iGain - Apriori: An interest gain based rule assessment for Association Rule Mining

Tumma Sailaja, A.Poongodai

Aurora’s Technological and Research Institute, Hyderabad, India

Abstract- At recent times many algorithms have evolved for excavating or mining the association rules. But only few of them got the significance. One among them that got prominence is the Apriori algorithm. The main ideology of this algorithm is to determine the continuously existing sets which allow deriving the important association rules and these rules must match up with the values of min support threshold and min confidence threshold. Keeping in view, the progress of effective mining and aim of fulfilling the customers desires our paper has put forward enhanced algorithms by presenting a new item measuring model interest gain along with interest items and frequency threshold. For validating these algorithms, effective testing has been done. The test considers the key values like the space and time complexity of the each algorithm and our algorithm has advanced many of the conventional ones.

Index Terms- apriori algorithm, dynamic mining, interest item, frequency threshold

I. INTRODUCTION

The world is advancing on the technical side, mostly in the field of artificial intelligence, progressing of database and statistical methodology likewise the development of the area of research, that is association rules is the most significant field for the new inventions of transaction database has also advanced the mining of the association rules especially in knowledge. The mining is done from the huge sized data through which the inter relation between the item sets and their connection can be known through these association rules [1, 2]. There are an unlimited technologies and methodologies at present for excavating the data or technically called mining. Some of them are association rules, clustering, neural networks, rough sets, genetic algorithms and so on. The most procedure in demand is the apriori algorithm and we utilize this algorithm which is a set of items that is connected as a relation among the rules.

The procedure consists of a categorization of solving two problems [3, 4]. They are:
1. We need to determine the item sets that occur recurrently which matches the minimum support.
2. Next is utilizing these determined item sets and relating them to minimum credibility to derive the association rules.

Among the two problems the initial one has the more concentration because it is transparency of time and space by which many association rules that are prevailing in recent times are working especially on determining the item sets that occur recurrently. To enhance the system of determining the item sets and expand the efficiency of association rules, there introduced some methodologies like the numeric association rules [5], multi-level association rules [6].

Following this methodology of improving the association rules a huge number of algorithms have been come into existence. But the algorithm that is above all the rest and that is utilized in our paper is the Apriori algorithm. Here we made an approach of using the apriori by presenting the new objects like the interest items and frequency threshold in order to dissolve the time complexity for the database to check for the items. Another advantage if this is that it can place the dynamic mining also in it for fulfilling the desires of the users and increase the performance of the algorithm to the maximum extent.

II. ASSOCIATION RULES

Let us assume the difficulties in a mall and the association rules are derived for those difficulties. These rules enable us to find out the connections among the various products available at the mall. The connections of the various products reveal the general interest of the consumer of the product and this methodology can be functionalized in consumer shopping, directory design, commercial publishing mail. The procedure of excavating the association rules was a concept of deep methodology mostly on the side of research which comprises of an uncountable amount of algorithms for the sake of excavation of the association rules. Due to the presence of these many algorithms the association rule excavation has become more profitable in all the research fields. The mining have a vast categorization and the data mining is highly established one [7, 8].

A. Related Concepts

Sequentially while starting the algorithm let us consider some assumptions for the clear understanding of the about the algorithm like:

Consider a transaction database $D = \{T_1, T_2, ..., T_n\}$ which constitutes of the following:

- For the above transaction database $i$ is a property assigned with some value.
- $I = \{i_1, i_2, ..., i_n\}$ is a set of all such properties existing consequently in the transaction.
- $T = \{i_1, i_2, ..., i_m\}$ ($m \leq n$) is the transaction.
- $|D|$ is the count of the occurrences of the transaction database transactions in the sense that how many times a particular transaction database $D = \{T_1, T_2, ..., T_n\}$ has occurred. This provides the facility of reoccurring iteratively. But there exists some obstacles due to this

www.ijsrp.org
Initial step: Determining the frequent or recurrent item sets

The process performance is mentioned in below steps:

1) Frequent 1 – itemsets
   \{\{A\},\{B\},\{C\},\{D\},\{E\}\}

2) Frequent 2 – itemsets
   \{\{AB\},\{AC\},\{AE\},\{BC\},\{BD\},\{CD\}\}

3) Frequent 3 – itemsets
   \{\{ABC\}\}

The entire item sets does not have the value of their supportive degree more than or equal to the value of the minSup. Only few of them have that value and they are classified as the frequent or recurrent item sets.

B. The introduction of traditional Apriori algorithm

At the outset, we excavate all the recurrent item sets, these are names as 1-itemsets, and this procedure is iterated recursively to excavate recurrent k-item sets (k>1). The same methodology is followed even in the case of association rules. Initially we excavate all the association rules and later we determine the recurrent association rules. In particular we carry out the excavation of individual recurrent k-itemsets by considering those, whose minimum confidence threshold minConf is the least value that should be attained by the association rules.

The Minimum confidence threshold minConf is the least value that should be attained by the association rules.

C. Illustration of the Apriori algorithms

A Transaction database is displaced in the table in given Table I, min Sup =50%, minConf = 70%. The recurrent association rules in transaction database D have been sent a request.

<table>
<thead>
<tr>
<th>Tid</th>
<th>Itemsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ABCDE</td>
</tr>
<tr>
<td>2</td>
<td>ABC</td>
</tr>
<tr>
<td>3</td>
<td>CDEF</td>
</tr>
<tr>
<td>4</td>
<td>ABE</td>
</tr>
</tbody>
</table>

TABLE: 1 TRANSACTION DATABASE

looping that is there is a chance of occurrences of the duplicate transactions. So to evade and resolve such kind of problems a distinct id has been given in our procedure for the individual transaction in the database namely TID.

- Sup(X) which is termed as the supportive degree is defined as the ratio of quantity of the X involved in the transaction database D.

- Conf (A ⇒ B) represents the confidence when we consider the association rule (A ⇒ B). This is defined as probability based on the condition that, existence of an item set B in the same constraint where the item set A exists.

- Coming to the Minimum support threshold minSup, it is defined as the least value that is to be attained by the item sets in processing the excavation.

- The Minimum confidence threshold minConf is the least value that should be attained by the association rules.

When we verify the exceeding performance of the procedure, the previously mentioned drawbacks are prevailing.

Property1. If \( A \rightarrow B \) exists in the association, then \( |AB| \geq \text{min Count} \). So if \(|T| < \text{min Count} \). This results in the relationships that comprise the any subset of T doesn’t match up with the minimum confidence.

Property2. If \( A \rightarrow B \) exists in the association, this proves that A and B are considered as frequent item sets, but there is no feasibility that AB compulsorily be a frequent item set

Definition1: A group of items which are concerned by the customers as this place a major role mining is defined as interested item I. The minimum confidence threshold minConf is the least value that should be attained by the association rules.

Definition2: The transaction database is numbered with some items set which are preferred with some concernment of subset of items that is termed as item frequency.

Definition3: The least count of items presented in the Association rules is termed as support of the frequency threshold shortly named as minSupCount.

\[ \text{min SupCount} = |D| \times \text{min Sup} \]

Definition4: The least count in the given rule set under is referred as interest gain in short iGain. iGain can measure as follow.

\[ iGain_i = \frac{{rC\over{\text{ov\ erase}\_i}}}{{|D\_left| + |D\_right|}} \times \text{support}_i \]

Here

\[ iGain_i \] is interest gain of itemset \( i \)

\[ rC\over{\text{ov\ erase}\_i} \] is coverage of itemset \( i \) in rule-set \( D \)

\[ |D\_left| \] is total number of left items in rule-set \( D \)

\[ |D\_right| \] is total number of right items in rule-set \( D \)

III. THE DESCRIPTION OF ENHANCED ALGORITHM

A. Introduction of enhanced algorithm

The major Drawbacks of conventional Data mining are

1. Too much recurrently examining the database.
2. The minSup and minConf values are predefined i.e. they do not alter. If the consumer wants to alter the values of minSup and minConf, the data should be mined again, which provides lot of ambiguity to the consumers.
3. In general there should be no focus of EF, i.e. the operation of mining should not take place on EF, which is feasible in conventional data mining.

When we verify the exceeding performance of the procedure, the previously mentioned drawbacks are prevailing.
support \_i \text{ support of itemset } i

Definition 6: Rule \( A \Rightarrow B \) is an interesting rule if \( iGain_A \geq iGain_B \)

B. Procedures in the Algorithm

Apriori algorithm has been developed by us by encapsulating three insufficient properties of the conventional mining algorithm and rest of the two properties. The development steps are as follows:

Initial step: The subset of items of frequency consists of variety of interest items are represented in this article.

1. The mining transactional database and interest items values are to be submitted
2. The transaction database is examined.
3. The database is stored with every sub-set of items. The frequency and the final count of attributes D are also be stored and subset is saved.

Secondary step: determine the association of items

1. The values of minSup and minConf are to be submitted and value of minSup in the minSupCount.
2. To determine the recurrent item sets and eliminate those subsets with frequency smaller than the minCount the saved subset with interest items is examined.
3. The output of the association rules is generated for which the confidence is more than the minConf

C. The Association Rule Mining for Interested Elements pseudo-code

Step 1: To calculate occurrences of the item for every subset of interested items.

Input: Transaction database \( D \); Interested items \( I \)
Output: The recurrence of Items for every subset of interested items and store the item id values in the database.

Considering every individual \( i \in I \) do // Negotiating each Interested items
Count++ // store the count of interested items
End for

\( I \text{ sum}\) = subset\( (I) \) // determine the no. of subsets of Interested items

Considering every log \( d \in D \) do
Considering every individual \( i \in d \) do // Negotiating each Interested items
If \( i \in I \) // whether the item is in the Interested items
\( d_I \cup = i \) // calculating every one of the Interested items
End if

\( I_d = \text{subset}(d_I) \)
End for

Considering every Item in \( I_d \) // Negotiating every item in the set.
\itemcount + + // for every item count is incremented in the set
End for
Count + + // store the count of the item id values in the database

End for

Pseudo-code of subset \( (I) \) is mentioned below:

\begin{verbatim}
foreach(i) where i = 1..Icount
Return \( I \text{ sum}\cup = I_1 \cup I_{i+1} \cup \ldots \cup I_{i+1} \ldots I\text{count} \)
End for
\end{verbatim}

Step 2: Calculating the association of \( I \text{ sum} \)

Input: Minimum support threshold \( (\text{min Sup}) \); Minimum confidence threshold \( (\text{min Conf}) \)
Output: The credibility and requirements to be matched up by the AR.

\( \text{min SupportCount} = \text{count} * \text{min Sup} * \text{min Conf} \)
Considering every individual \( l \in I \text{ sum} \) do
if \( (I_{\text{itemCount}} < \text{min SupportCount}) \)
Delete l // l is eliminated out of the \( I \text{ sum} \) if it’s frequency value is less than minSupportCount
Else
if \( (I_{\text{itemCount}} \geq \text{min Count}) \) do
\( L \cup = l \) // When the frequency value is more than the minCount then l is stored as a recurrent item \( l \text{count} + + \) // Store the count of the recurrent items.
End if
End if
End if
End for
Considering every individual \( l \in L \) do //Negotiating all the items of L.

\begin{verbatim}
foreach(i) where i = 1..lcount
Negotiating every one of the items of L

if\((L_i\text{.items} \cup L_k\text{.items})\) // if the items in \( L_i \) and \( L_k \) vary
if\((L_i\text{.items} \cup L_k\text{.items}) \in I \text{ sum}\) // if the combination of items in \( L_i \) and \( L_k \) are all present in \( I \text{ sum}\)

if\((|(L_i\text{.items} \cup L_k\text{.items})\text{.count}| \geq \text{min Conf})\) do
\( L_i\text{.items} \times \text{count}\)
if\((iGain_h \geq iGain_h)\)
Return \( L_i \Rightarrow L_k\text{, confidence}(L_i \Rightarrow L_k) \)
\end{verbatim

End if
End if
End if
End if
End for

Algorithm’s time complexity: The storage count of database \( m \) is constantly a huge value. But when we consider an invariable

www.ijsrp.org
value n, the interested items of the customer is very low. So, the algorithm complexity is derived as $O(m^*2^n - 1)$.

IV. EXPERIMENT AND EXPERIMENTAL ANALYSIS

The values of the objective data forms and interested items are given as the input for the enhanced algorithm flowchart and examine the every item sets in the database. By this examining we derive the value of frequency of the item sets and that is stores on the list-L. The threshold of credibility and support degree is given as the inputs and they are well examined to be considered as authorized input or not.

By validating the L, the frequencies of items with values less than the frequency threshold, are all eliminated.

Figure 1. The Interest gained Association rules - ARM for Interested Items algorithm flowchart

By considering all the associative rules we excavate the AR. Eventually, check whether more mining is necessary or not. And if the verification came true, loop the processes of delivering the input of extra threshold values related to support and credibility to persist the mining. Or if the verification is set false then terminate out of the program. The enhanced algorithm flowchart is illustrated in Fig.1.

V. PERFORMANCE EVALUATION OF IMPROVED APRIORI ALGORITHM

In our article, the evaluation of the performance of the algorithm resulted firstly in deriving the minimum support of degree and credibility that have the property of flexibility which provides basic demands of the users significantly. While considering the next step, there is an enhancement of effective mining by this refined Apriori mining algorithm which uses the précised attributes. This enhancement leads to the condensing of time and space complexity. In this article a comparison has been put forward between the conventional algorithm and our sophisticated Apriori algorithm by the author considering a similar situation for the two cases that is listed in the above Table I, and the subsequent outcome is illustrated in Fig. 2(a).

VI. CONCLUSION

Computational effectiveness, which is the primary concern faced by the Apriori algorithm is defined by our research. On analyzing of Apriori algorithm, we winded up our approach that the algorithm can trim down the frequency on examining the database and number of rules generated, shrink the set of items that is unfeasible for mining association, enhanced the elasticity for being an intelligible mining algorithm, which makes algorithm in well organized manner.

Some of the chief characteristics came into concern like developing effective mining when processing the huge quantity of information and setting up supportive degrees, admirable in several layered database where customers wants to excavate additional vital rules, mining of unusual item set in extra effective procedures[9, 10] can be considered in future research and development.

REFERENCES


First Author – Tumma Sailaja B. Tech, Pursuing M.Tech (Software Engineering) specialization at Aurora’s Technological and research institute, Hyderabad.

Second Author – A. Poongodai, M.Tech (CSE), Associate Professor in the Department of CSE at Aurora’s Technological Research Institute, Hyderabad.