SunithaVanamala, L.Padma sree, S.Durga Bhavani / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 3, May-Jun 2013, pp.753-757

Efficient Rare Association Rule Mining Algorithm

SunithaVanamala*, L.Padma sree **, S.Durga Bhavani***
*(Assistant Professor, Department of Information Technology, Kakatiya Institute of Technology and Science, Warangal,)
** (Professor, Department of ECE, VNR Vignan Jyothi Institute of Technology and science, HYD)
***(Professor, School of Information Technology , JNTUH ,HYD,)

Abstract

Data mining is the process of discovering correlations, patterns, trends or relationships by searching through a large amount of data stored in repositories, corporate databases, and data warehouses. In Data mining field, the primary task is to mine frequent item sets from a transaction database using Association Rule Mining (ARM).Whereas the extraction of frequent patterns has focused the major researches in association rule mining, the requirements of reliable rules that do not frequently appear is taking an increasing interest in a great number of areas. Rare association rule refers to an association rule forming between frequent and rare items or among rare items. In many cases, the contradictions or exceptions also offers useful associations. Recent researches focus on the discovery of such kind of associations called rare associations. The mining of associations involving rare items is referred as rare association rule mining. Approaches to association rule mining uses single minimum support for identifying frequent associations. To mine interesting rare association rules, single minimum support approaches are not useful. Hence, we propose an algorithm based on MSApriori ,we call this new algorithm as MSApriori_VDB which uses vertical database format. Experimental results shows that our algorithms out performs previous approaches in both memory requirement and execution time by reducing the number of database scans.

Keywords—*Data Mining, Association rules, Rare items, Rare Association rule mining, MSAPriri_vdb*

1. INTRODUCTION

Nowadays most research on Association Rule Mining (ARM) [1] [10][11] has been focused on discovering common patterns and rules in large datasets. In fact, ARM is in various application areas such as telecommunication networks, market and risk management, inventory control, mobile mining, graph mining, educational mining, etc. The patterns and rules discovered are based on the majority of commonly repeated items in the dataset, though some of these data can be either obvious or irrelevant. Unfortunately, not enough attention has been paid to the extraction process of rare association rules, also known as non-frequent, unusual, exceptional or sporadic rules, which provide valuable knowledge about non-frequent patterns. The goal of Rare Association Rule Mining (RARM) is to discover rare and low-rank item sets to generate meaningful rules from these items. This type of rule cannot be identified easily using traditional association mining algorithms.

Given a transaction database, the association rule mining task involves two steps.

1. Get frequent item sets and
2. Automatic generation of interesting association rules

The main problem in the association rule mining is setting the MST. A low support threshold generates too many itemsets sometimes uninteresting item sets but a high MST misses few rare itemsets. The real world database may have items that are of varying frequencies. Some items appear frequently in transactions and some of them appear rarely. The rare itemsets may also be interesting. A rare itemset is an itemset consisting of rare items. It may be found by setting a low support threshold but leads to large number association rules consisting of both interesting and uninteresting rules. But it is difficult to mine rare association rules using single support threshold based approaches like Apriori and Frequent Pattern-Growth (FP-Growth). The problem of specifying an appropriate support threshold causes rare item problem. rare associations are of two types.

1. Interesting rare association
2. Uninteresting rare association

Definition: Interesting rare association – An association is said to be interesting rare association if it has low support but the confidence of the association is high.

Definition : Uninteresting rare association – An association is said to be uninteresting rare association if it has low support and low confidence.

1.1 Problem statement

In this paper we first pre-process the dataset by converting into vertical data format before performing association rule mining. The rationale behind the conversion process prior to mining association rules is that to reduce the number of database scans performed on original

database. The second step deals with the mining of interesting rare associations using converted vertical data and to deal with the rare association problem, a multiple minimum support framework is being used. The framework assigns each item in the transaction dataset a minsup called Minimum Item Support (MIS).

Our process of generating rare associations consist of two phases 1. convert to vertical data format 2. finding interesting rare associations i,e rules with high confidence.

2. Related Work

Association Rule Mining algorithms [1] [2] [3] are well-known data mining methods for discovering interesting relationships between variables in transaction databases or other data repositories. An association rule is an implication X⇒Y, where X and Y are disjoint itemsets (i.e., sets with no items in common). The intuitive meaning of such a rule is that when X appears, Y also tends to appear. The two traditional measures for evaluating association rules are support and confidence. The confidence of an association rule X⇒Y is the proportion of the transactions containing X which also contain Y. The support of the rule is the fraction of the database that contains both X and Y. The problem of association rule mining is usually broken down into two subtasks. The first one is to discover those itemsets whose occurrences exceed a predefined support threshold(MST), and which are called frequent itemsets. A second task is to generate association rules from those large items constrained by minimal confidence(MCT). Though these algorithms are theoretically expected to be capable of finding rare association rules, they actually become intractable if the minimum level of support is set low enough to find rare rules.

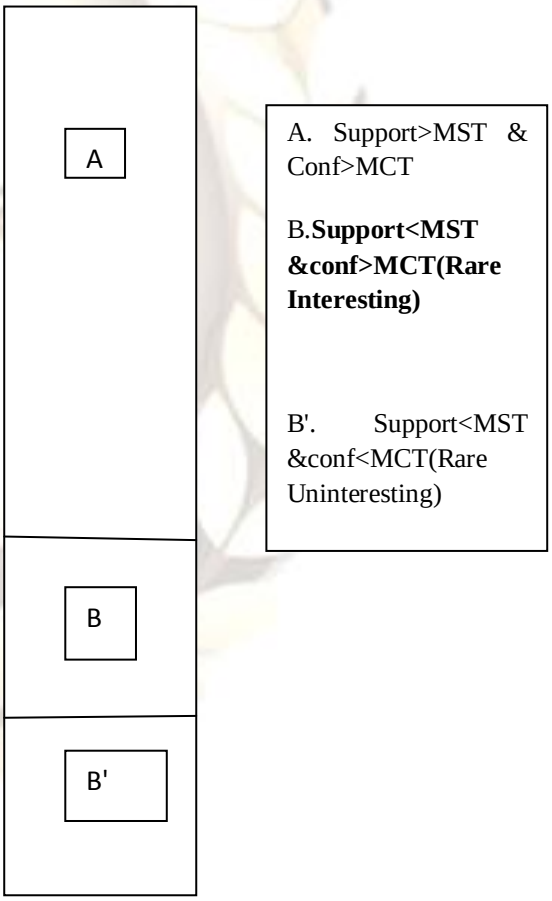
2.1 Data formats

There are two data formats to be adopted. One is the horizontal and other is vertical data format. The horizontal data format is the same as that stored in a database. The vertical data format is a conversion from the horizontal one, with the transaction identifiers grouped for each item. Algorithms for mining frequent item sets based on the vertical data format are usually more efficient than those based on the horizontal. Because the former often scan the database only once and compute the supports of item sets fast. The disadvantage is that it uses more memory for additional information, like Tidsets (Zaki et al., 1997; Zaki and Hsiao, 2005) or BitTable

Zaki et al. proposed for mining Frequent item sets by using additional data structure called TidSet that stores Tid for each data item in database. The support of item set x can be calculated by the cardinality of item in the corresponding Tidset. Thus support(X)=|Tidset(x)|, Tidset of two item sets can also be found easily by computing intersection of corresponding Tidset's (one items sets), i,e Tidset(xy)=Tidset(x)∩Tiidset(y).

The problem of discovering rare items has recently captured the interest of the data mining community [4]. As previously explained, rare itemsets are those that only appear together in very few transactions or some very small percentage of transactions in the database. Rare association rules have low support and high confidence in contrast to general association rules which are determined by high support and a high confidence level. Figure 1 illustrates how the support measure behaves in relation to the two types of rules.

Figure 1. Rules in a database.
The figure shows Interesting and uninteresting rare association rule



There are several different approaches to discover rare association rules. The simplest way is to directly apply the Apriori algorithm [1] by simply setting the minimum support threshold to a low value. However, this leads to a combinatorial explosion, which could produce a huge number of patterns, most of them frequent with only a small number of them actually rare. A different proposal, known as Apriori-Infrequent, involves the modification of the Apriori algorithm to use only the above-mentioned infrequent itemsets during rule generation. This simple change makes use of the maximum support measure, instead of the usual minimum support, to generate candidate itemsets, i.e., only items with a lower support than a given threshold are considered. Next, rules are yielded as generated by the Apriori algorithm. A totally different perspective consists of developing a new algorithm to tackle these new challenges. A first proposal is Apriori-Inverse [4], which can be seen as a more intricate variation of the traditional Apriori algorithm. It also uses the maximum support but proposes three different kinds of additions: fixed threshold, adaptive threshold and hill climbing. The main idea is that given a user-specified maximum support threshold, MaxSup, and a derived MinAbsSup value, a rule X is rare if Sup(X) < MaxSup and Sup(X)> MinAbsSup. A second proposal is the Apriori-Rare algorithm [11], also known as Arima, which is another variation of the Apriori approach. Arima is actually composed of two different algorithms: a naïve one, which relies on Apriori and hence enumerates all frequent itemsets; and MRG-Exp, which limits the considerations to frequent itemsets generators only. Finally, please notice that the first two approaches (Apriori-Frequent and Apriori-Infrequent) are taken to ensure that rare items are also considered during itemset generation, although the two latter approaches (Apriori-Inverse and Apriori-Rare) try to encourage low-support items to take part in candidate rule generation by imposing structural constraints.

Kanimozhi et al [7] proposed an approach based on multiple minimum support to efficiently discover rules that have a confidence value of hundred percent and also presented the calculation of the appropriate minimum support based on the frequency of the items.

Ravi Kumar [8], et al proposed efficient rare association rule generation by constructing the compact MIS-tree that uses the notion of "least minimum support" and "infrequent child node pruning. The approach is based on FP-Growth..

Cristóbal et al [9] proposed algorithm to extract rare associations from e_learing data by gathering student usage data from a Moodle system

3.Proposed work

Rare association rule mining is the process of finding the item sets involving low frequent items but at the same time the confidence of the association is strong i.e. above the minimum confidence level.

//finds all the rare patterns of interest and generates rules

Algorithm :MSApriori_VDB
Input : Transaction data Base D
Output :L, Interesting patterns in D
R, High confidence rules
procedure :
D'<=conv_vdb(D)
L1= find frequent 1 itemsets(D') //Remove any infrequent 1 itemsets

Calculate support of I for all I Є L1
MIS(I) <= support(I)
for (k = 2; Lk-1 ≠Ø, k++) do
Ck=candidate-gen (Lk-1)
end
for each candidate c Є Ck
c.count=get_count(c)
if c.count =0 then delete c;
MIS(c)= c.count | c.count is minimum for all Ck-1
Lk = {c Є_ Ck | c.count = MIS(c)}
If c.count=MIS(c) then
LHS = c | c.count is minimum for all Ck-1
Rk =Form_Rule(LHS,c)
Return ∪k Rk;
return ∪k Lk;
end
end
Procedure Candidate-gen(Lk-1)
for each itemset l1 Є tidset.[Lk-1]
for each itemset l2 Є tidset[Lk-1]
perform intersection operation l1, l2
if has_infrequent(c, Lk-1)
prune c;
else
tidset[k][i]].append(I1∩I2)
end if
end
end
return tidset[Ck];

Procedure has_infrequent(c, Lk-1)
for each (k-1) subset s of c
if s is in Lk-1
return false;
else
return true;
end if
end

Procedure Gen_Rule(LHS,c)
RHS=c minus LHS
Write rule in the format LHS⇒RHS
end

procedure conv_vdb(D)
for each tid in D
for each i in tid
tidset[1][pos(i)].appen**d(tid)**
end
end

3.1.Calculating MIS for item sets

Initially the dataset is scanned to calculate the support of the first level items. During the first level, the MIS value for each item is its support. 1-itemset is generated and arranged in the decreasing order of their support value.
Given an item X, at level 1, MIS(X)=sup(X)
According to the example given in Table1, the MIS values for the itemset{A, B,C,E, D} are {7, 6, 5, 3, 3}. All items in the first iteration is considered to be interesting and moved to the next iteration. In subsequent levels, the MIS value for each itemset is computed as the minimum support among the subset of items contained in the itemset. 2-itemset candidates are generated by intersecting corresponding 1-itemsets and support for each candidate is |TIDSLIST|. Itemsets with zero support are eliminated. Then the MIS value at the second level is calculated as follows:

Given an item set(P,Q), MIS(P,Q) = min(sup(P), sup(Q)) Rules which are not found to be interesting are pruned. Now, the association rule can be formed by finding the left hand side (LHS) of the pattern. The item with the lowest support value in the item set is considered as LHS.

Example:

The Table1 shows the original data set which is to be scanned once and it is converted to the vertical data format(item, Tidlist) as shown in Table2. MST is assumed as 30%. The length of Tid list gives the support of corresponding data item. Hence we don't require to scan the data base to calculate the support values. To calculate the support of two item sets we have to intersect the corresponding one item sets ,from this we get Table3. Table4 shows Rare Association rules for interesting two item sets. Following the similar procedure i.e. by intersecting two item sets we will get the corresponding 3 item sets. Similarly we can get n item set from (n-1) item set.

Table1.Dataset

Tid	**Items purchased**
1	A,B
2	A,D
3	B,C
4	B, E
5	A,B,C
6	A,D,B
7	A,D
8	A,C
9	B,E,C

Table2.Vertical data format

ITEM	TID_LIST	supp	MIS
A	1,2,5,6,7,8,9	7	7
B	1,3,4,5,6,9,10	7	7
C	3,5,8,9,10	5	5
D	2,6,7	3	3
E	4,9,10	3	3

Table3:TIDSET for 2 itemsets

ITEM	TID_LIST	**MIS**
A,B	1,5,6,9	6
A,C	5,8,9	5
A,D	2,6,7	3
A,E	7	3
B,C	3,5,9	5
B,D	6	3
B,E	4,9,10	3
C,E	9,10	3
A,B	1,5,6,9	6

Table4:Rare Association Rules with 2 item sets

Itemset	**Sup**	**MIS**	**LHS**	**Rule**
A, D	3	3	D	D → A
B,E	3	3	E	E → B

4. Conclusion

Our approach first performs conversion of original database into vertical data format and then generates the rare association rules from the

converted database. The approach is very simple to implement. The algorithm uses multiple minimum supports which are calculated based on the frequency of occurrence. Hence the approach reduces the burden of assumption about minimum support threshold. Advantage of vertical data format is that reduces the number database scans to one. One possible direction for future work would be to reduce the size of tidlist , which is proportional to number of transactions, so that additional memory requirements can be reduced. This would lead to the generation rare rules efficiently than would be possible with the MSAppriori_VDB

References

[1] Agrawal, R. and R. Srikanth, 1994. Fast algorithms for mining association rules. *VLDB.*

[2] Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large databases. *SIGMOD,* 207-216.

[3] Kanimozhi Selvi C.S, and A. Tamilarasi, 2009. An automated association rule mining technique with cumulative support thresholds. *Int. J. Open Problems Comput. Sci. Math., 2.*

[4] Koh, Y. and N. Rountree, 2005. Finding sporadic rules using apriori-inverse. *Proceeding Of PAKDD '05, Hanoi, Vietnam, LNCS, Springer, pp: 97-106.*

[5] Liu, B., Hsu, W. & Ma, Y. (1999), Mining associationrules with multiple minimum supports, *in 'Proceedingsof the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining',*

[6] Adda, M., Wu, L, Feng, Y. Rare itemset mining. *In Sixth Conference on Machine Learning and Applications. 2007, Cincinnati, Ohio, USA, pp-73-80.*

[7] C. S. Kanimozhi Selvi, A. Tamilarasi 2011 Mining of High Confidence Rare Association Rules .E*JSR ISSN*

[8]. T. Ravi Kumar , K. Raghava Rao 2011 Association Rule Mining using Improved FPGrowth .

[9] Cristóbal Romero, José Raúl Romero et al Mining Rare Association Rules from e-Learning Data

[10] B. Nath, D. K. Bhattacharyya, and A. Gosh. Faster generation of association rules. *volume 1, pages 267–279.IJITKM, 2008.*

[11] B. Nath and A. Ghosh. Multi-objective rule mining using genetic algorithm. *pages 123–133. Information Science 163, 2004.*

[12] R.Srikant and R. Agrawala. Mining generalized association rules. *pages 407–419. Proceedings of the 21st VLDB Conference Zurich, Swizerland, 1995.*