Mining Interesting Infrequent and Frequent Itemsets Based on MLMS Model

Xiangjun Dong[1,2], Zhendong Niu[3], Donghua Zhu[1],
Zhiyun Zheng[4], and Qiuting Jia[2]

[1] School of Management and Economics, Beijing Institute of Technology,
Beijing 100081, China
dxj@sdili.edu.cn, zhudh111@bit.edu.cn
[2] School of Information Science and Technology, Shandong Institute of Light Industry,
Jinan 250353, China
[3] School of Computer Science and Technology, Beijing Institute of Technology,
Beijing 100081, China
zniu@bit.edu.cn
[4] School of Information & Engineering, Zhengzhou University,
Zhengzhou 450052, China
iezyzheng@zzu.edu.cn

Abstract. MLMS (Multiple Level Minimum Supports) model which uses multiple level minimum supports to discover infrequent itemsets and frequent itemsets simultaneously is proposed in our previous work. The reason to discover infrequent itemsets is that there are many valued negative association rules in them. However, some of the itemsets discovered by the MLMS model are not interesting and ought to be pruned. In one of Xindong Wu's papers [1], a pruning strategy (we call it Wu's pruning strategy here) is used to prune uninteresting itemsets. But the pruning strategy is only applied to single minimum support. In this paper, we modify the Wu's pruning strategy to adapt to the MLMS model to prune uninteresting itemsets and we call the MLMS model with the modified Wu's pruning strategy IMLMS (Interesting MLMS) model. Based on the IMLMS model, we design an algorithm to discover simultaneously both interesting frequent itemsets and interesting infrequent itemsets. The experimental results show the validity of the model.

1 Introduction

Mining association rules in database has received much attention recently. Many efforts have been devoted to design algorithms for efficiently discovering association rules of the form $A \Rightarrow B$, whose support (s) and confidence (c) meet a minimum support (ms) threshold and a minimum confidence (mc) respectively. This is the famous support-confidence framework [2]. The rules of the form $A \Rightarrow B$ have been studied widely since then, while the rules of the form $A \Rightarrow \neg B$ have only been attracted a little attention until recently some papers proposed that the rules of the form $A \Rightarrow \neg B$ could play the same important roles as the rules of the form $A \Rightarrow B$. The rules of the form $A \Rightarrow \neg B$, which suggests that a customer may not buy item B after the customer buys

C. Tang et al. (Eds.): ADMA 2008, LNAI 5139, pp. 444–451, 2008.

item A, together with the rules of the forms $\neg A \Rightarrow B$ and $\neg A \Rightarrow \neg B$, are called ***negative association rules*** **(NARs)** and the rules of the form $A \Rightarrow B$ ***positive association rules*** **(PARs)** [1].

Discovering infrequent itemsets is one of the new problems that would occur when we study NARs. The reason to discover infrequent itemsets is that there are many valued negative association rules in them. How to discover infrequent itemsets, however, is still an open problem. Theoretically, infrequent itemsets are the complement of frequent itemsets and the number of infrequent itemsets is too large to be mined completely. So in real application, some constraints must be given so as to control the number of infrequent itemsets in a moderate degree. The constraints to infrequent itemsets in different applications are different. MLMS (Multiple Level Minimum Supports) model [3], which is proposed in our previous work, can be used to discover infrequent itemsets and frequent itemsets simultaneously. The details about the MLMS model are mentioned in section 3.1.

Another new problem is how to prune uninteresting itemsets because some itemsets are not interesting. In one of Xindong Wu's papers [1], a pruning strategy (we call it Wu's pruning strategy here) which concerns the support, the confidence, and interestingness of a rule $A \Rightarrow B$ is used to prune uninteresting itemsets. But the Wu's pruning strategy is only applied to single minimum support. In this paper, we modify the Wu's pruning strategy to adapt to the MLMS model to prune uninteresting itemsets and we call the MLMS model with the modified Wu's pruning strategy IMLMS (Interesting MLMS) model.

The main contributions of this paper are as follows:

1. We modify the Wu's pruning strategy to adapt to the MLMS model.
2. We design an algorithm *Apriori_IMLMS* to discover simultaneously both interesting infrequent itemsets and interesting frequent itemsets based on the IMLMS model.

The rest of the paper is organized as follows: Section 2 discusses the related works. Section 3 discusses the modified Wu's pruning strategy and the algorithm *Apriori_IMLMS*. Experiments are discussed in section 4. Section 5 is conclusions and future work.

2 Related Work

Negative relationships between two itemsets were first mentioned in [4]. [5] proposed a new approach for discovering strong NARs. [6] proposed another approach based on taxonomy for discovering PARs and NARs. In our previous work [7], we proposed an approach that can discover PARs and NARs by chi-squared test and multiple minimum confidences. [8] proposed an approach to discover confined PARs and NARs. In [9], a MMS (multiple minimum supports) model was proposed to discover frequent itemsets using multiple minimum supports to solve the problems caused by a single minimum support, [10] proposed an approach of mining multiple-level association rules based on taxonomy information.

These works, however, do not concentrate on discovering infrequent itemsets. There are many mature techniques to discover frequent itemsets, while few papers

study the approach of how to discover infrequent itemsets. In our previous work [11], a 2LS (2-Level Supports) model was proposed to discover infrequent itemsets and frequent itemsets, however, the 2LS model only used the same two constraints to divide the infrequent itemsets and frequent itemsets and did not prune uninteresting itemsets. In [1], the Wu's pruning strategy was used to prune uninteresting itemsets, however, the constrains to infrequent itemsets and frequent itemsets are still a single minimum support.

3 Mining Interesting Infrequent and Frequent Itemsets

3.1 Review of the MLMS Model

Let $I=\{i_1, i_2,\ldots, i_n\}$ be a set of n distinct literals called items, and TD a transaction database of variable-length transactions over I, and the number of transactions in TD is denoted as $|TD|$. Each transaction contains a set of item $i_1, i_2,\ldots,i_m \in I$ and each transaction is associated with a unique identifier TID. A set of distinct items from I is called an itemset. The number of items in an itemset A is the length of the itemset, denoted by $length(A)$. An itemset of length k are referred to as k-itemset. Each itemset has an associated statistical measure called support, denoted by s. For an itemset $A \subseteq I$, $s(A)=A.count / |TD|$, where $A.count$ is the number of transactions containing itemsets A in TD. The support of a rule $A \Rightarrow B$ is denoted as $s(A \cup B)$ or $s(A \Rightarrow B)$, where $A, B \subseteq I$,and $A \cap B = \Phi$. The confidence of the rule $A \Rightarrow B$ is defined as the ratio of $s(A \cup B)$ over $s(A)$, i.e., $c(A \Rightarrow B) = s(A \cup B) / s(A)$.

In the MLMS model, different minimum supports are assigned to itemsets with different lengths. Let $ms(k)$ be the minimum support of k-itemsets (k=1,2,…,n), $ms(0)$ be a threshold for infrequent itemsets, $ms(1) \geq ms(2) \geq,\ldots, \geq ms(n) \geq ms(0) > 0$, for any itemset A,

if $s(A) \geq ms(length(A))$, then A is a frequent itemset; and
if $s(A) < ms(length(A))$ and $s(A) \geq ms(0)$, then A is an infrequent itemset.

The MLMS model allows users to control the number of frequent and infrequent itemsets easily. The value of $ms(k)$ can be given by users or experts.

3.2 The Modified Wu's Pruning Strategy

The Wu's pruning strategy uses different methods to prune uninteresting itemsets from frequent and infrequent itemsets. The following equations are used to prune uninteresting frequent itemsets.

M is a *frequent itemset of potential interest* if

$$fipi(M) = s(M) \geq ms \wedge (\exists A, B: A \cup B = M) \wedge fipis(A, B), \quad (1)$$

where $fipis(A, B) = (A \cap B) = \Phi \quad \wedge \quad f(A, B, ms, mc, mi) = 1,$ (2)

$$f(A, B, ms, mc, mi) \quad (3)$$

$$= \frac{s(A \cup B) + c(A \Rightarrow B) + interest(A,B) - (ms + mc + mi) + 1}{|\, s(A \cup B) - ms \,| + |\, c(A \Rightarrow B) - mc \,| + |\, interest(A,B) - mi \,| + 1},$$

where *interest*(*A*,*B*) is a interestingness measure and *interest*(*A*,*B*)=|*s*(*A*∪*B*) - *s*(*A*)*s*(*B*)|, *mi* is a minimum interestingness threshold given by users or experts, and *f* () is a constraint function concerning the support, confidence, and interestingness of $A \Rightarrow B$.

Since $c(A \Rightarrow B)$ is used to mine association rules after discovering frequent itemsets and infrequent itemsets, here we replace *f* (*A*,*B*, *ms*, *mc*, *mi*) with *f* (*A*,*B*, *ms*, *mi*) as [1] did.

From the equations 1-3 we can see that the Wu's pruning strategy is only applied to single minimum support *ms*. Considering that there are different minimum support s *ms*(*k*) for *k*-itemsets in the MLMS model, we must modify the Wu's pruning strategy by replacing corresponding variables as follows.

M is a *frequent itemset of potential interest* in the MLMS model if

$$fipi(M) = s(M) \geq ms(length(M)) \wedge (\exists A, B: A \cup B = M) \wedge fipis(A, B), \tag{4}$$

where $fipis(A, B) = A \cap B = \Phi \wedge f(A,B, ms(length(A \cup B)), mi) = 1,$ (5)

$$f(A, B, ms(length(A \cup B)), mi)$$

$$= \frac{s(A \cup B) + interest(A,B) - (ms(length(A \cup B)) + mi) + 1}{|\, s(A \cup B) - ms(length(A \cup B)) \,| + |\, interest(A,B) - mi \,| + 1}. \tag{6}$$

Equations 4-6 can be used to prune uninteresting frequent itemsets in the MLMS model. As for pruning uninteresting infrequent itemsets in the MLMS model, we can not use the methods of pruning uninteresting infrequent itemsets in the Wu's pruning strategy because the constraints to infrequent itemsets in the MLMS model and in [1] are different. Fortunately, according to the above discussion, we can get the following equations by modifying the equations 4-6.

N is an infrequent itemset of potential interest if

$$iipi(N) = s(N) < ms(length(N)) \wedge s(N) \geq ms(0) \wedge$$

$$(\exists A, B: A \cup B = N) \wedge iipis(A, B), \tag{7}$$

where $iipis(A, B) = A \cap B = \Phi \wedge f(A,B, ms(0), mi) = 1,$ (8)

$$f(A, B, ms(0), mi) = \frac{s(A \cup B) + interest(A,B) - (ms(0) + mi) + 1}{|\, s(A \cup B) - ms(0) \,| + |\, interest(A,B) - mi \,| + 1}. \tag{9}$$

The IMLMS model is to use equation 4 -9 to prune the uninteresting frequent itemsets and the uninteresting infrequent itemsets discovered by the MLMS model.

3.3 Algorithm Design

```
Algorithm Apriori_IMLMS
Input: TD: Transaction Database; ms(k)(k=0,1,…,n):
       minimum support threshold;
Output: FIS: set of interesting frequent itemsets;inFIS:
        set of interesting infrequent itemsets;
(1) FIS=Φ; inFIS=Φ;
(2) temp1 = {A|A∈1-itemsets,s(A)≥ms(0)};
  FIS1 = {A|A∈temp1 ∧ s(A)≥ms(1)};
  inFIS1 = temp1-FIS1;
(3) for (k=2;tempk□1≠Φ;k++) do
  begin
  (3.1) Ck = apriori_gen(tempk□1, ms(0));
  (3.2) for each transaction t∈TD do
      begin
        /*scan transaction database TD*/
        Ct=subset(Ck, t);
        for each candidate c ∈ Ct
            c.count++;
       end
  (3.3) tempk = {c|c∈Ck  (c.count/|TD|)≥ms(0)};
      FISk = {A|A∈tempk ∧A.count/|TD|≥ms(k)};
      inFISk = tempk – FISk ;
  (3.4) /*prune all uninteresting k-itemsets in FISk */
      for each itemset  M in FISk do
          if NOT (fipi(M)) then
           FISk =FISk – { M }
  (3.5)/*prune all uninteresting k-itemsets in inFISk */
      for each itemset N in inFISk do
           if NOT (iipi(N)) then
           inFISk =inFISk – { N }
  end
(4) FIS = ∪FISk; inFIS = ∪inFISk;
(5) return FIS and inFIS;
```

The algorithm *Apriori_IMLMS* is used to generate all interesting frequent and interesting infrequent itemsets in a given database *TD*, where *FIS* is the set of all interesting frequent itemsets, and *inFIS* is the set of all interesting infrequent itemsets. There are four kinds of sets: FIS_k, $inFIS_k$, $temp_k$ and C_k. $Temp_k$ contains the itemsets whose support meets the constraint $ms(0)$. C_k contains the itemsets generated by the procedure *apriori_gen*. The procedure *apriori_gen* is the same as the procedure in traditional *Apriori* [2] and is omitted here. The main steps of the algorithm *Apriori_IMLMS* are the same as the *Apriori_MLMS* except steps (3.4) and (3.5).

In step (3.4), if an itemset M in FIS_k does not meet $fipi(M)$, then M is an uninteresting frequent itemset, and is removed from FIS_k. In step (3.5), if an itemset N in $inFIS_k$

does not meet *iipi*(*N*), then *N* is an uninteresting infrequent itemset, and is removed from $inFIS_k$.

3.4 An Example

The sample data is shown in Table 1. Let $ms(1)=0.5$, $ms(2)=0.4$, $ms(3)=0.3$, $ms(4)=0.2$ and $ms=0.15$. FIS_1, FIS_2, $inFIS_1$, and $inFIS_2$ generated by the MLMS model are shown in Table 2, where the itemsets with gray background are infrequent itemsets with non background are frequent itemsets and the itemsets.

Table 1. Sample Data

TID	itemsets
T_1	*A*, *B*, *D*
T_2	*A*, *B*, *C*, *D*
T_3	*B*, *D*
T_4	*B*, *C*, *D*, *E*
T_5	*A*, *C*, *E*
T_6	*B*, *D*, *F*
T_7	*A*, *E*, *F*
T_8	*C*, *F*
T_9	*B*, *C*, *F*
T_{10}	*A*, *B*, *C*, *D*, *F*

Table 2. FIS_1, FIS_2,$inFIS_1$, and $inFIS_2$ generated by the MLMS model from the data in table 1

1-itemsets		2- itemsets			
	s(*)		*s*(*)		*s*(*)
A	0.5	*BD*	0.6	*CD*	0.3
B	0.7	*BC*	0.4	*CF*	0.3
C	0.6	*AB*	0.3	*AE*	0.2
D	0.6	*AC*	0.3	*AF*	0.2
E	0.3	*AD*	0.3	*CE*	0.2
F	0.5	*BF*	0.3	*DF*	0.2

Take FIS_2 ={ *BD*, *BC* }and $inFIS_2$={*AB*,*AC*,*AD*,BF,*CD*,*CF*,*AE*,*AF*,*CE*,*DF*} for example. Suppose *mi*=0.05. The results are shown in Fig. 1.

$f(B, D, ms(length(B\cup D)), mi) = \frac{0.6+0.18-(0.4+0.05)+1}{\lvert 0.6-0.4\rvert+\lvert(0.18-0.05\rvert+1} = 1$	$f(B, C, ms(length(B\cup C)), mi) = \frac{0.4+0.02-(0.4+0.05)+1}{\lvert 0.4-0.4\rvert+\lvert 0.02-0.05\rvert+1} < 1$
$f(A, B, ms(0), mi)= \frac{0.3+0.05-(0.2+0.05)+1}{\lvert 0.3-0.2\rvert+\lvert 0.05-0.05\rvert+1} = 1$	$f(C, F, ms(0), mi)= \frac{0.3+0-(0.2+0.05)+1}{\lvert 0.3-0.2\rvert+\lvert 0-0.05\rvert+1} < 1$
$f(A, C, ms(0), mi)= \frac{0.3+0-(0.2+0.05)+1}{\lvert 0.3-0.2\rvert+\lvert 0-0.05\rvert+1} < 1$	$f(A, E, ms(0), mi) = \frac{0.2+0.05-(0.2+0.05)+1}{\lvert 0.2-0.2\rvert+\lvert 0.05-0.05\rvert+1} = 1$
$f(A, D, ms(0), mi)= \frac{0.3+0-(0.2+0.05)+1}{\lvert 0.3-0.2\rvert+\lvert 0-0.05\rvert+1} < 1$	$f(A, F, ms(0), mi) = \frac{0.2+0.05-(0.2+0.05)+1}{\lvert 0.2-0.2\rvert+\lvert 0.05-0.05\rvert+1} = 1$
$f(B, F, ms(0), mi)= \frac{0.3+0.05-(0.2+0.05)+1}{\lvert 0.3-0.2\rvert+\lvert 0.05-0.05\rvert+1} = 1$	$f(C, E, ms(0), mi) = \frac{0.2+0.02-(0.2+0.05)+1}{\lvert 0.2-0.2\rvert+\lvert 0.02-0.05\rvert+1} < 1$
$f(C, D, ms(0), mi)= \frac{0.3+0.06-(0.2+0.05)+1}{\lvert 0.3-0.2\rvert+\lvert 0.06-0.05\rvert+1} = 1$	$f(D, F, ms(0), mi) = \frac{0.2+0.1-(0.2+0.05)+1}{\lvert 0.2-0.2\rvert+\lvert 0.1-0.05\rvert+1} = 1$

Fig. 1. The results of calculation for the itemsets in FIS2 and inFIS2

BC is not interesting, and therefore is removed from FIS_2. *AC*, *AD*, *CF*, *CE* are not interesting, and therefore are removed from $inFIS_2$. Other itemsets can be analyzed in the same way.

4 Experiments

The real dataset records areas of www.microsoft.com that each user visited in a one-week timeframe in February 1998. Summary statistical information of the dataset is: 32711 training instances, 5000 testing instances, 294 attributes and the mean area visits per case is 3.0 (http://www.cse.ohio-state.edu/~yanghu /CIS788_dm_proj.htm# datasets).

Table 3. The number of the interesting inFISand the interesting FIS with different mi.(ms(1) =0.025, ms(2)=0.02, ms(3)=0.017, ms(4)=0.013, ms(0)=0.01)

mi		*k*=1	*k*=2	*k*=3	*k*=4	Total	
0	*FIS*	-	37	24	3	64	150
	inFIS	-	43	40	3	86	
0.005	*FIS*	-	30	24	3	57	119
	inFIS	-	21	38	3	62	
0.01	*FIS*	-	23	21	3	47	63
	inFIS	-	7	8	1	16	
0.015	*FIS*	-	20	13	1	34	34
	inFIS	-	0	0	0	0	

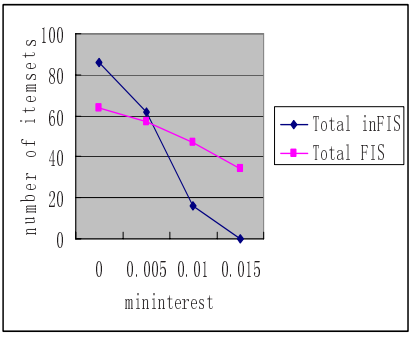

Fig. 2. The changes of total *inFIS* and total *FIS* with different *mi*

Table 3 shows the number of the interesting infrequent itemsets and the interesting frequent itemsets with different *mi* when *ms*(1)=0.025, *ms*(2)=0.02, *ms*(3)=0.017, *ms*(4)=0.013, and *ms*(0)=0.01. From table 3 we can see that the total number of *FIS* is 64,57,47,34, and total number of *inFIS* is 86,62,16,0 when *mi*=0, 0.005, 0.01, 0.015 respectively. With *mi* increasing, the total number decreases obviously. Figure 2 illustrates these changes. Table 3 also gives the number of *FIS* and *inFIS* in different *k*. These data show that *fipi*(*M*) and *iipi*(*N*) can efficiently prune the uninteresting itemsets in the IMLMS model, i.e., the modified Wu's pruning strategy can efficiently work on the MLMS model.

5 Conclusions and Future Work

When we study negative association rules, infrequent itemsets become very important because there are many valued negative association rules in them. In our previous work, a MLMS model is proposed to discover both infrequent itemsets and frequent itemsets. Some of the itemsets, however, are not interesting and ought to be pruned. In this paper, a new model IMLMS is proposed by modifying the Wu's pruning strategy to adapt to the MLMS model to prune the uninteresting itemsets. An algorithm *Apriori_IMLMS* is also proposed to discover simultaneously both interesting frequent itemsets and interesting infrequent itemsets based on the IMLMS model. The experimental results show the validity of the model.

For future work, we will work on finding interesting infrequent itemsets with some new measures and use some better algorithms than *Apriori*, FP-growth, for example, to discover infrequent itemsets and frequent itemsets.

Acknowledgements

This work was supported by Program for New Century Excellent Talents in University, China; Huo Ying-dong Education Foundation, China (91101); Excellent Young Scientist Foundation of Shandong Province, China (2006BS01017); and Natural Science Foundation of Shandong Province, China (Y2007G25).

References

1. Wu, X., Zhang, C., Zhang, S.: Efficient Mining of both Positive and Negative Association Rules. ACM Transactions on Information Systems 22, 381–405 (2004)
2. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Database. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216. ACM Press, New York (1993)
3. Dong, X., Zheng, Z., Niu, Z., Jia, Q.: Mining Infrequent Itemsets based on Multiple Level Minimum Supports. In: Proceedings of the Second International Conference on Innovative Computing, Information and Control (ICICIC 2007), Kumamoto, Japan (September 2007)
4. Brin, S., Motwani, R., Silverstein, C.: Beyond Market: Generalizing Association Rules to Correlations. In: Processing of the ACM SIGMOD Conference, pp. 265–276 (1997)
5. Savasere, A., Omiecinski, E., Navathe, S.: Mining for Strong Negative Associations in a Large Database of Customer Transaction. In: Proceedings of the 1998 International Conference on Data Engineering, pp. 494–502 (1998)
6. Yuan, X., Buckles, B.P., Yuan, Z., Zhang, J.: Mining Negative Association Rules. In: Proceedings of The Seventh IEEE Symposium on Computers and Communications, pp. 623–629 (2002)
7. Dong, X., Sun, F., Han, X., Hou, R.: Study of Positive and Negative Association Rules Based on Multi-confidence and Chi-Squared Test. In: Li, X., Zaïane, O.R., Li, Z. (eds.) ADMA 2006. LNCS (LNAI), vol. 4093, pp. 100–109. Springer, Heidelberg (2006)
8. Antonie, M.-L., Zaiane, O.: Mining Positive and Negative Association Rules: An Approach for Confined Rules. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 27–38. Springer, Heidelberg (2004)
9. Liu, B., Hsu, W., Ma, Y.: Mining Association Rules with Multiple Minimum Supports. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 1999), San Diego, CA, USA, August 15-18, 1999, pp. 337–341 (1999)
10. Han, J., Fu, Y.: Mining Multiple-Level Association Rules in Large Databases. IEEE Transactions on Knowledge and Engineering 11(5), 798–805 (1999)
11. Dong, X., Wang, S., Song, H.: 2-level Support based Approach for Mining Positive & Negative Association Rules. Computer Engineering 31(10), 16–18 (2005)