Design & Analysis of Fuzzy based Association Rule Mining

Anubha Sharma
Department of Computer Science & Engineering
Shriram College of engineering & Management [SRCEM]
Banmore, Gwalior (MP), India

Nirupama Tiwari
Department Of computer Science & Engineering
Shriram College Of Engineering & Management[SRCEM]
Manmore, Gwalior(MP), India

ABSTRACT

Data mining is sorting through data to identify patterns and establish relationships. Association rule mining is a well established method of data mining that identifies significant correlations between items in transactional data. Measures like support count, comprehensibility and interestingness, used for evaluating a rule can be thought of as different objectives of association rule mining problem. In this paper we proposed efficient fuzzy apriori association rule mining technique to find all co-occurrence relationships among data items. Our technique has three steps. Firstly, Transform the quantitative values of each transaction into fuzzy sets. calculate the membership values of each attribute in a transaction by applying the fuzzy membership function. In third step employs techniques for mining of fuzzy Apriori Associate rules. We also find fuzzy Apriori Association rule measured by Leverage. Our experimental results showed better performance than previous work.

KEYWORDS: Data Mining, Association Rule Mining, Fuzzy Apriori Association, Fuzzy Apriori Association with Leverage.

1.INTRODUCTION

In Association, the relationship of a particular item in a data transaction on other items in the same transaction is used to predict patterns. In Classification, the methods are intended for learning different functions that map each item of the selected data into one of a predefined set of classes. Given the set of predefined classes, a number of attributes, and a “learning (or training) set,” the classification methods can automatically predict the class of other unclassified data of the learning set. Cluster analysis takes ungrouped data and uses automatic techniques to put this data into groups. Clustering is unsupervised, and does not require a learning set. It shares a common methodological ground with Classification [2].

Prediction analysis is related to regression techniques. The key idea of prediction analysis is to discover the relationship between the dependent and independent variables, the relationship between the independent variables. Sequential Pattern analysis seeks to find similar patterns in data transaction over a business period. Existing algorithms [2] and [10-11] for mining association rules are mainly worked on a binary database, termed as market basket database. On preparing the market basket database, every record of the original database is represented as a binary record where the fields are defined by a unique value of each attribute in the original database. The fields of this binary database are often termed as an item. For a database having a huge number of attributes and each attribute containing a lot of distinct values, the total number of items will be huge. Storing of this binary database, to be used by the rule mining algorithms, is one of the limitations of the existing algorithms. Another aspect of these algorithms is that they work in two phases. The first phase is for frequent item-set generation. Frequent item-sets are detected from all-possible item-sets by using a measure called support count (SUP) and a user-defined parameter called minimum support. Support count of an item set is defined by the number of records in the database that contains all the items of that set. Selecting better rules from them may be another problem. After detecting the frequent item-sets in the first phase, the second phase generates the rules using another user-defined parameter called minimum confidence [2] and support [3-5]. In this paper we proposed efficient fuzzy apriori leverage association rule mining technique to find all co-occurrence relationships among data items.

1.1 Association Rules

Association rules are if and then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts, an antecedent (if) and a consequent (then). Association rule is expressed as X=>Y, where X is the antecedent and Y is the consequent. Each association rule has two quality measurements, support and confidence. Support implies frequency of occurring patterns, and confidence means the strength of implication [1-3] and [7-10].

Associations: Conceptually, associations are sets of objects describing the relationship between some items (e.g., as an itemset or a rule) which have assigned values for different measures of quality. Such measures can be measures of significance (e.g., support), or measures of interestingness (e.g., confidence, lift), or other measures (e.g., revenue covered by the association).

1.2 Rules interestingness measures

The aim of the association rules is to reveal interesting relations between data. For that reason certain are used which evaluate the level of importance of each rule. These are:

Confidence: The confidence of an association rule is the proportion of the isolates that are covered by the LHS of the rule that are also covered by the RHS. Values of confidence near value 1 are expected for an important association rule [8-11].

Leverage: The leverage of an association rule is the proportion of additional isolates covered by both the LHS and RHS above those expected if the LHS and RHS were independent of each other. Leverage takes values inside [-1, 1]. Values equal or under value 0, indicate a strong independence between LHS and RHS. On the other hand values near 1 are expected for an important association rule [1].

Lift: The lift of an association rule is the confidence divided by the proportion of all isolates that are covered by the RHS. This is a measure of the importance of the association that is independent of coverage.

2.RELATED WORK

Farah Hanna AL-Zawaidah and Yosef Hasan Jbara and Marwan AL-Abed Abu-Zanona et. al. presented a novel association rule mining approach that can efficiently discover the association rules in large databases. The proposed approach is derived from the conventional Apriori approach with features added to improve data mining performance. They had performed extensive experiments and compared the performance of the algorithm with existing algorithms found in the literature. They developed a visualization module to provide users the useful information regarding the database to be mined and to help the user manage and understand the association rules. Future work includes: 1) Applying the proposed algorithm to more extensive empirical evaluation; 2) applying the developed approach to real data like retail sales transaction and medical transactions to confirm the experimental results in the real life domain; 3) Mining multidimensional association rules from relational databases and data warehouses (these rules involve more than one dimension or predicate, e.g. rules relating what a customer shopper buy as well as shopper's occupation); 4) Mining multilevel association rules from transaction databases [1].

Praveen Arora, R. K. Chauhan and Ashwani Kush et. al. is to find Association rules from large Data warehouses are becoming increasingly important. In support of this trend, the paper proposes a new model for finding frequent itemsets from large databases that contain tables organized in a star schema with fuzzy taxonomic structures. The paper focuses on finding rules from multiple tables that contain fuzzy data and are arranged in star schema. If traditional data mining algorithms are used to discover association rules in such a case then the join of these tables will affect the efficiency and cost of the algorithm used. To overcome this problem the study focuses on extending the traditional algorithms in such a way that the mining of multi level fuzzy association rules become fairly simple [3].

K.Sangeetha , Dr.P.S.Periasamy , S.Prakash et. al. present the main focus of this research work is to propose an improved association rule mining algorithm to minimize the number of candidate sets while generating association rules with efficient pruning time and search space optimization. The relative association with reduced candidate item set reduces the overall execution time. The scalability of this work is measured with number of item sets used in the transaction and size of the data set. Further Fuzzy based rule mining principle is adapted in this work to obtain more informative associative rules and frequent items with increased sensitive. The requirement for sensitive items is to have a semantic connection between the components of the item-value pairs. The effectiveness of item-value pairs minimizes the search space to its optimality. Optimality of the search space indicates the tradeoff between pruning time and size of the data set. [4].

In this paper The problem of mining quantitative data from large transaction database is considered to be an important critical task.Researchers have proposed efficient algorithms for mining of frequent itemsets based on Frequent Pattern (FP) tree like structure which outperforms Apriori like algorithms by its compact structure and less generation of candidate itemsets mostly for binary data items from huge transaction database.Fuzzy logic softens the effect of sharp boundary intervals and solves the problem of uncertainty present in data relationships.This proposed approach integrates the fuzzy logic in the newly invented tree-based algorithm by constructing a compact sub-tree for a fuzzy frequent item significantly efficient than other algorithms in terms of execution times, memory usages and reducing the search space resulting in the discovery of fuzzy frequent itemsets [5].

This work presents a new foundational approach to Fuzzy Weighted Associative Classifiers where quantitative attributes are discritized to get transformed binary database. In such data base each record fully belongs to only one fuzzy set [6].

In this paper we proposed the efficient algorithm for mining fuzzy association rules. The FCBAR algorithm creates cluster table to aid discovery of fuzzy large itemsets. Contrasts are performed only against the partial cluster tables that were created in advance. In this paper we proposed the efficient algorithm for mining fuzzy association rules. The FCBAR algorithm creates cluster table to aid discovery of fuzzy large itemsets. Contrasts are performed only against the partial cluster tables that were created in advance [7].

3.PROPOSED TECHNIQUES

In this paper we proposed efficient fuzzy apriori association rule mining technique to find all co-occurrence relationships among data items. Our technique has three steps. Firstly, Transform the quantitative values of each transaction into fuzzy sets. calculate the membership values of each attribute in a transaction by applying the fuzzy membership function. In third step employs techniques for mining of fuzzy Apriori Associate rules. We also find fuzzy Apriori Association rule measured by Leverage.

Definitions:

Support

The rule X ⇒ Y holds with support s if s% of transactions in D contains X ∪ Y. Rules that have a s greater than a user-specified support is said to have minimum support.

Confidence

The rule X ⇒ Y holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence is said to have minimum confidence.

- ***Itemset:*** An itemset is a set of items. A k-itemset is an itemset that contains k number of items.
- ***Frequent itemset:*** This is an itemset that has minimum support.
- ***Candidate set:*** This is the name given to a set of itemsets that require testing to see if they fit a certain requirement [1] and [5].

3.1 Discovering Frequent Itemsets using Fuzzy Apriori Algorithm

The proposed of our method is the Fuzzy APriori algorithm. Our contributions are in providing novel scalable approaches for each building block. We start by counting the support of every item in the dataset and sort them in decreasing order of their frequencies. Next, we sort each transaction with respect to the frequency order of their items. We call this a horizontal sort. We also keep the generated candidate itemsets in horizontal sort. Furthermore, we are careful to generate the candidate itemsets in sorted order with respect to each other. We call this a down word sort. When itemsets are both horizontally and down word sorted, we call them fully sorted. As we show, generating sorted candidate itemsets (for any size k), both horizontally and down word, is computationally free and maintaining that sort order for all subsequent candidate and frequent itemsets requires careful implementation, but no cost in execution time. This conceptually simple sorting idea has implications for every subsequent part of the algorithm.

3.2 Candidate Generation

Candidate generation is the important first step in each iteration of Algorithm. Typically it has not been considered a bottleneck in the algorithm and so most of the literature focuses on the support counting. However, it is worth pausing on that for a moment. Modern processors usually manage about thirty million elementary instructions per second. We devote considerable attention to improving the efficiency of candidate generation, too.

Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers, clusters. Apriori, and improve it quite significantly by introducing what we call a down word sort.

3.3 Discovering Fuzzy Sets using Membership function

The traditional way to discover the fuzzy sets needed for a certain data set is to consult a domain expert who will define the sets and their membership functions. This requires access to domain knowledge which can be difficult or expensive to acquire. In order to make an automatic discovery of fuzzy sets possible, an approach has been developed which generates fuzzy sets automatically by clustering. This method can be used to divide quantitative attributes into fuzzy sets, which deals with the problem that it is not always easy do define the sets a priori. The proposed method uses a clustering method to find the *k* clusters. The whole process of automatically discovering fuzzy sets can be subdivided into four steps:

1. Transform the database to make clustering possible (the value of all the attributes has to be positive integer).
2. Find the *clusters* of the transformed database using a clustering method.
3. For each quantitative attribute, fuzzy sets are constructed using the medoids.
4. Generate the associated membership functions.

After discovering *k* medoids, we can compute *k* fuzzy sets out of them. We define $\{m1, m2, \ldots, mk\}$ as the *k* medoids from a database. The *i* -th medoid can be defined as $mi=\{ai1, ai2, \ldots, ai\ n\}$. If we want to discover the fuzzy sets for the *j* -th attribute, ranging from *min j* to *max j* , our mid-points will be $\{ai1, ai2, \ldots, ai\ n\}$. The fuzzy sets will then show the following ranges: $\{min_j - a2_j\}$, $\{a_{1j} - a_{3j}\}$, $\{a_{(i-1)j} - a_{(i+1)j}\}$, ... ,$\{a_{(k-1)j} - max_j\}$. Finally, the membership functions for the fuzzy sets have to be computed.

We can get our membership function looking at the definition of the sets above. For the fuzzy set with mid-point *akj* , the membership function looks as follows: If $x \leq a_{(k-1)j}$, the membership of *x* is 0. Also for $x \geq a_{(k-1)j}$, $\mu_x=0$ because in both cases, the value lies outside the range of the

fuzzy set. If x takes exactly the value of the mid-point a_{kj}, the membership is 1. For all other cases, we have to use a formula in order to compute the specific membership.
Generate membership functions (triangular function):

$$f_{ij}(x : \min_j, a_{\frac{k}{2}j}, \max_j) = \begin{cases} 1, & \text{if } a_{\frac{k}{2}j} = x \\ \dfrac{x - \min_j}{a_{\frac{k}{2}j} - \min_j}, & \text{if } \min_j \le x \le a_{\frac{k}{2}j} \\ \dfrac{\max_j - x}{\max_j - a_{(\frac{k}{2})j}}, & \text{if } a_{(\frac{k}{2})j} \le x \le \max_j \\ 0, & \textit{ortherwise} \end{cases}$$

A distinction between two types of fuzzy sets has been introduced. These two types are called equal space fuzzy sets and equal data points fuzzy sets. Equal space fuzzy sets are symmetrical and all occupy the same range in the universal set. In contrary, equal data points fuzzy sets cover a certain number of instances and thus are not symmetrical.

3.5 Algorithm for Fuzzy Apriori Association Rule Mining

The algorithm first searches the database and returns the complete set containing all attributes of the database. In a second step, a transformed fuzzy database is created from the original one. The user has to define the sets to which the items in the original database will be mapped. After generating the candidate itemsets, the transformed database is scanned in order to evaluate the support and after comparing the support to the predefined minimum support, the items with a too low support are deleted. The frequent itemsets F_k will be created from the candidate itemesets C_k. New candidates are being generated from the old ones in a subsequent step. Ck is generated from C_{k-1} as described for the Apriori algorithm in step 1. The following pruning step deletes all itemsets of Ck if any of its subsets does not appear in C_{k-1}.

The Fuzzy apriori mining Associate rules are composed of two steps:

1. Find all itemsets that have fuzzy support (FS<X,A>) above the user specified minimum support. These itemsets are called frequent itemsets.

2. Use the frequent itemsets to generate the desired rules. Let X and Y be frequent itemsets. We can determine if the rule X => Y holds by computing the fuzzy confidence FC<<X, A>,<Y,B>> and this value is larger than the user specified minimum confidence value.

$$FS_{<X,A>} = \frac{\sum_{t_i \in T} \Pi x_{j \in X} \mu_{xj}(a_j \in A, t_i.x_j)}{|D|}$$

D = {t_1, t_2, …, t_n}: transactions, <X,A> with X and Y is attributes and A and B is the corresponding fuzzy sets in X.

$$FC_{<<X,A>,<Y,B>>} = \frac{\sum_{t_i \in T} \prod_{z_j \in Z} m_{zj}(c_j \in C, t_i.z_j)}{\sum_{t_i \in T} \prod_{x_j \in X} m_{xj}(a_j \in A, t_i.z_j)}$$

Fuzzy Apriori Association rule measured by Leverage:

$$\boxed{\textbf{\textit{Leverage}} = \textbf{Pr}\ (\textbf{\textit{L}}, \textbf{\textit{R}}) - \textbf{Pr}\ (\textbf{\textit{L}}).\textbf{Pr}\ (\textbf{\textit{R}})}$$

$$FL_{<<X,A>,<Y,B>>} = \sum_{t_i \in T} \prod_{z_j \in Z} m_{zi}(c_i \in C, t_i.z_i) - \sum_{t_i \in T} \prod_{x_j \in X} \mu_{xj}(a_j \in A, t_i \cdot x^j) \cdot \sum_{t_i \in T} \prod_{y_j \in Y} \mu_{yj}(b_j \in B, t_i \cdot y_j)$$

The rules are discovered based on the specified threshold values for support. For each rule, the frequency counts for the LHS and RHS of each rule is given, as well as the values for confidence, lift, leverage, and conviction. Note that leverage and lift measure similar things, except that leverage measures the difference between the probability of co-occurrence of L and R as the independent probabilities of each of L and R, i.e.In other words, leverage measures the proportion of additional cases covered by both L and R above those expected if L and R were independent of each other.

Thus by applying the above rules and finding the interesting rules in a very compact and with accuracy.

4 EXPERIMENTAL RASULTS

We implement our proposed fuzzy apriori association rule mining technique using JAVA platform in which jdk1.6 version is used and NETBEANS IDE6.9 is used for the graphics and analysis. To evaluate our proposed mechanism for efficient association rule mining we first find the fuzzy apriori association rule, after then we measured fuzzy association rule using leverage. Table 1 represent the fuzzy apriori association rule(Leverage) with support, confidence, frequent set, generation time, fuzzy apriori of rules, total fuzzy apriori of nodes created,whenrun on Mushroom Dataset Having 8124 records.

Support/Confidence	**30/50**	**40/60**	**50/80**	**56/70**	**60/50**
No. Of Frequent Sets	**646**	**303**	**110**	**67**	**43**

#No.Of Rules	2977	1240	363	186	90
Generation Time	0.17	0.15	0.08	0.06	0.05
Total# Nodes Created	1017	544	260	199	166

Table 1: Fuzzy Apriori Association Rule With Leverage

Table 2 represent the Comparison of fuzzy aprori association leverage rule with Fuzzy Apriori(Conf) support, confidence, frequent set, generation time, fuzzy aprori of rules with leverage.

Support/Confidence	30/50	40/60	50/80	56/70	60/50
Apriori Like algorithms	13901	4570	633	371	266
Fuzzy Apriori(Confidence)	3056	1304	418	276	138
Fuzzy Apriori(Leverage)	2977	1240	363	186	90

Table 2: Comparison Of ARs On Diff algorithms.

Table 3 represents the comparison association rule value of previously proposed Algo., Fuzzy Apriori and our proposed Fuzzy Apriori with leverage techniques when run On Chess Datasets having records of 3124.

Support/Confidence	20/90	30/50	50/50	60/40	80/60
#Association Rule in Fuzzy apriori(Conf)	19778	9450	25062	14728	3678
#Association Rule in Fuzzy Apriori(Leverage)	19695	9371	25007	14680	3650
Generation Time In Fuzzy Apriori(Conf)	0.76	0.69	0.41	0.38	0.16
Generation time in Fuzzy Apriori(Leverage)	0.64	0.67	0.37	0.35	0.14

Table 3: Comparison Association Rule of previous work and proposed Fuzzy Apriori with leverage.

Figure 1 represents the comparison association rules graph of previously proposed Algorithms, and our proposed Fuzzy Apriori with leverage techniques when run with Mushroom Dataset. With respect to all the support values (30,40,50,56 and 60) our proposed Fuzzy Apriori and Fuzzy Apriori with leverage techniques shows better performance than other existing work.

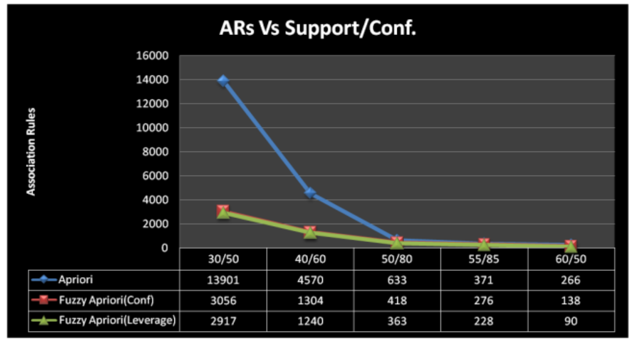

Figure 1 Comparison Association Rules graph of previous work and Fuzzy Apriori and proposed Fuzzy Apriori with leverage.

Figure 2 represents the comparison association rules graph of previously proposed Algorithms, and our proposed Fuzzy Apriori with leverage techniques when run with Mushroom Dataset. With respect to all the support values (30,40,50,56 and 60) and generation time our proposed Fuzzy Apriori and Fuzzy Apriori with leverage techniques shows better performance than other existing work.

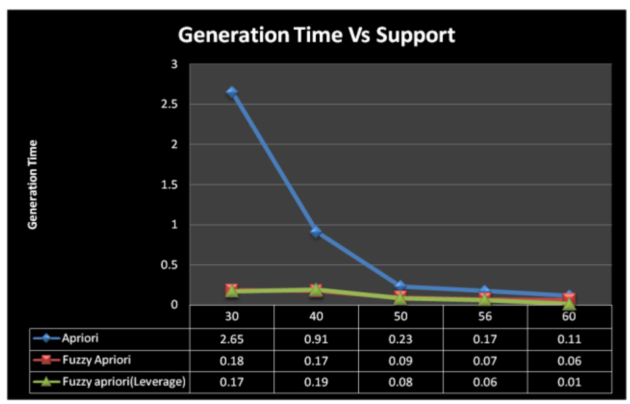

Figure 2 Comparison Generation Time graph of previous work and Fuzzy Apriori and proposed Fuzzy Apriori with leverage.

5 CONCLUSION AND FUTURE WORKS

Discovering association rules is at the heart of data mining. Mining for association rules between items in large database of sales transactions has been recognized as an important area of database research. These rules can be effectively used to uncover unknown relationships, producing results that can provide a basis for forecasting and decision making. In this paper we proposed efficient fuzzy apriori association rule mining technique to find all co-occurrence relationships among data items. Our technique has three steps. Firstly, the Apriori principle which allows considerably reduced the search space with discover the frequent item set. Secondly proposed an approach for finding fuzzy sets for quantitative attributes in a database and finally employs techniques for mining of fuzzy Apriori Associate rules and also find fuzzy Apriori Association rule measured by Leverage. We compare association rules of previously proposed Algorithms with our proposed Fuzzy Apriori and Fuzzy Apriori leverage techniques. With respect to all the support values our proposed Fuzzy Apriori and Fuzzy Apriori with leverage techniques shows better performance. The work presented in the paper can be extended for multi-level association rule mining.

6 ACKNOWLEDGEMENT

We wish to thank Dr. Anil Kishore Saxena (Director) & Prof. Manoj Ramiya (HOD of CSE), Shriram College of Engineering & Management, Gwalior, M.P., India, for his insightful comments and suggestions.

7 REFERENCES

[1] Farah Hanna AL-Zawaidah, Yosef Hasan Jbara and Marwan AL-Abed Abu-Zanona, "An Improved Algorithm for Mining Association Rules in Large Databases", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 7, 2011, pp. 311-316.

[2] Han, J., Kamber, M. (2001). "Data Mining: Concepts and Techniques", Harcourt India Pvt. Ltd.

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

[4] K.Sangeetha , Dr.P.S.Periasamy , S.Prakash," Efficient Parallel Pruning of Associative Rules with Optimized Search", IOSR Journal of Computer Engineering (IOSRJCE) ,volume no.3,pp.26-30.

[5] K.Suriya Prabha, R.Lawrance," Mining Fuzzy Frequent itemset using Compact Frequent Pattern(CFP) tree Algorithm" International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.

[6] Reza Sheibani , Amir Ebrahimzadeh ,Member, IAUM," An Algorithm For Mining FuzzyAssociation Rules", Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol I, March, 2008,pp.486-490.

[7] G Vijay Krishna,P Radha Krishna,"A Novel Approach for Statistical and Fuzzy association Rule Mining on Quantiative Data ",Journal of scientific and industrial Reasearch ,vol no.67,jul2008,pp.512-517.

[8] Ashish Mangalampalli, Vikram Pudi," FPrep: Fuzzy Clustering driven Efficient Automated Pre-processing for Fuzzy Association Rule Mining", IEEE Intl Conference on Fuzzy Systems (FUZZ-IEEE), July 2010.

[9] Sunita Soni, O.P.Vyas,"Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care Data Mining", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.1, February 2012.

[10] Duke Hyun Choi , Byeong Seok Ahn , Soung Hie Kim, Prioritization of association rules in data mining:

Multiple criteria decision approach, Expert Systems with Applications: An International Journal, v.29 n.4, p.867-878, November, 2005.

[11] Choi et al., (2005).Prioritization of association rules in data mining: Multiple criteria decision approach. Expert Systems with Applications, v29, 867-878.