in Table IV.

TABLE III. SHOWS DATASET COMPRISES OF NINE ATTRIBUTES

RI	refractive index	K	Potassium	Al	Aluminum
Na	Sodium	Ca	Calcium	Si	Silicon
Mg	Magnesium	Ba	Barium	Fe	Iron

TABLE IV. RESULTS

Rules	**Confidence**
Cluster 1 → Sodium(Medium),Magnesium (High) Aluminum (Medium), Calcium(Medium), Barium (Low)	0.805362
Cluster 1 → Refractive index (Low), Sodium (Medium), Aluminum (Medium), Calcium (Medium), Barium (Low)	0.833583
Cluster 1 → Refractive index(Low), Sodium (Medium (Medium), Magnesium (High), Aluminum (Low), Calcium (Low), Barium (Low)	0.745075
Cluster 2 → Refractive index (Low), Sodium (medium), Potassium (Low), Barium(Low)	0.656585
Cluster 3 --> Refractive index(Low),Sodium(medium),Magnesium(High), Aluminum (Low), Calcium (Low), Barium (Low), Iron (Low)	0.827017
Cluster 3 --> Refractive) index(Low),Sodium (medium), Magnesium(High), Aluminum(Medium), Silicon(Medium),Potacium(Medium),Calcium(Low),Borium(Low),Iron(Low)	0.787036

For Glass data although we have taken only three clusters within cluster 1 & 3 we found three and two different groups which indicate that there can be in all 6 clusters from dataset. Table 6 shows Rules having confidence more than minimum threshold confidence of 0.6. In Cluster formation we found that Refractive index (Attributes 1), Barium (Attribute 8) and Iron (Attribute 9) are attributes which are having similar values in all clusters indicating that those are attributes not playing role in cluster formation. Important attributes on the basis of whom clusters are formulated are Sodium, Magnesium, Aluminum Silicon, Potassium, Calcium (attribute 2, 3, 4, 5, 6, 7). These are prime attributes whose values categorize type of Glass. Results are verified with Glass data description given in WEKA, We found the results are almost matching with description. As described, there are in all seven classes of data. For one class data is absent. Hence only six classes are present in data which is same as the number of clusters that we have obtained. In results there are three clusters. Cluster 1 comprises of three clusters, cluster 2 is single cluster and cluster 3 has two subgroups which can comprise into two clusters. Hence total number of clusters is 6 which match according to description in WEKA for Glass data.

C. Experiment 3: Real data set

We used our proposed algorithm on weather data set which contains real time data having the attributes specified in Table V.

TABLE V. ATTRIBUTES

Air Tempreture(C)	Solar Radiation(W/m2)	Wind Speed(m/s)
Humidity(%Rh)	Precipitation(mm)	Wind Direction(D)
Pressure(hPa)	Rain Duration(min)	Date, Time

We preprocessed data by removing attributes such as date and time. Applying the proposed algorithm we found results listed in following table. We divided each attribute into three fuzzy attributes namely Low, Medium and High. Whole data is clustered into three clusters. After execution our algorithm we found the results as shown in table VI.

TABLE VI. RESULTS

Rules	**Confidence**
Cluster 1->Temperature (Medium), Humidity(Medium), Pressure (Medium), Radiation (High)	1
Cluster 2->Temperature (High), Humidity (Medium), Pressure(Low), Radiation(Low)	0.8368
Cluster 3->Temperature (Low), Humidity (High), Pressure (High), Radiation (Medium)	0.7619

Results from above table show that Temperature, Pressure & Radiation are the important attributes in deciding type of days.

D. Experiment 4

Proposed algorithm is useful in identifying irrelevant attributes which do not play vital role in cluster formation. Identification of relevant attributes becomes important to improve the quality of clusters. We have executed the clustering algorithm from well-known data mining tool WEKA. We applied simple k means algorithm on various datasets to get initial clusters initially with all attributes. After applying our proposed algorithm which provides relevant attributes, we removed irrelevant attributes from data set. Again applying the clustering algorithm on this reduced data set we found that cluster sum of squared error has been reduced significantly as shown in Table VII.

TABLE VII. SQUARED ERROR

Data Set	**Initial Attributes**	**No. of Clusters**	**Relevant Attributes**	**Sum of Squared Errors**	**Sum of squared errors (After removal of irrelevant attributes)**
Iris	Sepal Length,Sepal Width, Petal Length,Petal Width	3	Petal Length Petal Width	6.982	1.7018
Glass	RI, Na, Mg, Al, Si, K, Ca, Ba, Fe	6	Na, Mg , Al K, Ca , Ba	77.124	12.92
Liver	Mean corpuscular volume (mcv), Alkaline phosphotase (alkphos), Alamine Aminotransferase (sgpt), Aspartate Aminotransferase (sgot), Gamma-glutamyl anspeptidase (gammagt)	3	mcv, sgpt, sgot, gammagt	373.73	291
Wine	Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulphates, Alcohol	4	Fixed acidity, Volatile acidity, Citric acid, Residual sugar, pH	425.48	120.18

VI. Conclusion

Fuzzy partitioning of data can be carried in unsupervised manner using the cluster centers generated by k means clustering algorithm. Fuzzy Association Rules have better interpretability because of use of linguistic variables and can be combined with clustering which is one of the important data mining techniques which suffer from interpretability problem. The results are satisfactory for data sets whose classifications are known. The method can be applied to arbitrary data sets. In this paper we have used three linguistic variables but it can be easily generalized to more than three variables such as poor, satisfactory, good, very good, Excellent depending on application requirement. The limitations of Apriori algorithm are evident when the number of attributes increase and so also the linguistic variables. It is necessary to adapt a faster Association rule mining algorithm for generating large item sets and the one that generates only required association rules so that the algorithm can be readily applied for any data set.

References

[1] Agarwal, R.C., Aggarwal, C.C. and Prasad, V.V.V., “A tree projection algorithm for generation of frequent item sets”, Journal of parallel and Distributed Computing, 61(3), pp.350-371, 2001.

[2] Agrawal, R. and Srikant, R.,” March. Mining sequential patterns. In Data Engineering”,. Proceedings of the Eleventh International Conference on, pp. 3-14. IEEE, 1995.

[3] Agrawal, R., Imielinski, T. and Swami, A., “June. Mining association rules between sets of items in large databases”, In Acm sigmod record, Vol. 22, No. 2, pp. 207-216., 1993.

[4] Ashrafi, M.Z., Taniar, D. and Smith, K., “A new approach of eliminating redundant association rules”, In International Conference on Database and Expert Systems Applications, Springer Berlin Heidelberg, pp. 465-474., August 2004.

[5] Ashrafi, M.Z., Taniar, D. and Smith, K., “Redundant association rules reduction techniques”, In Australasian Joint Conference on Artificial Intelligence, Springer Berlin Heidelberg, pp. 254-263, December 2005.

[6] Bai, V.M.A. and Manimegalai, D., “An analysis of document clustering algorithms”, In Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on, pp. 402-406, IEEE, October 2010.

[7] Baralis, E. and Psaila, G., “Designing templates for mining association rules”, Journal of Intelligent Information Systems, 9(1), pp.7-32, 1997.

[8] Bezdek, J.C., “Pattern recognition with fuzzy objective function algorithms”, Springer Science & Business Media, 2013.

[9] Brin, S., Motwani, R. and Silverstein, C., “June. Beyond market baskets: Generalizing association rules to correlations”, In ACM SIGMOD Record, Vol. 26, No. 2, pp. 265-276, 1997.

[10] Chaturvedi, S.K., Richariya, V. and Tiwari, N., “Anomaly detection in network using data mining techniques”, International Journal of Emerging Technology and Advanced Engineering, ISSN, pp.2250-2459, 2012.

[11] Chien, B.C., Lin, Z.L. and Hong, T.P., “An efficient clustering algorithm for mining fuzzy quantitative association rules”, In IFSA World Congress and 20th NAFIPS International Conference, Joint 9th, Vol. 3, pp. 1306-1311, IEEE. July 2001.

[12] Cristofor, L. and Simovici, D., “Generating an informative cover for association rules”, In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on, pp. 597-600, IEEE, 2002.

[13] Das, A., Ng, W.K. and Woon, Y.K., “Rapid association rule mining” In Proceedings of the tenth international conference on Information and knowledge management, pp. 474-481, October 2001.

[14] Delen, D., Walker, G. and Kadam, A., “Predicting breast cancer survivability: a comparison of three data mining methods”, Artificial intelligence in medicine, 34(2), pp.113-127, 2005.

[15] Delgado; M., Marín; N., Sánchez; D; Vila, M.A., “Fuzzy association rules: general model and applications”, IEEE Transactions on fuzzy systems, 11(2), pp.214-225, 2003.

[16] Dinh Manh Tuong,” Artificial Intelligence”, Faculty of Technology, Vietnam National University, Hanoi – 2003.

[17] Do, T.D., Hui, S.C. and Fong, A., “Mining frequent itemsets with category-based constraints”, In International Conference on Discovery Science, Springer Berlin Heidelberg, pp. 76-86, October 2003.

[18] Dunn, J.C., “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, 1973.

[19] Gagolewski, M. and Caha, J., “A Guide to the FuzzyNumbers Package for R”, 2015.

[20] Gupta, S., Kumar, D. and Sharma, A.,” Data mining classification techniques applied for breast cancer diagnosis and prognosis”, Indian Journal of Computer Science and Engineering (IJCSE), 2(2), pp.188-195, 2011.

[21] Gyenesei, A., “A Fuzzy Approach for Mining Quantitative Association Rules” Acta Cybern., 15(2), pp.305-320, 2001.

[22] Gyenesei, A., “Determining fuzzy sets for quantitative attributes in data mining problems. Proc. of Advances in Fuzzy Systems and Evol”, Comp, pp.48-53, 2001.

[23] Han, J. and Pei, J., “Mining frequent patterns by pattern-growth: methodology and implications”, ACM SIGKDD explorations newsletter, 2(2), pp.14-20, 2000.

[24] Han, J., Pei, J. and Yin, Y., “Mining frequent patterns without candidate generation”, In ACM Sigmod Record, Vol. 29, No. 2, pp. 1-12, May 2000.

[25] Hegland, M., “Algorithms for association rules In Advanced lectures on machine learning”, Springer Berlin Heidelberg, pp. 226-234, 2003.

[26] Hilderman, R.J. and Hamilton, H.J., “Knowledge Discovery and Interest Measures”. Kluwer Academic, Boston, 18, pp.135-142, 2002.

[27] Hirota, K. and Pedrycz, W., “Linguistic data mining and fuzzy modelling”, In Fuzzy Systems, 1996., Proceedings of the Fifth IEEE International Conference on, Vol. 2, pp. 1488-1492, September1996.

[28] Hong, T.P., Lin, K.Y. and Chien, B.C., “Mining fuzzy multiple-level association rules from quantitative data”, Applied Intelligence, 18(1), pp.79-90, 2003.

[29] Jang, J.S.R., Sun, C.T. and Mizutani, E., “Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence, 1997.

[30] Jaroszewicz, S. and Simovici, D.A., “Pruning redundant association rules using maximum entropy principle”, In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, pp. 135-147, May 2002.

[31] Kaur, H. and Wasan, S.K., “Empirical study on applications of data mining techniques in healthcare”, Journal of Computer Science, 2(2), pp.194-200, 2006.

[32] Kaya, M. and Alhajj, R., “Genetic algorithm based framework for mining fuzzy association rules”, Fuzzy sets and systems, 152(3), pp.587-601, 2005.

[33] Klawonn, F. and Keller, A., “Fuzzy clustering based on modified distance measures”, In International Symposium on Intelligent Data Analysis, Springer Berlin Heidelberg, pp. 291-301, August1999.

[34] Koh, H.C. and Tan, G., “Data mining applications in healthcare”, Journal of healthcare information management, 19(2), p.65, 2011.

[35] Lopez, F.J., Blanco, A., Garcia, F., Cano, C. and Marin, A., “Fuzzy association rules for biological data analysis: a case study on yeast”, BMC bioinformatics, 9(1), p.1, 2008.

[36] Mangalampalli, A. and Pudi, V., “Fuzzy association rule mining algorithm for fast and efficient performance on very large datasets”, In Fuzzy Systems, FUZZ-IEEE 2009. IEEE International Conference on, pp. 1163-1168, August 2009.

[37] Margahny, M.H. and Mitwaly, A.A., “Fast algorithm for mining association rules”, In the conference proceedings of AIML, CICC, pp (36-40) Cairo, Egypt (pp. 19-21), December 2005.

[38] Oladipupo, O.O., Ayo, C.K. and Uwadia, C.O., “A Fuzzy Association Rule Mining Expert-Driven (FARME-D) approach to Knowledge acquisition”, African Journal of Computing & ICT, 5(5), pp.53-60, 2012.

[39] Pavel Berkhin, “A survey of Data Mining Techniques Grouping Multidimenstional Data –Recent advances in Clustering”, Springer, ISBN 978-3-540-28348-5, pg.no.25-71, 2006.

[40] Pi, D., Qin, X. and Wang, Q., “Fuzzy Clustering Algorithm Based on Tree for Association Rules”, International Journal of Information Technology, 12(3), pp.43-53, 2006.

[41] Popescu, B.; Popescu, A; Brezovan, M.; Ganea, E.; Decebal, B. and Romania, C., “Unsupervised Partitioning of Numerical Attributes Using Fuzzy Sets”, In Fed CSIS, pp. 751-754, September 2012.

[42] Ramakrishnan, N., Hanauer, D. and Keller, B., “Mining electronic health records”, Computer, 43(10), pp.77-81, 2010.

[43] Sarawagi, S. and Thomas, S., “Mining Generalized Association Rules and Sequential Patterns Using SQL Queries”, In Proc. of KDD Conference, 1998.

[44] Shakir Khan, D.; Sharma, A.; Zamani, A.S. and Akhtar, A., “Data Mining For Security Purpose & Its Solitude Suggestions”.

[45] Srikant, R. and Agrawal, R., “Mining quantitative association rules in large relational tables”, In Acm Sigmod Record, Vol. 25, No. 2, pp. 1-12, June 1996.

[46] Tian Zhang, Raghu Ramakrishanan and Miron LivnyBIRCH, “An efficient Data Clustering Method for Very Large Databases”, Technical Report, Computer Science Dept science, University of Wisconsin Madison, 1995.

[47] Tien Dung Do, Siu Cheung Hui and Alvis Fong, “Mining Frequent Itemsets with Category- Based Constraints”, Lecture Notes in Computer Science, Volume 2843, pp. 76 - 86, 2003.

[48] Toshihiko Watanabe, “Fuzzy Association Rules Mining Algorithm Based on Output Specification and Redundancy of Rules”, IEEE, 978-1-4577-0653-0/11/$26.00 ©, 2011.

[49] Tseng, M., Lin, W. and Jeng, R., “Maintenance of Generalized Association Rules Under Transaction Update and Taxonomy volution”, Lecture Notes in Computer Science, Volume 3589, Pages 336 – 345, Sep 2005.

[50] V. Mary, Amala Bai and Dr. D. Manimegalai, “An Analysis of Document Clustering Algorithms”, ICCCCT-10.

[51] Virender Kumar Malhotra, Harleen Kaur and M.Afshar Alam, “An Analysis of Fuzzy Clustering Methods”, International Journal of computer Applications, pp. 0975 – 8887, Volume 94 – No. 19, May 2014.

[52] Wang, C., Tjortjis, C. and PRICES, “An Efficient Algorithm for Mining Association Rules”, Lecture Notes in Computer Science, Volume 3177, Pages 352 – 358, Jan 2004.

[53] Watanabe, T., “Fuzzy association rules mining algorithm based on output specification and redundancy of rules”, In Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, pp. 283-289, October 2011.

[54] Wojciechowski, M., Zakrzewicz, M., “Dataset Filtering Techniques in Constraint- Based Frequent Pattern Mining”, Lecture Notes in Computer Science, Volume 2447, pp.77-83, 2002.

[55] Yager, R.R., “Fuzzy Summaries in Database Mining”, Proc. Conf. Artif. Intell. Appl., pp. 265–269, 1995.

[56] Yanling Li.,Gang Li,”Fast Fuzzy c-Means Clustering Algorithm with spatial constraints for Image Segmentation”, Advances in Neural Network Research and Applications Lecture Notes in Electrical Engineering Volume 67,pp 431-438, 2010.

[57] Yu Steck, M. Lobur, Faisal M.E. Sardieh, M. Dombrova, V. Artsibasov, “Development and Study of Clustering Algorithms for Large sets of Data”, CADSM’2011, 23-25, Polyana-Svalyava (Zakarpattya), UKRAINE, pp. 202-204, February 2011.

[58] Yuan, Y., Huang, T., “A Matrix Algorithm for Mining Association Rules”, Lecture Notes in Computer Science, Volume 3644, Pages 370 – 379, Sep 2005.

[59] Yuepeng Cheng, Tong Li, Song Zhu, “Document Clustering Technique based on Term Clustering and Association Rules”, IEEE, 978-1-4244-6977-2/10 © 2010.

[60] Zadeh, L.A., “Fuzzy sets”, Information and control, 8(3), pp.338-353, 1965.

[61] Zadeh, L.A., “The concept of a linguistic variable and its application to approximate reasoning—I”, Information sciences, 8(3), pp.199-249, 1975.

[62] Zahra Farzanyar, Mohammadreza Kangavari, “Efficient mining of Fuzzy Association rules from the preprocessing dataset”, Computing and Informatics, Vol. 31, 331–347, 2012.

[63] Zhang, T., Ramakrishnan, R. and Livny, M., BIRCH, “an efficient data clustering method for very large databases”, In ACM Sigmod Record, Vol. 25, No. 2, pp. 103-114, June 1996.

[64] Zimmermann H. J., “Fuzzy Set Theory and Its Applications”, Kluwer Academic Publishers, 1991.