Assignment NO:-8

Title:-

Case study (Market Basket Analysis)

Problem statement:- A mall has no of items for sale. Build a required Database to develop BA & I tool for considering one aspect of growth to the business such as organization of products based on demand & patterns.

Input:- Transaction Database & minimum support.

Output:- Frequent item sets, Association Rules & graphical representation of rules as per confidence & & lift

Pre-Lab:- 1. Knowledge of R programming Language.
2. Concept & theory of Apriori algorithm.

Theory:- 1) Itemsets are used to explore K itemsets. First, it employes an iterative approach known as a level-wise search.
2-itemsets, which is then used to find L3, & so on, until no more frequent k-itemsets can be found.
To improve the efficient of the level-wise generation of frequency itemsets, an important property called Apriori property is used to reduce the search space.

Apriori property:-
All nonempty subsets of a frequent itemset must also be frequent.

1) The join step:- That is, members l_1 & l_2 are joined if $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \ldots \wedge (l_1[K-2] = l_2[K-2]) \wedge (l_1[K-1] < l_2[K-1])$. The condition $l_1[K-1] < l_2[K-1]$ simply ensures that no duplicates are generated. The resulting itemset formed by joining l_1 & l_2 is $\{l_1[1], l_1[2], \ldots, l_1[K-2], l_1[K-1], l_2[K-1]\}$

2) The prune step:- set C_K is a superset of L_K, then the candidate cannot be frequent either & so can be removed from C_K. This subset testing can be done quickly by maintaining a hash tree of all frequent itemset.

An Example of the use of The Apriori Algorithm:-

C_1 itemset	Support count.
{1}	6
{2}	7
{3}	6
{4}	2
{5}	2
{6}	1

The set of frequent 1-itemsets, L_1, consists of the candidate itemsets satisfying the minimum support count of 2. Thus all the candidates in C_1, except for {6}, are in L_1.

L_1 itemset	Support count.
{1}	6
{2}	7
{3}	6
{4}	2
{5}	2

To discove the set of frequent 2-itemsets, L_2, the algorithm joins L_1 with itself to generate a candidate set of 2-itemsets, C_2. Note that no candidates are removed from C_2 during the pruning step since each subset of the candidates is also frequent.

C_2 itemset
{1,2}
{1,3}
{1,4}
{1,5}
{2,3}
{2,4}
{2,5}
{3,4}
{3,5}
{4,5}

Next the transactions in D are scanned & the support count of each candidate itemset in C_2 is accumulated.

C_2 itemset	Support count
{1,2}	4
{1,3}	4
{1,4}	1
{1,5}	2
{2,3}	4
{2,4}	2
{2,5}	2
{3,4}	0
{3,5}	1
{4,5}	0

The set of frequent 2-itemsets, L_2, is the determined, consisting of those candidate 2-itemsets in C_2 having minimum support.

L_2 itemset	Support count
{1,2}	4
{1,3}	4
{1,5}	2
{2,3}	4
{2,4}	2
{2,5}	2

since {2,3} is a frequent itemset, we keep {1,2,3} in C_3.
since {2,5} is a frequent itemset, we keep {1,2,5} in C_3.
since {3,5} is not a frequent itemset, we remove {1,3,5} from C_3.
since {3,4} is not a frequent itemset, we remove {2,3,4} from C_3.
since {3,5} is not a frequent itemset, we remove {2,3,5} from C_3.
since {4,5} is not a frequent itemset, we remove {2,4,5} from C_3.
∴ after pruning C_3 given by:

C_3
itemset
{1,2,3}
{1,2,5}

The transactions in D are scanned to determine L_3, consisting of those candidates 3-itemsets in C_3 having at least minimum support.

C_3 itemset	Support Count
{1,2,3}	2
{1,2,5}	2

since both 3-itemsets in C_3 have the least minimum support, L_3 is therefore given by:

L_3 itemset	Support Count
{1,2,3}	2
{1,2,5}	2

Finally, L_3 is joined with itself to generate a candidate set of 4-itemsets, C_4. This results in a single itemset {1,2,3,5}. However this itemset is pruned since its subset {3,5} is not frequent. Thus, $C_4 = \phi$ & the algorithm terminates, having found all of the frequent itemsets.

Execution Guidlines:—

1) Install packages 'arules', 'arulesViz' from cran mirror through HTTP.

2) Use data set 'Groceries'.
3) Use apriori function in R to get item sets providing length of itemset & support.
4) Generate rules using apriori fun in R to get item sets providing length of item set & support.
5) Plot rules for given confidence.
6) Plot graph for visualizing the high lift rules.

Analysis :—

1) Observe the graphs for generated rules with different support confidence & lift.
2) Observe top rules & use this patterns for organizar of products.

Conclusion:—

Thus the Groceries dataset is used to generate rules and applied rules for organization of products based on demand & patterns frequent itemsets are found using apriori algorithm based on association rules data mining technique. Observations are recorded in terms of graph.