An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules

by

R. Uday kiran, P Krishna Reddy

in

2009 IEEE Symposium on Computational Intelligence and Data Mining (IEEE CIDM 2009)

Report No: IIIT/TR/2009/24

Centre for Data Engineering
International Institute of Information Technology
Hyderabad - 500 032, INDIA
January 2009

An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules

R. Uday Kiran and P. Krishna Reddy

***Abstract*—In this paper we have proposed an improved approach to extract rare association rules. Rare association rules are the association rules containing rare items. Rare items are less frequent items. For extracting rare itemsets, the single minimum support (minsup) based approaches like Apriori approach suffer from "rare item problem" dilemma. At high minsup value, rare itemsets are missed, and at low minsup value, the number of frequent itemsets explodes. To extract rare itemsets, an effort has been made in the literature in which minsup of each item is fixed equal to the percentage of its support. Even though this approach improves the performance over single minsup based approaches, it still suffers from "rare item problem" dilemma. If minsup for the item is fixed by setting the percentage value high, the rare itemsets are missed as the minsup for the rare items becomes close to their support, and if minsup for the item is fixed by setting the percentage value low, the number of frequent itemsets explodes. In this paper, we propose an improved approach in which minsup is fixed for each item based on the notion of "support difference". The proposed approach assigns appropriate minsup values for frequent as well as rare items based on their item supports and reduces both "rule missing" and "rule explosion" problems. Experimental results on both synthetic and real world datasets show that the proposed approach improves performance over existing approaches by minimizing the explosion of number of frequent itemsets involving frequent items and without missing the frequent itemsets involving rare items.**

I. Introduction

Data mining represents techniques for discovering knowledge patterns hidden in large databases. Several data mining approaches are being used to extract interesting knowledge [1]. Like, association rule mining techniques [2] [3] discover association between the entities, clustering techniques [4] group the unlabeled data into clusters such that there exists high inter similarity and low intra similarity between the clusters, classification techniques [5] to identify the different classes existing in categorical labeled data.

It can be observed that most of the data mining approaches discover the knowledge pertaining to frequently occurring entities. However, real-world datasets are mostly non-uniform in nature containing both frequently and relatively rarely occurring entities. In literature, it has been reported that there exists useful knowledge pertaining to rare entities [6] [10].

The rare cases are more difficult to detect and generalize from because they contain fewer data. Realizing the importance of rare knowledge patterns pertaining to rare events, research efforts are going on to investigate improved approaches to extract rare knowledge patterns like rare association rules and rare class identification [6].

In this paper, we are investigating an improved approach to extract rare association rules. Association rule mining [2] is a popular knowledge pattern and has been extensively studied [7] [8]. The basic terminology about association rules is as follows. Let I = $\{i_1, i_2, ..., i_n\}$ be a set of items and T be a set of transactions. Each transaction T_i (i = 0, 1, ..., m) is a set of items such that $T_i \subseteq I$. An itemset X is a set of items $\{i_1, i_2, ..., i_k\}$ ($1 \leq k \leq n$) such that $X \subseteq I$. An itemset containing k number of items is called k-itemset. An association rule is an implication of the form, $A \Rightarrow B$, where $A \subset I$, $B \subset I$ and $A \cap B = \emptyset$. The rule $A \Rightarrow B$ holds in T with *support s* if s% of the transactions in T contain $A \cup B$. Similarly rule $A \Rightarrow B$ holds in T with *confidence c* if c% of transactions in T that support A also support B. Given T, the objective of association rule mining is to discover all association rules that have support and confidence greater than the user-specified minimum support $minsup$ and minimum confidence $minconf$.

A rare association rule refers to an association rule forming between either frequent and rare items or among rare items. There exist a useful knowledge in rare associations.

> **Example 1:** Consider the set of items {bread, jam, bed, pillow} selling in a super market. It can be observed that the items in the set {bread, jam} are frequently purchased items while the items in the set {bed, pillow} are infrequently or rarely purchased items. Even though, {bed, pillow} contains rare items, it is interesting as it may generate more revenue in this case.

The popular approaches like Apriori discover association rules based on frequent itemsets which are extracted by fixing the same (or single) minimum support for all items. For extracting rare itemsets, the single minimum support (minsup) based approaches suffer from "rare item problem" dilemma [9]. At high *minsup* value, rare itemsets could not be extracted as rare items fail to satisfy the minsup criteria. Also, if low minsup is set to extract rare itemsets, the number of itemsets explode as huge number of frequent itemsets satisfy the minsup criteria.

In the literature, efforts are being made to propose improved approaches to mine rare associations [10] [11] [12] [13]. In [10], instead of fixing single minsup value for all items, the minsup value is calculated for each item based on the percentage of its support and frequent itemsets are extracted if an itemset satisfied the lowest minsup value of the items in it. (The support of an item is the ratio of frequency of an item by transaction dataset size.) In [11], items are categorized into frequent and rare items. Frequent itemsets involving only frequent items are extracted using single minsup approaches and frequent itemsets involving rare items are extracted using

a concept of "relative support". In [12], a stochastic mixture model know as Negative-Binomial distribution is utilized to known the process of generating transaction data and to find all Negative-Binomial frequent itemsets. The approach proposed in [13] extracts the association rules by considering only infrequent items (i.e., items having support less than minsup).

Among the existing approaches, the percentage-based approach improves the performance over single minsup based approaches. However, it has been observed that it still suffers from both "rule missing" and "rule explosion" problems. If minsup for the item is fixed by setting the percentage value high, the rare itemsets are missed as the minsup for the rare items becomes close to their support, and if minsup for the item is fixed by setting the percentage value low, the number of frequent itemsets explodes. The drawbacks of the other approaches are mentioned in the related work section.

In this paper, we propose an improved approach by using the notion of "support difference" (SD) in calculating minsup value for each item. Through the notion of SD, a constant difference is ensured between the item support and the corresponding minsup value for each item. So, by using the notion of SD, the proposed approach successfully extracts the frequent itemsets involving rare items and limits the explosion of frequent itemsets involving frequent items. The proposed approach also extracts frequent itemsets involving both frequent and rare items. Experimental results on synthetic and real-world dataset show that the proposed approach discovers frequent itemsets involving rare items in an efficient manner as compared to the existing approaches.

The paper is organized as follows. In Section 2, we discuss the related work. In Section 3, we explain the proposed approach and algorithm. In Section 4, experiment results are presented. The last section contains conclusions and future work.

II. Related Work

In this section we briefly discuss the related approaches for the extraction of rare associations.

A. Apriori algorithm

In the literature, Apriori [3], Frequent-Pattern Tree approach [7] and related approaches have been proposed for finding frequent itemsets. These approaches use single minsup value for all data items to extract frequent itemsets. Here we explain the "rare item problem" dilemma faced by single minsup based approaches by considering Apriori approach.

The Apriori algorithm employs an iterative level-wise search for generating frequent itemsets. An itemset is a set of items. A candidate k-itemset refers to an itemset having 'k' number of items and frequent k-itemset refers to subset of candidate k-itemsets whose support is greater than or equal to user-specified *minsup*. The Apriori algorithm repeats the steps from (i) to (iii) starting with k=1 till no more frequent itemsets are found [3]: (i) C_k is generated. (ii) L_k is generated from C_k by pruning the itemsets whose support is greater than minsup value. (iii) C_{k+1} is generated by joining L_k with itself.

TABLE I
Transaction dataset.

TID	Items
1	bread, jam
2	bread, jam, pencils
3	bread, jam, pen
4	bread, jam, ball
5	bread, ball
6	bed, pillow
7	bed, pillow
8	ball, bat
9	ball, bat
10	ball, bat

The extraction of frequent itemsets using Apriori is illustrated in Example 2.

Example 2: Table I shows the dataset of 10 transactions with the itemset I = {bread, jam, pillow, bed, ball, bat}. The working of Apriori with minsup=40% is depicted in Figure 1. C_1 is generated after the first scan. From C_1, frequent 1-itemset L_1 (itemsets represented in bold) is generated with the items whose support is greater than or equal to *minsup*. C_2 is generated by joining L_1 with itself. From C_2, L_2 (itemsets represented in bold) is generated if an itemset support is greater than or equal to *minsup*. The generated L_2 is {{bread, jam}, {bat, ball}}.

Mining of rare associations (association rules involving rare items) with single minsup approaches may cause "rare item problem" dilemma. The "rare item problem" is as follows: if *minsup* is high, frequent itemsets involving rare items are missed as the support of the rare items is less than the given minsup. In order to find frequent itemsets involving rare items, the minsup value should be fixed at low value. As a result, the number of frequent itemsets explode.

Example 3: In Example 2, frequent itemsets {bread, jam} and {bat, ball} are extracted at *minsup* = 40%. However, it can be observed that frequent itemset {pillow, bed} involving rare items is missed. In order to extract frequent itemset {pillow, bed}, the *minsup* has to be set to low value, equal to 20%. With minsup=20%, the process of finding frequent itemsets is depicted in Figure 2. It can be observed that out of four frequent 2-itemsets only three frequent 2-itemsets {{bread, jam}, {bat, ball}, {bed, pillow}} are interesting. The other frequent 2-itemset {bread, ball} is uninteresting because items {bread} and {ball} are frequently purchased in a supermarket. For {bread, ball} to be useful, it should have satisfied much higher minsup (say 40%).

B. Multiple minimum support approaches

To improve the performance of extracting frequent itemsets involving rare items, an approach known as Multiple Support Apriori (MSApriori) has been proposed in [10]. In this approach, each item is assigned with a minsup value known as "Minimum Item Support" (MIS) and frequent itemsets are generated if an itemset satisfies the lowest MIS value among the respective items. The MIS value is assigned to each item

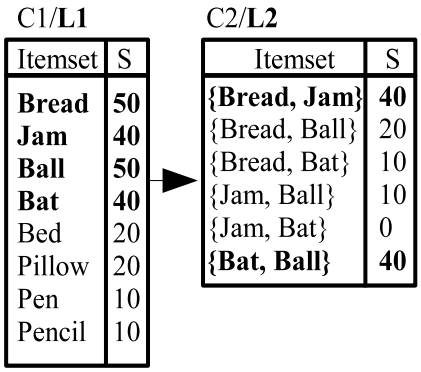


Fig. 1. Working of Apriori Algorithm at minsup=40%. Here, the notation 'S' indicates the support percentage.

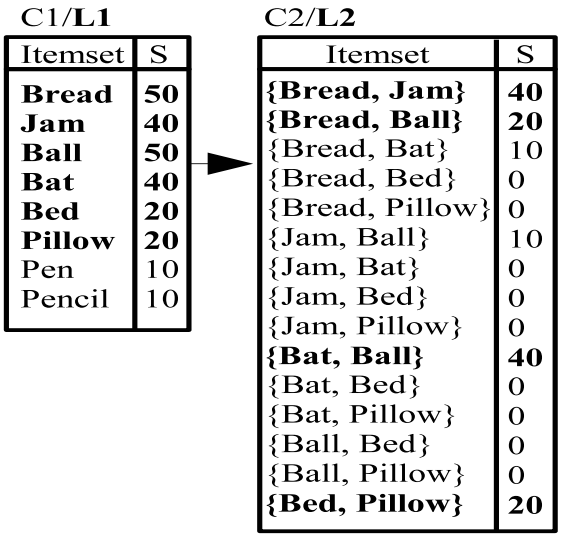


Fig. 2. Working of Apriori Algorithm at minsup=20%. Here, the notation 'S' indicates the support percentage.

equal to a percentage of its support. For every item $i_j \in$ I, the MIS(i_j) is calculated as per Equation 1.

$$MIS(i_j) = \beta S(i_j)),\ if\ \beta S(i_j) > LS \\ = LS \quad else \tag{1}$$

where, β is a user-specified proportional value which can be varied between 0 to 1, $S(i_j)$ refers to support of an item equal to f(i_j)/N, (f(i_j) represents frequency of i_j and N is the number of transactions in a transaction dataset) and LS corresponds to user-specified least support value.

In this method, as the MIS values for the items are derived based on the percentage of their supports frequent items are assigned with higher MIS values and rare items are assigned with relatively lower MIS values. For an itemset containing only frequent items to be a frequent itemset, it has to satisfy relatively higher minsup than the itemset containing frequent and rare items or only rare items. So the MSApriori approach improves the performance over single minsup based approaches by addressing the "rare item problem". The working of MSApriori is illustrated in Example 4.

Example 4: The working of MSApriori for the transaction dataset shown in Table I is depicted in Figure 3, by considering β=0.75 and LS=20%. After calculating C_1, MIS is calculated by using the Equation 1. From C_1, the L_1 is generated with the items whose support is greater than or equal to their respective MIS. From L_1, C_2 is generated by joining L_1 with itself. From C_2, L_2 is generated if an itemset satisfies the lowest MIS value of the items in it. The generated L_2 is {{bread, jam}, {bat, ball}, {pillow, bed}}. Note that, the uninteresting frequent 2-itemset {bread, ball} which had been generated using Apriori algorithm (when *minsup* is set very low) fails to get generated in MSApriori, since candidate itemset {bread, ball} do not satisfy the low MIS value (40%) among the items in the itemset.

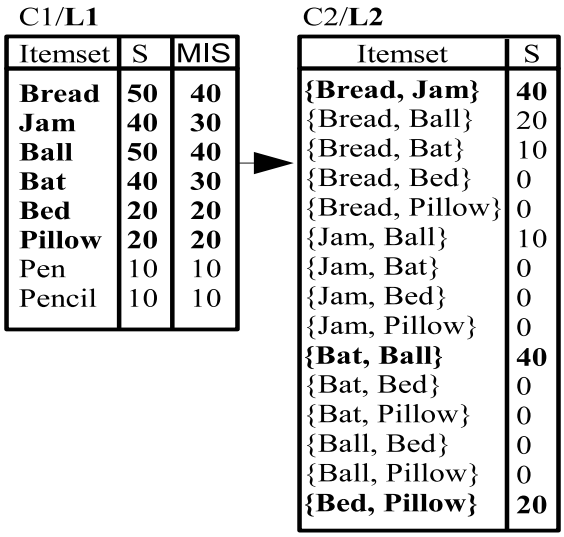


Fig. 3. Working of MSApriori algorithm with β=0.75 and LS=20%. Here, the notations 'S' and 'mis' indicates the support percentage and minimum item support percentage values respectively.

It was observed that MSApriori still suffers from the "rare item problem" dilemma, if items support vary widely. The reason is as follows: for a user-specified proportion β value, as we move from frequent items to relatively rare items the difference between the support of an item and its MIS is not constant, rather decreases. As a result, MSApriori suffers from the following problems.

1) If β is set high, it can be observed that MIS for rare items will be relatively more close (almost equivalent) to their supports as compared with frequent items. As a result, itemsets containing rare items fail to satisfy the support of (lowest MIS) rare item in that itemset. So, frequent itemsets involving rare items are missed.

   Example 5: Consider four items Bread, Beer, Pillow, and Bed having supports 44%, 36%, 10% and 5% re-

spectively. With β=0.9 and LS=1, the MIS values for Bread, Beer, Pillow and Bed comes to approximately 40%, 32%, 9% and 4.5% respectively. Note that our interest is to extract rare itemsets which contain the item "Bed". It can be observed that, for the rare item like "Bed", the difference between its support and MIS is relatively less (5% - 4.5% = 0.5%) as compared with relatively frequent item like Bread (44% - 40% = 4%). Hence as the itemset length increases (more items are combined), the itemsets containing item "Bed" fail to satisfy the support equivalent to support of the item "Bed". So, frequent itemsets involving "Bed" are missed.

2) To facilitate participation of rare items, MIS for the rare items have to be less than their support values. It can be achieved by setting low β value. However, this may cause frequent items to set very low MIS values. With low MIS values for frequent items, the items will be associating with one another in all possible ways, thereby generating a huge number of frequent itemsets.

   Example 6: Continuing from Example 5, to facilitate participation of rare items like 'Bed', β value is set to low which is equal to 0.3. When β=0.3 and LS=1%, the MIS(Bread) $\approx$ 13%, MIS(Beer) $\approx$12%, MIS(pillow)=3% and MIS(Bed)=1.5%. For the rare item 'Bed', the difference between its support and MIS is (5% - 1.5% = 3.5%) is relatively higher than the previous value. As a result, frequent itemset like {Pillow, Bed} which contains the rare item "bed" is generated. However, it can be observed that, due to low β value, the uninteresting 2-itemset {bread, ball} also becomes frequent. The itemset {bread, ball} is uninteresting because knowing the fact that 20% of the customers bought the two items together is useless. The reason is, the items 'bread' and 'ball' are frequently purchased items in a supermarket. For this itemset to be useful, it should have satisfied higher minimum support i.e., 32%.

In the literature, another approach known as Relative Support Apriori Algorithm (RSAA) has been proposed [11] for discovering frequent itemsets involving both frequent and rare items. In discovering frequent itemsets, RSAA uses three user-specified measures: first support s_1, second support s_2 ($s_1 > s_2$) and relative support $Rsup$. Items having support greater than or equal to s_1 are considered frequent items and the items having support less than s_1 and greater than or equal to s_2 are considered as rare items. If an itemset contains only frequent items, it has to satisfy s_1 to be a frequent itemset. If an itemset contains rare items, its support has to satisfy s_2 and its $Rsup$ should satisfy user specified minimum relative support $mRsup$ to be a frequent itemset. The main issue in this algorithm is providing values to these parameters for the given dataset.

C. Other approaches

An approach is proposed in [12] by utilizing the knowledge of the process for generating transaction data by applying a stochastic mixture model known as negative binomial (NB) distribution. This model along with a user-specified precision threshold, finds local frequency thresholds for groups of itemsets based on which algorithm finds all NB-frequent itemsets in a dataset. This model considers highly skewed data (skewed towards right) with the underlying assumption of Poisson processes and Gamma mixing distribution. This model can effectively be implemented in the datasets like general web logs etc, which are exponentially distributed. However, this approach is not effective on other datasets like general super market datasets which are generally not exponentially distributed. The reason is frequent items will distort the mean and the variance and thus will lead to a model which grossly overestimates the probability of seeing items with high frequencies. If we remove items of high frequencies as suggested in the approach, we may miss some interesting frequent itemsets pertaining to frequent items.

An approach has been suggested to mine the association rules by considering only infrequent items i.e., items having support less than *minsup* [13]. However, this approach fails to discover associations between frequent and rare items.

In this paper we have proposed an improved approach by using the notion of *support difference*. The proposed approach is different from MSApriori as it follows a different method to assign minsup for each items. It also differs from RSAA approach as it requires only two measures instead of three measures used in RSAA. The approaches proposed in [12] and [13] deal with frequent itemsets involving only rare items whereas the proposed approach extracts frequent itemsets involving both frequent and rare items.

III. Proposed approach

A. Basic Idea

In this approach, we use the notion of **support difference** *SD* to specify the minimum supports to items. Support difference (SD) refers to the acceptable deviation of an item from its frequency (or support) so that an itemset involving that item can be considered as a frequent itemset. For each item 'i_j', calculation of minsup known as minimum item support (MIS(i_j)) is as follows:

$$MIS(i_j) = S(i_j) - SD \ when \ (S(i_j) - SD) > LS \quad (2)$$
$$= LS \quad otherwise$$

where, S(i_j) refers to support of item 'i_j' and LS refer to the user-specified least support. Using SD, MIS for the items range from (-∞, +∞). To prevent MIS values of the items in reaching 0 or lower, we use concept of least support (LS). Least support refers the lowest minimum support an item or itemset should satisfy to become a frequent itemset. The LS takes a value in the range [0%, 100%].

For a given dataset, the value of SD can be calculated as per Equation 3.

$$SD = \lambda(1 - \alpha) \quad (3)$$

where λ represents the parameter like mean, median, mode, maximum support of the item supports and α is the parameter ranging between 0 to 1. SD takes values from (0, λ).

1) If α = 0 then SD = λ. Higher the SD value, minimum supports for the items will be relatively less than their corresponding supports.
2) If α = 1 then SD = 0. When SD = 0, minimum support for an item is equivalent to their corresponding support values.

Both λ and α play a major role in determining SD. The value of λ can be calculated by determining appropriate statistical parameter suitable to data set. If data set size is huge, sampling methods can be employed. The value of α is set by the user.

After specifying MIS values for each item as per Equation 3, the frequent itemsets are generated using Equation 4.

$$S(i_1, i_2, ..., i_k) \geq min \begin{pmatrix} MIS(i_1), MIS(i_2) \\ ..., MIS(i_k) \end{pmatrix} \quad (4)$$

where S(i_1, i_2, $\cdots$, i_k) represents the support for an itemset $\{i_1, i_2, \cdots, i_k\}$.

The Equation 4 ensures the extraction of frequent itemsets involving frequent items, rare items and both frequent and rare items efficiently. Note that the MIS value for each item depends upon its support. If an itemset contains all frequent items, it has to satisfy the lowest MIS of frequent items. Similarly, if an itemset contain frequent and rare items, it has to satisfy the lowest MIS of rare items in it.

In this approach, a constant difference between support of an item and the MIS can be ensured irrespective of item support values. The proposed approach is not prone to spread or gap among the item supports so it can be used in all kinds of datasets including the datasets where item frequencies (or supports) vary widely. The advantage of proposed approach over MSApriori is illustrated using Example 7.

Example 7: Consider four items Bread, Beer, Pillow, and Bed having supports 44%, 36%, 10% and 5% respectively. We explain the processing under proposed approach by fixing SD = 4% and LS = 1%. Let us consider two frequent itemsets: {Bread, Beer} containing frequent items and {Pillow, Bed} containing rare items. Under MSApriori, two different values have to be fixed to extract both itemsets. To extract {Bread, Beer}, minimum support should be set to 32% (β=0.9). Similarly to extract {Pillow, Bed}, the minimum support should be set to 1% (β=0.3). Whereas in the proposed approach, using single SD value (SD=4%), the MIS values are as follows: MIS(Bread) = 40%, MIS(Beer) = 32%, MIS(Pillow) = 6% and MIS(Bed) = 1%.

B. Algorithm

The proposed algorithm generalizes the Apriori algorithm for finding frequent itemsets given in [2]. We call the new algorithm as Improved Multiple Support Apriori Algorithm (IMSApriori).

The proposed approach generates frequent itemsets by making multiple passes over the data. However, in contrary to single minsup approaches which follow "downward closure property" (all the subsets of a frequent itemset are frequent), multiple minimum support approaches follow "sorted closure property". That is, if an itemset is not frequent at (k-1) itemset, it can not be discarded as any addition of items to it can be frequent.

Let C_k denote the set of candidate k-itemset and L_k denote the set of frequent k-itemset. Note that the items in an itemset are arranged in increasing order of their MIS (frequency) values from left to right. The algorithm for finding frequent itemsets is given in Algorithm 1.

Algorithm 1 Generate-Large-itemsets (T:transaction dataset, SD: support difference, LS:least support)

```
1: Generate candidate 1-itemset C1;
2: Calculate support S, for itemsets in C1.
3: MIS=calculate-MIS(S, SD, LS);
4: L1={<i> |i∈C1, S(i)≥MIS(i)};
5: L1 = sort(L1,MIS);
6: for k=2;Lk−1≠∅;++k do
7:    Ck=candidate-gen(Lk−1);
8:    for transaction t∈D do
9:       Ct=subset(Ck,t);
10:      for each candidate c∈Ct do
11:         c.count++;
12:      end for
13:   end for
14:   Lk=c∈C| c.count∗100/|T| ≥min(MIS(i)|∀i∈c)
15: end for
16: Answer=∪kLk;
```

In IMSApriori approach, the generation of frequent itemsets is shown in Algoirthm 1. The procedure for specifying MIS values for each item which is shown in Procedure 2. After generating frequent itemsets, the approach given in [3] should be employed to generate association rules.

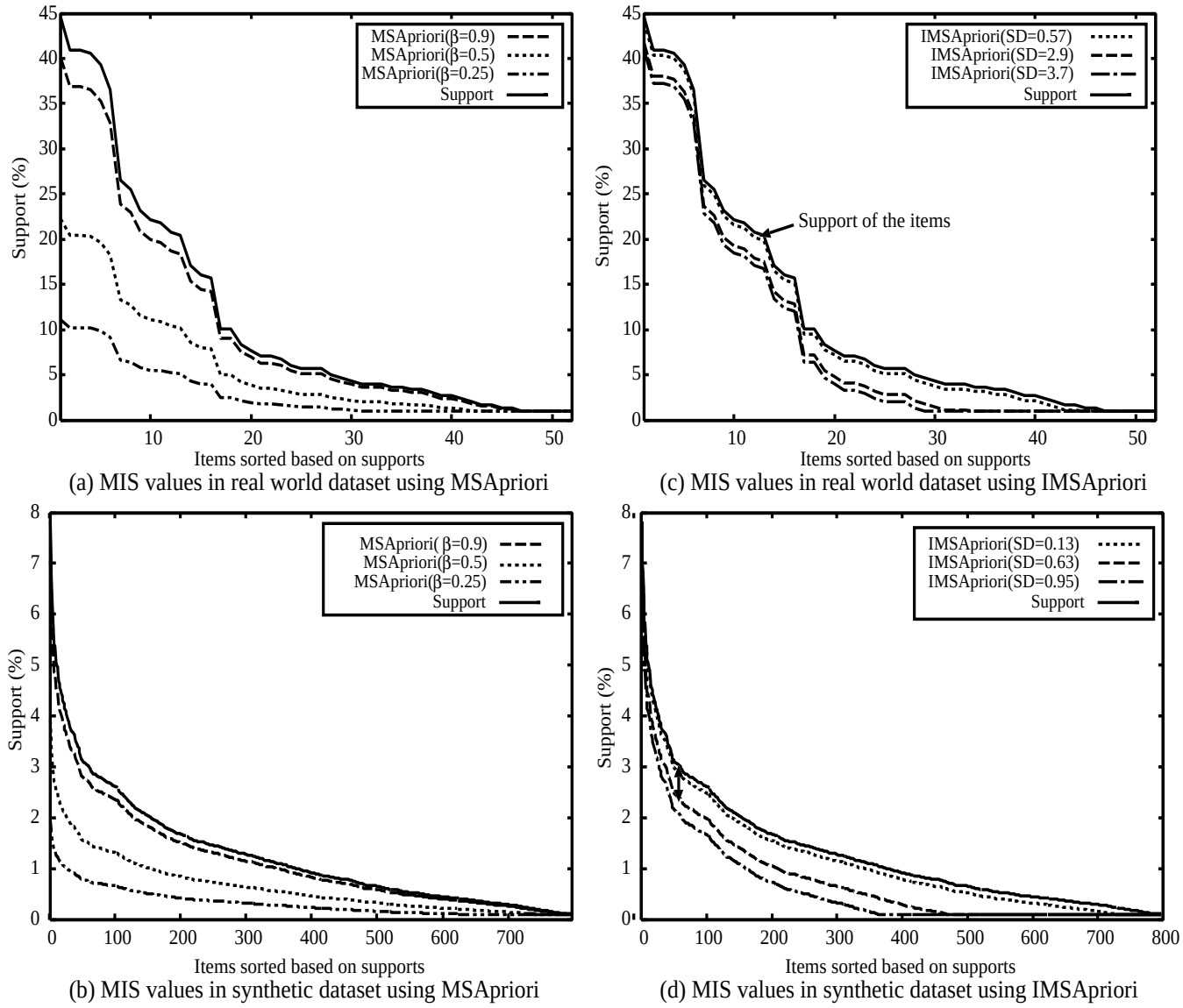
Procedure 2 calculate-MIS (S: supports for the candidate 1-itemset, SD:support difference, LS: least support)

```
1: for i = 1; i ≤ | C1 |; ++i do
2:    M(i) = S(i) − SD(n);
3:    if M(i) < LS then
4:       MIS(i) = LS;
5:    else
6:       MIS(i) = M(i);
7:    end if
8: end for
9: return MIS;
```

IV. Evaluation

A. Experimental details

The performance of the proposed approach is evaluated by considering two kinds of datasets: synthetic and real world



(a) MIS values in real world dataset using MSApriori

(c) MIS values in real world dataset using IMSApriori

(b) MIS values in synthetic dataset using MSApriori

(d) MIS values in synthetic dataset using IMSApriori

Fig. 4. Minimum item supports.

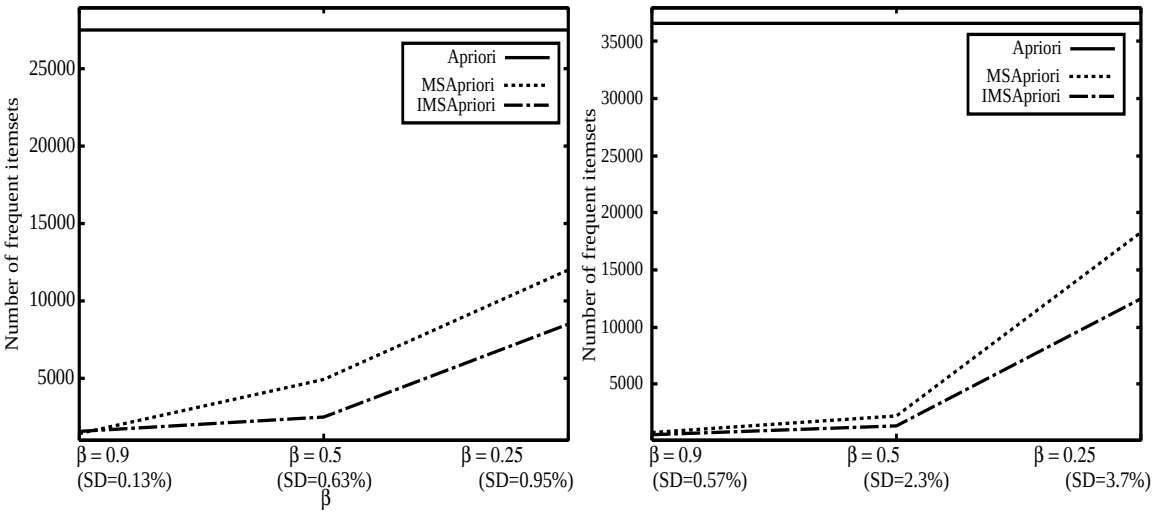
datasets. The synthetic dataset is generated with the data generator [3] which is widely used for evaluating association rule mining algorithms. The synthetic dataset is T10I4D100K containing 870 items with 100K transactions and is available at [18]. The real world dataset contain 88 items with 298 transactions having an average transaction of 10 items. The details of the datasets are given in Table II.

In the experimental results, we have compared the proposed approach with Apriori and MSApriori approaches. We have not considered [11] [12] [13] approaches for comparison due to the following reasons.

- In generating rare association rules, the approach proposed in [11] requires three parameters whereas the proposed approach and MSApriori requires only two parameters.
- The approach proposed in [12] is suitable for skewed distributions and requires the removal of frequent items for extracting rare association rules. As a result, the rare association rules involving frequent and rare items cannot be extracted.
- The approach proposed in [13] extracts only association rules involving only rare items and do not extract the

TABLE II
DETAILS OF THE DATASETS.

	Synthetic dataset	Real world dataset
Total items	870	83
minsup (%)	0.1	1
Items participated	798	52
Mean	1.26	11.49
Median	0.83	5.7
Standard deviation	1.22	12.78



(a) Frequent itemsets generated in synthetic dataset using MSApriori

(b) Frequent itemsets generated in real world dataset using IMSApriori

Fig. 5. Frequent itemsets generated.

TABLE III
SUPPORT DIFFERENCE VALUES USED IN SYNTHETIC AND REAL WORLD DATASETS.

	Synthetic dataset SD=λ(1-β)	Real world dataset SD=λ(1-β)
β = 0.9	0.13%	0.57%
β = 0.5	0.63%	2.9%
β = 0.25	0.95%	3.7%

association rules involving frequent and rare items.

The proposed IMSApriori approach requires two parameters, LS and SD. Similarly, MSApriori approach requires two parameters, LS and β. The parameter LS is common in both the approaches. To compare the performance of IMSApriori and MSApriori approaches, SD in IMSApriori is calculated by using Equation 3. (Note that SD can be direct value specified by the user.)

In the experimental results, the performance of IMSApriori and MSApriori is evaluated by varying the β value and by fixing the LS value at 1% and 0.1% in real world and synthetic datasets respectively. The value of λ is fixed as the mean of the datasets (Table II).

B. Performance results

In the performance graphs, the items frequencies (high to low) are represented on X-axis and the corresponding support values are represented on the Y-axis. On the same graph, by varying the β (SD) values, we have plotted minimum item supports for the items using MSApriori (IMSApriori) approach.

Figure 4(a) shows the performance of MSApriori at various β values for synthetic dataset. The thick line shows the support of the items. It can be observed that at low β values the difference between the support of an item and its minimum item support is very high for the frequent items which leads to combinatorial explosion of association rules. When β is high, the difference between the support of an item and minimum item support is very less for the low frequent (rare) items which leads to missing of rare association rules. Figure 4(b) shows the performance of MSApriori at various β values for real world dataset.

Figure 4(c) shows the performance of proposed IMSApriori approach at various SD values for synthetic dataset. It can be observed that for a given SD value, the difference between the support of an item and its minimum item support remains constant for both high frequency and low frequency (rare) items. As a result, the proposed IMSApriori approach can extract associations including rare associations in a more efficient manner as compared to MSApriori. Figure 4(d) shows the performance of proposed IMSApriori approach at various SD values for real world dataset.

Figure 5(a) shows how the number of frequent itemsets varied with different β (corresponding SD) values for Apriori, MSApriori and IMSApriori approaches in synthetic dataset. (For Apriori, $minsup$ is fixed as LS.) It can be observed that the number of frequent itemsets generated using Apriori algorithm is very high. Regarding MSApriori at high β values the number of frequent itemsets generated are very less. As β value decreases, the number of frequent itemsets extracted also increase. Regarding the proposed IMSApriori approach, it can be observed that it extracts less number of frequent itemsets as compared to MSApriori. At high β value (corresponding SD value), the proposed approach extracts more frequent itemsets pertaining to rare items. As we decrease β, the proposed approach extracts less number of frequent itemsets through efficient pruning. Figure 5(b) shows how the number of frequent itemsets varied with different β values for Apriori, MSApriori and proposed IMSApriori approaches in real world dataset.

V. Conclusions and future work

In this paper, we have proposed an improved approach to extract frequent itemsets involving rare items to discover rare association rules. In the proposed approach, the minimum support for each item is calculated using the notion of "support difference". The proposed approach dynamically assigns appropriate minimum support to each item so that frequent itemsets involving rare items can be extracted in a more efficient manner as compared to the existing approaches. Most important, the proposed approach ensures that the difference between the support of an item and the corresponding minimum support remains constant for all items including rare items. As a result, it efficiently reduces the explosion of frequent itemsets involving frequent items without affecting the extraction of frequent itemsets involving rare items. We have evaluated the performance of the proposed approach by conducting experimental results on both synthetic and real world datasets. The results show that, as compared to existing approaches, the proposed approach prunes frequent itemsets involving frequent items in a more efficient manner and without missing the frequent itemsets involving rare items.

As a part of future work, we are planning to investigate the following.

- In this paper we have investigated how to improve the performance of extracting frequent itemsets involving rare items. After extracting frequent itemsets, association rules are discovered by fixing the same minimum confidence value for all the rules. As a part of future work, we are going to investigate appropriate methodology for assigning confidence values in a dynamic manner to generate rare association rules in an efficient manner.
- The existing approaches extract frequent itemsets involving rare items by assigning minimum support to each item and employing an iterative process to discover frequent itemsets. We are going to investigate how popular non-iterative approaches like frequent-pattern growth (FP-growth) approaches can be extended to assign minimum support to each item to extract frequent itemsets involving rare items.

Acknowledgement

This work is supported by "Nokia", Findland and "Media Lab Asia", India. The authors are thankful to the anonymous referees for their useful comments. This work was done while R. Uday Kiran was at International Institute of Information Technology-Hyderabad, India.

References

[1] Melli, G., Osmar, R. Z., and Kitts, B. "Introduction to the Special Issue on Successful Real-World Data Mining Applications.", SIGKDD Explorations, Volume 8, Issue 1, 2006.

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

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

[4] Xu, R. "Survey of Clustering Algorithms.", IEEE Transactions on Neural Networks, Volume. 16, Number 3, May 2005.

[5] Weiss, S., and Kulikowski, C. A. "Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.", Morgan Kaufmann, 1991.

[6] Weiss, G. M. "Mining With Rarity: A Unifying Framework.", SIGKDD Explorations, 2004, Vol. 6, Issue 1, pp. 7 - 19.

[7] Jiawei, H., Jian, P., Yiwen, Y., and Runying, M. "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach*.", DMKD, 2004, pp. 53-87.

[8] Rajanish, D., and Ambuj, M. "Fast Frequent Pattern Mining in Real-Time.", CSI, 2005, pp. 156 - 167.

[9] Mannila, H. "Methods and Problems in Data Mining.", ICDT, 1997.

[10] Liu, B., Hsu, W., and Ma, Y. "Mining Association Rules with Multiple Minimum Supports." SIGKDD Explorations, 1999.

[11] Yun, H., Ha, D., Hwang, B., and Ryu K. H. "Mining association rules on significant rare data using relative support.", The Journal of Systems and Software 67, 2003, pp. 181-191.

[12] Hahsler, M. "A Model-Based Frequency Constraint for Mining Associations from Transaction Data.", DMKD, Springer Science and Business Media, 2006, pp. 137-166.

[13] Zhou, L., and Yau, S. "Association Rule and Quantitative Association Rule Mining among Infrequent Items.", MDM, 2007. pp. 156-167.

[14] Hipp, J., Guntzer, U, Nakhaeizadeh, G. "Algorithms for Association Rule Mining A General Survey and Comparision.", SIGKDD, 2000, Volume 2, Issue 1, pp. 50-64.'

[15] Lee, W., Stolfo, S., and Mok, K. W. "Mining audit data to build intrusion detection models." SIGKDD, 1998.

[16] Tan, P., and Kumar, V. "Interestingness Measures for Association Patterns : A Perspective", SIGKDD, 2002.

[17] Ross, S. M. "Introductory Statistics", Second Edition, Morgan kaufmann publishers, 2005.

[18] Synthetic Dataset, T10I4D100K, 31^{th} Dec. 2008, <http://fimi.cs.helsinki.fi/data/>.