A Survey on Association Rule Mining Algorithms

Yamee Patel[1], Maitrey Patel[2]
[2] Assistant Prof,
[1, 2] G.M.F.E, Himatnagar, Gujarat

Abstract- *In this paper, a review of four different association rule mining algorithms Apriori, AprioriTid, Apriori hybrid , I2:Apriori and Fp-growth algorithms and their drawbacks which would be helpful to find new solution for the Problems found in these algorithms and also presents a comparison between different association mining algorithms. Association rule mining is the one of the most important technique of the data mining. Its aim is to extract interesting correlations, frequent patterns and association among set of items in the transaction database.*

Keywords- Data mining, Association rule algorithms, Apriori, AprioriTid,Apriori hybrid,I2:Apriori and Fp-growth algorithms

I. INTRODUCTION

Data mining is the process of extracting useful information from large amount of data. In Data mining, user can analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. It is also known as "Knowledge mining from data". Data mining techniques are the result of a long process of research and product development. The main types of task performed by DM techniques are Classification, Dependence Modeling, Clustering, Regression, Prediction and Association [4].

Figure 1 show Knowledge Discovery in Database processes where it takes data from various repositories like data warehouse, database, information repositories, relational database etc. It performs various operations like data cleaning, integration, transformation etc. and produces useful information from that [1].

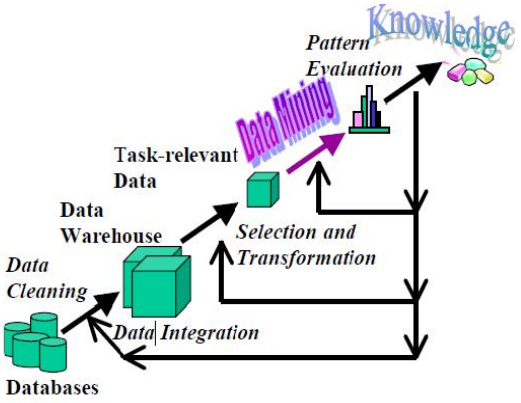


Figure 1: Data mining steps in the process of knowledge discovery

1.1 DATA MINING TECHNIQUES

- **Characterization:** Summarization of general features of objects in a target class, data relevant to a user-specified class.
- **Discrimination:** The comparison of the general features of objects between two classes referred to as the target class and the contrasting class; final result include comparative measures.
- **Association analysis:** study & prediction related to the *co-occurrence* of subset: "How Subset influences the presence of another subset."
- **Classification:** Classification analysis is the organization of data in given classes / supervised classification, the classification uses given class labels to order the objects in the data collection.
- **Prediction:** Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. Prediction is to forecast of missing numerical values, or increase/ decrease trends in time related data.
- **Clustering:** Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. Clustering approaches all based on the principle of "maximizing the intra-class similarity and inter-class dissimilarity." Outlier analysis: Outliers are data elements that cannot be grouped in a given class or cluster.

II. ASSOCIATION RULE MINING:

In data mining, Association rule mining is a popular and well researched method for discovering interesting relations between variables in large databases. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence. Association rule mining is the process of two steps. In first step frequent itemsets are generated and in second step rules are discovered.

There are two association rules mentioned in Example, The first one states that when peanut butter is purchased, bread is purchased 30% of the time. The second one states that 40% of the time when peanut butter is purchased so is jelly. Association rules are often used by retail stores to analyze market basket transactions. The task of mining association

rules over market basket data is considered a core knowledge discovery activity.

Association rule mining provides a useful mechanism for discovering correlations among items belonging to customer transactions in a market basket database.

Association rules are also used for other applications such as prediction of failure in telecommunications networks by identifying what events occur before a failure.

Let D be the database of transactions and J = $\{J_1...J_n\}$ be the set of items. A transaction
T includes one or more items in J (i.e., T⊆ J). An association rule has the form X⇒Y, where X and Y are non-empty sets of items (i.e. X ⊆J, Y⊆ J) such that X ∩ Y = Ø.

- Support and Confidence are two important measures of association rules.
- Support: If there are two items then support is defined as the ratio of occurrence of that two items and total number of transactions.
  For rule P ⇒ Q, Support = frq (P,Q) / N
- Confidence: The possibility of seeing the rule's consequent under the condition that the transactions also contain the antecedent is called confidence.

Confidence is defined as: For rule X ⇒ Y,

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$

For Association rule mining, many different algorithms are used like Apriori, FP-Growth, etc. Many existing algorithms scan database many times, so the I/O overhead and computational complexity becomes very high. They cannot meet the requirements of large-scale database mining. Genetic algorithm is based on biological mechanism. It works in iterative manner. Genetic algorithm is an optimization technique. To overcome the problem of Association rule mining algorithms, We can use Genetic algorithm.

III. APRIORI ALGORITHM

Apriori Algorithm is used to mine all frequent itemsets in database. It uses level wise and breadth first search to find frequent itemsets where k-itemsets are used to generate k+1- itemsets. In the first scan of dataset frequent-1-itemset (L1) is determined from candidate-1-itemset (C1) based on their frequency. Then this frequent-1-itemset (L1) is used to form candidate-2-itemset (C2) and then this candidate-2-itemset is used to determine frequent-2-itemset (L2) and so on until there are not any more k-itemsets[10].

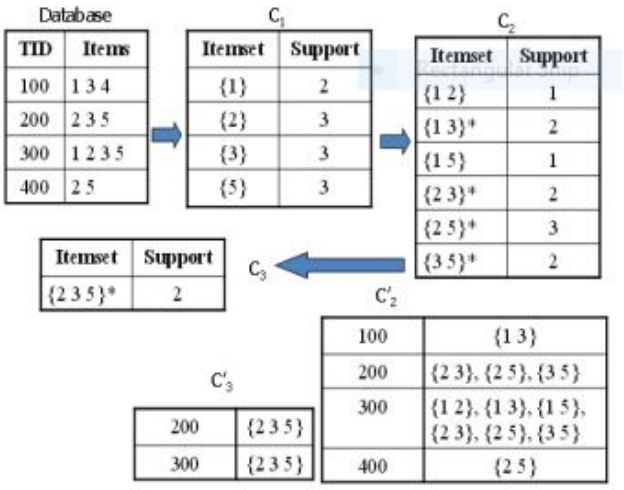
Apriori Algorithm

1. Initially Ck: Candidate itemset of size k, Lk : Frequent itemset of size k.
2. Procedure APRIORI ALGORITHM
3. for (k = 1; Lk ! = ; k++) do begin
a) Ck+1 = Candidates generated from Lk;
b) Prune (Ck+1)
c) for each transaction t in database do increment the count of all candidates in (Ck+1) that are contained in t
d) (Lk+1) = candidates in (Ck+1) with min-support
4. end
5. return UkLk
6. end procedure

There are two drawbacks of the Apriori algorithm. First is the complex candidate generation process which uses most of the time, space and memory. Another drawback is it requires multiple scans of the database.

III. APRIORITID ALGORITHM

In this algorithm [5], database is not used for counting the support of candidate itemsets after the first pass. The process of candidate itemset generation is same like the Apriori algorithm. Another set C' is generated of which each member has the TID of each transaction and the large itemsets present in this transaction. The set generated i.e. C' is used to count the support of each candidate itemset. The advantage of this algorithm is that, in the later passes the performance of Aprioritid is better thanApriori.



1. Aprioritid example

IV. APRIORIHYBRID ALGORITHM

As Apriori does better than Aprioritid in the earlier passes and Aprioritid does better than Apriori in the later passes. A new algorithm [5] is designed that is Apriorihybrid which uses features of both the above algorithms. It uses Apriori algorithm in earlier passes and Aprioritid algorithm in later passes.

DRAWBACKS

a) An extra cost is incurred when shifting from Apriori to AprioriTid.
b) Suppose at the end of K th pass we decide to switch from Apriori to AprioriTid. Then in the (k+1) pass, after having generated the candidate sets we also have to add the Tids to C'k+1.

VII. I2APRIORI ALGORITHM

The main idea of this method is to reduce the number of transaction scan. To do so 3 methods are included: Based on Support, InfrequentCount, 2-way searching that will help to reduces CPU computation time by reducing transaction scan. The Concept infrequent count is based on minimum threshold support and 2- way searching to reduce execution time during scanning of transaction. It also reduces the time by traversing the dataset from both sides. When the minimum threshold support is low then based on support the scan is reduced. When the minimum threshold is high then scans is reduced based on InfrequentSupport[7].

I2Apriori Algorithm :

1. Initially Ck: Candidate itemset of size k, Lk : Frequent itemset of size k.
2. Procedure apriori algorithm
3. InfrequentSupport = Total transaction - min-sup+1;
4. for (k = 1; Lk ! = ; k++) do begin
a) Ck+1 = Candidates generated from Lk;
b) Prune (Ck+1)
c) Scan transaction top to bottom and bottom to top increment the support count and InfrequentCount of all candidates in (Ck+1)
d) Stop scanning if ((supportcount=minimum threshold)or (InfrequentCount = Total transaction – (min-sup) +1)
e) Declare the items as frequent or infrequent based on supportcount or InfrequentSupport respectively
f) (Lk+1) = candidates in (Ck+1)
4. end
5. return UkLk
6. end procedure

VIII . FP-GROWTH ALGORITHM

To break the two drawbacks [8] of Apriori algorithm, FP-growth algorithm is used. FP-growth requires constructing FP-tree. For that, it requires two passes. FPgrowth uses divide and conquer strategy. It requires two scans on the database. It first computes a list of frequent items sorted by frequency in descending order (F-List) and during its first database scan. In the second scan, the database is compressed into a FP-tree [9]. This algorithm performs mining on FP-tree recursively. There is a problem of finding frequent itemsets which is converted to searching and constructing trees recursively. The frequent itemsets are generated with only two passes over the database and without any candidate generation process. There are two sub processes of frequent patterns generation process which includes: construction of the FP-tree, and generation of the frequent patterns from the FP-tree.
FP-tree is constructed over the data-set using 2 passes are as follows:
Pass 1:

1) Scan the data and find support for each item.
2) Discard infrequent items.
3) Sort frequent items in descending order which is based on their support. By using this order we can build FP-tree, so that common prefixes can be shared.

Characteristics	Apriori	Aprioritid	Apriori hybrid	I2:Apriori	FP-growth
Data support	Limited	large	Very Large	Very Large	Very Large
Speed in initial phase	High	Slow	High	High	High
Speed in later phase	Slow	High	High	High	High
Execution time	More time	More time	Comparitively slow	less time	less time

Table 1. Comparison of Association Rule Mining Algorithms

IX. CONCLUSION

There are various association rule mining algorithms. In this paper we have discussed six association rule mining algorithms: Apriori, Aprioritid, Apriorihybrid, I2 Apriori, FP-growth. Comparison is done based on the above performance criteria. Each algorithm has some advantages and disadvantages. From the above comparison we can conclude that, FP-growth performs better than all other algorithms discussed here.

REFERENCES

[1] Trupti A. Kumbhare , Prof. Santosh V. Chobe, "An Overview of Association Rule Mining Algorithms" IJCSIT,2014.

[2] Jyoti Arora1 , Nidhi Bhalla , Sanjeev Rao, "A REVIEW ON ASSOCIATION RULE MINING ALGORITHMS",IJIRCCE, Vol. 1, Issue 5, July 2013.
[3] K.Saravana Kumar, R.Manicka Chezian, "A Survey on Association Rule Mining using Apriori Algorithm", IJCA ,Volume 45– No.5, May 2012.
[4] Jeetesh Kumar Jain, Nirupama Tiwari, Manoj Ramaiya, "A Survey: On Association Rule Mining", IJERA, Vol. 3, Issue 1, January -February 2013.
[5] Manisha Girotra, Kanika Nagpal Saloni inocha Neha Sharma Comparative Survey on Association Rule Mining Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 84 – No 10, December 2013
[6] Mihir R. Patel, Dipti P. Rana, Rupa G. Mehta, "FApriori: A Modified Apriori Algorithm Based on Checkpoint", IEEE, 2013.
[7] Shyam Kumar Singh, Preetham Kumar, "I2Apriori: An Improved Apriori Algorithm based on Infrequent Count", IEEE, 2016.
[8] Sotiris Kotsiantis, Dimitris Kanellopoulos, AssociationRules Mining: A Recent Overview, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82
[9] Gagandeep Kaur, Shruti Aggarwal , Performance Analysis of Association Rule Mining Algorithms, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013, ISSN: 2277 128X
[10] J. Han, M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, San Francisco, USA, 2001, ISBN 1558604898.