Association Rule Mining with Data Item including Independency based on Enhanced Confidence Factor

Yingquan Wang,Tomohiro Murata,

***Abstract*—Along with the development of data collection and various storage technology, the large data of users activities in economy is stored. Extracting valuable information or knowledge regarding behavior of user from these data is becoming more and more important for marketing strategies of sales and commerce. Association rule mining is one of useful techniques in this application field and widely studied. But sometimes too many rules that generated by association rule mining usually caused the wrong decisions made by manager, parts of generated rules are meaningful and useful, but other generated rules are unnecessary for manager to make the right decisionsIn this paper, in order to extract useful rules efficiently, we proposed a new framework of association rule mining based on enhanced confidence factor. Thus, the certainty factor was introduced to identify different situations and analysis the accuracy of association rule mining respectively. We illustrate some merits of our proposed method by theoretical analysis. Our experiment results show that the sets of useful rules can be generated in a more efficient way by using our method, which means less and more accurate rules could be used to make the proper decisions by manager.**

***Index Terms*—Association rule mining, Independency, Negative dependence, Certainty factor.**

I. Introduction

Because of the broadly using of information technologies, the conflict between data explosion and poorness of knowledge has become more and more acute, and the necessity of data mining is also becoming urgent. Among all branches of data mining, research of association rule mining was deeply studied and the applications in this field were widely used.

The motivation of association rules was for the Market Basket Analysis problem. Suppose that the store manager wants to know more about the customers shopping habits. In particular, would you like to know which product customer may purchase at the same time? To answer this question, we can analysis the quantity and category of purchasing items in the customers shopping basket. Association rule mining analyzes the customers shopping habits by identifying the associations between different items placed in the shopping basket. These association rules can help retailers to understand which products are frequently purchased by customers at the same time. Thus, according to these information the better marketing strategies could be made.

In this paper, we proposed an improved association rule mining based on enhanced confidence factor method to generate real strong rules. Too many misleading rules generated by conventional methods is noise for the result in our experiment. Configuration of the paper is as below. Some related research and work will be introduced in Section 2. In section 3, we will explain the drawbacks by using the conventional association rule mining. In section 4 a new framework of association rule mining is proposed to solve the problem which we mentioned in section 3. In section 5, some merits are shown in experiment results of comparing between the conventional method and our proposed method. Finally the conclusion will be summarized in Section 6.

Manuscript received December 27, 2016; revised December 30, 2016.

Yingquan Wang with the master student at the School of Information, Production and System in Waseda university, Kitakyushu, Fukuoka province 8080135 Japan (email: wangyingquan@fuji.waseda.jp).

Tomohiro Murata is with the professor at the School of Information, Production and System in Waseda university, Kitakyushu, Fukuoka province 8080135 Japan (email: t-murata@waseda.jp).

II. Related Works

Several authors have proposed some new method based on conventional association rule mining support-confidence framework. In this section we give brief introduction about that.

A. Association rule mining

In 1993, Agrawal et al. First proposed the concept of association rule mining, and gave the corresponding mining algorithm AIS [1], but the performance was not very well. In 1994, they established the project set grid space theory, and according to the above two theory, propose a new method which is the famous Apriori algorithm, so far Apriori still as a classical algorithm of association rules mining is widely discussed.

Assume I = $\{I_1$,I_2,I_3,..,$I_M\}$ is the assemblage of items. Given a transaction dataset D, in this dataset each transaction t is a nonempty subset of I, and each transaction corresponds to a unique identifier TID (Transaction ID). In dataset D, the support function contains two items: X and Y, the pattern of manifestation is shown as a probability(Eq.1).

$$Support(X \to Y) = Support(X \cup Y) = \frac{P(X \cap Y)}{M} \quad (1)$$

The confidence function obtains the probability of Y when we have already known that X is contained in transaction in the dataset D, the pattern of manifestation is shown as the conditional probability(Eq.2).

$$\begin{aligned} Confidence(X \to Y) &= \frac{support(X \cup Y)}{support(X)} \\ &= \frac{support(X \to Y)}{support(X)} \end{aligned} \quad (2)$$

The association rule is considered to be interesting if the rules support value and confidence value over the minimum

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

support threshold and the minimum confidence threshold respectively. These thresholds are artificially set according to mining requirements, usually the setting of those thresholds is based on the experience of the manager.

The process of association rules mining consists of two phases: the first phase must find all the high frequency itemsets from the dataset, and the second phase generate all association rules from the group of high frequency itemsets.

In the first step of association rule mining, all the high frequency itemsets must be found in the original dataset. High frequency means that the occurrence frequency of an itemset which is relative to all records in the dataset has reached a certain level. The occurrence frequency of an itemset is called support. Use a 2-itemset containing X and Y as an example, we can obtain the support value of itemset which contains X, Y through Eq.1. If the support is greater than or equal to the minimum support threshold, then X, Y are called high frequency itemset. A k-itemset that satisfies the minimum support is called a high frequency k-itemset, it is generally expressed as Frequent-k. Algorithm and from Frequent k of the itemset to produce Frequent k+1. The calculations ends when any high frequency itemset can no longer be found.

The second step of association rule mining is to generate association rules from the high frequency itemsets, actually we generate association rules after the calculation of the high-frequency k-itemset. If the confidence value of a rule that we obtained is greater than or equal to the minimum confidence value, it is defined as an association rule. For example, the rule X, Y generated by the high-frequency k-itemset X, Y can be obtained by the Eq.2. If the confidence is greater than or equal to the minimum confidence, X, Y is called the association rule.

B. Support-Confidence framework

In 1995, Pack et al. propose DHP algorithm (Dynamic Hash Prune algorithm). Use hash prune on all dataset, eliminate the transaction which contain less item, thus decrease the amount of transaction for traversing [2].Savasere et al. propose use partition algorithm to separate mage transaction set, generate local frequent itemset in every part, than do process generate whole frequent itemset [3].In 1997, Zaki et al. Research on new algorithm about quick association rule mining, propose four new method which is Ecalt, MaxEclat, Clique, MaxClique respectively. The main idea of those method is adopting itemset clustering, transaction clustering and traversing grid and so on [4].In 1997, S. Brin, et al. use dynamic counting on itemset to generate frequent itemset, set different threshold to overcome the drawback of support-confidence framework.In 2000, Jiawei Han propose a method FP-tree algorithm, use the FP-tree reserve all transaction set, generate the frequent itemset without candidate itemset [5].

III. Problem Description

A. Drawbacks of conventional in the conventional association rule mining

consider that one rule is good or not, usually base on two parameters. One is support value which is an indication of how frequently the itemset appears in the dataset, the other one is confidence value which is an indication of how often the rule has been found to be true. Because the definition of the confidence is a conditional probability, so here we can consider that $Confidence(X \rightarrow Y)$ means the probability of item Y can be obtained when we have already known the item X occurred.

Then we make a comparison between $Confidence(X \rightarrow Y)$ and $Support(Y)$.

1) If $Confidence(X \rightarrow Y) > Support(Y)$, it means that the condition of item X have positive effect on the item Y;

2) If $Confidence(X \rightarrow Y) = Support(Y)$, it means that the condition of item X have no effect on the item Y, we can consider that the relationship between the item X and item Y is independent;

3) If $Confidence(X \rightarrow Y) < Support(Y)$, it means that the condition of item X have negative effect on the item Y;

TABLE I
SMALL DATASET FOR ILLUSTRATE THE PROBLEM

ITEM	1	2	3	4	5	6	7	8	9	10
COMPUTER	1	1	1	1	1	1	0	0	0	0
GAMES	0	0	1	1	1	1	1	1	1	0
KEYBOARD	1	1	1	0	0	0	0	0	1	1
MOUSE	1	1	1	1	0	0	0	1	0	0

TABLE II
CALCULATED RESULT ABOUT ITEM/ITEMSET

ITEM/ITEMSET	Support	Confidence
COMPUTER	0.6	-
GAMES	0.7	-
KEYBOARD	0.5	-
MOUSE	0.5	-
COMPUTER→GAMES	0.4	0.667
COMPUTER→KEYBOARD	0.3	0.5
COMPUTER→MOUSE	0.4	0.667

B. An example

Here, we use a small dataset to illustrate those kinds of situation. The rows represent items, and columns represent transactions. The more detail of item data is shown in the TABLE.I.

After the calculation of the support and confidence, we set $Minsupport = 0.3$ and $Minconfidence = 0.5$, the result is shown in the TABLE.II.

This rule of $computer - games$ suggests that a person who bought a computer would buy games at the same time. As the following:
$Supp(computer \rightarrow games) = 0.4 > Minsupp$,
$Conf(computer \rightarrow games) = 0.667 > Minconf$.

However, $Supp(games) = 0.7$, actually the probability of a person buying games decreases (from0.7 to 0.667) when we know this person has already bought a computer. We consider the condition of a person buying a computer for buying games is a negative dependence, hence this rule generated by conventional association rule mining is negative and misleading.

Regarding the rule of $computer - keyboard$ we can easily see that:
$Supp(keyboard) = 0.5 = Conf(computer \rightarrow keyboard)$,

it means the condition of a person buying computer has no effect with a person buying a keyboard, so we can consider

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the item computer and item keyboard is independent. This rule is also misleading and unnecessary.

Regarding the rule of $computer - mouse$ shown in the TABLEII, we make 3 comparisons about the rule (COMPUTER→MOUSE).
$Supp(computer \rightarrow mouse) = 0.4 > Minsupp$,
$Conf(computer \rightarrow mouse) = 0.667 > Minconf$,
$Conf(computer \rightarrow mouse) = 0.667 > Supp(mouse)$.

So it means a person buying a computer has positive effect with buying a mouse, this kind of positive-dependence rule of items is our wanted.

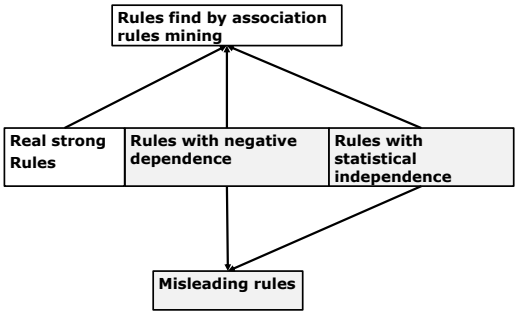

Fig. 1. The all kinds of rules generated by conventional association rule mining.

From Fig.1, we could see some problems by using conventional methods, which contains too many misleading rules. This causes unnecessary calculation and inaccurate results. Usually, when we consider using association rule mining to find rules to help us making decisions, the most useful rule is the rule with positive dependence between items, so the rule with negative dependence or independency is useless for manager to make decisions ,we can consider those kinds of rules is noise for our result.

IV. Proposed Association Rule Mining Method based on Enhanced Confidence Factor

In conventional association rule mining, confidence value is the accuracy measure. Some authors think that the support-confidence framework has some drawbacks [7], [8], [10]. Therefore, in our proposed method, consider the problem which we mentioned in Section 3, we choose the Certainty factor instead of confidence, as the new accuracy measure for our proposed method [6].

We define value CF as the certainty factor of $X \rightarrow Y$ (X, Y means different items which exists in the same itemset.)

When $Confidence(X \rightarrow Y) > Support(Y)$:

$$CF(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y) - Support(Y)}{1 - Support(Y)} \tag{3}$$

When $Confidence(X \rightarrow Y) < Support(Y)$:

$$CF(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y) - Support(Y)}{Support(Y)} \tag{4}$$

When $Confidence(X \rightarrow Y) = Support(Y)$:

TABLE III
Description of Datasets

Dataset	Number of transaction	Number of attribute
Belgian retail store	75000	49
Online retail	541909	2603

$$CF(X \rightarrow Y) = 0 \tag{5}$$

The certainty factor as a new accuracy measure replaces the confidence in our proposed method. In different conditions the certainty factor is different. When $Confidence(X \rightarrow Y) > Support(Y)$, it means the relationship between item X and item Y is positive dependence, so we can calculate the value of CF which is positive, it means the accuracy of $Itemset(X \rightarrow Y)$; When $Confidence(X \rightarrow Y) < Support(Y)$, it means the relationship between item X and item Y is negative dependence; When $Confidence(X \rightarrow Y) = Support(Y)$, it means the relationship between item X and item Y is independent. So here value of CF should be equal to 0.

Hence, by setting the threshold, we could calculate the value of CF to eliminate the negative dependence and independence in itemset, this can generate less rules than using the confidence in conventional way. Indeed, the result of rules generated without the misleading rules which cant be achieved by using conventional method.

Based on the FP-growth algorithm, a new accuracy measure CF is introduced, the flowchart of association rule mining with data item including independency based on enhanced confidence factor is shown in Fig.2.

V. Experiments and Evaluation

We prepare two different datasets for illustrating the problems we mentioned before, and show the performance of our proposed algorithm. Also we make comparison of accuracy and efficiency between our proposed algorithm and Apriori algorithm, which is classic association rule mining method. The first dataset is customer transactions from an anonymous Belgian retail store which contains 75,000 transactions and 49 different kind of attribute [9], and the second dataset is transformed from the online retail dataset, the transformed dataset contains 541,909 transactions and 2603 attributes. TABLE.III lists all the features of two datasets.

Our program coding is written by python, running on PyCharm 2016.3, and all the experiments running environment is window 10 machine with a 2.1 GHz Intel and 4GB RAM.

A. Improvement of accuracy

In this phase, we make comparison about the number of rules generated by proposed method, conventional method and the managers objective result. Here, we consider the manager's objective result in the real strong rule. In this experiment, we illustrate the improvement by using Belgian retail store dataset. Usually how to choose the threshold is depending on the experience, for more close to the reality here we choose $Minsupport = 0.003$ and $Minconfidence$ or $MinCF = 0.4$. In the following figure Fig.3 we could get that, although few misleading rules still exists after evaluation by using our method, most of the unnecessary

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

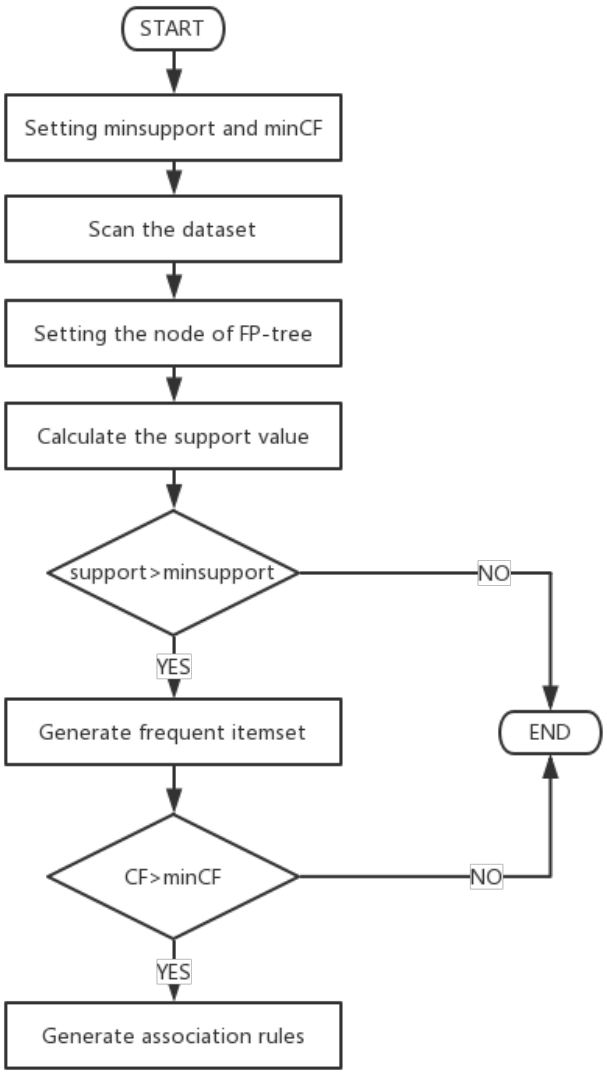

Fig. 2. The flowchart of association rule mining with data item including independency based on enhanced confidence factor.

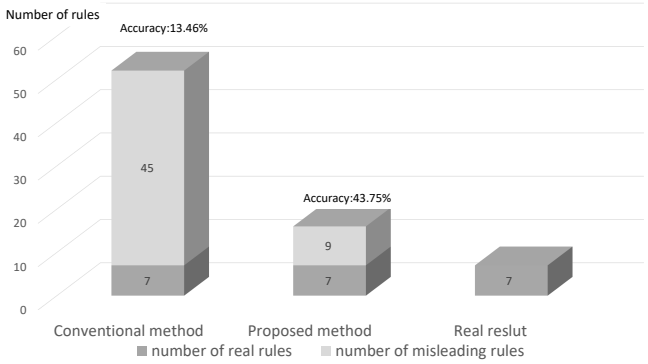

Fig. 3. Evaluate accuracy improvement of proposed method.

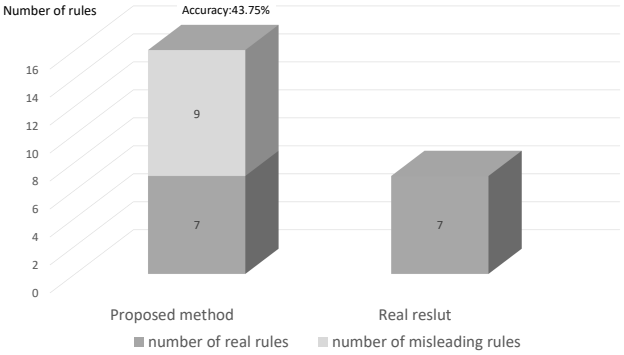

Fig. 4. Test in data of Belgian retail store (contains 75,000 transactions and 49 items).

TABLE IV
SENSITIVITY ANALYSIS AGAINST MINCF AND MINSUPPORT

Minsuppot \ MinCF	Low MinCF ($MinCF = 0.3$)	High MinCF ($MinCF = 0.7$)
Low Minsupport ($Minsupport = 0.002$)	217	79
High Minsupport ($Minsupport = 0.005$)	42	15

rules have been already deleted, the accuracy of result has improved from 13.4% to 43.75%.

B. B. Improvement of stability

Through running the proposed method on two different datasets, comparing the number of rules generated and the managers objective result, we have got the following two figures.

Fig.4 shows the test result of Belgian retail store (contains 75,000 transaction and 49 items), and Fig.5 shows the test result of Online retail(contains 541,909 transaction and 2603 items). We could see that stable accuracy result is obtained by using our method in two different datasets.

C. Sensitivity analysis against the MinCF and Minsupport.

We fix the $MinCF$ and change the $Minsupport$ from low to high, also we fix the $Minsupport$ and change the $MinCF$ from low to high. The result is shown in TABLE.IV.

When we set the high $MinCF$ and high $Minsupport$, there have 15 rules generated, this kind of rules means those combinations of item have a lot of customer willing to buy. The sales of those item should be very large. When we set the high $MinCF$ and Low $Minsupport$, there have 79 rules generated, this kind of rules means those combinations of item have some kinds of specific customer willing to buy. Even if the sales of those combinations can not to be too much, but focus on some specific customers, those kinds of rules are very meaningful and useful.

VI. CONCLUSION

In this paper, we propose a new framework of association rule mining, by using certainty factor as the new accuracy measure to replace the confidence. Based on different relationships between the items which in the same itemset, we calculate the certainty factor in different situations.

From the theoretical analysis and experiments result, we can clearly see that our proposed association rule mining by enhanced confidence factor is an effective method. It

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

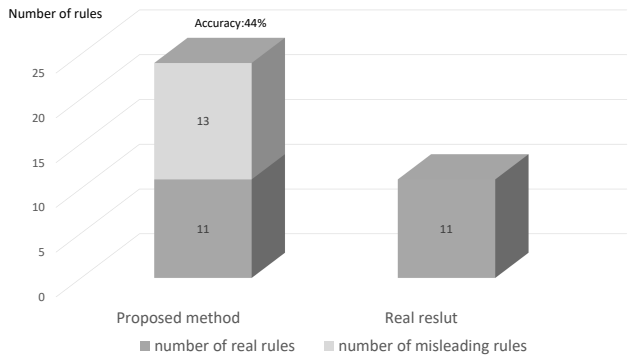

Fig. 5. Test in data of Online retail (contains 541,909 transactions and 2603 items).

can avoid to obtain misleading rules comparing with the conventional way, and provide more accurate result, satisfy the requirements of the decision maker. Our experiments in both of two databsets could confirm this point. And using the FP-growth algorithm to generate frequent itemset is another advantage of our new framework, it can help to reduce the calculation time and space expended in the first step of the discovery process.

But from the result of experiment we also can see that even if we avoid to obtain the rules which has negative dependence and independence on items in the same itemset, there still generates some noise rules. Considering this part, we shall follow this research avenue in the future.

REFERENCES

[1] Agrawal R, Imieliski T, Swami A., "Mining association rules between sets of items in large databases," *Acm sigmod record, ACM,*, 22(2): 207-216, 1993.
[2] Park, J. S., Chen, M. S., Yu, P. S., "An effective hash-based algorithm for mining association rules," *ACM.*, vol. 24, No. 2, pp. 175-186, 1995.
[3] Savasere, Ashok, Edward Robert Omiecinski, and Shamkant B. Navathe, "An efficient algorithm for mining association rules in large databases," 1995.
[4] Zaki M J, Parthasarathy S, Ogihara M, et al., "New Algorithms for Fast Discovery of Association Rules," *KDD.*, 97: 283-286, 1997.
[5] Han J, Kamber M., "Data mining concept and technology," *Publishing House of Mechanism Industry.*, : 70-72, 2001.
[6] Wei C P, Piramuthu S, Shaw M J., "Knowledge discovery and data mining," *Handbook on Knowledge Management. Springer Berlin Heidelberg.*, : 157-189, 2003.
[7] Piatetsky-Shapiro G., "Discovery, analysis, and presentation of strong rules," *Knowledge discovery in databases.*, : 229-238, 1991.
[8] Shortliffe E H, Buchanan B G., "A model of inexact reasoning in medicine," *Mathematical biosciences.*, 23(3): 351-379, 1975.
[9] de Moerloose C, Antioco M, Lindgreen A, et al., "Information kiosks: the case of the Belgian retail sector," *International Journal of Retail and Distribution Management.*, 33(6): 472-490, 2005.
[10] Berzal F, Blanco I, Snchez D, et al., "A new framework to assess association rules," *International Symposium on Intelligent Data Analysis. Springer Berlin Heidelberg.*, : 95-104, 2001.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)