USER CONSTRAINTS IN DISCOVERING ASSOCIATION RULES MINING

Hassan M. Najadat*, Mohammad K. Kharabsheh**, Ismail Hmeidi*

* Faculty of Computer and Information Technology, Jordan University of Science and Technology, Irbid, Jordan
najadat@just.edu.jo, hmeidi@just.edu.jo

** Al-Huson College, Al-Balqa Applied University, Jordan
mohkh86@yahoo.com

ABSTRACT

This paper introduces a new algorithm called User Association Rules Mining (UARM) for solving the problem of generating inadequate large number of rules in mining association technique using a fuzzy logic method [1, 2]. In order to avoid user's defined threshold mistakes, the user has flexibility to determine constraints based on a set of features.

In comparison with other well-known and widely used association rules algorithms, such as Apriori algorithm, UARM attempts to give more enhancements to the problem and adopts significance of association rules mining to enhance the quality of the application by providing insightful clues to more effective decision-making.

Keywords: *Data Mining, Association Rules, Minimum Support, Fuzzy Logic*.

1. INTRODUCTION

Recently, a new paradigm of computer science has been revealed in order to find hidden patterns inside huge amount of data. This paradigm has been referred to as the Data Mining paradigm (DM); it is also known as the science of knowledge discovery [8].

Association rules are a collection of rules; each of which reflects a certain system behavior or decision. Algorithms that have the potential to form precise rules have pursuit the thinking of numerous researchers. Introduced association rules techniques involve discovering the relations between the database items. Such algorithms require the database be scanned in order to find the association among its items [10].

The mining process for association rules explores interesting relationships between data items that occur frequently together [19]. Many of research papers have been published presenting new algorithms or improvements on existing algorithms to solve such mining issue in an efficient and fast manner.

This paper introduces a new association rules algorithm that applies a fuzzy logic method to avoid using thresholds for the minimum support and confidence percentages. The proposed algorithms gather the whole data according to some user defined features which called UARM. The work is introduced to enhance the quality of generated rules by reducing the redundant rules and generating rules for rare data, and the method provides a good prediction levels when the application has different user defined features.

The paper is organized as follows. Section 2 presents the theoretical background of mining association rules. The problem statement is presented in section 3. User association rules algorithm is provided in section 4. Experiments and future work are provided in sections 5 and 6 respectively.

2. RELATED WORK

Mining association rules is the process of discovering the relations between the items in a database to help decision-making. This approach requires database scans to find the associated items. Many algorithms for mining association rules have been introduced since their introduction in 1993 [11].

The Apriori [8] algorithm is an influential algorithm for finding frequent itemsets using candidate generation. Apriori algorithm performs multiple passes over the database to guarantee all non-empty subsets of frequency itemsets. In the kth pass, the algorithm counts the support of the k-itemsets. After that, the candidate k+1-itemsets are generated using the previously induced frequent k-itemsets. Other itemsets are discarded and the process continues until no more candidate itemsets can be explored.

To attain the usefulness of association rules, a fuzzy approach [2, 3] is used to mine association rules in an efficient manner; this approach builds the fuzzy decision trees to discover the changes in the association rules by using linguistic variables and the change in the discovered rules. Association rules can benefit from the theory of fuzzy subsets by computing fuzzy approximate dependencies to detect possible existing relations between attributes levels in the database [4, 5].

Many association rules algorithms use only a single minimum support for the whole database, assuming that all items have the same frequency in the data. However, in many applications, some items appear very frequently in the data, while others would rarely appear. Therefore, selecting the best minimum support to find rules that involve rare items is considered as a challenge. Many algorithms that allow users to specify multiple minimum supports were also proposed [5,7].

3. PROBLEM STATEMENT

When studying association rules, several statistical measures are involved [8]. This includes support (strength of rule): number or percentage of transactions

that satisfy the rule, and confidence (significance of rule): number or percentage of transactions that satisfy the consequent of the rule given that they contain the antecedent part of the rule.

The process of discovering all association rules can be divided into two subdivisions: the first problem is how to find all itemsets that have a support greater than the minimum support, which will be used subsequently in any future scan to find the frequency of item in transactions; the second problem tries to find how to generate desired rules according to the confidence measurement.

We apply a fuzzy logic method in generating association rules to avoid using thresholds for the minimum support and confidence percentages. The new algorithm groups the whole data according to some user defined feature. Furthermore, this work employs a fuzzy logic technique with a clustering technique that is based on similarity of data to obtain the most efficient desired rules.

The major goal in this work is to enhance the quality of application; which has chosen from the King Abdullah University Hospital in Jordan (distributing the quantities of medicine and medical supply in the hospital).The experiments confirmed that association rules mining is desirable to develop the decision making process in this application.

4. USER ASSOCIATION RULE MINING (UARM)

It is believed that the problem in generating ARM is to avoid redundant rules and control the quality of them. Our algorithm grants the user ability to define an interested set of features in the preprocessing phase.

Interesting association rules are the ones that satisfy the minimum support and the minimum confidence. The support and confidence of a rule ($A \Rightarrow B$), A and B are set of features, can be defined as

$$support(A \Rightarrow B) = support(A \cup B)$$

and

$$confidence\ (A \Rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(A)}$$

Where support ($A \cup B$) is the number of transactions containing the itemset A and B, and support (A) is the number of transactions containing the itemset A [9].

The algorithm constructs a candidate set of itemsets based on user defined features, and then scans the database several times until all possible frequent itemsets are discovered.

The fuzzy logic technique is applied to provide a good man-machine interface that allows users to give a suitable minimum support. The Fuzzy logic techniques map the user minimum support in real support for every group of transactions independently.

UARM consists of two phases which include grouping the database D based on user defined feature, and then the algorithm predicates the suitable minimum support threshold based on user threshold. The step 11 and the second phase require the Apriori algorithm which can be found in [9].

UARM Algorithm
Input: Database, D of transactions; number of grouping ng; minimum support threshold, min-sup; confidence threshold, conf.
Output: Rules for each group, R;
Main Procedure of Phase 1:

1. Group the D to ng based on user defined feature;
2. Scan Dg for each group to get support of every 1-itemsets;
3. Let a ← 1/|Dg| {Dg for every group};
4. Let b ← {maximum support of all 1-itemsets};
5. Let Lean ← {the degree of lean};
6. Set objective functions of min-sup, lean, and real-sup, as FuncMin_Sup, FuncLean, FuncReal_sup, respectively;
7. Get two fuzzy concepts to describe min_sup and lean using objective function of min-sup and lean;
8. Generate several fuzzy rules based on the values of input parameters;
9. Select the desired fuzzy rules;
10. Get real support for each group by defuzzfing the fuzzy rules;
11. Call Apriori Algorithm [8] for each group to find frequent itemsets in Dg

Phase 2:
C*all Association Rules Algorithm* [8] for each group to find rules.

5. EXPERIMENTS AND RESULTS

A real datasets was gathered from King Abdullah University Hospital in Jordan. The dataset consist of 41249 transactions which list distributing the quantities of medicine and medical supply in the hospital. Hospital datasets contain the set of attributes which include (Item Code, Item Name, from store, To Store, Date: Issuing date and Qty: Quantity issued to store). The dataset was considered to test our algorithm, since each attribute gives the user a chance to group the data based on her preferred feature.

The experiments were done using Pentium 4 computer with a clock rate of 3000 MHz and 512 Mbytes of main memory.

The following is a sample of generated rules:

Rule 1:
Gauze abdominal 30cmx30cm → gauze, swabs,4"x4" mesh 24x20
Rule 2:
Bandage zinc oxide plaster 2 → gloves, examination latex med
Rule 3:
Gloves, examination latex med & syrng disp str 10ml w-need 2pc → syrng disp str 5ml w-need 2pc

Rule 4:

Syrng disp str 10ml w-need 2pc & syrng disp str 2ml w-need 2pc→ syrng disp str 5ml w-need 2pc

The rules' degree quality is used to compare UARM algorithm and the other association rules algorithms. To calculate the accuracy, we also divided the datasets randomly into two parts: the training datasets which contains 75% of datasets (transactions) and the testing datasets which contains the remaining 25% of datasets.

We computed the rule counter by comparing every rule with every transaction in the test dataset, so if the rule is identical and justified the transaction then the counter of the rule is increased by one, otherwise the process of comparison is continuing to the next transaction until all the transactions and rules are compared. The accuracy is calculated according to the following equation:

Accuracy =

∑ Rule counter for all rules / number of transactions in test datasets.

Figure 1 shows the accuracy of the UARM algorithm on hospital dataset (which grouped the transactions monthly in 12 groups) with support threshold (0.6) and confidence threshold (0.75). Each group in this figure has some of the rules generated from training datasets identical to the transactions in the test datasets. As shown in the figure, the rules in the UARM did not go to the accuracy value equal to zero.

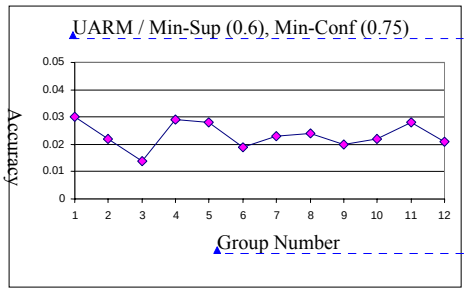

Figure 1: Accuracy UARM using 12 groups (Hospital datasets)

Figure 2 shows the accuracy of UARM1 algorithm on hospital dataset (which grouped the transactions monthly in 12 groups) with support threshold (0.6) and confidence threshold (0.3). This figure shows the group number and the accuracy of each group. In this figure the number of groups and the support threshold value are equal to the figure 4.3, and the confidence threshold decreases to 0.3. As we can see all the accuracy values of the groups are increased because the number of generated rules is increased.

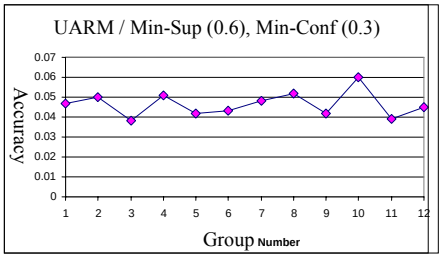

Figure 2: Accuracy UARM1 using 12 groups (Hospital datasets)

Through our experiments, it has noticed that the performance of our work is better than previous work especially, when using different minimum support and minimum confidence. Furthermore, extensive experiments have shown a better improvement achieved than the Fuzzy Approach for identifying Association Rules with Database-Independent Minimum-Support (FARDIMS) and Apriori algorithms, especially, when there is increased minimum support. Experimental results of our work show big improvement in mining association rules of the distribution of medicine and medical supply in the King Abdullah University Hospital.

Figures 3, 4, and 5 show the accuracy values of UARM, FARDIMS, and Apriori. The accuracy values are shown with support thresholds 0.5 and 0.6, respectively and with confidence values equal to 0.3, 0.5, and 0.7, respectively. As obvious from the figures, the UARM has a best accuracy value for two minimum supports and all minimum confidence. The UARM, FARDIMS, and Apriori algorithms have the same relation, as the minimum support increases, the accuracy value decreases; but in UARM algorithm the accuracy values can not decrease to reach the zero value as other algorithms. As described above, the UARM used the user defined feature to group the transactions, and then employed the advantages from the FARDIMS for each group to convert minimum support to real support. Converting the minimum support to good real support happens in UARM by collecting the same observations in the same group (grouped transactions timely, geographically, etc.). Then generate rules that are not redundant and generate rules for rare data.

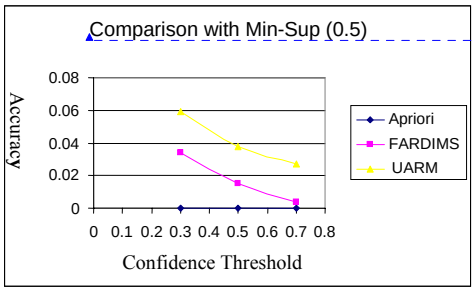

Figure 3: Accuracy for UARM, FARDIMS, Apriori (Hospital datasets)

Mis en forme : Police :Non Gras, Police de script complexe :Non Gras

Mis en forme : Police :9 pt, Police de script complexe :9 pt

Mis en forme : Police :Non Gras, Police de script complexe :Non Gras

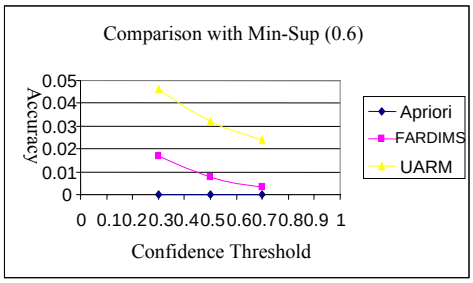

Figure 4: Accuracy for UARM, FARDIMS, Apriori (Hospital datasets).

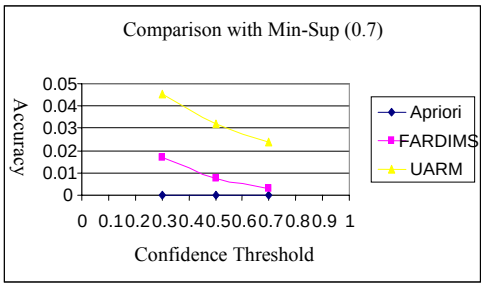

Figure 5: Accuracy for UARM, FARDIMS, Apriori (Hospital datasets).

6. CONCLUSIONS AND FUTURE WORK

Several algorithms for mining association rules have been introduced. When large number of rules is generated, these rules can be tedious and redundant. Additionally, they cannot find the correlation between rare data and other data with high frequency in transactions. Our work has given a possible solution to this problem. Through the experiments, we observed that the performance of our UARM algorithm that is a better performance achieved than FARDIMS and Apriori algorithms, particularly, when increasing minimum support. The performance experiments shows that UARM algorithm runs sequentially when grouping the transactions based on a user defined feature. Instead of this, each group can be run in parallel to reduce the execution time and reach more enhancements. This is left for future work.

REFERENCES

[1] Savasere A, Omiecinski E, Navathe S. "An Efficient Algorithm for Mining Association Rules In Large Databases", *Proceeding of the 21st VLDB Conference*, Zurich, Switzerland, pp. 432-444, September, 1995.

[2] Wai-Ho Au, Keith C.C.Chan. "Mining Changes in Association Rules: A Fuzzy Approach", *Fuzzy Sets Systems,* vol. 149 pp. 87-104, 2005.

[3] Wai-Ho Au, Keith C.C. Chan. "Fuzzy Data Mining for Discovering Changes in Association Rules over Time", *The research was supported in part by PolyU Grant*, AP209 and G-V918, 2001.

[4] Liu B, Hsu W, Ma Y. "Mining Association Rules With Multiple Minimum Supports", *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, USA, pp. 337-341, August, 1999.

[5] F.Berzal, I.Blanco, D.Sanchez, J.M.Serrano, M.A.Vila. "A Definition for Fuzzy Approximate Dependencies", *Fuzzy Sets Systems,* vol. 149, pp. 105-129, 2005.

[6] Billsus D, Pazzani M.J. "Learning Collaborative Information Filters", *In Proc. Fifteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers, Madison, Wisconsin, 1998.

[7] Han J, Pei J, Yin Y., "Mining Frequent Patterns without Candidate Generation". *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, TX , 1-12, May 2000.

[8] Han J, Kamber M. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, pp. 21-26, 2001.

[9] Agrawal R, Srikant R., "Fast Algorithms for Mining Association Rules", *Proceedings of the 20th VLDB Conference*, Santiago, Chile pp. 487-499, September 1994.

[10] Agrawal R, Imilienski T, Swami A. "Mining Association Rules Between Sets of Items in Large Databases", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington, DC, 207–216, May 1993.

[11] Hyunyoon Yun, Danshim Ha, Buhyun Hwang, Keun Ho Ryu. "Mining Association Rules On Significant Rare Data Using Relative Support", *The Journal of Systems and Software*, vol. 67 pp. 181–191. 2003.