
A Novel Approach for Finding Rare Items Based on Multiple Minimum Support Framework

Urvi Bhatt, Pratik Patel
urviybhatt@gmail.com, patelpratik1@live.in
Department of Computer Science and Engineering, Gujarat Technological University, Gujarat, India

Abstract

Pattern mining methods describe valuable and advantageous items from a large amount of records stored in the corporate datasets and repositories. While mining, literature has almost singularly focused on frequent itemset but in many applications rare ones are of higher interest. For Example medical dataset can be considered, where rare combination of prodrome plays a vital role for the physicians. As rare items contain worthwhile information, researchers are making efforts to examine effective methodologies to extract the same. In this paper, an effort is made to analyze the complete set of rare items for finding almost all possible rare association rules from the dataset. The Proposed approach makes use of Maximum constraint model for extracting the rare items. A new approach is efficient to mine rare association rules which can be defined as rules containing the rare items. Based on the study of relevant data structures of the mining space, this approach utilizes a tree structure to ascertain the rare items. Finally, it is demonstrated that this new approach is more virtuous and robust than the existing algorithms.

© 2015 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: RP-Tree; Maximum Constraint Mode; Rare Itemset

1. Introduction

In order to mine useful information, it is necessary to perform processing on a large amount of data for which Data mining methods are widely used. Processed data is directly interpreted or exploited in order to feed into other processes. Pattern analysis has recently attracted a considerable attention to researchers and practitioners because it has been evident in many areas such as decision support, market strategy, financial forecasts. Association Rule mining [1, 2] is one of the algorithms, used to obtain valuable and useful patterns from large dataset.
In many cases, Pattern analysis techniques are used to find relations that can be of interest to a particular application domain. Itemsets, graph, sequences and association rules are the various categories of patterns that exist in databases. Selection of the method depends not only on what should be obtained but also depends on the nature of input data and view of the information, required to be represented in a more concise and intelligible way. Confidence and Support are the criteria used to filter out the patterns [3]. While the confidence shows how frequently items occur among all records, support represents frequency of patterns occurring in the dataset.

Association rules discover all rules which satisfy the user-defined minimum confidence and minimum support constraints. Market Basket Analysis is an application of association rule mining. In this application, correlations between the purchased items are analyzed. An example of the association rule is as follows,

\[ \text{cheese} \rightarrow \text{beer} \ [\text{support} = 10\%, \text{confidence} = 80\%] \]

This rule shows that 10% of consumers purchase cheese and beer together and consumers who purchase cheese also purchase beer 80% of the time.

Association rules provide beneficial and interesting itemsets from the dataset. Among them, frequent itemset is consistent with the expectations of the researchers and they are the witnesses of recurrent phenomena. However in many applications, search for rare items are more interesting [4]. Rare items are contrasting to frequent itemsets i.e. they do not occur frequently in the dataset. They possibly contradict the beliefs of domain experts and correspond to unexpected phenomena. Rare items convey highly interesting information to many domains including medicine or biology.

Single minimum support constraint model assumes that all items have similar frequency or nature in the dataset. In many real life applications, we will encounter following problems [5]:

- If the minimum support is set to a higher value, we cannot find those rules that consist of rare items.
- For finding the rule which consist of both rare and frequent items, we have to set low minimum support value. But it generates a large amount of frequent patterns which are not useful.

Example 1: In order to find rules which consist of rare purchased items such as cooking pan and food processor from the supermarket dataset [6], low minimum support value (i.e. 0.4%) is needed. We can discover the following important rule:

\[ \text{FoodProcessor} \rightarrow \text{CookingPan} \ [\text{support} = 0.4\%, \text{confidence} = 70\%] \]

However, these low minimum support value discover the following uninteresting rules:

\[ \text{Milk, Bread, Cheese} \rightarrow \text{CookingPan} \ [\text{support} = 0.4\%, \text{confidence} = 70\%] \]

As we know that, 0.4% of the consumers purchase those 4 items together is not useful because all these items are frequently sold out in a supermarket. In order to find useful rules, the value of support must be higher. This dilemma is called the rare item problem.

To emphasize this problem, efforts have been made to discover patterns using “multiple minimum support framework”. In this paper, we discuss issues related to discovering the rare items and present a new approach, named Maximum Constraint based Rare Pattern Tree (MCRP-Tree) which uses multiple minimum support framework. The number of rare items obtained by maximum constraint model is higher compared to minimum constraint model. 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 importantly, the proposed approach ensures that the rule will include rare items in it. Result shows that as compared to a single minsup model, the proposed model prunes more number of uninteresting rules while mining rare association rules.

The rest of the paper is organized as follows. In Section 2, we first summarize the basic algorithms of rare association rule mining followed by the description of different models used in multiple minimum support
framework. The letter Section, describes a proposed approach (MCRP-Tree) with example. In Section 3, we provide results which prove the better performance of proposed approach than existing algorithms. Finally, we conclude the paper in Section 5.

2. Literature Survey

2.1. Overview of The Existing Algorithm

Existing rare itemset mining algorithms are based on a level-wise approach, similar to the Apriori algorithm [7]. Apriori Algorithm employs iterative level-wise search for frequent itemset generation which uses a single minsup value at all levels to find frequent itemsets. This algorithm is more useful for finding the frequent itemsets and not the rare itemsets, except the case where value of minsup is set to a lower value. It inherits the drawback of explosion of frequent itemsets generation and also consumes too much time, space and memory for candidate generation process. It is a bottom-up approach.

In [8] Liu et al. proposed MS-Apriori which is an extension of Apriori algorithm. MS-Apriori tries to mine frequent itemsets involving rare items. Troiano et al. analyzed the problem of bottom-up approach algorithms that searches through many levels. For reducing the number of searches, author proposed the Rarity algorithm [9]. In [10] Adda et al. proposed AfRIM algorithm that also uses a top-down approach which is similar to the Rarity Algorithm. In [11] Szathmary et al. proposed MRG-Exp Algorithm and ARIMA algorithms that can mine rare itemset. In those algorithms, three type of itemsets were defined: minimal generators (MG), minimal rare generators (MRG), and minimal zero generators (MZG).

In [12], Koh et al. proposed Apriori Inverse used to mine perfect rare itemsets. For allowing Apriori Inverse to find near prefect rare itemsets, Koh et al. also proposed several modifications. Han et al. [13] proposed FP-Growth Algorithm which used frequent pattern tree (FP-tree) for storing transactions of database and reduce database scanning. It performs one scan for finding the items which satisfied minimum frequency support threshold; another scan is for an initial FP-tree construction. Maximum Constraint Based Conditional Frequent Growth (MCCFP-Growth) Algorithm [14] is extension of FP-Growth algorithm. This algorithm took more time for database scan because of item pruning. It also occupies more memory space.

RP-Tree Algorithm [15] is a modification of the FP-Growth algorithm. Similar to FP-Growth algorithm, this algorithm performs database scan for counting the support. In the second scan for building initial tree, RP-Tree used the transactions having at least one rare item. In this way, the transactions having non-rare items were not included in RP-Tree construction. This algorithm tried to provide a complete set of rare-item itemset. RP-Tree is an efficient algorithm that uses the tree data structure and identifies most of the rare association rules. Following example shows the working of RP-Tree Algorithms

Example 2: Let us consider a transactional dataset as shown in Figure 1 (a). Items A, C, B, D, E, F, G have support count 6, 6, 4, 1, 4, 2, 2 respectively. Value of Minimum Frequent Support is set to 5 (minFreqSup=5) and Minimum Rare Support is set to 3 (minFreqSup=3). An item is said to be rare if its support is less than the Minimum Frequent Support and greater than the Minimum Rare Support. So, \{B, E\} are the rare items identified by this algorithm. For RP-Tree construction, select the transaction which has least one rare item. Remove the items which having lower support value than Minimum Rare Support from the selected transactions. Finally, construct RP-Tree with help of selected transactions.

2.2. Multiple Minimum Support Framework

Most of the algorithms have used the fundamental Apriori approach which is single minsup based frequent pattern mining technique. It consists of potentially expensive pruning steps and candidate generation. Those algorithms tried to find out all rare itemsets, but they spent most of the time in searching for non-rare itemsets which tends to provide uninteresting association rules. To emphasize the “rare item problem”, efforts have been made in the literature to discover frequent patterns using “multiple minimum support framework” [4, 10, 16–20].
Fig. 1. Execution of RP-Tree Algorithm (a) Transactional Dataset; (b) Support calculation for each unique Item; (c) Select the Transactions which having at least one Rare Item from transactional Dataset; (d) Construction of RP-Tree

As per various user and application requirement, different models have been suggested in this framework. They are: (i) minimum constraint model [8, 16-19] (ii) maximum constraint model [20].

- **Minimum Constraint Model**: In this model, every item has one minimum item support (MIS). With the use of minimal MIS value among all items, minsup of pattern is represented. In this way, each pattern satisfies a different minsup value among the respected items within it. Instead of satisfying downward closure property, all patterns are satisfying sorted closure property. As per sorted closure property, “all non-empty sub sets of a frequent pattern need not be frequent, only the subsets consisting of the item having lowest MIS value within it should be frequent”. Hence, based on this model Apriori-like [8, 16] or FP-growth-like [17-19] approaches consider frequent and infrequent patterns. The sorted closure property was briefly explored in [8].

- **Maximum Constraint Model**: Maximum constraint model has been proposed in [20]. In this model MIS values were given to each item and it is known as frequent pattern only if it satisfies MIS values of all the items. This model was capable to mine uninteresting patterns but issue was that, only Apriori-like approach was used with this model. As this approach was having performance problem, we cannot extend it. With this motivation, we propose tree like approach that uses this model for finding rare patterns. The numbers of Rare association rules and Rare items obtained by using the maximum constraint are more than those using the minimum constraint. So, this is an efficient model to identify rare Items.

The proposed Maximum Constraint based Rare Pattern Tree (MCRP-Tree) is an improvement over existing algorithm in many ways. First, it avoids expensive pruning step and item generation by using tree data structure based on FP-tree for finding rare items. Secondly, MCRP-Tree focuses on rare-item itemset which gives interesting rules and does not spend more time in finding non-rare itemsets which are uninteresting. Third is that, MCRP-Tree contains only rare items by excluding the transactions which does not have rare items and also eliminate frequent items from the selected transactions. In the next section, the proposed approach is presented which uses maximum constraint model for finding rare items.

### 3. Maximum Constraint Based Rare Pattern Tree

MCRP-Tree Algorithm is the modification of RP-Tree algorithm. It accepts the transactional dataset and similarly as RP-Tree, it performs only one database scan for counting support. With a prior knowledge of item’s MIS value, this approach discovers the rare items from the dataset. MCRP-Tree selects only those transactions which have at least one rare item in it. At the time of tree construction, this approach takes only rare items and
prunes the other items from the transaction. In this way, this proposed approach constructs only a rare item tree. For calculating MIS value for each item in the transaction dataset,

\[
MIS(i_k) = \beta \times S(i_k) \quad \text{if} \quad \beta \times S(i_k) > S_{\text{lowest}} \\
S_{\text{lowest}} \quad \text{Else}
\]

Where, \(S(i_k)\) is the support of item \(i_k\) in the dataset. \(S_{\text{lowest}}\) is the user specified lowest minimum item support allowed. \(\beta\) is used to control MIS value of items according to their support. Range of \(\beta\) is 0 to 1.

During the insertion of items into a tree, order of the item should be according to the frequency of the items in original dataset and not according to the pruned transaction dataset. It may be possible that rare items have higher support value than the frequent items. This is the main reason behind not considering the support of pruned dataset. MCRP-Tree constructs conditional pattern base and conditional trees for each item existing in the tree. In this way, MCRP-Tree generates a complete set of rare items. Figure 2 shows the working of proposed algorithm.

MCRP-Tree Example:

Let us consider a transactional dataset as shown in Figure 3 (a) Items A, C, B, D, E, F, G have support count 6, 6, 4, 1, 4, 2, 2 and MIS values 4, 4, 3, 3, 3, 2, 2 respectively. Value of Minimum Frequent Support is set to 5 (\(\text{minFreqSup}=5\)). An item is said to be rare if its support is less than the Minimum Frequent Support and greater than or equal to the respective MIS value of that item. So, \{B, E, F, G\} are the rare items identified by this algorithm. During the construction of MCRP-Tree, all the transactions are selected because each transaction has at least one rare item among rare items. If the transaction does not have any rare item, it cannot contribute into the construction of a tree. While construction of MCRP-Tree, it inserts only rare items as shown in Figure 3 (d).

The initial tree is constructed using RP-tree, which only ignores the items that fall below \(\text{minRareSup}=3\), as shown in Figure 1. This RP-tree has 3 additional nodes that show frequent items as compared to the tree building using MCRP-Tree (shown in Figure 3). In addition, RP-tree does not contain F, G rare items as it removes the items having support less than \(\text{minRareSup}\) which is constant for each item in the database. MCRP-Tree builds conditional pattern bases and conditional tree to find the rare-item itemsets for each rare item.
Fig. 3. Execution of MCRP-Tree algorithm (a) Transactional Dataset; (b) Support and MIS Value calculation for Each unique Item; (c) Select the Transactions which having at least one Rare Item from transactional Dataset and remove non-rare items; (d) Construction of MCRP-Tree

4. Results

From the literature survey, we can determine that all association rule mining algorithms can be implemented in Java with sequential pattern mining framework. Experimental results for a number of itemset generated and the Time taken for itemset generation by FP-Growth and RP-Tree algorithm can be found with a brief survey. Time taken by ARIMA is significantly greater because of computationally expensive pruning steps and candidate generation. With the help of a literature review, we have observed that runtime of ARIMA is 32 times greater than FP-Growth in many datasets. RP-Tree mines less number of items in comparison to FP-Growth in majority of cases [11, 13, 15].

Maximum Constraint based Rare Pattern Tree is an extension of RP-Tree as it builds reduced tree with more number of rare items and prunes all frequent nodes. Proposed approach minimizes the time taken for itemset generation and finds more number of rare items very efficiently as compared to the FP-Growth and RP-Tree algorithm. All existing algorithm spends most of the time for discovering rare items and as a result they identify non-rare or frequent items. Proposed approach provides two constraints i.e. minimum frequent support and MIS value for finding rare items. It removes the frequent items from the transactions during tree generation and builds a tree with rare items only, known as MCRP-Tree. As an insertion of node in the tree is computationally expensive, this approach significantly reduces the execution time as compared to others.

Fig. 4. Number of Itemsets Generated
With the help of literature survey, we can provide results that prove the efficiency of proposed approach. It provides a comparison of Time taken for Itemset generation and number of itemset generated by proposed algorithm with existing algorithm. From results generated on the basis of literature survey, we can deduce that Maximum Constraint based Rare Pattern Tree requires less than half the time for execution and generates more number of rare items.

5. Conclusion

Rare items are those that appear uncommonly even though they are highly related with very specific data. The maximum constraint model, uses “multiple minimum support framework” for finding rare items. If the items’ supports in a database differ widely, it generates useful patterns very efficiently. An effective method is presented for finding rare items which uses tree structure. Maximum Constraint based Rare Pattern Tree (MCRP-Tree) effectively prunes more number of frequent items as compared to a single minimum support model. As the name suggests, this algorithm inserts only rare items while constructing the tree. The time for inserting the frequent items in the tree is avoided and as a result, processing time for mining the rare items is minimized. An effectiveness of the algorithm is shown by the results. As a part of future work, we can apply proposed approach on Big data for mining rare items.

References