Summary By Deeksha Ramakrishna

Research Paper 1 - Mining Rare Patterns by Using Automated Threshold Support

Detecting those rare patterns which are small in number among huge pair of data is the most challenging part.

By giving various ranges to the patterns, and by utilizing the support threshold, patterns could be well classified easily. If the max threshold is defined, we are able to obtain the rare patterns or the patterns which occur less frequently. High utility is an extension to the problem caused by frequent pattern mining.

The objective of association rule is to obtain the correlation between items. Here complexity is major challenge. It is improved by utilizing a minimum support threshold.

Infrequent Weighted Itemset (IWI) is the idea where each and every item is assigned particular weight which is beneficial to mine them accordingly [8][9].

Problem Statement: In some situations, e.g when the requirement is always to minimize a specific cost function, discovering rare data correlations is more interesting than mining frequent ones. Considers the two drawbacks of traditional sequential pattern ,i.e consideration of only frequent sequence and other is restricted to the static environment of data set only. This study emphasizes an approach to obtain the infrequent itemsets involving rare items by setting the support thresholds automatically by using logical itemset mining.

Proposed System: proposed method combines Apri-ori and MS-Apriori to mine logically rare itemsets among huge amount of data. Three different groups namely Most interesting group(MIG), Rare interesting group(RIG) and some what interesting groups (SIG)are minned to obtain rare item-sets.

Unique items are identified from these transaction,support count is calculated by using the standard formula. High occurrence noise is removed and is further processed for preserving low occurrence signals. From this, only rare items are inserted after finding rare transaction. Hence MRCP tree is generated and as a result, set of rare items are obtained.MRCP tree scans the data once which reduces the space as well as the time complexity of the system.

The system will cover all the drawbacks of Fp Growth algorithm,sliding window,sequential algorithm.etc. One most important benefit of this system is that it will decide the value of the threshold on its own and hence the result obtained will be accurate accordingly.

Research paper 2 - Rare Association Rules Mining of Diabetic Complications Based on Improved Rarity Algorithm

It is more valuable to study the rare pattern mining, because it tends to find some unknown and unexpected associations.

Proposed System: So based on the Rarity algorithm, this paper presents an improved top-down approach to efficiently mine all rare itemsets and their association rules, which uses the graph structure to indicate all combinations of existing items in the database, defines a pattern matrix to record all itemsets and the support_count, and combines the hash table to accelerate support calculation to quickly find all rare itemsets, and then generate all patterns to choose useful rules according to their interesting rate. In the experiment, this paper uses the real diabetic clinical data to verify this improved approach and mines some useful rules among the diabetic complications. Moreover, compared with the two methods mentioned above, this method decreases much time and space complexity in the association rules mining.

Drawback of the existing Apriori and FP-Growth Algorithm - the Apriori algorithm that uses a bottom-up strategy to search frequent itemsets in the database, which is easy to understand, but generates many candidate subsets and needs to repeatedly scan the database.

FP- growth algorithm based on the FP-Tree, a novel frequent pattern tree structure, to solve above problems. It only scans the database twice and avoids costly generating lots of candidate sets, however, if the database is large, the tree structure will be too big to store in the memory.

This paper presents an efficient method of rare pattern mining that can be divided into two steps, i.e., find all rare itemsets based on the improved Rarity, and rely on these rare itemsets to mine their association rules.

Experiment Results: The experiment totally generates 180 rare association rules, and via comparing their interesting rate, this paper choose the top 5 rules, it can be concluded that if a diabetic patient with complications of retinopathy and diabetic foot, this patient is very likely to have complications of renal, neurological, and cardiovascular, and if a diabetic patient has cardiovascular and diabetic foot, this patient may have retinopathy, renal, and neurological at same time. And the first rule is more universal and useful, according to the interesting rate.

Comparing with Arima, the method in this paper has a significant improvement in the space and time complexity. And comparing with Rarity, this method saves much memory, and solves the biggest memory problem of Rarity algorithm. Moreover, if adding or deleting one type of item, the cost of Rarity algorithm is huge because of rebuilding the tree of full combination, but it's very easy for our method to modify records in the directed graph and hash table.

Conclusion and future work: So this paper proposes an improved method based on Rarity algorithm of mining rare association rules. This method uses paths in a directed graph to

represent every itemsets in the database, generates a pattern matrix, and stores each metavector and its corresponding support_count in a hash table.

First, continue exploring how to decrease the mining time to improve this method. Second, adjust more different hash functions to optimize the space utilization.