The Survey on Rare Association Rule Mining

Isha u. Acharya[1], Mr. Ankur Shah[2],
[1]Student, [2]Ass.Prof,
[1, 2] DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,
[1,2] *PARUL INSTITUTE OF TECHNOLOGY*, Vadodara, India.
[1]ishaacharya1@gmail.com, [2]ankur11586@gmail.com

Abstract - To Analyze, manage and make a decision of such type of huge amount of data we need techniques called Data mining. Association rule mining is the most important technique in the field of data mining. The main task of association rule mining is to mine association rules by using minimum support thresholds decided by the user, to find the frequent patterns. Association rule mining approach is suffered from "rare item problem". A number of methods and techniques have been developed for rare association rule mining in data mining like Apriori-algorithm, MSapriori algorithm, FP-growth, RSAA algorithm used for retrieving mining frequent patterns. While MIS-tree is widely crucial information about frequent pattern & finds different frequent item sets. Moreover, discuss CP-tree is efficient for frequent pattern mining & interactive and incremental mining.

Index Terms - Data mining, association rule mining, rare item sets, frequent pattern, FP-growth, MIS-tree, CP-tree.

I. INTRODUCTION

There are different data available in the different formats so that the proper action to be taken. Not only to analyzed those data but also take a good decision and maintain the data . As and when the customer will required t h e data should be retrieved from the database and make the better decision. This technique is actually we called as a data mining or Knowledge H u b o r KDD (Knowledge Discovery Process). The perception of *"**we are data rich but information poor".*** With the enormous amount of data stored in files, databases, and other repositories in data warehouse and for the extraction of interesting knowledge that could help in decision-making. The only answer t o all above is **'Data Mining'**.
So , Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses.

Association rule mining is the efficient method which is used in finding the association rules. These association rules describe the associations between the attribute values of any item set. They can be found by means of various methods among which support and confidence & it will be considered as the optimized methods in finding them.
The key to find the association rules is to find all the frequent item sets present in the given transactional record by means of the minimum support threshold. An association rule is best expressed by means of the expression X -> Y. here, X is called as antecedent and Y is called as the consequent.

Support &Confidence

Support of an association rule is defined as the percentage/fraction of records that contain X ,Y to the total number of records in the database. Support(s) is calculated by the following formula & Confidence is another approach for finding the association rules. Confidence of an association rule is defined as the fraction of the number of transactions that contain X Y to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule X=>Y can be generated.

$$Support = \frac{(X \cup Y).count}{N}$$

&&

$$Confidence = \frac{(X \cup Y).count}{X.count}$$

Rare association rules can provide useful knowledge about those relatively infrequent or rare item sets. IN "Rare Item Problem", rare association rules are usually required to satisfy a user specified minimum support and a user specified minimum confidence at the same time. For using this main goal is to generate the rare rules, which might be given valuable information. Rare Association rule is an association rule consisting of rare items. MIS-tree like called multiple item support considers.

FP-growth like approach for utilizes the prior knowledge provided by the user at the time of input and discovers frequent patterns with a single scan on the transactional dataset. We are presented a novel tree structure, called CP-tree (Compact-pattern tree), CP-tree capture database information with one scan (Insertion phase) and provided the same mining performance as the FP-growth method (Restructuring phase) by dynamic tree restructuring process's-tree can give functionalities for interactive and incremental mining with single database scan with our CP-tree outperforms in terms of both execution time and memory requirements. Hence, we are going to

 Issues in Finding Rare Association Rules

A. Minimum Support Threshold

Many algorithms such as Apriori, FP-tree etc. Use this minimum support threshold in finding the frequent item sets. This threshold value is pre-set by the users. This value is set by user only. When user set high threshold value any infrequent item sets will lost. And if it is set low, many infrequent item sets will come into consideration.

Due to this problem an optimized decision cannot be taken. So threshold should be set very precisely.

B. Multiple Scans across the Transactional Database

While finding any frequent item sets, we have to scan whole database many times.

This multiple scan will lead to, following problems:

i) Wastage of time, because searching entire database for any item takes lot of time.

ii) Wastage of space, because lot of memory is needed.

II. LITERATURE SURVEY

A. Apriori Algorithm:

Apriori algorithm was first proposed by Agrawal, in Apriori is more efficient during the candidate generation process. It uses a breadth-first search strategy to count the support of item sets and uses a candidate generation function which exploits the downward closure property of support. Apriori uses pruning techniques to avoid measuring certain item sets, while guaranteeing completeness. The Apriori algorithm is based on the Apriori principle, which says that the item set X' containing item set X is never large if item set X is not large. Based on this principle, the Apriori algorithm generates a set of candidate large item sets whose lengths are (k+1) from the large k item sets (for $k \geq 1$) and eliminates those candidates, which contain not large subset. Then, for the rest candidates, only those with support over minimum support threshold are taken to be large (k+1)-item sets. The Apriori generate item sets by using only the large item sets found in the previous pass, without considering the transactions. The algorithm uses Apriori principle to generate candidate k-item sets from frequent (k-1)-item sets, and prunes candidate item sets. Through the support counting, get candidate k-item sets. Then the candidate k-item sets generate frequent (k-1)-item sets, so back and forth, until the frequent item sets cannot be produced.

B. MSApriori Algorithm:

Association rule mining is an important model in data mining. Its mining algorithms discover all item associations (or rules) in the data that satisfy the user-specified minimum support (min_sup) and minimum confidence (min_conf) constraints. 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 minsup is set too high, those rules that involve rare items will not be found. To find rules that involve both frequent and rare items, minus has to be set very low. This may cause combinatorial explosion because those frequent items will be associated with one another in all possible ways. This dilemma is called the rare item problem. The MSapriori Technique allows the user to specify multiple minimum supports to reflect the natures of the items and their varied frequencies in the database. In this model, the minimum support of a rule is expressed in terms of 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 specified by the user. By providing different MIS values for different items, the user effectively expresses different support requirements for different rules.

C. FP-Growth Algorithm:

FP-tree is a frequent pattern tree. An efficient mining method of frequent patterns in large

Database: using a highly compact FP-tree, divide-and-conquer method in nature. Formally, Scan the DB only twice and twice only.

FP-tree is a tree structure defined below:

1. One root labelled as "null", a set of *item prefix sub-trees* as the children of the root, and a *Frequent-item header table*.

2. Each node in *the item prefix sub-trees* has three fields:

– Item-name: register which item this node represents,

– Count, the number of transactions represented by the portion of the path reaching this node,

– Node-link that links to the next node in the FP-tree carrying the same item-name, or Null if there is none.

3. Each entry in the *frequent-item header table* has two fields,

– Item-name

– Head of node-link those points to the first node in the FP-tree carrying the item- name.

D. RSAA algorithms:

The Relative Support Apriori Algorithm (RSAA) generates rules which involve significant rare item sets. The main objective of this algorithm is to increase the support threshold values for the items having lower frequency and decrease the support threshold for items having higher frequency of occurrences. Like Apriori and MSapriori, RSAA is exhaustive in its generation of rules, so it spends a significant amount of time looking for the rules which are not rare. If the minimum permissible relative support count is set close to zero, then RSAA takes a similar amount of time to that taken by Apriori to generate low support rules. To generate candidate item sets in RSAA, we should be able to construct the candidate item set that contains rare data. The set of candidate item sets in RSAA

consists of two groups. One group includes the frequent items that satisfy the first support, and the other group includes the rare items that do not satisfy the first support count but satisfies the second support count. The former group is the same set as the one computed by Apriori. RSAA is exhaustive in its generation of rules, so it spends significant amount of time looking for rules which are not sporadic. However, it uses two thresholds, one is Minsup and the other is Minsup.

E. MIS algorithms:

The MSapriori algorithm can find rare item rules without producing a huge number of meaningless rules. In this model; the definition of the minimum support is changed. Each item in the database can have its minsup, which is expressed in terms of minimum item support (MIS). In other words, users can specify different MIS values for different items. By assigning different MIS values to different items, we can reflect the natures of the items and their varied frequencies in the database. a new tree structure, named the MIS-tree, is proposed for mining frequent pattern with multiple MS. It is an extended version of the FP-tree structure.

F. CP-Tree algorithms:

CP-Tree that capture one scan for database information & same mining performance as FP-growth method by Restructure process. It's functionality for interactive and incremental mining with single database scan's tree constructs a compact prefix-tree with one scan & performance as FP-growth technique by efficient tree restructuring process's tree increase outperforms on execution time and memory requirement.

Construct the CP-tree based on following Steps are:

1] Insertion phase: Staring this phases into it inserts transection into CP-tree according to sort order of I-list and update freq_count of respectively items in I-list. process into that item having higher count value are rearranging at upper most portion of the tree.

2] Restructuring phase: Finishing with this phase into that rearranges the I-list before frequency descending order of items and restructure the tree node in to new I-list & the end of database. Lexicographical Can-Tree containing more nodes with respect to CP-tree for same dataset. Any value of support threshold ∂ by starting from the bottom most item in I-list having count value $\geq \partial$ .Using tree Restructuring mechanism, Existing Path Adjusting Method (PAM) & Branch sort method (BSM) proposed based on the value of DD (degree of displacement). It is easy to maintain feature & structuring Cost & their full database information in highly Compact fashion facilities providing in interactive, incremental and stream data.

III. CONCLUSIONS

Mining frequent item sets for the association rule mining from the large transactional database is a very crucial task. There are many approaches that have been discussed; nearly all of the previous studies were using Apriority approach, According to our observations, the performances of the algorithms are strongly depends on the support levels. For discover rare association rules, Maximum constraint model using raises performance problem shown by conducting experiments on synthetic dataset. As a part of future work, we are going to analyses the behavior of CP-tree and MIS-tree for various interesting measures on mining rare association rules.

IV. REFERENCES

1) Neelamadhab Padhy, Dr.Pragnyaben Mishra, "The Survey of Data Mining Applications and Feature Scope", DOI: 10.5121, IJCSEIT-2012.Vol.2, No.3.
2) R. Uday Kiran, P. Krishna Reddy, IN: DASFAA-2010 , "Mining Rare Association Rules In the Datasets with Widely Varying Items' Frequencies".
3) Weimin Ouyang , Qinhua Huang ,"Mining Direct and Indirect Association Patterns With Multiple Minimum Supports",H:IEEE 2010.
4) Azadeh Soltani and M. R. Akbarzadeh T.,"Confabulation-Inspired Association Rule Mining for Rare and Frequent Item sets ",IN:IEEE TRANSACTION ON NETWORKS AND LEARNING SYSTEM-2014.
5) M. Sinthuja, S. Sheeba Rachel and G. Janani, "MIS-Tree Algorithm for Mining Association Rules with Multiple Minimum Supports"-Bonfring International Journal of Data Mining, Vol-1, December-2011.
6) Sandeep Singh Rawat, Lakshmi Rajamani,"Probability Apriori based Approach to Mine Rare Association Rule."IN:IEEE-2011,3rd Conference on Data Mining and Organization(DMO).
7) R. Uday Kiran, P. Krishna Reddy, IN:IEEE-CIDM-2009 , "An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules".
8) N. Hoque, B. Nath, D. K Bhattacharyya, "A New Approch on Rare Association Rule Mining "International Journal of Computer Application(0975-8887)vol-53,no.-3.September-2010.
9) Syed Khairuzzaman Tanbeer, Chowdhury Farhan Ahmed, Young -Koo Lee, "CP-Tree: A Tree Structure for Single-Pass Frequent Pattern Mining"-Springer-2008.
10) R. Uday Kiran and Polepalli Krishna Reddy, "An Efficient Approach to Mine Rare Association Rules Using Maximum Items' Support Constraints"-Springer-Verlag Berlin Heidelberg-2012.
11) B. Liu, W. Hsu, and Y. Ma. Mining association rules with multiple minimum supports. Pages 337–341. ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations, 1999.
12) B. Nath, D. K. Bhattacharyya, and A. Gosh. Faster generation of association rules,volume 1, pages 267–279. IJITKM, 2008.
13) B. Nath and A. Ghosh. Multi-objective rule mining using genetic algorithm. pages 123– 133. Information Science 163, 2004.

14) A. Savesere, E. Omiecinski, and S. Navathe. An effective algorithm for mining association rules in large database. pages 432–443. In proceedings of International Conference on VLDB95, 1995.