Aim: To implement Apriori Algorithm on large dataset using Open source tool WEKA.

Objective: To make students well versed with open source tool like WEKA to implement Apriori algorithm.

Theory:

- Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an itemset occurs in a transaction.
- A typical example is a Market Based Analysis. Market Based Analysis is one of the key techniques used by large relations to show associations between items.
- It allows retailers to identify relationships between the items that people buy together frequently.
- Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Support Count () – Frequency of occurrence of a itemset.
Here ({Milk, Bread, Diaper})= 2

Frequent Itemset – An itemset whose support is greater than or equal to minsup threshold.

Association Rule – An implication expression of the form X Y, where X and Y are any 2 itemsets.

Example: {Milk, Diaper}{Beer}

- WEKA contains an implementation of the Apriori algorithm. The algorithm works only with discrete data.
- It can identify statistical dependencies between groups of attributes.
- Apriori algorithm can compute all rules that have a given minimum support and exceed a given confidence.
- Clicking on the "Associate" tab will bring up the interface for the association rule algorithms.
- The Apriori algorithm which we will use is the default algorithm selected. However, in order to change the parameters for this run (e.g., support, confidence, etc.) we click on the text box immediately to the right of the "Choose" button. Note that this box, at any given time, shows the specific command line arguments that are to be used for the algorithm.
- WEKA allows the resulting rules to be sorted according to different metrics such as confidence, leverage, and lift. We can also change the default value of rules (10) to be 20; this indicates that the program will report no more than the top 20 rules. The upper bound for minimum support is set to 1.0 (100%) and the lower bound to 0.1 (10%).
- Apriori in WEKA starts with the upper bound support and incrementally decreases support (by delta increments which by default is set to 0.05 or 5%). The algorithm halts when either the specified number of rules are generated, or the lower bound for min. support is reached. Once the parameters have been set, the command line text box will show the new command line. We now click on start to run the program. This results in a set of rules. The panel on the left ("Result list") now shows an item indicating the algorithm that was run and the time of the run. You can perform multiple runs in the same session each time with different parameters. Each run will appear as an item in the Result list panel. Clicking on one of the results in this list will bring up the details of the run, including the discovered rules in the right panel. In addition, right-clicking on the result set allows us to save the result buffer into a separate file. Note that the rules were discovered based on the specified threshold values for support and lift. For each rule, the frequency counts for the LHS and RHS of each rule is given, as well as the values for confidence, lift, leverage, and conviction. In most cases, it is sufficient to focus on a combination of support, confidence, and either lift or leverage to quantitatively measure the "quality" of the rule. However, the real value of a rule, in terms of usefulness and action ability is subjective and depends heavily of the particular domain and business objectives.

OUTPUT:

Dataset :

@relation car

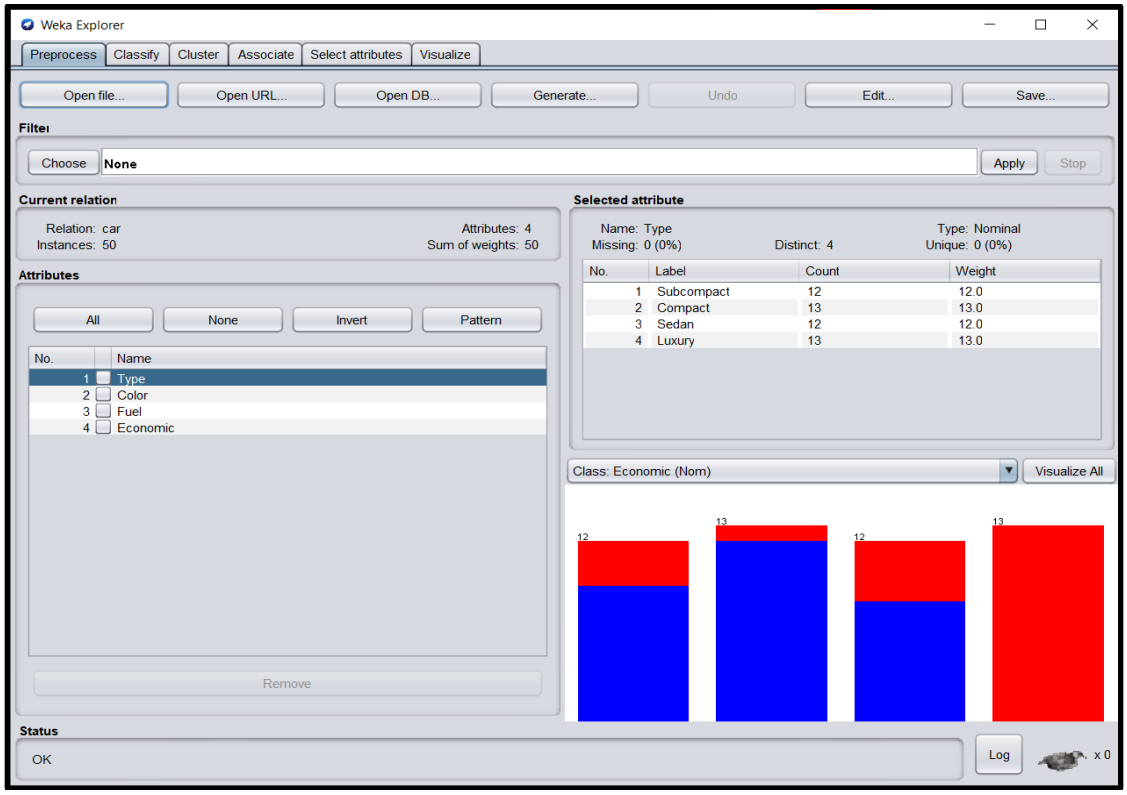
@attribute Type {Subcompact, Compact, Sedan, Luxury}
@attribute Color {Red,Silver,Black}
@attribute Fuel {Petrol,Diesel,CNG,Electric}
@attribute Economic {Yes,No}

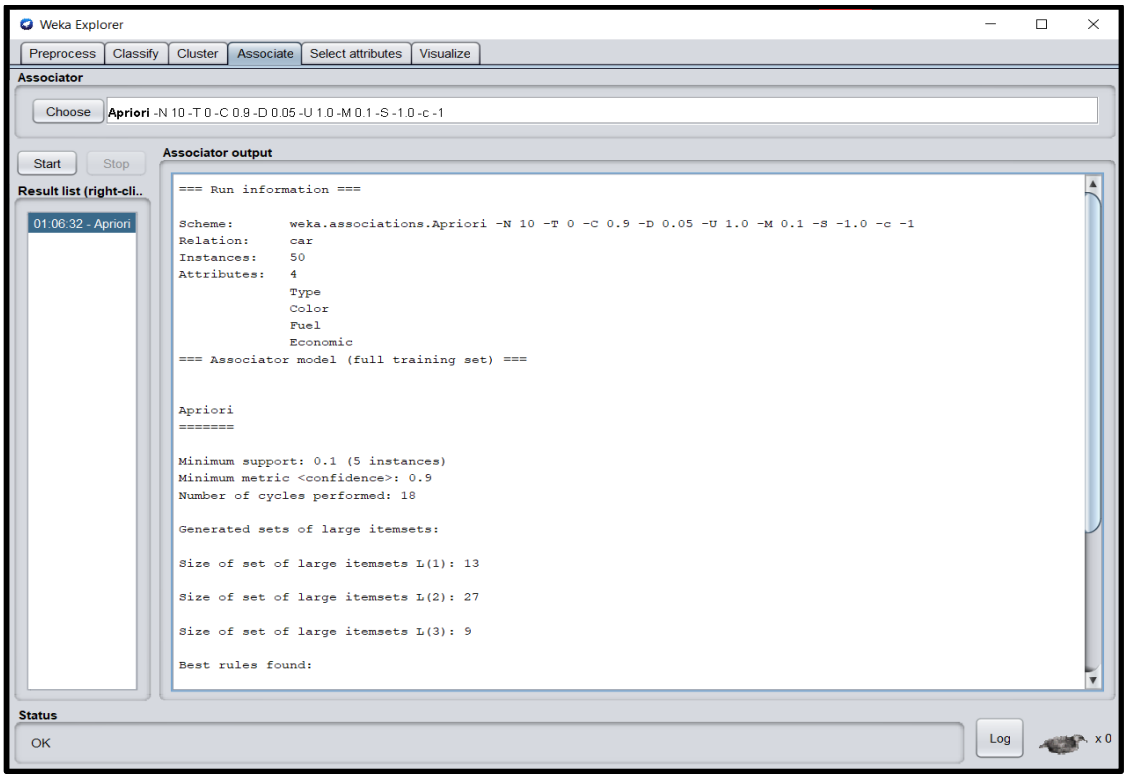
@data

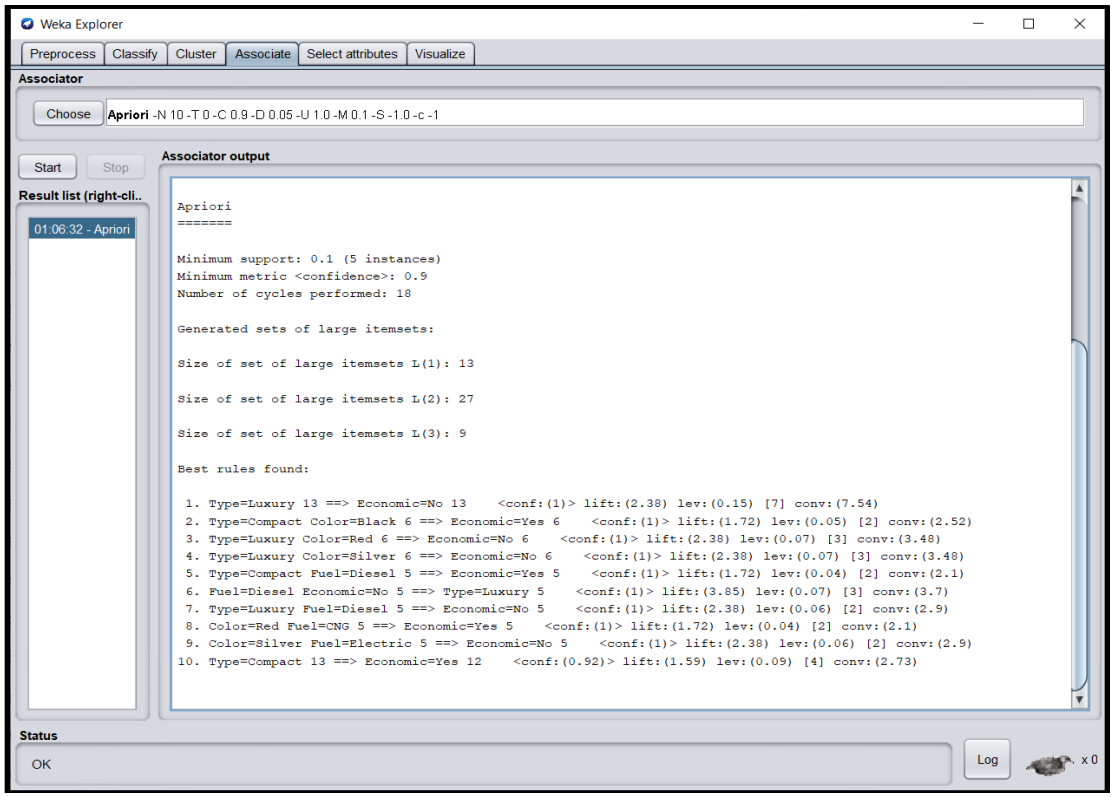
Subcompact	Red	Petrol	Yes
Compact	Black	Diesel	Yes
Compact	Silver	Petrol	Yes
Luxury	Red	Electric	No
Subcompact	Silver	CNG	Yes
Luxury	Red	Petrol	No
Sedan	Silver	Electric	No
Sedan	Black	Diesel	Yes
Subcompact	Black	Electric	Yes
Compact	Red	CNG	Yes
Sedan	Silver	Petrol	Yes
Luxury	Red	Diesel	No
Luxury	Silver	Electric	No
Sedan	Black	CNG	Yes
Compact	Black	Diesel	Yes
Compact	Red	Electric	No
Subcompact	Black	Petrol	Yes
Luxury	Silver	CNG	No
Sedan	Red	Diesel	Yes
Sedan	Silver	Electric	No
Subcompact	Red	CNG	Yes
Compact	Black	Petrol	Yes
Subcompact	Red	Petrol	Yes
Luxury	Silver	CNG	No
Sedan	Black	Diesel	Yes
Sedan	Red	Electric	No
Subcompact	Red	Diesel	Yes
Compact	Silver	Petrol	Yes
Luxury	Red	Diesel	No
Sedan	Black	Petrol	Yes

Subcompact	Black	Petrol	Yes
Subcompact	Red	Electric	No
Compact	Red	Petrol	Yes
Luxury	Silver	CNG	No
Compact	Black	Diesel	Yes
Luxury	Silver	Electric	No
Subcompact	Red	CNG	Yes
Luxury	Red	Diesel	No
Sedan	Black	Electric	No
Compact	Black	CNG	Yes
Compact	Red	CNG	Yes
Sedan	Silver	Petrol	Yes
Luxury	Black	Diesel	No
Subcompact	Silver	Electric	No
Sedan	Red	CNG	Yes
Compact	Black	Diesel	Yes
Luxury	Red	Petrol	No
Subcompact	Black	Electric	No
Luxury	Silver	Diesel	No
Compact	Red	Diesel	Yes

Weka Output




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Start
Stop
Associator output
Result list (right-cli..
01:06:32 - Apriori
=== Run information ===
Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: car
Instances: 50
Attributes: 4
Type
Color
Fuel
Economic
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.1 (5 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 13
Size of set of large itemsets L(2): 27
Size of set of large itemsets L(3): 9
Best rules found:
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Start
Stop
Associator output
Result list (right-cli..
01:06:32 - Apriori
Apriori
=======
Minimum support: 0.1 (5 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 13
Size of set of large itemsets L(2): 27
Size of set of large itemsets L(3): 9
Best rules found:
1. Type=Luxury 13 ==> Economic=No 13 <conf:(1)> lift:(2.38) lev:(0.15) [7] conv:(7.54)
2. Type=Compact Color=Black 6 ==> Economic=Yes 6 <conf:(1)> lift:(1.72) lev:(0.05) [2] conv:(2.52)
3. Type=Luxury Color=Red 6 ==> Economic=No 6 <conf:(1)> lift:(2.38) lev:(0.07) [3] conv:(3.48)
4. Type=Luxury Color=Silver 6 ==> Economic=No 6 <conf:(1)> lift:(2.38) lev:(0.07) [3] conv:(3.48)
5. Type=Compact Fuel=Diesel 5 ==> Economic=Yes 5 <conf:(1)> lift:(1.72) lev:(0.04) [2] conv:(2.1)
6. Fuel=Diesel Economic=No 5 ==> Type=Luxury 5 <conf:(1)> lift:(3.85) lev:(0.07) [3] conv:(3.7)
7. Type=Luxury Fuel=Diesel 5 ==> Economic=No 5 <conf:(1)> lift:(2.38) lev:(0.06) [2] conv:(2.9)
8. Color=Red Fuel=CNG 5 ==> Economic=Yes 5 <conf:(1)> lift:(1.72) lev:(0.04) [2] conv:(2.1)
9. Color=Silver Fuel=Electric 5 ==> Economic=No 5 <conf:(1)> lift:(2.38) lev:(0.06) [2] conv:(2.9)
10. Type=Compact 13 ==> Economic=Yes 12 <conf:(0.92)> lift:(1.59) lev:(0.09) [4] conv:(2.73)
Status
OK
Log
x 0

Conclusion:

From the given supermarket dataset we have successfully implemented an apriori algorithm on it using the weka tool. Apriori is an algorithm for frequent itemset mining and association rule learning over relational 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.