An Novel Approach to Mine Rare Association Rules Based on Multiple Minimum Support Approach

Anubha Bansal, Neelima Baghel & Shruti Tiwari

Department of Electronics & Communication, Malwa Institute of Technology, Indore, M.P., India
E-mail : bansal24.anubha@gmail.com, neelima.baghel@gmail.com, myec074@rediffmail.com

Abstract – The paper discusses a new approach to extract rare association rules that can be defined as the rules containing the rare items or infrequent items in the dataset. For extracting rare itemsets, various efforts have been done. One of them is MSApriori method which finds the rare association rule. This method works up to some extent, which suffers from “rare item problem” dilemma. To extract rare itemsets effectively, an approach has been proposed that would allow user to specify multiple minimum support that reflects the nature of items and their varied frequencies in database. The proposed approach improves the existing MSApriori approach to mine the rare association rules. Experimental results show how the technique proves to be better for mining rare association rule.

Keywords – Apriori, MSApriori, Rare association rules, rare item problem

1. INTRODUCTION

Association rule mining is an important model in data mining, that was introduced in[1], which discovers all items associations (or rules) in the data that satisfy the user-specified minimum support and minimum confidence constraints. Minsup controls the minimum number of data cases that a rule must cover. Minconf controls the predictive strength of the rule. The classic application is market basket analysis [1]. It analyzes how the items purchased by customers are associated. An example of an association rule is as follows,

\[ \text{cheese} \rightarrow \text{beer} \ [\text{sup} = 10\%, \text{conf} = 80\%] \]

This rule says that 10% of customers buy cheese and beer together, and those who buy cheese also buy beer 80% of the time.

Since only one minsup is used for the whole database, the model implicitly assumes that all items in the data are of the same nature and/or have similar frequencies in the data. This is, however, seldom the case in real life applications. In many applications, some items appear very frequently in the data, while others rarely appear. If the frequencies of items vary a great deal, we will encounter two problems:

1. If minsup is set too high, we will not find those rules that involve infrequent items or rare items in the data.

2. In order to find rules that involve both frequent and rare items, we have to set minsup very low [4]. However, this may cause combinatorial explosion, producing too many rules, because those frequent items will be associated with one another in all possible ways and many of them are meaningless.

Example 1: In a supermarket transaction data, in order to find rules involving those infrequently purchased items such as food processor and cooking pan (they generate more profits per item), we need to set the minsup to very low (say, 0.5%) [4]. We may find the following useful rule:

\[ \text{foodProcessor} \rightarrow \text{cookingPan} \ [\text{sup} = 0.5\%, \text{conf} = 60\%] \]

However, this low minsup may also cause the following meaningless rule to be found:

\[ \text{bread}, \text{cheese}, \text{milk} \rightarrow \text{beer} \ [\text{sup} = 0.5\%, \text{conf} = 60\%] \]

Knowing that 0.5% of the customers buy the 4 items together is useless because all these items are frequently purchased in a supermarket. For this rule to be useful, the support needs to be much higher.

This dilemma is called the rare item problem [2].

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 [3].

A rare association rule refers to an association rule forming between either frequent or rare items or among rare items. In the literature, efforts are being made to propose improved approaches to mine rare associations [4], [5], [6], [7]. In [4], 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 [5], 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 [6], a stochastic mixture model knows 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 [7] extracts the association rules by considering only infrequent items (i.e., items having support less than minsup).

In this paper, we propose an improved approach by calculating minsup value for each item. 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 show that the proposed approach discovers frequent itemsets involving rare items in an efficient manner as compared to 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

The Apriori algorithm [9] [10] 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:

(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.

The extraction of frequent itemsets using Apriori is illustrated in Example2.

Example2: Table 1 shows the transaction dataset. The working of Apriori with minimum support = 20% is as shown in Fig. 1. C1 is generated after the first scan. From C1, frequent 1-itemset L1 (itemsets represented in bold) is generated with the items whose support (Number of times that item occur/Total number of transactions in database) is greater than or equal to minsup. C2 is generated by joining L1 with itself. From C2, L2 (itemsets represented in bold) is generated if an itemset support is greater than or equal to minsup. Similarly, C3, L3 is generated with itemsets \{(I1, I2, I3), (I1, I2, I5)\}.

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I1, I2, I5</td>
</tr>
<tr>
<td>2</td>
<td>I2, I4</td>
</tr>
<tr>
<td>3</td>
<td>I2, I3</td>
</tr>
<tr>
<td>4</td>
<td>I1, I2, I4</td>
</tr>
<tr>
<td>5</td>
<td>I1, I3, I6</td>
</tr>
<tr>
<td>6</td>
<td>I2, I3</td>
</tr>
<tr>
<td>7</td>
<td>I1, I3</td>
</tr>
<tr>
<td>8</td>
<td>I1, I2, I3, I5</td>
</tr>
<tr>
<td>9</td>
<td>I1, I2, I3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>$S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>66.67</td>
</tr>
<tr>
<td>I2</td>
<td>77.78</td>
</tr>
<tr>
<td>I3</td>
<td>66.67</td>
</tr>
<tr>
<td>I4</td>
<td>22.22</td>
</tr>
<tr>
<td>I5</td>
<td>22.22</td>
</tr>
<tr>
<td>I6</td>
<td>11.11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>$S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1, I2</td>
<td>44.44</td>
</tr>
<tr>
<td>I1, I3</td>
<td>44.44</td>
</tr>
<tr>
<td>I2, I3</td>
<td>44.44</td>
</tr>
<tr>
<td>I1, I4</td>
<td>22.22</td>
</tr>
<tr>
<td>I2, I4</td>
<td>22.22</td>
</tr>
<tr>
<td>I1, I5</td>
<td>22.22</td>
</tr>
<tr>
<td>I2, I5</td>
<td>0</td>
</tr>
<tr>
<td>I3, I4</td>
<td>0</td>
</tr>
<tr>
<td>I3, I5</td>
<td>11.11</td>
</tr>
<tr>
<td>I4, I5</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig.1: Working of Apriori Algorithm at minsup = 20 %. Here, the notation ‘S’ indicates the support percentage.

Mining of rare associations (association rules involving rare items) with single minsup approaches may cause “rare item problem” [14] dilemma. The “rare item problem” is as follows:
If \textit{minsup} is high, frequent itemsets involving rare items are missed as the support of the rare items is less than the given \textit{minsup}. In order to find frequent itemsets involving rare items, the \textit{minsup} value should be fixed at low value. As a result, the number of frequent itemsets explodes.

\textit{Example 3:} In the transaction dataset shown in Table 1\{(I1, I2, I3), (I1, I2, I5)\} are extracted at \textit{minsup} = 20\%. However, it can be observed that frequent itemset \{I1, I3, I6\} involving rare items is missed. In order to extract frequent itemset \{I1, I3, I6\}, the \textit{minsup} has to be set to low value, equal to 10\%. With \textit{minsup}=10\%, the process of finding frequent itemsets is depicted in Fig. 2.

It can be observed from Figure 2 that while decreasing the minimum support value to 10\% number of frequent 2-itemsets increases (as compared to the frequent 2-itemsets when support threshold is set to 20\%). Then also the frequent itemset does not include the rare item (I6) in it. Hence decreasing the support value is not a solution to mine the rare association rules.

\textbf{B. Multiple minimum support approach}

To improve the performance of extracting frequent itemsets involving rare items, an approach known as Multiple Minimum Support Model \[4\] is used.

In this model, each item is assigned with a \textit{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 equal to a percentage of its support.

In this model, the definition of association rules remains the same. The definition of minimum support is changed.

The model, defines the minimum support of a rule in terms of \textit{minimum item supports} (MIS) of the items that appear in the rule. That is, each item in the database can have a minimum item support that can be calculated using some formula or can be specified by the user. By providing different MIS values for different items, the user effectively expresses different support requirements for different rules.

Let MIS \((i)\) denote the MIS value of item \(i\). The minimum support of a rule \(R\) is the lowest MIS value among the items in the rule. That is, a rule \(R_{a1, a2, \ldots, ak \rightarrow ak+1, \ldots, ar}\)

where, \(a_j \in I\), satisfies its minimum support if the rule’s actual support in the data is greater than or equal to: min (MIS\((a1)\), MIS\((a2)\), \ldots, MIS\((ar)\)).

The minimum Item Support (MIS) is calculated as described below. For every item \(i \in I\), the MIS \((i)\) is calculated as per Equation (1).

\[
\text{MIS} (i_j) = \beta S (i_j); \quad \text{if } \beta S (i_j) > LS
\]

\[
= LS \quad \text{else}
\]

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

In this method, 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 Multiple Minimum Support model improves the performance over single minsup based approaches by addressing the “rare item problem”.

---

**Fig. 2**: Working of Apriori Algorithm at \textit{minsup} = 10\%. Here, the notation 'S' indicates the support percentage.
Example 4: The working of the MS Apriori (Multiple Minimum Support) for the transaction dataset shown in Table 1 is depicted in Figure 3, by considering $\beta=0.7$ and LS=20%.

After calculating C1, MIS is calculated by using the Equation 1. From C1, the L1 is generated with the items whose support is greater than or equal to their respective MIS. From L1, C2 is generated by joining L1 with itself. From C2, L2 is generated if an itemset satisfies the lowest MIS value of the items in it. Similarly C3/L3 is generated with frequent itemset $\{(11, 12, 15)\}$.

It was observed that MSApriori still suffers from the “rare item problem” dilemma, as can be seen from the Example 4 that the rare item (I6) is not included in the frequent itemset. Hence this algorithm also fails to discover the rare association rules.

III. PROPOSED APPROACH

In this approach, we would be able to extract the rare association rules by using the previous knowledge used in MSApriori algorithm. For each item ‘$i$’, calculation of minimum support known as minimum item support $\text{MIS}(i)$ is as follows:

$$\text{MIS}(i) = \begin{cases} \beta S(i) & \text{if } \beta S(i) > \text{LS} \\ S(i) & \text{else} \end{cases}$$

Equation (2) ensures the extraction of frequent itemsets involving frequent items, rare items and both frequent and rare items efficiently. All the notations used in the proposed approach are same as that used in MSApriori. The advantage of the proposed approach over MSApriori is as illustrated using Example 5.

Example 5: Consider the dataset as illustrated in Table 1, with minimum support (LS) set at 20% and $\beta=0.7$. The results with the proposed algorithm are as depicted in Fig. 4.

After calculating C1, MIS is calculated by using the Equation 2. From C1, the L1 is generated with the items whose support is greater than or equal to their respective MIS. From L1, C2 is generated by joining L1 with itself. From C2, L2 is generated if an itemset satisfies the lowest MIS value of the items in it. Similarly C3/L3 is generated with frequent itemset $\{(11, 12, 15), (11, 13, 16)\}$. It can be observed from the Fig.4 that the frequent itemset found after 3rd iteration includes the rare item (I6) in it; hence the proposed algorithm is able to extract the rare association rule.

\[
\begin{array}{|c|c|}
\hline
\text{Item} & S \\
\hline
11 & 66.67 \\
12 & 77.78 \\
13 & 66.67 \\
14 & 22.22 \\
15 & 22.22 \\
16 & 11.11 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
\text{Item} & \text{MIS}(i) \\
\hline
11 & 44.45 \\
12 & 22.22 \\
13 & 44.45 \\
14 & 22.22 \\
15 & 44.45 \\
16 & 11.11 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
\text{Itemset} & \text{S} \\
\hline
11,12 & 44.45 \\
11,13 & 44.45 \\
11,14 & 11.11 \\
12,13 & 44.45 \\
12,14 & 22.22 \\
13,14 & 44.45 \\
11,12,13 & 22.22 \\
11,12,14 & 11.11 \\
11,13,14 & 11.11 \\
12,13,14 & 11.11 \\
12,13,15 & 11.11 \\
12,14,15 & 11.11 \\
\hline
\end{array}
\]

Fig.3: Working of MSApriori with $\beta=0.7$, minsup=20 %. Here, the notation ‘S’ indicates the support percentage.

IV. EVALUATION

A. Experimental details

The performance of the proposed approach is evaluated by considering the real world dataset. Due to confidentiality agreement, we are unable to provide the details of the application. Here, we only give the characteristics of the data. The real world dataset contain 50 items with 200 transactions. In the experimental results, we have compared the proposed approach with Apriori and MSApriori approaches.

B. Performance Results

For this application, the user sets the different values of support (LS). The results are shown in Fig. 4, 5, 6 which include both the numbers of rules on Y-axis and the support on X-axis. The performance results shown in Fig. 4, 5, 6 compares the result obtained by existing methods of finding rules i.e. Apriori, MSApriori and the proposed approach. Our new method increases the number of rules as compared to MSApriori but have less number of rules as compared to Apriori for the same support values, which means that the proposed approach includes the rare association rules.

V. CONCLUSION

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 equation 2. 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 rule would include the rare items in it. We have evaluated the performance of the proposed approach by conducting experimental results on real world dataset. The results show that, as compared to existing approaches, the proposed approach prunes frequent itemsets involving rare items.

<table>
<thead>
<tr>
<th>Item</th>
<th>MIS(β)</th>
<th>Item</th>
<th>MIS(β)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>66.67</td>
<td>I1</td>
<td>44.45</td>
</tr>
<tr>
<td>I2</td>
<td>77.78</td>
<td>I2</td>
<td>53.46</td>
</tr>
<tr>
<td>I3</td>
<td>86.67</td>
<td>I3</td>
<td>44.45</td>
</tr>
<tr>
<td>I4</td>
<td>22.22</td>
<td>I4</td>
<td>22.22</td>
</tr>
<tr>
<td>I5</td>
<td>22.22</td>
<td>I5</td>
<td>22.22</td>
</tr>
<tr>
<td>I6</td>
<td>11.11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>S</th>
<th>Itemset</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1,I2</td>
<td>22</td>
<td>I1,I2,I3</td>
<td>22</td>
</tr>
<tr>
<td>I1,I4</td>
<td>11</td>
<td>I1,I3</td>
<td>11</td>
</tr>
<tr>
<td>I1,I5</td>
<td>11</td>
<td>I1,I3,I5</td>
<td>11</td>
</tr>
<tr>
<td>I2,I3</td>
<td>22</td>
<td>I1,I3,I6</td>
<td>11</td>
</tr>
<tr>
<td>I2,I5</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I2,I6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I3,I4</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I3,I5</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I3,I6</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I4,I5</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I4,I6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4: Working of proposed approach with β=0.7, minsup=20%.

Fig. 5: Number of Rules versus Support in Apriori algorithm

Fig. 6: Number of Rules versus Support in multiple minimum support model

VI. REFERENCES


[13] Yun Sing Koh, Nathan Rountree- “Rare Association Rule Mining: An Overview”


+++