A comparative study of fuzzy logic and weighted Association rule mining in frequent datasets

K. Rubia
MPhil, Research Scholar,
Department of M.C.A
Sree Saraswathi Thyagaraja College
Pollachi. Coimbatore Dist,
Tamilnadu,India.
rubiyasuresh@gmail.com

S. Sasikala MCA., MPhil, [PhD]
HOD, UG Computer Science
Sree Saraswathi Thyagaraja College
Pollachi, Coimbatore Dist,
Tamilnadu, India.
Sasivenkatesh04@gmail.com

Abstract---Mining of frequent patterns in transaction databases has been a fashionable area of research. Many methods are being used to solve the trouble of discovering association rules among items in large databases [8]. Transaction pattern base have been introduced to reduce the number of passes over the database. Here considers the problem of using support for generating association rule. The classical associations rule mining frameworks assume that all items have the same significance that their weight within a transaction is the same which is not always the case. Here proposed the use of weighted support along with the transaction pattern which increases the efficiency in generating association rule using matrix manipulation[5] [3].

Keywords: Data mining, Associations rule mining, FP Tree growth algorithm, WARM, Fuzzy Logic.

I. INTRODUCTION

Data mining tools expect prospect trends and behaviors, allowing businesses to make practical, knowledge-driven decisions. The computerized, probable analysis offered by data mining move clear of the analyses of history events provided by presentation tools representative of decision support systems. Data mining tools can answer business question that conventionally were time consuming to determination [6]. They clean databases for unseen patterns, finding analytical information that expert may miss because it lies external their expectations. Data mining is the process that attempt to discover pattern in large data sets. It utilizes methods at the connection of artificial intelligence, statistics, machine learning and database systems. The generally goal of the data mining process is to take out in sequence from a data set and transform it into logical structure for supplementary use [6].

II .ASSOCIATION RULES

Association rules are used to learn basics that co-occur normally within a dataset consisting of numerous autonomous selections of elements (such as purchasing transactions), and to discover rules [6]. Association rules are extensively used in various areas such as telecommunication networking, business and risk management, inventory control etc. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository [7]. An association rule has two parts, an antecedent which represent if part and a consequent which represents the then part. A predecessor is an item found in the data. A consequent is an item that is establish in combination with the predecessor [8].

Association rules uses two criteria support and confidence to identify the relationships and rules are generated by analyzing data for frequent if/then patterns. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rules are frequently required to gratify a user-specified weighted support and a user-specified minimum confidence at the same occasion. Association rule generation is usually ripped up into two separate steps [7]:

- Foremost, minimum support is useful to find all frequent item sets in a database.
- Next, these frequent item sets and the minimum confidence restriction are used to shape rules. While the next step is straightforward, the primary step needs more attention.

A support(S)

Support(S) of an association rule is distinct as the percentage/portion of records that contain X ∪Y to the total number of account in the database. Expect the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchase of this item [3].

Support= (Support count up of (XY))/ (Total amount of transaction in D)

B confidence(C)

Confidence(C) of an association rule is distinct as the percentage/portion of the number of transactions that hold X ∪Y to the total numeral of records that contain X. Confidence is a calculate of strength of the association rules, expect the confidence of the association rule X⇒Y is 80%, it means that 80% of the transactions that contain X also contain Y jointly [4].

Confidence (X | Y) = (Support (XY))/ (Support(X))

III. ASSOCIATION RULE MINING

In data mining, association rule mining is a trendy and well research method for discover motivating relations between variables in huge databases [3]. Association rules are introduced for discovering regularities between goods in large scale subject data recorded by point-of-sale systems in supermarkets. Association rule mining is to find out association rules that assure the predefined minimum support and confidence from a given database [6]. The difficulty is usually decayed into two sub problems. One is to find those item sets whose occurrences go above a predefined threshold in the database; those item sets are called common or large item sets. The second problem is to produce association rules from those huge item sets with the constraints of least confidence. Suppose one of the great item sets is Lk, Lk = {I1, I2… Ik}, association rules with this entry sets are generated in the following way: the first rule is {I1, I2, and Ik-1}! {Ik}, by checking the confidence this rule can be resolute as motivating or not. Then other rule are generated by delete the last items in the precursor and inserting it to the resulting, further the confidences of the new rules are checked to conclude the interestingness of them. Those processes iterated until the precursor becomes empty [6]. Since the second sub-problem is quite straight forward, most of the researches focus on the first sub-problem. The first sub-problem can be supplementary divided into two sub-problems: candidate big item sets generation process and frequent item sets production process. The item sets whose bear exceed the support entry is referred as large or frequent item-sets, those item sets that are probable or have the hope to be large or frequent are called contender item sets.

Normally, an association rules mining algorithm contains the following stepladder [7] [4].

- The set of candidate k-item sets is generated by 1-extensions of the large (k -1)-Item sets generated in the earlier iteration.
- Supports for the candidate k-item sets are generated by exceed more than the database.

Item sets that do not have the minimum support are useless and the remaining item sets are called large k-item sets Generally, an association rules mining algorithm contains the following steps [7] [4].

- The set of candidate k-item sets is generated by 1-extensions of the large (k -1)-Item sets generated in the previous iteration.
- Supports for the candidate k-item sets are generated by a pass over the database.

Item sets that do not have the minimum support are discarded and the enduring item sets are called large k-item sets.

IV .VARIOUS ASSOCIATION RULE MINING ALGORITHMS

A.AIS Algorithm

The AIS algorithm was the initial algorithm proposed for mining association rule. In this algorithm only one item following association rules are generated, which means that the resulting of those rules only contain one item [3]. The main problem of the AIS algorithm is too many nominee item sets that lastly turned out to be small are generated, which requires more space and wastes much effort that twisted out to be useless. At the same time this algorithm requires too many passes over the complete database.

B .Apriori Algorithm

Apriori is designed to activate on databases containing transactions. Additional algorithms are planned for sentence association rules in data having no transactions or having no time stamps [5]. Apriori attitude states that, if an item set is frequent, then all of its subsets must be frequent.

In A priori algorithm, the primary pass of the algorithm basically counts item occurrences to conclude the large 1-itemsets. A subsequent pass, say pass k, consists of two phases [5]. Primary the large item sets Lk-1 found in the (k-1)th pass are use to generate the candidate item sets Ck, using the A priori-generate function subsequently, the database is scanned and the support of candidates in Ck is counted.

C. Fp-Tree

FP-Tree, frequent pattern mining, is an additional milestone in the expansion of association rule mining, which breaks the main problem of the Apriori [6]. The regular item sets are generated with only two passes over the database and without any candidate generation process. FP-tree is an unlimited prefix-tree structure storing crucial, quantitative in turn about frequent patterns. Only frequent length-1 items will have nodes in the tree, and the tree nodes are agreed in such a way that more regularly occurring nodes will have better probability of sharing nodes than less regularly occurring ones.

V. WEIGHTED ASSOCIATION RULE MINING ALGORITHM

In Association Rule Mining through Matrix Manipulation using Transaction pattern base, transactional pattern base where transactions with identical pattern are added as their frequency is increased [8]. Thus succeeding scanning requires only scanning this compressed dataset which increases efficiency of the respective methods. This method is implemented using two-dimensional matrix as a substitute of using FP-Growth method, as used by most of the algorithms.

The computational cost of association rules mining can be summary by reducing the quantity of passes over the database. With the intention of necessary transactional pattern base every transaction of D is sorted to reduce the transactional base. This pruned transactional base is then altered to Transactional pattern base [7].

A Transaction

The following is a recognized statement of the problem [4] Let I = {i_1, i_2 … i m} be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that T⊆*I*. Associated with each transaction is a distinctive identifier, called its TID. We say that a transaction T contains X, a set of some items in I, if X ⊆ T. An association rule is an allusion of the form X⇒ Y, where X ⊂ I, Y ⊂ I, and X ∩ Y = ∅.The rule X⇒Y holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y. The rule X ⇒ Y has support s in the transaction set D, if s% of transactions in D contains X∪Y. Given a set of transactions D, the problem of mining association rules is to generate all association rules that have support and confidence larger than the user-specified minimum support (called min sup) and minimum confidence (called min confidence) correspondingly[6].

B Transactional Database

A transactional database is a DBMS where write operations on the database are capable to be *rolled back* if they are not concluded accurately. A transactional database TDB is a set of transactions [1]. A transaction T = (t id, X) is a tuple where t id is a transaction-id and X is an item set. The item set is called transactional pattern which has t id in the transactional database. The item sets Y and Z (such that Y! X! Z!) Are not transactional pattern if they do not have their own t id in TDB [3]. A pattern may belong to multiple transactions with different t ids

C Transactional Pattern

A transactional database is the set of transactions consisting of set of items I [4]. A transaction T = (t id, X) is a tuple where t id is a transaction-id A transaction pattern base is a set of patterns P. A Pattern P = (p id, X, f) is a tuple where p id is a pattern id, X is a transactional Pattern and f is the frequency of the transactional pattern, where f is the occurrence of the particular pattern in the transactional database [1].

f = Frequency(X, TDB):= count {t id | (t id, X) ∈TDB, X ⊆I}

It is experimental that the amount of the transactions in the transactional database is matching to the sum of frequencies of the transactional pattern base. All transactional patterns in the transactional pattern base will be exclusive.

D Transactional Pattern base

The transactional database is altered into a condensed data structure called transactional pattern base. Transactional pattern base consists of p id, item set X and frequency f of the patterns. Where pid is the inimitable identifier each pattern X and frequency is the amount of occurrences of the item set X in the transactional database, where from we assemble transactional pattern base[4] [6].

In sequence from transactional databases is necessary for generating association rules. If we can assemble transactional pattern base from the transaction database in the first scan, it may condense frequent scanning of database due its compressed size as compared to transactional database which contain surplus transactional pattern.

First scan the database to find the different items going on in the database and then make the transactional matrix by lettering all the transaction patterns next to the row side and all the frequencies taking place in the database along the column side[3] [6]. Now complete the Transactional matrix, if the transaction contains the item mentioned in the column then write 1 if not 0 in the row corresponding to that operation

VI. WARM CONCEPT

Weighted Association Rule Mining The idea of association rule mining, it proposed the support-confidence measurement framework and compact association rule mining to the discovery of frequent item sets. Much effort has been fanatical to the classical (binary) association rule mining problem since then. These algorithms strictly follow the classical measurement framework and produce the similar results once the minimum support and minimum confidence are given. WARM generalizes the traditional model to the case where items have weights. WARM Introduced weighted support of association rules based on the costs assign to both items as well as transactions. An algorithm called WIS was anticipated to derive the rules that have a weighted support larger than a given threshold. Weighted support in a similar way except that they only took item weights into account. The definitions bust the downward closure property. As a result, the planned mining algorithm became more difficult and time consuming, invent new measures (weighted support) based on these weights, and build up the corresponding mining algorithms. A directed graph is formed where nodes denote items and links represent association rules [2]. A generalized version of HITS is applied to the graph to grade the items, where all nodes and links are permitted to have weights. Anyway, it may be the first successful attempt to apply link-based models to association rule mining. However, the model has a limitation that it only ranks items but does not provide a measure like weighted support to evaluate an arbitrary item set. To overcome this we will go with fuzzy association rule mining.

VII .ASSOCIATION RULE MINING THROUGH FUZZY LOGIC

A Definition of fuzzy logic

A mathematical logic that attempts to solve troubles by transmission values to an inaccurate spectrum of data in order to turn up at the most correct conclusion possible. Fuzzy logic is designed to answer problems in the similar way that humans do: by allowing for all available in sequence and making the best probable decision given the input [1].

The conception of fuzzy association rule mining approach generated from the necessity to efficiently mine quantitative data frequently in attendance in databases. Algorithms for mining quantitative association rules have already been projected in classical association rule mining. Separating an attribute of data into sets cover certain ranges of values, engages the sharp frontier setback [1]. To overcome this problem fuzzy logic has been introduced in association rule mining. But fuzzy association rule mining also have some troubles.

During the fuzzy association rule mining method, the original data set is absolute with attribute values within the range (0, 1) due to the large number of fuzzy partitions are being done on each of the quantitative influence. To process this extended fuzzy dataset, some procedures are needed which are based on t-norms [4].In this way the fuzzy dataset E is shaped upon which the projected algorithm will work. The dataset is reasonably divided into p disjoint horizontal partitions$P1, P2,\ldots$,. Each partition is as large as it can vigorous in the available main memory. They have used the following notations,

- E=fuzzy dataset generated earlier than pre-processing
- Set of partitions$P= \{P1, P2,\ldots, PP\}$
- $Td\ [it]$ = t id list of itemset.
- μ = fuzzy relationship of any itemset
- $count[it]$ = growing μ of itemset *it* over all partitions in which *it* has been processed
- d = number of partitions (for any particular itemset *it*) that have been processed since the partition in which *it* was added.

VIII. ADVANTAGE OF FUZZY LOGIC

Permits fuzzy threshold, fuzzy sets are easily modified, relates input and output in linguistic term, allows rapid prototyping cheaper because easier to design, increase robustness[3].

Experiments were performed in two singular machine configurations. One with advanced configuration and the other one with minor configuration [4]. Developers have used the USCensus1990unrefined dataset which has 2.5 million transactions. And their fuzzy dataset can developed10 million transactions. So, it is obvious this amount of dataset handling in the accessible main memory is easy. In their dataset 6 are quantitative and 4 are binary. Results obtained on higher configuration system on USCensus1990unrefined dataset as the experimental dataset, using multiple support values range between 0.075 to 0.600. From their experimental results they have shown that their proposed algorithm is fast and efficient than the previous weighted association rule mining in ARM algorithms.

An Intelligent system solves domain specific problems. Association rule mining is basically of two types. One is traditional or crisp association rule mining and the other one is fuzzy association rule mining. Classical association rule mining make use of Boolean logic to translate numerical attributes into binary attributes by facilitate of sharp crunchy partitions .In which valuable data may become conflicting over these sharp partitions

[2]. Another difficulty with classical association rule mining is, here a user has to present a minimum support value for the mining purpose. And as we know that we humans are mistake prone. Any wrong setting of minimum support could end up in incorrect results. This can even cause the generation of huge number of surplus rules as well as useless rules. So, it is very a difficult task of surroundings an accurate minimum support rate by hand. That is why classical association rule mining is time consuming and fewer accurate process. Hence we go with fuzzy ARM.

Fuzzy association rule mining is reasonably a newer concept as we already said it was easy to design and too cheaper as compared transaction WARM method in which is applied to calculate minimum support and confidence between frequent dataset in association rule mining.

IX. AN REAL TIME EXAMPLE

In ARM model when the data are quantitative such as income, age, price, etc., which are very familiar in many real applications, in association rule mining. And ultimately apply the Apriori-type method. Thus, association rule like X −−> Y reflects association among nominal values of data items. Examples of such rules are "(Age, old), (diabetes, high) [3] (Leukemia, yes), "(profits, small) −−> (Age, medium)", and so on. Such the mining results are exaggerated by how the intervals are partitioned, mainly for data values around interval boundaries. That is the so-called "sharp boundary" problem.

Subsequently, the result of associative classification may also be affected in terms of correctness and understandability.

In medical domain there are number of quantitative attributes suffers from hard boundary problem. For example attributes for example if in a particular record the BMI (Body mass Index) is 41 them according to following discritization rule [3].

BMI [20-30] Obesity="kind"

BMI [31-45] Obesity="moderate"

BMI [40-65] Obesity="sever"

The patient is considered to be strictly obese. This may not give a correct result because of sharp frontier problem. As an alternative by applying fuzzy logic the patient is partially belonging to each fuzzy set. Hence the patient membership value to the fuzzy set should be(μ(Obesity, "gentle") = 0.1 , μ(Obesity, "temperate") = 0.3, μ(Obesity, "disunite") = 0.6). To deal with crisp boundary problem of quantitative attribute in ARM model the [2] proposed the Fuzzy WARM (FWARM) Algorithm and redefine the weighted support and weighted confidence to adapt in Fuzzy environment. In Fuzzy Weighted Association Rule Mining (FWARM) model the Fuzzy Weighted Support (FWS) and Fuzzy Weighted Confidence framework is proposed to mine Fuzzy Weighted Association Rule. The algorithm FWARM can be used to generate the CAR rules in Fuzzy Weighted environment. \

X .CONCLUSION

This paper focused on bringing a comparative study between WARM and fuzzy. On comparing these two association rule mining technique fuzzy provide higher performance on frequent datasets, low time consuming and less memory use .and it is cost less to develop because the advantages of fuzzy rule mining itself says that it is easier to design .

REFERENCES

[1] K.Sotiris, and D.Kanellopoulos, "association rules mining: A Recent Overview. GESTS International transactions on computer science and engineering", vol.32 (1), 2006, pp.7182.

[2] M.Suleman Khan, Maybin Muybea, M.Frans coenen, fuzzyweighted association rule mining with weighted support and confidence framework.2009.

[3] J.Han and M.kamber , data mining: concepts and techniques: the Morgan Kaufmann series,2001

[4] K.Sun and F.Bai, "mining associatin rule without preassigned weights," IEEE Trans.Knowledge and Data Eng., vol.20, no. 4, pp.489-495, Apr. 2008.

[5] R.Agarwal and R.Srikant, "fast algorithm for mining association rules," Proc.20th Int"I Conf.very large databases (VLDB "94),pp. 487-499,1994.

[6] Mining most interesting rules,R J Bayardo Jr. and R.Agarwal,Published by IEEE computer society, Proc. ACM SIGKDD 99,pp.145-154,1999.

[7] R.Agarwal,T.Imielinski, and Swami, " mining association rules between set of items in large dataases," Proc.ACM SIGMOD Int'1 Conf.Management of Data (SIGMOD'93) PP.207-216,19.

[8] Ashok Savasere, Edward Omiecinski,Shamkant Navathe[1995]-"an efficient algorithm for mining rules in large databases"-proceeedings of the 21st VLDB Conferene.