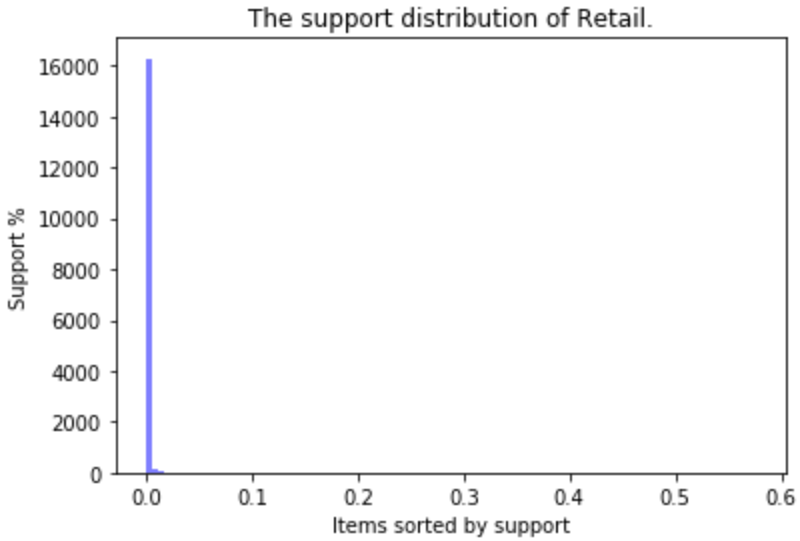
Data Mining

Assignment - 3

Given Algorithm solves the problem of mining association patterns from the data sets with skewed support distribution. Since setting the support threshold to high or low could cause problem in mining patterns. Some concepts like hconf or all confidence and cross support were introduced in the paper. The have formally described property of cross support which helps in efficiently eliminating patterns.

Dataset - Retail
I observed skewed nature of the support distribution in Retail Dataset

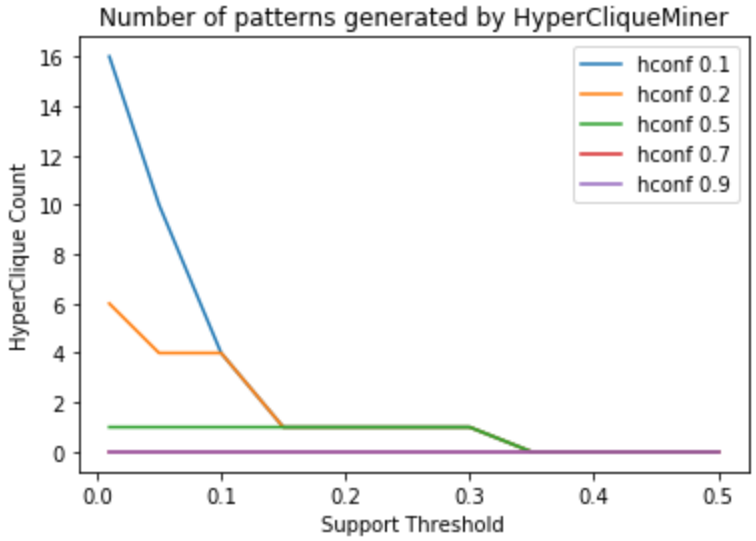


Algorithm Details:
To Find hyper clique patterns in the dataset, I used a library called apyori. Which already has some couple of functions already implemented, which makes it easier to implement the algorithm mentioned in the paper. Some of the Api functions used - create_next_candidates, calc_support, TransactionManager Object. Some of the helper functions I created like Prune A, B, GetSupport, Get C1 candidates.

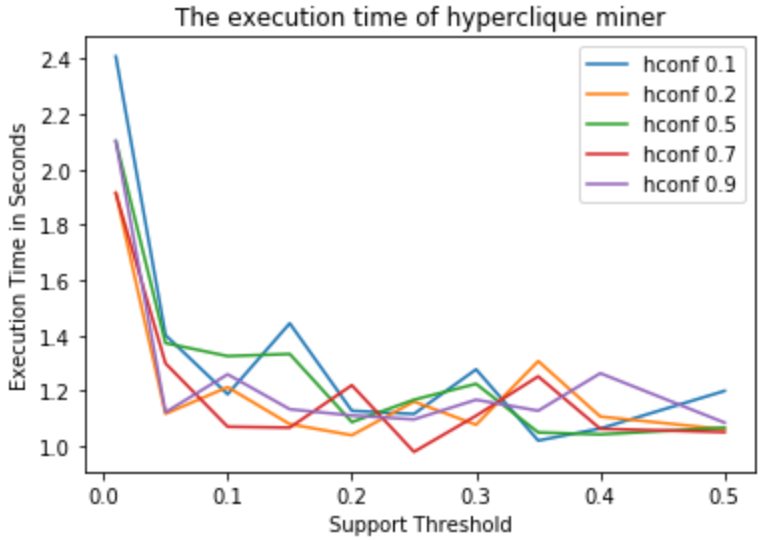
Initially the algorithm starts by calculating support of size 1 candidates and filtering and filtering the ones which has support greater than a threshold, which are then later used to create c2 candidates.
New candidate generation and Prune A is called Apriori Algorithms.
For Prune B step, if item set satisfy this condition then its pruned (hc > min(Arr)/max(Arr)).
Where Min and Max Arr are the minimum support of items in the itemset and maximum support of the items in the itemset.

Prune C simply remove the itemset based on threshold of min_supp and min_hc.

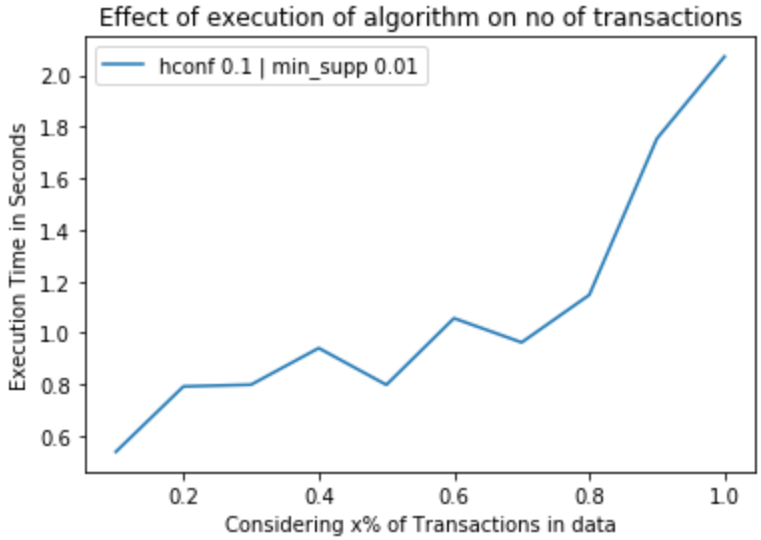
Graphs are mostly self explanatory



We can see that more no of HyperClique patterns are generated for Low support and hconfidence threshold. For hconf threshold >= 0.7 there are zero patterns generated. We can see a significant change in no of patterns for hconf threshold.



We can see that execution time of the algorithm is close for all hconf thresholds. Where as for lower threshold, algorithm required more time to generate patterns since more no of candidates are formed



This Graph represent the execution time on subsets of the dataset. We can see that more processing time is required for data with more no of transactions.