Effective measures in Association Rule Mining

Dr.S.P.Victor, J.R.Jeba

Abstract -Association rule mining is one of the most popular data mining techniques to find associations among items in a set by mining necessary patterns in a large database. However, not all of the generated rules are interesting, and some unapparent rules may be ignored. Efficient discovery of such rules has been a major focus in the data mining research. In this paper, we analyze the measure of support and confidence for mining association rules, from which we find many redundant or unrelated rules besides the interesting ones. We introduce a new proposed framework in order to obtain misleading rules, improving the manageability and quality of the results.

Index Terms— Association Rule Mining,Support, Modified support, Confidence, Modified Confidence, Data Mining

1 Introduction

Association rules were presented by R. Agrawal and others in 1993[8], it is an important research issue in data mining. Mining association rules aims at finding the correlation between the different items in a database. It can be used to find the purchase patterns of customers such as how the transaction of buying some goods will impact on the transaction of buying others. Association rules is composed of the following two steps: 1) Find the large item sets that have transaction support above a minimum support and 2) From the discovered large item sets generate the desired association rules. This phase is to find the effective ways needed to select interesting rules from discovered rules.

Let I = {I_1, I_2, ….I_n} be a set of items. Let D, the task-relevant data, be a set of transactions in a supermarket, where each transaction T is a set of items, such that T$\subseteq$ I. Each transaction is assigned an identifier called TID. Let A be a set of items, a transaction T is said to contain A if and only if A$\subseteq$T. An association rule is an implication of the form A$\Rightarrow$B, where A$\subset$I, B$\subset$I, and A$\cap$B=$\varnothing$. The rule A$\Rightarrow$B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain A$\cup$B (i.e., both A and B). This is taken to be the probability P(A$\cup$B). The rule A$\Rightarrow$B has confidence c in the transaction set D if c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P(B|A). That is, Support (A$\Rightarrow$B) = P(A$\cup$B) = S, Confidence (A$\Rightarrow$B) = P(B|A) =Support (A$\Rightarrow$B)/Support (A)=C.

Support and Confidence are the usual measures to assess the association Rules. Sup-

- *Dr.S.P.Victor,Associate Professor & Head,Department of Computer Science, St.Xavier's College(Autonomous), PalayamKottai. Email :victorsp@rediffmail.com*
- *J.R.Jeba,Associate Professor & Head, Department of Computer Applications,Noorul Islam Centre for higher Education, Email : jrjeba@rediffmail.com*

port is the percentage of transactions where the rule holds. Confidence is the conditional probability of C with respect to A or, in other words, the relative cardinality of C with respect to A. The techniques of mining association rules attempt to discover the rules whose support and confidence are greater than user-defined thresholds called minsupp and minconf respectively. These are called strong rules.

However, several authors have pointed out some drawbacks of this framework that lead to find many more rules than it should[1,9]. In section 2, we describe some drawbacks of the support/confidence framework. Section 3 contains some related work. Section 4 is devoted to describe our new proposal. Experiments and conclusions are summarized in sections 5 and 6 respectively.

1. Drawbacks of the Support/Confidence Framework

A Common principle in association rule mining is “ the greater the support, the better the itemset”, but we think this is only true to some extent. Researchers have been applying the framework of support-confidence to set up association rules in the process of producing association rules. Indeed, itemsets with very high support are a source of misleading rules because they appear in most of the transactions and hence any itemset seems to be a good predictor of the presence of the high support itemset. A lot of redundant and unrelated rules are generated when the framework of support-confidence is applied to find rules.

Table :1 Transaction set R1

A	**B**	**C**	**D**	**E**
1	0	1	0	1
0	0	0	1	0
0	1	1	1	1
0	1	1	1	0
1	1	1	1	1
1	1	1	1	1

An example is {C} in Table 1. It is easy to verify that any itemset involving only A and B is a perfect predictor of {C} (any rule with {C} in the consequent has total accuracy, that is, confidence is 1 for all such rules). Also, $\text{Conf}(\{D\} \Rightarrow \{C\}) = 0.8$, that is pretty high. But we cannot be sure that these associations hold in real world. In fact what holds most times is negative independence or independence. When Supp(C) is very high and $\text{Conf}(A \Rightarrow C) > \text{Supp}(C)$, we can obtain a high accuracy. However there is a lack of variability in the presence of C in data that does not allow us to be sure about the rule.

Confidence denotes the probability that the emergence of some item sets will lead to the occurrence of the others. However we notice that the confidence of association rule $A \Rightarrow C$ only takes into consideration the possibility of the case when only C occur and the case whether A and C are correlated. Hence many association rules obtained using support and confidence tend to be ineffective.

2. Related Work

Some authors have proposed alternatives to confidence[2,4,5,6,3,7,9]. In this section we briefly describe two of them.

2.1 Conviction

Conviction was introduced in[9] to be

$$Conv(A \Rightarrow C) = \frac{supp(A)\ supp(\neg C)}{supp(A \cup \neg C)}$$

where $\neg C$ means the absence of *C*. Its domain is (0,∞), 1 meaning independence. Values in (0, 1) mean negative dependence. In our opinion, the main drawback of this measure is that its range is not bounded, so it is not easy to compare the conviction of rules because differences between them are not meaningful and, much more important, it is difficult to define a conviction threshold.

2.2 Lift

In [3, 7] Lift measures how far from independence are A and C. Values close to 1 imply that A and C are independent and the rule is not interesting. Values far from 1 indicate that the evidence of A provides information about C. Lift measures co-occurrence only.

$$\text{Lift}(A \Rightarrow C) = \frac{\text{Conf}(A \Rightarrow C)}{supp(C)}$$

3. A New Framework to asses Association Rules

According to those problems discussed above we suggest some modification in the support and Confidence. The rule *$A \Rightarrow C$ is very strong* if both $A \Rightarrow C$ and $\neg C \Rightarrow \neg A$ are strong rules. The rationale behind this definition is that $A \Rightarrow C$ and $\neg C \Rightarrow \neg A$ are logically equivalent, so we should look for strong evidence of both rules to believe that they are interesting. This definition can help us to solve the support drawback since when supp(C) (or supp(A)) is very high, $Supp(\neg C \Rightarrow \neg A)$ is very low, and hence the rule $\neg C \Rightarrow \neg A$ won't be strong and $A \Rightarrow C$ won't be very strong.

By definition, a very strong rule must verify:

Modified Support conditions:

(a) $Supp(A \Rightarrow C) >$ minsupp

(b)If $supp(A)+Supp(C)<=1$, then it should satisfy $\text{Supp}(\neg C \Rightarrow \neg A) >$ minsupp

$$\text{Supp}(\neg C \Rightarrow \neg A) = 1 - \text{supp(C)} - \text{supp(A)} + \text{Supp}(A \Rightarrow C)$$

Modified Confidence :

we notice that the confidence of association rule $A \Rightarrow C$ only takes into consideration the possibility of the case when only C occur and the case whether A and C are correlated. In order to diminish the drawbacks of confidence, we introduce the concept of "Modified Confidence". Assume that P(A) denotes the probability of A's occurrence, P(C) denotes the probability of C's occurrence, P(AUC) denotes the probability of the case when A and C occur simultaneously. P(¬AUC) denotes the probability of the case when C occurs but A does not occur, P(¬A) denotes the probability of A's non-occurrence. Then, we define :

Modified-Confidence of the rule =
$\mathrm{Conf}\,(A \Rightarrow C) - \mathrm{Conf}\,(\neg A \Rightarrow C)$. *(i.e)*

Modified- Confidence

$$= \frac{\mathrm{Supp}(AUC)}{\mathrm{Supp}(A)} - \frac{\mathrm{Supp}((\neg A\,UC)}{\mathrm{Supp}(\neg A)}$$

Where

$$\mathrm{Supp}(\neg A\,UC) = \mathrm{Supp}(C)\text{-}\mathrm{Supp}(AUC)$$

And

$$\mathrm{Supp}(\neg A)=1\text{-}\ \mathrm{Supp}(A)$$

We make the declaration that an item set is irrelated to any other item set when its support is 1, in which case we can overlook the modified confidence of this item set with other item sets. Consequently, the probability for any item set A's occurrence is $0<P(A)<1$. The range of modified Confidence is [-1,1] because $0 < P(AUC) / P(A) \leq 1$ and $0 < P(\neg AUC)/P(\neg A) \leq 1$. If modified-Confidence > 0, we have $P(AUC) > P(A) * P(C)$, which proves that A and B are correlated. If modified-Confidence=1, we have $P(AUC) = P(A) = P(C)$, which indicates the case that A and C appear simultaneously.

Thus the new proposed framework "Modified Support-Modified Confidence" leads to find Strong rules and it avoids redundant and irrelated rules.

3.1 Algorithm using Modified-Support and Modified-Confidence

```
Input : R1,R2,R3….. Rk // a set of association rules.;
        Min-support, min-conf,
Output : RS// asset of final effective association rules.
(1) For i=1 to k    // R1,R2,……..Rk , each of association rule L→R
(2)     If ( Support (LR)>=min-support) && (Support(L)+Support(R) >1) then goto step 5
(3)     If (Support(L)+Support(R) ≤1) then
            Calculate Support($\neg R \Rightarrow \neg L$) =
          $1 -$ Support(R)−Support(L)+Support(L$\Rightarrow R$).
(4)     If ( Support(LR)>=min-support) && (Support($\neg R \Rightarrow \neg L$) >=min-support) then
(5)     Conf(L→R)=Support(LR)/Support(L);
(6)     Conf($\neg R \Rightarrow L$)= Support(¬LR)/
                              Support(¬L);
(7)     If(Conf(L→R)-Conf($\neg R \Rightarrow L$))>=min-conf) then
```

(8) $RS \leftarrow Ri$

4. Experiments :

Implementing the concept of modified support and modified Confidence can reduce the occurrence of redundant rules. Frequent item sets are produced using Frequent item set mining algorithm. Association rules under modified Support-Modified Confidence framework are compared with Support-Confidence framework. The comparison is based on the number of association rules produced and the generation of effective rules. Parameters for Comparison is : min-sup =0.4 and min-confidence=0.6.

Table 2 : Implementation Results

NUMBER OF RECORDS	NUMBER OF RULES IN MODIFIED SUPPORT-MODIFIED CONFIDENCE
1000	10
2000	20
3000	15
4000	13
5000	15

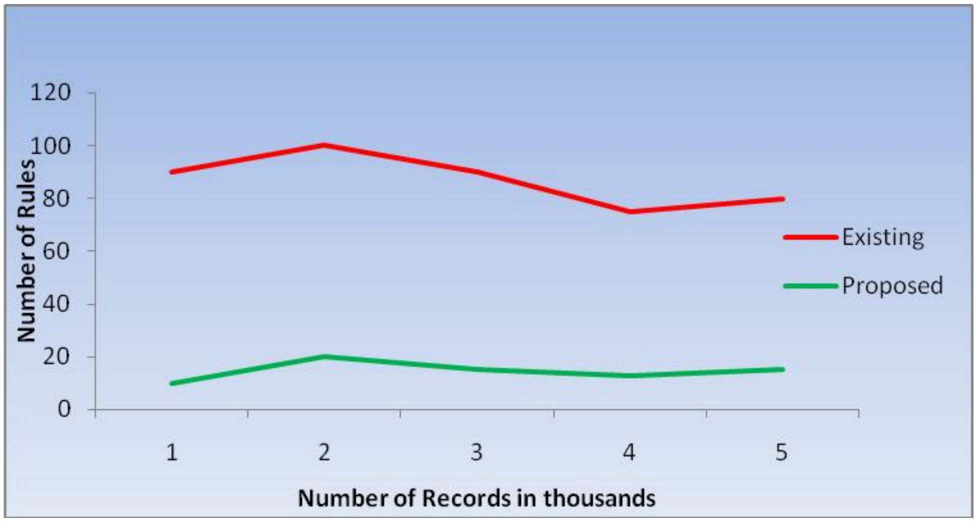

Fig : 1 Comparison in Adult Database

The result of the experiment is shown in Fig 1 and table 2. It shows the proposed modified support--modified Confidence framework can generate effective association rules.

5. Conclusions

In this paper, we discuss and analyze the measure support-confidence framework as well as the limitation of it and propose the modified-support and modified- confidence framework. The proposed measure standard has the advantages of reducing the creation of redundant and irrelated rules.

References

[1] C.Silverstein, S. Brin and R. Motwani, Beyond market baskets: Generalizing association rules to dependence rules,Data Mining and Knowledge Discovery 2 (1998), 39–68.

[2] CH.Sandeep Kumar, K.Srinivas, Peddi Kishor, T.Bhaskar, “An Alternative Approach to Mine Association Rules”, Electronics Computer Technology(ICECT), 2011 3rd International Conference , April 2011,420-424

[3] Ghada Almodaifer, Alaadin Hafez and Hassan Mathkour,”Discovering Medical Association Rules from Medical Datasets”, IT in Medicine and Education (ITME), 2011 ,Volume:2, Page(s): 43 – 47

[4] Izwan Nizal Mohd. Shaharanee ⇑, Fedja Hadzic, Tharam S. Dillon,”Interestingness measures for association rules based on statistical validity”, Knowledge-Based Systems 24 (2011), 386–392

[5] Luo Ke,Wu Jie,How to get valid association rules,Mini-micro System 23 (6) (2002) 711-713

[6] M. Sulaiman Khan, Maybin Muyeba, Frans Coenen,”A Weighted Utility Framework for Mining Association Rules”, EEE Computer Society Washington, DC, USA 2008, 87-92

[7] Michael Hahsler, Kurt Hornik ,“ New Probabilistic Interest Measures For Association rules”, Journal of Intelligent Data Analysis, Volume 11 Issue 5, October 2007 ,Pages 437-455

[8] Ramesh Agrawal, Tomasz Imielinski, and A. Swami, “Mining association rules between sets of items in large databases”, ACM-SIGMOD Int. Conf. Management of Data, Washington,D.C., May 1993, pp 207–216.

[9] S.Brin, R. Motwani, J.D. Ullman and S. Tsur, Dynamic itemset counting and implication rules for market basket data,SIGMOD Record 26(2) (1997), 255–264.