Optimized Weighted Association Rule Mining using Mutual Information on Fuzzy Data

[1]D.Sathyavani and [2]D.Sharmila

[1]Department of IT, Maharaja Engineering College, Avinashi
sathyavani.it@gmail.com
[2]Proffesor & Head - EIE, Bannari Amman Institute of Technology, Sathy
sharmiramesh@rediffmail.com

Abstract: *Association rule mining is used to find the frequent item sets in large database. Generally Apriori algorithm used to find association rules for single dimensional database. Due to the candidate set generation in large database, it decreases the mining efficiency. As well as different items having different weight on its transaction. In this paper, we introduce a novel technique, called Weighted Fuzzy Association Rule Mining using Mutual Information (WFARMI), for mining association rules using fuzzy set theory. This algorithm avoids the costly generation of a large number of candidate sets. Mutual Information is used to provide the strong relationship among the attributes and assigning weight on the fuzzy data. The paper concludes with shows that the proposed algorithm is capable of discovering meaningful and useful weighted fuzzy association rules in an effective manner. Also speeding up the mining process and obtaining most of the high confidence.*

Keywords: Fuzzy Association Rules, Mutual Information, fuzzy theory, frequent Item set, Weighted Fuzzy Association Rule

1. Introduction

Data mining is a technique to extracting the data's from the large database. Association rule mining discovers the large transaction databases for association rules which provide the implicit relationship among data attributes. Association Rule Mining is used to find the interestingness relationship among the items and generate the association rules. The database DB consists of a set of transactions. In each transaction I contain many subsets. An association rule is an inference of the form XY with, and X ∩ Y = 0. The meaning of the rule is likely containing items in Y. Two methods are support and confidence used to determine whether an association rule is interesting. An association rule, AB, has support S% in DB if S% of transactions in DB contains items in A U B. The association rule is said to have confidence c% among the transactions containing items in A, there are c% of problem is to find all association rules which satisfies predefined minimum support and minimum confidence constraints [3].

In the existing system, Quantitative Association Rule mining, quantitative attributes must be discretized into number of intervals to determine the quantitative association rule. In Boolean, items are assumed to have only two values as 0 and 1. And those are referred as Boolean attributes. If the item is available in a transaction, then the attribute value will be 1; otherwise the value will be 0. Many interesting and efficient algorithms have been proposed for mining association rules for these Boolean attributes. Most data items do not come with preassigned weights [1]. The weighted items and its transaction shown in below table (Table 1 and Table 2).

Table 1. Weighted items database

ID	Item	Profit	Weight
1	Blazer	30	0.3
2	Shirt	40	0.4
3	T-shirt	20	0.2

Table 2. Transactions

TID	Items
1	1,2
2	2,3
3	1,2,3

In the proposed system, WFARMI algorithm has been used to avoid the costly generation of candidate sets. The fuzzy association rules which provide a smooth boundary, where each attribute will have a fuzzy set. The potential frequent item sets also discovered. The clustering technique is used to find the similar group of items among the frequent item set. Based upon the minimum support and minimum confidence, weights will be applied. Since it reduces the costly generation and increase the speed of the generation and performance.

2. Related Works

The QAR mining algorithm is to find the association rules by partitioning the attribute domain combining adjacent partitions and then transforming the problem into a binary state. It suffers from two problems. Mining fuzzy association rules for quantitative values has been considered based on the Apriori algorithm [12]. First problem is, it must be combined the consecutive intervals of a quantitative attribute to gain sufficient support. Second problem is suffered by the sharp boundary between intervals. In order to overcome this problem, Mining fuzzy association rules for quantitative values has been used by a number of researchers [8][13]. Then proposed a method to find the fuzzy sets based on many clustering methods. Each of these researchers considered all attributes as same.

Gyenesei et al. introduces the problem of mining weighted quantitative association rules based on fuzzy approach. Gyenesei [6] considered this issue and used weighted quantitative association rule mining based on fuzzy concept and proposed two methods with and without normalization. He assigns weights to the fuzzy sets to reflect their importance to the user and proposes two different definitions of weighted support: with and without normalization similar to his previous method.

Clustering is the process of grouping the similar item in a dataset. Similarities are commonly said in terms of how close to the item in their characteristics such as space and distance function. The quality of the cluster may be represented by its diameter and the maximum distance between the items in the cluster group. This method is used as preprocessing step for generating mining association rule in large database [16].

In this paper we used k-means clustering algorithm to group the similar items. The Mutual Information [3] is used to represent a majority of frequent item sets. Also by utilizing the cliques in the MI graph, frequent item set will be computed. This method automatically clusters the values of a given quantitative attribute in order to obtain large number of large item sets in short duration. By applying the weight on the item, it does not require the support and confidence value by the user. But finding all frequent item sets in large databases with this algorithm requires multiple database scans. So using complicated data structures that requires extra space, leads to computation and time complexity. The Weighted Fuzzy Association Rule mining using Mutual Information is used to increase the performance of the mining process. Also generate the rule efficiently in short duration.

3. Basic Concepts

In this section, we present the basic concepts of Fuzzy Weighted Association Rule mining using Mutual Information.

1. Definitions

A Fuzzy Weighted Association Rule mining(FWAR), R, is an implication of the form *XY*, where X and Y are item sets, and *attr(X)* ∩ *attr(Y)* = . *X* and *Y* are called the antecedent and the consequent of R, respectively. We define the attribute set of R as *attr(R)* = *attr(X)attr(Y)* [3].

2. Entropy and Mutual Information

A) Entropy

Entropy is an information theory, which is used to measures the uncertainty in a random variable. Entropy and mutual information are closely related in rule mining concept [3].

B) Mutual Information

Mutual information describes that how much information one random variable tells about another one. The MI graph represents a highest priority of the frequent item sets. Clustering is a concept used in many applications. It uses mutual information (MI) as a similarity measure and discovers its grouping property: The Mutual Information between three objects such as X, Y, and Z is equal to the sum of the MI between X and Y, and the MI between Z and the combined object (XY) [3].

4. Rule Construction

To find the fuzzy association rules, a mining algorithm was proposed based on the concept of large item sets [4][8]. It transfers each quantitative item into fuzzy membership values and uses fuzzy operations to find fuzzy rules. It extracts the association rules in two ways. In the first phase, candidate item sets are generated, and counts by scanning the transaction in large database. In the second phase, an association rules are generated from the large item sets found in the first phase. This approach consists of below steps.

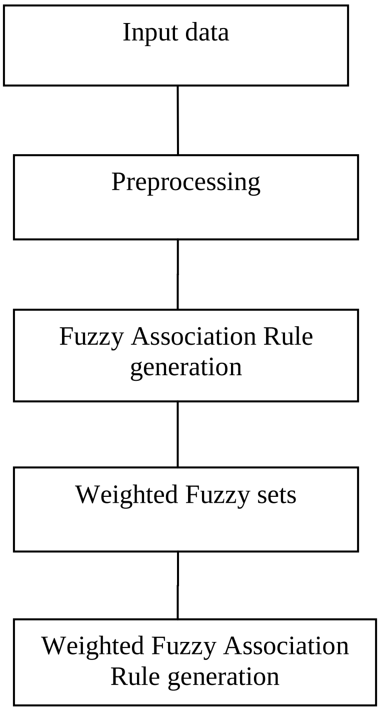
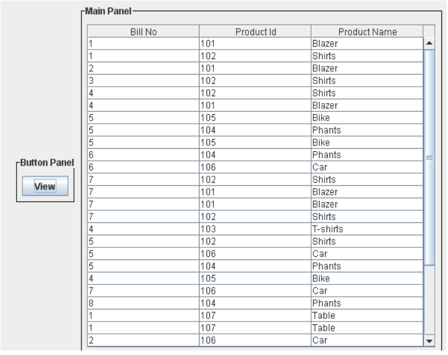

Input data
Preprocessing
Fuzzy Association Rule generation
Weighted Fuzzy sets
Weighted Fuzzy Association Rule generation

Fig1: Steps of proposed algorithm

All the possible combination ways of association rules for each large item set are formed, and the ones with their calculated confidence values larger than a predefined threshold (*minconf*) are output as desired association rules[6][8]. Figure 2 shows the input data collected from the transaction database.



Bill No	Product Id	Product Name
1	101	Blazer
1	102	Shirts
2	101	Blazer
3	102	Shirts
4	102	Shirts
4	101	Blazer
5	105	Bike
5	104	Phants
5	105	Bike
6	104	Phants
6	106	Car
7	102	Shirts
7	101	Blazer
7	101	Blazer
7	102	Shirts
4	103	T-shirts
5	102	Shirts
5	106	Car
5	104	Phants
4	105	Bike
7	106	Car
8	104	Phants
1	107	Table
1	107	Table
2	106	Car

Fig 2. Data View

The proposed method for discovering the association rules from preprocessing in terms of fuzzy terms from quantitative values and weight calculation is shown in above figure.

4.1 Fuzzy preprocessing

In this section, Lotfi et al. describes the fuzzy pre-processing methodology and fuzzy measure that are used for the actual fuzzy ARM process. We use a pre-processing approach which makes the fuzzy partitions into numerical attributes. This approach requires very less manual intervention for very large datasets. The fuzzy sets used in most real-life datasets. Those are differed in heterogeneous datasets. But, in our preprocessing technique is able to generate such Gaussian-like fuzzy datasets from any real-life dataset. Numerical data present in most real-life datasets translate into Gaussian like fuzzy sets, where in a particular data point can belong to two or more fuzzy sets simultaneously. And, this simultaneous membership of any data point in more than two fuzzy sets can affect the quality and accuracy of the fuzzy association rules generated using these data points and fuzzy sets [5].

Table 3. T-norms in fuzzy sets

T-norm
TM(*x*, *y*) = min(*x*, *y*)
TP(*x*, *y*) = *xy*
TW(*x*, *y*) = max(*x* + *y* - 1, 0)

The amount of fuzziness and Gaussian nature of fuzzy sets can be controlled using an appropriate value (-2) of the fuzziness parameter *m*.

4.1.1 Pre-processing Methodology

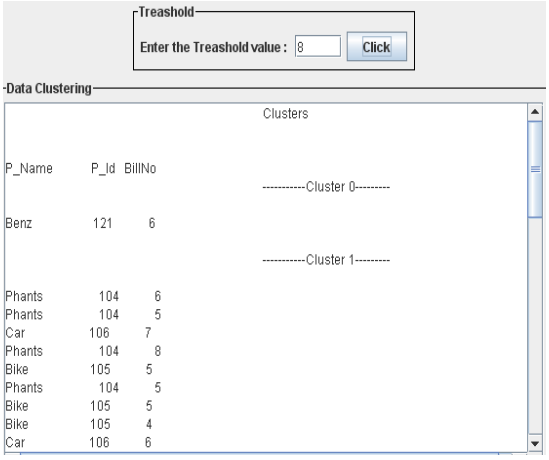
This pre-processing approach consists of two phases:

1. Generation of fuzzy partitions for numerical attributes
2. Conversion of a crisp dataset into a fuzzy dataset using a standard way of fuzzy data representation.

As part of pre-processing, we have used fuzzy k-means clustering in order to create fuzzy partitions from the dataset, such that every data point belongs to every cluster to a certain degree *μ* in the range [0, 1]. The algorithm tries to minimize the objective function:

$$\sum_{i=1}^{N}\sum_{j=1}^{C}\boldsymbol{\mu_{ij}}\,\|\mathbf{x_i}-\mathbf{c_j}\|^2$$

where *m* is any real number such that $1 \leq m < \infty$, μ_o is the degree of membership of *xi* in the cluster *j*, *xi* is the *ith* dimensional measured data, *cj* is the *d*-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. The fuzziness parameter *m* is an arbitrary real number (*m* > 1) [5].



Treashold

Enter the Treashold value : 8 Click

Data Clustering

```
                              Clusters


P_Name      P_Id  BillNo
                              -----------Cluster 0---------

Benz         121      6

                              -----------Cluster 1---------

Phants        104      6
Phants        104      5
Car          106      7
Phants        104      8
Bike         105      5
Phants        104      5
Bike         105      5
Bike         105      4
Car          106      6
```

Fig 3. Clustered Item sets

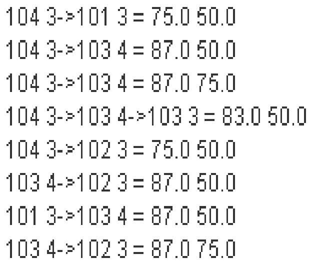
4.1.2 Clustering

Clustering can be used as a preprocessing step for mining association rules. There are many techniques for finding the clusters. In proposed method we have used fuzzy k-means clustering. It has been used for finding the membership for each attribute value, divides the values of each attribute into k-clusters (Figure 3).

4.2 Rule Generation and Rule selection

This phase deals with the generation and optimization of the rules. The combination of a pair of rules must be followed below conditions:

1) The consequent of two rules must be identical.
2) The rules must not contain similar antecedents on their left-hand sides.
3) The normalized mutual information of antecedents of two rules is greater than μ.



```
104 3->101 3 = 75.0 50.0
104 3->103 4 = 87.0 50.0
104 3->103 4 = 87.0 75.0
104 3->103 4->103 3 = 83.0 50.0
104 3->102 3 = 75.0 50.0
103 4->102 3 = 87.0 50.0
101 3->103 4 = 87.0 50.0
103 4->102 3 = 87.0 75.0
```

Fig 4. Rule generation

4.3 Algorithm in details

The proposed algorithm converts each quantitative value into a fuzzy set with linguistic terms using fuzzy membership functions. And then it calculates the *Normalized Mutual Information* of each attribute on all the transaction data. Using these NMI extracts search space after assigning the weights on each attribute. The detail of the proposed mining algorithm is described as follows:

1. Computation all the values of normalized mutual information between each distinct pair of attributes.

2. If a and b are two adjusted attributes then represents the strong information relationship between the attributes in a WFAR mining problem. We also given the user with the flexibility to specify the threshold value μ to assigns the

value in the range between [0, 1]. Based upon to the user's requirement, this provides the strongest relationship between the attributes.

4.4 Experimental Result

We evaluated the performance of our algorithm with real datasets. We use Mining Weighted Fuzzy Association Rules using Mutual Information (MFARMI) algorithm for comparison on the efficiency of the algorithms. The datasets are chosen from the commonly used transaction database.

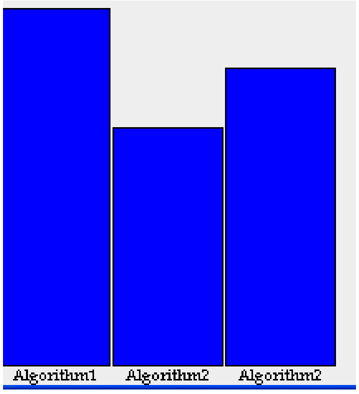


Fig 5. Comparison of algorithms

In proposed algorithm does not require minsup and minconf measures for evaluation the rules. But to compare the runtime of algorithms, we generate various sets of WFARMIs at the minimum confidence and minimum support thresholds. The number of association rules decreases along with an increase in minsup under a given specific minconf, which shows an appropriate minsup [3], in this approach it can constraint the number of association rules. And avoids the maximum occurrence of some association rules.

5. Conclusion

In order to overcome a QAR problem we can use weighted fuzzy association rules using mutual information(WFARMI) which provide a smooth boundary, where each attribute will have a fuzzy set. To get exact fuzzy association rules, information-theoric measure will be used. Hence new measure discovers potential frequent item sets. Fuzzy logic also employed and assigns weight on each item. These new measures, will lead to an efficient and scalable algorithm. Also speeding up of the rule generation.

References

[1] Ke Sun and Fengshan Bai, "Mining Weighted Association Rules without Preassigned Weights", IEEE transaction on knowledge and data engineering, Vol.20, no.2.pp 489-495, 2008.

[2] D.Saravana Kumar, N.Ananthi, D.Yadavaram, "New Approach to Weighted Pattern Sequential Mining-Dataset", International Journal of Scientific & Engineering Research Volume 2, Issue 5, 2011.

[3] S.Lotfi, M.H.Sadreddini, "Mining Fuzzy Association Rules using Mutual Information", In proceeding of international multi conference in Engineers and Compter Scientists 2009 Vol I.

[4] Praveen Arora, R.K.Chauhan, Ashwani Kush, "Frequent Itemsets from Multiple Datasets with Fuzzy data", International Journal of Computer Theory and Engineering, Vol.3, No. 2, 2011.

[5] Ashish Mangalampalli, Vikram Pudi, "Fuzzy Association Rule Mining Algorithm for Fast and Efficient Performance on Very Large Datasets", IEEE International Conference on Fuzzy Systems, Report No:IIIT/TR/2009/173, 2009.

[6] Gyenesei A, "Mining weighted association rules for fuzzy quantitative items", TUCS Technical Report No. 346, 2000.

[7] P. Bosc and O. Pivert, "On some fuzzy extensions of association rules", Proceedings of IFSA-NAFIPS 2001, Piscataway, NJ, IEEE Press, 2001.

[8] C. Kuok, A. Fu and H. Wong, "Mining fuzzy association rules in databases", ACM SIGMOD Record, 27, 1998.

[9] R. Ladner, F.E. Petry and M.A. Cobb, "Fuzzy set approaches to spatial data mining of association rules", Transactions in GIS, 7, 2003.

[10] J. Shu, E. Tsang and D. Yeung, "Query fuzzy association rules in relational databases", In Proceedings of IFSANAFIPS 2001 Piscataway, NJ, IEEE Press.

[11] Kaya M, Alhajj R, "Facilitating fuzzy association rules mining by using multi-objective genetic algorithms for automated clustering", In: Proceedings of the third IEEE international conference on data mining (ICDM'03),2003.

[12] D.L. Olson, Yanhong Li, "Mining Fuzzy Weighted Association Rules", Proceedings of the 40th Hawaii International Conference on System Sciences, 2007.

[13] Sanobe Shaikh, Madhri Rao and S.S.Mantha, "A new Association Rule Mining Based on frequent Item set" , David Bracewell, AIAA 2011, CS & IT 03, pp.81-95, 2011.

[14] G.Vijay Krishna and P.Radha Krihna, "A novel approach for statistical and fuzzy association rule mining on quantitative data," Journal of Scientific and Industrial Research, Vol.67, pp.512-517, 2008.

[15] M. Sulaiman Khan, Maybin Muyeba and Frans Conen, "Weighted Association Rule Mining from Binary and Fuzzy Data," P.Perner(Ed): ICDM LNAI 5077, pp.200-212, 2008.

[16] Tzung-Pei Hong, Ming-Jer Chiang and Shyue-Liang Wang, "Data Mining with Linguistic Thresholds," Int. J. Contemp.Math. Sciences, Vol.7, no.35, 1711-1725, 2012.

[17] Saket Agawal and Leena Singh, "Mining Fuzzy Association Rule using Fuzzy ARTMAP for Clusteing ," JERS/Vol.II/Issue I/pp.76-80, 2011.

[18] Partima Gautam, Neelu Khare and K.R.Pardasani, “A model for mining multilevel fuzzy association rule in database,” Journal of Computing, Vol. 2, Issue I, Jan 2010.