Frequent Pattern Mining based on Multiple Minimum Support using Uncertain Dataset

Meenu Dave, Ph.D
Principal
Jagannath Gupta Institute of Engineering & Technology,
Jaipur, India

Hitesh Maharwal
M.Tech. Scholar
Department of Computer Science, Jagan Nath University,
Jaipur, India

ABSTRACT
Association rule mining plays a major role in decision making in the production and sales business area. It uses minimum support (minsup) and support confidence (supconf) as a base to generate the frequent patterns and strong association rules. Setting a single value of minsup for a transaction set doesn't seem feasible for some real life applications. Similarly the probabilistic value of items in the transaction set may be acceptable. So generating the frequent pattern from the uncertain dataset becomes a concern factor. This research work details the aforesaid problem and proposes a solution for the same.

Keywords
Association Rule Mining, Minimum Support (minsup), Support Confidence (supconf), Uncertain Dataset

1. INTRODUCTION
Consumables and low-price products are bought frequently, while the luxury goods, electric appliance and high-price products infrequently. In such a situation, if we set minimum support (minsup) too high, all the discovered patterns are concerned with those low-price products, which only contribute a small portion of the profit to the business and we miss the frequent itemsets involving rare items because rare items fail to satisfy high minsup. On the other hand, if we set minsup too low, we will generate too many meaningless frequent patterns and they will overload the decision makers, who may find it difficult to understand the patterns generated by data mining algorithms.

Rare associative rule mining is used for the infrequent data or rare items, in order to satisfy a user specified minimum support and a user specified support confidence at the same time [1]. Additionally, rare associative rule mining is provided with the use of multiple minimum support which is an important generalization of mining problem. In this we have to provide a multiple minimum support instead of setting a single support value for itemsets. Since each item can have its own minimum support, it is very difficult for users to set the appropriate thresholds for all items at a time. In practice, users need to tune items' supports and run the mining algorithm repeatedly until a satisfactory end is reached.

Certain dataset is the dataset that does not have probability as 0 or 1 (i.e. either they are present or not present). While in the case of uncertain dataset, there is certain probability of occurrence of the particular event. In uncertain dataset, little modification is made to the apriori algorithm as the existential probability is added instead of certain probability. The support count of each item is the product of existential probability corresponding to each itemset in which item is occurring. But if the existential probability of an item is very less, it does not contribute significantly to the support count. On the contrary, it increases the number of scans which result in increased CPU cost. As a solution to this, data trimming process is used which trims the items that have very low existential probability.

In order to provide a solution for the initially mentioned problem, the proposed algorithm for generating the frequent pattern, takes help from Apriori algorithm [2] [3]. For the rare association rule mining with multiple minimum support one of the optimized algorithm which has been in use is CFPGrwoth++ algorithm and to manipulate the probabilistic value of itemset, concept of expected support of UApriori algorithm is used. These two algorithms work individually in an efficient manner, one finds out the rare and frequent patterns, and the other generates frequent patterns from the uncertain dataset. But if one wants to generate rare patterns from the uncertain dataset, then use of any one algorithm does not yield proper results. In this research paper, this problem has been dealt with and a new algorithm has been proposed which makes use of both CFPGrowth++ and UApriori algorithm.

2. MULTIPLE MINIMUM SUPPORT USING UNCERTAIN DATASET: A FEW BASICS

2.1. Multiple Minimum Support
An itemset is a collection of items. Minsup is used to cut the search space and to limit the number of rules generated. Setting of single minsup value for itemset does not work in real-life application. In many applications, some items appear very frequently in the data, while others rarely appear. There are two problems when frequencies of items vary a great deal,

1. If minsup is set too high, one loses rules that have frequent items or rare items in the data.
2. In order to find rules that have both frequent and rare items, it is necessary to set minsup very low. This will produce explosion of too many meaningless rules.

Rather setting up single minsup value for the whole transaction set, multiple values of minsup are specified for each item in the transaction set [4].

2.2. Uncertain Data Model
Traditional frequent itmeset mining provides certainty in whether or not an item occurs within a particular transaction or not. On the other hand, probabilistic frequent itemset mining have the knowledge of probability of an item occurring within a particular transaction. The probabilistic values of each item is associated with it in the transaction dataset. The probabilistic values of the items are concerned where the products are customized on customers demand.

Table 1 shows an example of uncertain dataset, where E1 has two uncertain items, laptop and cleaning kit with existential

probabilities 0.93 and 0.85, respectively, and one certain item pen drive. Transactions E2, E3 and E4 both have two uncertain items and no certain items.

Table 1. Example Uncertain Dataset

TID	Uncertain Itemset
E_1	laptop(0.9), cleaning kit(0.8), anti-virus(1.0)
E_2	laptop(0.9), external hard disk(0.4)
E_3	laptop(0.9), anti-virus(0.8)
E_4	pen drive(0.8), antivirus(0.6)

To generate frequent pattern from the uncertain dataset, the concept of expected support value is adopted.

2.3. Expected Support

Given a world Wi and an itemset X , then P(Wi) be the probability of world Pi and S(X, Wi) be the support count of X in world Wi. Furthermore, Ti,j to denote the set of items that the jth transaction, i.e., tj , contains in the world Wi. If items' existential probabilities in transactions are determined through independent observations2, then P (Wi) and the expected support Se (X) of X are given by the following formulae:

$$P(Wi) = \prod_{j=1}^{d}(\prod_{x\in Ti,j} Ptj(x).\prod_{y\notin Ti,j}(1 - Ptj(y))) \quad (1)$$

$$\mathrm{Se}(X) = \sum_{i=1}^{|W|} \mathrm{P(Wi)xS(X, Wi)} \quad (2)$$

where W is the set of possible worlds derived from an uncertain dataset D [5].

Computing Se(X) according to equation 2 requires enumerating all possible worlds and finding the support count of X in each world. This is computationally infeasible since there are 2m possible worlds where m is the total number of items that occur in all transactions of D. Fortunately, we can show that

$$\mathrm{Se(X)} = \sum_{j=1}^{|D|} \prod_{x\in X} Ptj(x\ \mathrm{Ptj(x)}) \quad (3)$$

Thus, Se (X) can be computed by a single scan through the dataset D [5].

3. RELATED WORK

The occurrence of rare item problem with the usage of traditional data mining techniques to discover knowledge involving rare items was introduced by Weiss [6] in 2004. In [7], Liu et.al. introduced "multiple minsups framework" to address rare item problem and MSApriori algorithm for extracting frequent patterns. An FP-growth-like algorithm [8], called CFP-growth [9], has been proposed to mine frequent patterns. It was shown that the performance of CFP-growth is better than the MSApriori algorithm. In 2009, Kiran et.al. [4], proposed a preliminary algorithm to improve the performance of CFP-growth by suggesting two pruning techniques for reducing the size of constructed tree structure. CFPGrowth++ algorithm work with the multiple support value to find out the rare and frequent patterns. This algorithm takes certain value of the items appear in the transaction set. Each item is associated with the individual support value. It only works with the certain value of the itemsets. But, it doesn't deal with the probabilistic value of the itemset or uncertain dataset. In 2007, Chun et.al. [5], proposed to manipulate the probabilistic value of items to find out the frequent pattern using the expected support concept. UApriori algorithm works with probabilistic value of the items from the uncertain dataset. It can not use to find out the rare items in the transaction set.

4. WORKFLOW OF THE PROPOSED WORK

The workflow followed in the proposed work for the generation of frequent patterns is shown in figure 1.

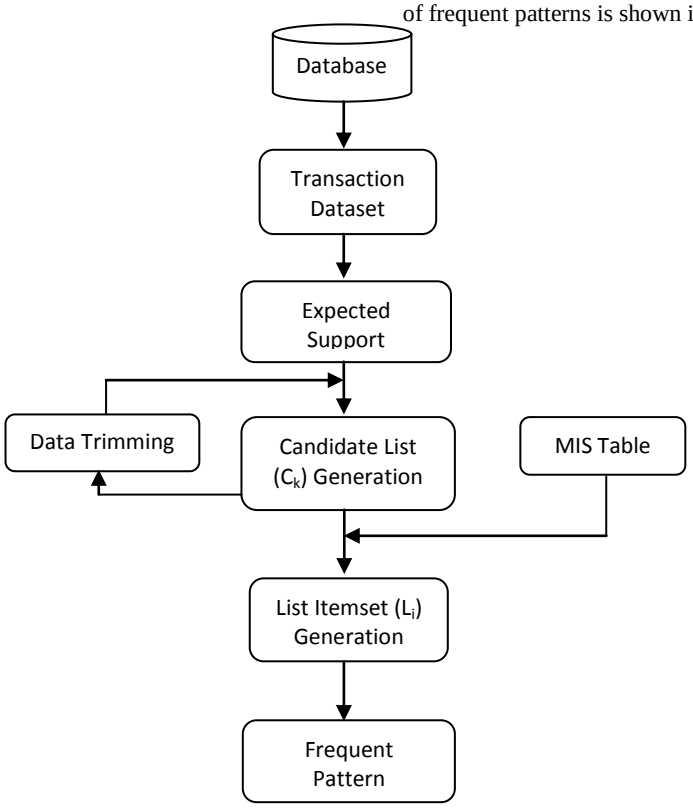

Figure 1. Workflow of process for generating frequent patterns

- Initially transaction sets are fetched from the database.
- These transaction sets are scan to trim the itemsets that have very low existential probabibility.
- Then Expected support value for the particular itemssets in the transaction is calculated.
- Candidate list is generated on the basis of itemsets present in the transaction set.
- Data trimming is used to trim the items or itemsets, which have very low existential probabilistic value.
- MIS table, which contains multiple support value for each item in the transaction set is used to generate the list itemset.
- Finally frequent patterns are generated.

5. DATA TRIMMING

To improve the efficiency of the U-Apriori algorithm, we use a data trimming technique to avoid insignificant candidate support increments performed in the Subset-Function. The basic idea is to trim away items with low existential probabilities from the original dataset and to mine the trimmed dataset instead.

Hence the computational cost of those insignificant candidate increments can be reduced. In addition, the I/O cost can be greatly reduced since the size of the trimmed dataset is much smaller than the original one. More specifically, the data trimming technique works under a framework that consists three modules: the trimming module (Local Trimming Strategy), pruning module (Global Pruning Strategy) and patch up module (Single-pass Patch Up Strategy) [4].

6. PROPOSED ALGORITHM

1. input, transaction sets from the uncertain dataset and MIS table
2. **for** each x ⊆T in transaction set **do**
   /scan D(database)
3. **if** existential probability < λ (user specified threshold) */if very low existential probability*
4. remove that item from the list */trimming of items having very low probability*
5. **endif**
6. now the trimmed dataset is D^T
7. **end for**
8. **for** each x ⊆ T in transaction set **do**
   /scan D^T
9. calculate expected support(expsup) for x
10. $Se(x) = \sum_{j=1}^{|D|} \prod_{x \in X} Ptj(x\ Ptj(x))$ */Se(X)=expected support for x, Ptj=probability*
11. **end for**
12. **for** each item in candidate list(C_k) **do**
    /k-iteration
13. **if** x ⊆ C_k **x.expsup <= MIS(x_i)** /comparing expsup(x) with MIS value of item
    then add x in list item(L_i)
14. **else** remove from the list
15. **end if**
16. **end for**

7. RESULTS ANALYSIS AND PERFORMANCE

Using the above algorithm, the following results have been achieved. Table 2, shown below contains all the transaction sets that are fetched from the uncertain dataset. Items contains the probabilistic value associated with them. This table initially scans to trim the items that may have very low existential probability and may cause to generate meaningless frequent patterns. E_1, E_2, E_3 and E_4 are corresponding transaction ids of the transaction set.

Table 2. Input transaction dataset

Tid	Items
E_1	laptop(0.9), cleaning kit(0.8), anti-virus(1.0)
E_2	laptop(0.9), external hard disk(0.4)
E_3	laptop(0.9), anti-virus(0.8)
E_4	pen drive(0.8), antivirus(0.6)

Table 3 shows the multiple minimum support value for each item. This table is used to generate the list item(L_i) to generate the frequent patterns. This MIS table is directly fetched from the database. It also contains the probabilistic MIS value for the items.

Here L: laptop, CK: cleaning kit, AV: anti-virus, E-HD: external hard disk, PD: pen drive.

Table 3. MIS value for items

Item	L	CK	AV	E-HD	PD
MIS value	0.8	0.6	0.6	0.3	0.7

Table 4, contains the frequent pattern generated from the proposed algorithm. Frequent pattern column contains all the combinations of the itemsets that followed the minimum support constraint to generate the frequent patterns. Corresponding support column shows the associated support value for the items and itemsets in the frequent pattern list.

Table 4. Output frequent pattern generated

Frequent Pattern	Support
L(0.9)	2.7
AV(1.0)	1.8
CK(0.8)	0.8
E-HD(0.4)	0.4
PD(0.8)	0.8
CK(0.8) AV(1.0)	0.8
L(0.9) AV(1.0)	1.62
L(0.9) CK(0.8)	0.7200000000000001
L(0.9) E-HD(0.4)	0.36000000000000004
L(0.9) CK(0.8) AV(1.0)	0.7200000000000001

8. CONCLUSION

A single minsup is insufficient for association rule mining since it cannot reflect the nature and frequency differences of the items in the database. It is neither satisfactory to set the minsup too high, nor is it satisfactory to set it too low. For some real life applications customized value of items may be considered. This paper proposes a more flexible and powerful algorithm. It allows the user to specify multiple minimum item support corresponding to their existential probability. This algorithm enables us to find rare item rules without producing a large number of meaningless rules with frequent items from the uncertain datasets.

9. REFERENCES

[1] Laszlo Szathmary, Amedeo Napoli, Petko Valtchev: Towards Rare Itemset Mining. ICTAI (1) 2007: 305-312

[2] Rakesh Agrawal, Tomasz lmielinski, and Arun Swami. Mining association rules between sets of items in large databases. In Proc. Of the ACM SIGMOD Conference on Management of Data, pages 207-216, Washington, D.C., May 1993.

[3] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In Advances in knowledge discovery and data mining, pages 307–328. AAAI, 1996.

[4] R. U. Kiran and P. K. Reddy. An improved frequent pattern-growth approach to discover rare association rules. In Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, pages 43–52, 2009.

[5] Chun-Kit Chui, Ben Kao, and Edward Hung(2007), Mining Frequent Itemsets from Uncertain Data, PAKDD 2007, LNAI 4426, pp. 47–58, 2007. © Springer-Verlag Berlin Heidelberg 2007, pp 47-58

[6] G. M. Weiss. Mining with rarity: a unifying framework. SIGKDD Explor. Newsl., 6(1):7–19, 2004.

[7] B. Liu, W. Hsu, and Y. Ma. Mining association rules with multiple minimum supports. In KDD ’99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 337–341. ACM, 1999.

[8] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Min. Knowl. Discov., 8(1):53–87, 2004.

[9] Y.-H. Hu and Y.-L. Chen. Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. Decis. Support Syst., 42(1):1–24, 2006.