PROJECT REPORT (PROJECT 2)

Description: In project2, we are focusing on implementing Apriori algorithm. We have to work on two datasets and compare effectiveness and efficiency of the algorithm

Apriori Algorithm:
Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

Datasets:
Balance Scale Data Set:
This data set was generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The attributes are the left weight, the left distance, the right weight, and the right distance. This dataset is well suited for implementing data mining algorithms on it. There are 625 instances, 4 attributes and no missing values in this dataset.

Contraceptive Method Choice Data Set
This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview. The problem is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics. There are 1473 attributes, 9 instances and no missing values in this dataset.

Analysis:
The Apriori Algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. The data sets, support and the total execution time will tell us the efficiency and effectiveness of the algorithm. The efficiency can be improved by various methods like Hash-based item set counting, Transaction reduction, Partitioning and Sampling.

Implementation:
We have implemented Apriori algorithm in Java. We have chosen java because of the run time constrains and better accessibility of object oriented concepts. As a part of implementation, we have read the datasets using "," as delimiter. Dataset file's path, support, confidence are passed through command prompt to the algorithm. Then we generate frequent item sets for the data set. Association rules for the data are also built based on the dataset.

Analysis of the algorithm on our datasets:
After performing the prior analysis on the two datasets, their where considerable changes in the result thus obtained. Execution time for the both the data sets varied and this shows that the time complexity varies as per the selected data set. These datasets have been examined clearly with the association rules and these association rules will allow us to examine the operation of the data based on the minimum support and confidence considered. Association rules are also generated whose confidence is greater than the minimum confidence.
For contraceptive method dataset:
In this dataset, Frequent4 item sets are generated. 20 Item sets candidates on the whole satisfying minimum support of 30 % are generated. The execution time for the dataset taken is 0.089 seconds. In this dataset, it explains the information regarding women and their health condition. This says that the algorithm is efficient on the dataset by executing it in 0.089 seconds. Extracting all the possible frequent item sets for a data set can be considered as effectiveness. Apriori algorithm has extracted 20 frequent patterns on contraceptive method dataset. Association rules are also generated whose confidence is greater than the minimum confidence.

For Balance scale dataset:
In this dataset, Frequent2 item sets are generated. 15 Item sets candidates on the whole satisfying minimum support of 30 % are generated. The execution time for the dataset taken is 0.013 seconds. In this dataset, it explains the information regarding women and their health condition. This says that the algorithm is efficient on the dataset by executing it in 0.013 seconds. 15 item sets are generated on Balance scale dataset. We can say that the algorithm.

By this we can say that the algorithm's performance is good on balance scale dataset. Effectiveness is more on contraceptive method dataset. However the performance of the algorithm depends on the support and confidence value we give, this algorithm is good in both effectiveness and efficiency.

In conclusion, Apriori will use the large datasets properly when compared to other algorithms. The algorithm can be easily parallelized and easy to implement. It will reduce the number of candidates being considered by only exploring the item sets whose support count is greater than the minimum support count. Confidence plays a very crucial role in creating association rules.