e-ISSN: 2349-9745
p-ISSN: 2393-8161

Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

A Novel Apriori Algorithm for Association Rules Mining

Raj Jaiswal[1] , Ranu Soni [2]
[1]RCPIT, Shirpur
[2]RCPIT, Shirpur

Abstract - In today’s comparative business world vast amount of data is stored within many database and datasets. This data is used is inputs for the knowledge discovery process. This requires automated data mining tools and they need to be effective and efficient. Association rule mining is the process of finding hidden previously unknown and valid and useful information for large database. Association rule mining use data sets and database as inputs and discover the required knowledge which is used for decision making. In this paper, we will describe the most popular classical Apriori algorithm and the problems of this algorithm. And then we describe the new algorithm that overcomes the problems of the classical Apriori algorithm. At the end of this paper we will discuss the results.

Key words - data mining, KDD Association Rule Mining, Apriori Algorithm, frequent Item-sets.

1. INTRODUCTION

In today's world of competitive business environment and popularization of computer and development of database, more and more data are stored in the large database. There is a great need to extract hidden and potentially meaningful information from large database. It is impossible to find useful information with traditional method. And then, data mining techniques have come as a reflection of this problem. Association rule mining is an important research area in data mining field. Is aim to find out the remarkable association or relationship between a large set of data items.

Data mining techniques are very easy to handle. It is very easy to implement this technique on the existing software and hardware platforms to improve the quality of the information resources. It can also be integrated with new techniques and systems which are newly introduced in the market.

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. Association rules provide information of this type in the form of "if-then" statements. These rules are computed from the data and, unlike the if- then rules of logic, association rules are probabilistic in nature. In addition to the antecedent (the "if" part) and the consequent (the "then" part), an association rule has two numbers that express the degree of uncertainty about the rule. In association analysis the antecedent and consequent are sets of items (called itemsets) that are disjoint (do not have any items in common).

Support: The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. (The support is sometimes expressed as a percentage of the total number of records in the database.)

Confidence: Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent.

Lift: Lift is nothing but the ratio of confidence to expected confidence. Lift is a value that gives us information about the increase in probability of the "then" (consequent) given the "if" (antecedent) part.

1.1. Knowledge Discovery in database

Knowledge Discovery in Database is the process of identifying a valid, potentially useful and ultimately understandable structure in data. This process involves selection, cleaning or pre-processing it, transforming or reducing it, applying a data mining component to produce a structure, and then evaluating the derived structure.

1.1.1. Necessity of KDD

Recently, the progress of bar code technology, most of the product will come with bar-code which was possible the organization to store the huge amount of product sales data referred to as a basket data. A transaction will store the basket data brought by the customer and the size of the database goes increased. Today millions of databases have been used by business management, government administration, engineering data management and many other applications. Rapid use of database technology the data mining will become popular research area. [11]

1.1.2. Stages of KDD

The stage of KDD, start with the row data and generate the knowledge from row data [12].

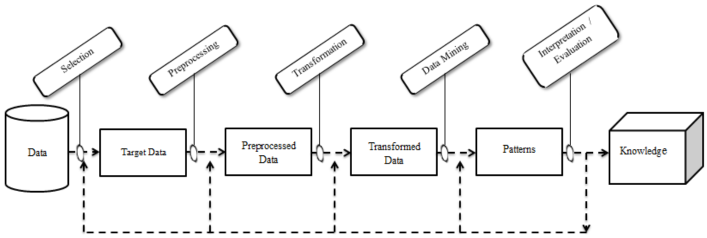

Figure 1 Knowledge discovery process

- **Selection:** In this stage data will be selected or sampled from row data that are relevant to the some criteria. Example, for credit card customer profiling, we extract the type of transaction for each type of customers and we do not select in detail of the shop where the transaction takes place.
- **Preprocessing:** In this stage unnecessary information is removed which is not match in the given criteria.
- **Transformation:** In this stage the data is not merely transferred across, but transformed in order to be suitable for the operation of the data mining.
- **Data Mining:** In this stage the actual patterns is generated form row data according to the user criteria. For this process it will use the mining approaches like association rule mining, clustering, classification and the algorithms.
- **Interpretation and Evaluation:** The patterns obtained in the data mining stage are converted in to knowledge, which is used to support decision making.

1.2. Classical Apriori algorithm[1]

Apriori employs an iterative approach known as a level wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L_1. Next, L_1 is used to find L_2, the set of frequent 2- itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database. To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property, presented is used to reduce the search space.

Apriori property: All nonempty subsets of a frequent itemset must also be frequent. A two-step process is used to find the frequent itemsets: join and prune actions.

The join step: To find L_k a set of candidate k-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_k.

The prune step: The members of C_k may or may not be frequent, but all of the frequent k-itemsets are included in C_k. A scan of the database to determine the count of each candidate in C_k would result in the determination of L_k (i.e., all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to L_k). To reduce the size of C_k, the Apriori property is used as follows. Any (K-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (K-1)-subset of a candidate k-itemset is not in L_{k-1}, then the candidate cannot be frequent either and so can be removed from C_k.

1.3. Apriori Algorithm

Input: transactional database D and minimum support threshold min_sup.

Output: L, frequent item-sets in D.

Method:

```
1.    $L_1$ = Frequent items of length 1.
2.    For (k=1;$L_k$!=ϕ;K++) do,
3.    Ck+1 = Candidates generated from $L_k$.
4.    For each transaction t in database do,
5.    Increment the count of all candidates in $C_{k+1}$ that are contained in t.
6.    $L_{k+1}$= Candidates in $C_{k+1}$ with minimum support.
7.    End do.
8.    Return the $L_k$ as the set of all possible frequent item-sets.
```

1.4. Limitations of Apriori algorithm

Apriori algorithm, despite its simple logic and inherent pruning advantage, suffers from limitations of a huge number of repeated scans of entire transaction database. Since it is a level wise algorithm hence it requires separate scans of the database and over the entire frequent item-set mining process, this become tedious and is a serious limitation.

Another limitation of the Apriori algorithm is the generation of candidate sets which can become cumbersome and time consuming when the number of frequent 1 item-set is large.

2. APRIORI ALGORITHM IMPROVEMENTS

2.1. The ideology of Apriori algorithm improvement

To enhance the efficiency of production of the frequent itemsets, this paper discusses two problems of the Apriori algorithm. First, we need to scan the database multiple times and Second, it will generate large candidate itemsets, which will increase the time and space complexity. To overcome these defects we first find frequent_one_itemset of database then generate power set of the frequent_one_itemset and initialized itemset count=0. Cal l this power set as global power set. When we scan database for itemset counting, first we delete items from transaction which is not present in frequent_one_itemset list. This step will reduce the extra generation of candidate itemsets. After delete process we generate local power set of remaining items of the transaction and compare with the global power set. When match fund increase the itemset count by one. This step will reduce the multiple scan of database. These steps will use for increase the efficiency of the algorithm.

2.2. Improved algorithm description

Input:

1)Database D with the format (Tid, itemset), where Tid is a business id and itemset is the itemset corresponding to the business.

2)Minimum support threshold: min-sup;

Output: Li, Frequent itemset in D;

Here is the flow chart:

1)L1= find frequent_one_itemset(D);

2)Generate power set of L1(frequent_one_itemset(D)) and initialize itemset count=0, and called it Global power set;

3)Scan the database D till End

i) Read itemset from transaction and delete items which are not in L1 and then generate local power set of remaining items of the transaction.

ii) Compare local power set with Global power set one by one and if itemset is match then increase the itemset count by one of the Global power set.

Prune the acquired candidate itemset.

4) Scan Global power set and test each item set count of candidate itemset;

i) If itemset count of candidate itemset is less then min-sup then delete that item set from Global power set.

5) Remaining itemset of the Global power set will be our required frequent itemset.

3. ANALYSIS

Improved algorithm first finds the one item-set list by scanning the database once and then generates the power set of the one item-set list. After that it will scan transaction one by one from database and then compare the item set list of the transaction to one item-set list and remove the items from transaction which is not available in one item-set list. And of the remaining item-set of the transaction generate the power set and compare with the global power set and increment the item-set count of the item-set which is match. This step will remove the unnecessary candidate generation and comparison. At the end when all the transaction is process and compared the global power-set will store the all possible candidate item-sets. According to the user define support threshold value, item-set will be remove from global power-set which not hold the support value. And the final the after pruning the global power-set will hold the only the required item-sets.

This algorithm will scan the entire database twice only. At the start when algorithm scan database first round to generate one item-set list, infrequent item-set will be removed at this step and not further generated and counted. Hence we can say that is this step is used to remove the unnecessary generation of candidate which is major problem of the classical Apriori algorithm.

This improved algorithm will eliminate the problem of the classical Apriori algorithm, that is: multiple time scan of entire database and the unnecessary generation of candidate item-set. Hence we can say that our improved algorithm improve the performance of system and data mining tool.

4. CONCLUSION

Today information is collected almost everywhere in our daily lives. This leads to the huge amount of data available. The analysis of this data by the classical techniques is not possible. Data mining provides tools to find useful information from large database which are stored already. A well known technique is association rule mining to finding the related information in the large database. In this paper we provide new improved Apriori algorithm of association rule mining which remove the problem of classical Apriori algorithm

REFERENCES

[1] Agrawal; Rakesh;, “Fast Algorithms for Mining Association Rules in Large Databases”, Proceedings of the ACM SIGMOD International Conference Management of Data, Washington, 1993, pp.207-216.

[2] AgrawaI.R; Imielinski.T; Swami.A; Mining Association rules between Sets of Items in large Databases [Cl In Proceedings of the ACM-SIGMOD Conference on Management of Data,1993:207-2l6

[3] Sheng Chai, Jia Yang; Yang Cheng;,” The Research of Improved Apriori Algorithm for Mining Association Rules,” Service System and Service Management, 2007 International Conference on, vol., no.,pp.1-4, 9-11 June 2007

[4] Wanjun Yu; Xiaochun Wang; Erkang Wang; Bowen Chen;, “The research of improved apriori algorithm for mining association rules,” Communication Technology, 2008. ICCT 2008 11th IEEE International Conference on, vol., no., pp.513-516, 10-12 Nov. 2008.

[5] Yubo Jia; Guanghu Xia; Hongdan Fan; Qian Zhang; Xu Li;,”An Improved Apriori Alogirhm Based on Association Analysis,” Third International conference on networking and distributed computing, 2012,pp.208-211.

[6] Rui Chang; Zhiyi Liu; , "An improved apriori algorithm," Electronics and Optoelectronics (ICEOE), 2011 International Conference on , vol.1, no., pp.V1-476-V1-478, 29-31 July 2011

[7] Sheila A. Abaya;,” Association rule mining based on apriori algorithm in minimizing candidate generation,” International journal of scientific and engineering research volumev3, issue 7, july 2012

[8] Jaishree Singh; Hari Ram; Dr. J.S.Sodhi;, “Improving Efficiency of apriori algorithm using transaction reduction,” International journal of scientific and research publication, volume 3, issue 1, January 2013.

[9] Libing Wu, Kui Gong, Fuliang Guo, Xiaolua Ge, Yilei Shan,” Research on improving apriori algorithm based on Interested Table,” pp 422-426, IEEE 2010.

[10] Yanfei Zhou; Wanggen Wan; Junwei Liu; Long Cai;,” Mining association rule based on an improved apriori algorithm,” ICALIP 2010.

[11] Chen, M. S.; Han, J.; and Yu, P.S. Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6): 866-883, 1996.

[12] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, *Advances in Knowledge Discovery and Data Mining,* AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34