AN EFFICIENT TECHNIQUE FOR FREQUENT ITEMSET GENERATION USING THE SIGNIFICANCE DEGREE OF ITEMS

Wael Ahmad AlZoubi[1], Khairuddin Omar[1], Azuraliza Abu Bakar[1]

Faculty of Computer and Information Technology

University Kebangsaan Malaysia

alzoubi_wael@yahoo.com, ko@ftsm.ukm.my, aab@ftsm.ukm.my

Abstract

Mining association rules is one of the most important tasks in data mining. The classical model of association rules mining is support-confidence. The support-confidence model concentrates only on the existence or absence of an item in transaction records and does not take into account the products' prices and quantities and how such these detailed information can affect the overall performance of rule mining process. In this paper a new measure for mining association rules is proposed based on the quantity of each itemset bought in a transaction; which is the significant degree measure to improve the classical method of mining association rules. The property of the new interestingness measures is analyzed, which validity has been tested in this paper.

1. INTRODUCTION

Data Mining has been applied to broad range of activities that attempt to discover new implicit patterns and rules from existing information, and obtain interesting and useful knowledge. However, no each pattern and rule has significance. In order to determine whether a rule or pattern is interesting, it needs a suitable metric to measure the degree of rule or pattern that customer is interested in. According to the quantified value, the uninteresting rules are crossed out. Therefore, the research on interestingness measure of rule is very important of data mining.

Association rules are one of well-studied problems in data mining. The problem of finding association rule $X \longrightarrow Y$ was first introduced in 1993 by Agrawal et al [3]. In the classical framework, technical model of association rules mining is *Support-confidence* [4], which is described as following:

Suppose that $I = \{I_1, I_2, \ldots, I_n\}$ be a dataset of items in a market basket database of transactions, and $D = \{T_1, T_2 \ldots T_m\}$ is the set of all transactions in the database, where each transaction is a set of items, i.e. $T_i = \{I_{i1}, I_{i2}, \ldots, I_{ik}\}$ and $I_{ij} \in I$ is a transaction. (Such that $T \subseteq I$), each transaction has a unique identifier *TID*. An association rule is an expression $X \longrightarrow Y$ where X and Y are disjoint sets of items. The meaning of such a rule is that transactions in the database which contain the items in X also tend to contain the items in Y. the support and confidence are two main measures mostly used to determine the interestingness of such a rule. They are given by formula 1 and 2 respectively.

$$\text{Support } (X \longrightarrow Y) = P(X \cup Y) \quad (1)$$

$$\text{Confidence } (X \longrightarrow Y) = \frac{P(X \cup Y)}{P(X)} \quad (2)$$

Association rules are those that have support and confidence not less than the

minimum support and *minimum confidenc*e that costumers provide.
The classical model mentioned above is named as *support-confidence* framework, where *conf(X* $\longrightarrow$ *Y)* measures interesting of the rule *X* $\longrightarrow$ *Y*. In fact, this classical interestingness measure has much shortage. Several authors [1, 2, 8] criticized the use of *support confidence* framework. In this paper, an improvement on the *support-confidence* framework have been proposed based on *significance measure* of itemsets to improve the classical method of mining association rules. The property of the new interestingness measure is discussed. Its strength has been analyzed in this paper.
The rest of this paper is organized as following: In Section 2, the problem definition is analyzed. A new interestingness measure is proposed in Section 3, and the results analysis and conclusion are presented in section 4.

2. PROBLEM DEFINITION

The process of association rule mining consists of two step process[5].

I. Generate all frequent itemsets. This is both computation and I/O intensive.

II. Generating strong rules, i.e. Rules of the form X/Y $\longrightarrow$ Y where Y $\subset$ X are generated for all frequent itemsets obtained in step I provided they satisfy the minimum confidence. Our focus is on the generation of frequent itemsets in an efficient manner.

The first step constitutes an area where significant research findings have been reported, as if there are *m* items then there can be potentially 2^m frequent itemsets. The main goal of this papers is to improve the process of generating frequent itemsets by minimizing the dataset size without any loss of information.

The measure “support” indicates the frequencies of occurrence times for both *X* and *Y* in the whole dataset, the “confidence” measure denotes the strength of implication of the rule. The above rule will hold if its support and confidence are equal to or greater than the user specified minimum support (*S*), and minimum confidence (*C*), respectively.

A new measure was proposed in this paper to overcome the problems in support confidence framework, the main problem in the support measure is choosing an proper value of minimum support since choosing a high value of minimum support may cause valuable but rarely used items to be ignored from further processing while choosing some low value of support results in considering too many items as frequent items although some of these items may be not important. The proposed measurement, significance of items, takes in consideration the quantity of each item bought in a transaction to determine the degree of importance according to some predefined thresholds assigned by the user. There are – in this paper – three classifications of items according to the significance value, i.e. highly important items which represent most frequent itemsets, items of medium importance which may or may not considered as frequent itemsets and unimportant items that have low significance value that represent infrequent itemsets and they will be eliminated from the dataset of transaction records.

3. SIGNIFICANCE DEGREE OF ITEMS

Association rule mining algorithms that adopt the support-confidence didn't take in consideration the price and quantity of itemsets bought in a transaction while finding the frequent itemsets. Although the support confidence model gives some accuracy, it still gives the wrong impression about the interesting rules. The following example illustrates some shortages in *support-confidence* framework.

Example 1: [9] Suppose we have market basket data from a grocery store, consisting of *n* baskets. Let us focus on the purchase of tea and coffee. In the following table, rows *t* and $\overline{t}$ correspond to baskets that do and do not, respectively, contain tea, and similarly columns *c* and $\overline{c}$ correspond to coffee. The numbers represent proportion of baskets.

	C	c	Total
t	20	$\overline{5}$	25
t	70	5	75
Total	90	10	100

Table 1: Number of transactions that do and don't have tea and coffee

Suppose that the minimum support threshold = 0.1 and the minimum confidence threshold = 0.5 and we have the following rule; *r:* $t \rightarrow c$, sup (*r*) =20/100 = 0.2, confidence (*r*) = 0.2/0.25 = 0.8; this means that *r* is confident. While correlation (*r*) = P ($t \cup c$)/ (P (*t*) * P (*c*)) = 0.2 / (0.25*0.9) = 0.89 < 1. This means that tea and coffee are negatively correlated and so *r* is a misleading association rule. This example gives a clear impression that the *support – confidence* model has a shortage and it requires to be improved.

In the rest of this paper, we will propose a new model to measure the degree of itemsets' significance, this proposed measure takes in consideration – together with the classical model – the quantity of each item bought in a transaction to determine the degree of importance according to some predefined thresholds assigned by the user. There are – in this paper – three classifications of items according to the significance value, i.e. highly important items which represent most frequent itemsets, items of medium importance which may or may not considered as frequent itemsets and unimportant items that have low significance value that represent infrequent itemsets and they will be eliminated from the dataset of transaction records.

Table 2 shown below displays an example of transaction database, the new here is accompanying each item in all transaction records with its quantity.

TID	Items & Quantities
1	{Milk, 2}, {Bread, 1}, {Butter, 3}, {Beef, 5}
2	{Milk, 3}, {Bread, 4}
3	{Milk, 1}, {Bread, 2}, {Butter, 4}
4	{Milk, 1}, {Bread, 2}, {Beef, 2}
5	{Milk, 1}, {Butter, 5}
6	{Milk, 3}

Table 2: En example of a transaction database

Table 3 illustrates the proposed method used to compute the degree of item's significance by computing first the quantity bought for each item through the whole database of transactions.

TID	Milk	Bread	Butter	Beef	Overall Total
1	2	1	3	5	
2	3	1	0	0	
3	1	2	4	0	
4	1	2	0	2	
5	1	0	5	0	
6	3	0	0	0	
Total	*11*	*6*	*12*	*7*	*36*

Table 3: The quantity bought for each item in the dataset of transactions

The significance degree of an itemset (x) is given by formula 3 given below:

$$Significance(x) = \sup(x) * (\sum_{t=1}^{n} Qty(x) / Total) \quad (3)$$

Where Qty (x) is the number of pieces bought of item x in all recorded transactions, and so, Qty (Milk) = 2+3+1+1+1+3 = 11, Qty (Bread) = 15, Qty (Butter) = 11, and Qty (Beef) = 7. Depending on table 3, the *support* (Milk) = 6/6 = 1, and in the fashion *support* (Bread) = 0.67, *support* (Butter) = 0.5, and *support* (Beef) = 0.33. Given that the minimum support threshold is 0.5 and minimum confidence threshold is 0.7. According to the support – confidence model, the set of frequent itemsets are: {Milk, Bread, Butter}; and so we have six association rules, these rules and their corresponding confidence are given bellow:

1. Confidence (Milk ⟶ Bread) = 0.67
2. Confidence (Milk ⟶ Butter) = 0.5
3. Confidence (Bread ⟶ Butter) = 0.5
4. Confidence (Bread ⟶ Milk) = 1
5. Confidence (Butter ⟶ Milk) = 1
6. Confidence (Butter ⟶ Bread) = 0.67

Only two of these rules are confident since their confidence is not less than the pre-specified minimum confidence threshold; i.e. rule 3 and 4. The significance of the four items are calculated as in formula 3 and given below:

Significance (Milk) =1.0 * 11/36 = 30.6 %
Significance (Bread) =0.67* 6/36 = 11.2%
Significance (Butter) =0.5*12/36 = 16.7%
Significance (Beef) = 0.33* 7/36 = 6.5 %

Suppose that the minimum significance threshold is 0.15, then {Bread, Beef} are considered as insignificant and they will be eliminated from the dataset. So, we have – in this case – only one rule to be checked, it is: *if* Milk *then* Butter with 32% degree of significance. The proposed measure doesn't need to change the consequent and the result as was done with respect to confidence, which may reduce the time required to generate the association rules by 50%.

4. RESULTS ANALYSIS AND CONCLUSION

To evaluate the efficiency of the proposed measurement in pruning the huge dataset of transactions without any loss of information, we had been re-implemented the Apriori algorithm, using Microsoft Visual Basic 6.0 on a Pentium IV 2400 MHz PC with 1024MB of available physical memory. The test database is the Chess transaction database. In this experiment, the efficiency of the Apriori algorithm was increased as the time required in order to find the association rules had been decreased when the proposed significant measure was applied on the itemsets. Table 4 displays the improvement achieved with respect to the time required to generate association rules at different values of minimum support

threshold before and after using the proposed significance measurement.

Minimum Support	Time 1 (Seconds)	Time 2 (Seconds)
0.1	22.9	23.5
0.2	20.1	22.6
0.3	19.5	21.8
0.4	18.2	21.0
0.5	15.3	19.8
0.6	13.4	18.9
0.7	12.2	17.3

Table 4: The execution time before and after applying the significance measure

In table 4, *time1* represents the time required to generate confident rules after using the significance degree of itemsets using Apriori algorithm on the Chess dataset, where *time2* represents the time needed to generate rules after using the proposed measure. It is clear that the time required to generate association rules was decreased especially at high values of minimum supports.

5. REFERENCES

[1] J. Han & M. Kamber. "Data mining: concepts and techniques". *Morgan Kaufmann*, Academic, San Fransisco, NewYork, 2001.

[2] A. Silberschatz & A. Tuzhilin. "What makes pattern interesting in knowledge discovery systems." IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6), pp: 970-974.

[3] R.agrawal, T. Imielinski & A. Swami. "Mining Association Rules between Sets of Items in Large Databases." *Proceedings of the 1993 ACM SIGMOD International conference on Management of Data*, Washington DC (USA), 1993, pp: 207-216.

[4] R. Agrawal & R. Srikant. "Fast Algorithms for Mining Association Rules." *Proc. 20th Int. Conf. on Very Large Data Bases*, 1994, pp: 487 – 499.

[5] B. Kalpana & R.Nadarajan. 2007, Optimizing Search Space Pruning in Frequent Itemset Mining With Hybrid Traversal Strategies-A Comparative Performance on Different Data Organizations, IAENG International Journal of Computer Science, 2007.

[6] A. Ceglar & J.F. Roddick. Association Mining, ACM Computing Surveys, vol.38, No.2, July, 2006.

[7] Y.-J. Tsay & J.Y. Chiang. CBAR: an efficient method for mining association rules. Knowledge-Based Systems 18 (2005) 99–105.

[8]S.Kotsiantis & D. Kanellopoulos. "Association Rules Mining: A Recent Overview" *GESTS International Transactions on Computer Science and Engineering*, 2006, Vol. 32(1), pp: 71-82.

[9] S. Brin, R. Motwani & C. Silverstein. "Beyond market basket: Generalizing association rules to correlation." In Proc.ACM SIGMOD Intl.Conf.Management of Data, 1997, pp: 265 – 276.