Data Mining and Data Warehousing
Frequent Pattern Mining-Frequent Pattern Mining Algorithms

Tasks

- Prove the anti-monotone property with an example.

- Given a census (For your convenience you can get them inside Self Assessment Quadrant) dataset, generate the frequent itemsets and extract the association rules using the Apriorialgorithm. Consider the $\text{minsup} = 0.3$ and $\text{min\_conf} = 0.6$ U3_M2_T2_Hints (For your convenience you can get them inside Self Assessment Quadrant)

- Calculate the complexity of the Apriori algorithm given $n$ transactions and a support count of $m$ where $n >> m$.

- State the major problems associated with the Apriori algorithm? U3_M2_T4_Hints (For your convenience you can get them inside Self Assessment Quadrant)

- The file Transactions (For your convenience you can get them inside Self Assessment Quadrant) contains the transaction database as an m x n matrix. Transaction 1 appears in row one. Columns are separated by a space and represent items. A 1 indicates that item is present in the transaction and a 0 indicates it is not. Using a minsupport of 20% (0.2) use the FP-Growth algorithm to determine the frequent itemsets.

- Most frequent pattern mining algorithms consider only distinct items in a transaction. However, multiple occurrences of an item in the same shopping basket, such as four cakes and three jugs of milk, can be important in transaction data analysis. How can one mine frequent itemsets efficiently considering multiple occurrences of items? Propose modifications to FP-Growth to adapt to such a situation.
Suppose you are given a set of transactions T from the FIMI itemset repository. A Maximal Frequent Itemset is defined as one none of whose supersets are frequent. Come up with an extension of the FP-Growth algorithm called FP-Max to effectively mine maximal frequent itemsets. Write down the pseudo code of this algorithm and highlight the important procedures involved. Similarly go through the document provided in the resources and understand the notion of Closed Frequent Itemsets. Once this is done repeat the above process and suggest an algorithm FP-Close which is a variant of the FP-Growth algorithm to mine the closed frequent itemsets only.

The general association rule mining algorithms we have looked at like Apriori and FP-Growth algorithms use only one support value throughout the dataset. However in certain real life datasets it is generally observed that data distribution can be highly skewed and normal ARM algorithms cannot mine all the association rules directly. So to tackle such situations multiple support association rule mining algorithms were developed. Now think of the real life scenarios around you and come up with the description of a scenario where multiple support association rule mining can be applied effectively.

Project

You are to assume the role of a data scientist of Walmart, which is a multi-national company for shopping and marketing. The company is planning to open a new market in Kansas City. They want you to analyze the purchasing trend of people of Kansas and help the company to successfully set the market according to the consumer needs of Kansas. This will tremendously help the company to make huge profits and perform risk analysis.

Given the data set of wholesale customers purchase Wholesale Customers Data, (For your convenience you can get them inside Self Assessment Quadrant)
your task is to extract items which are purchased together more frequently using any one of the frequent pattern mining algorithms covered in the module.