PRACTICAL 7

Aim: Implement Apriori Algorithm.

Theory:

In general association rule mining can be viewed as a two-step process:
1. Find all frequent itemsets: By definition each of these itemsets will occur at least as frequently as a predetermined minimum support count min sup .
2. Generate strong association rules from the frequent itemsets: By definition these rules must satisfy minimum support and minimum confidence

Let I={i1,i2,i3,…,in} be a set of n attributes called items and D={t1,t2,…,tn} be the set of transactions. It is called a database. Every transaction, ti in D has a unique transaction ID, and it consists of a subset of itemsets in I.
A rule can be defined as an implication, $X \longrightarrow Y$ where X and Y are subsets of $I(X,Y \subseteq I)$, and they have no element in common, i.e., $X \cap Y$. X and Y are the antecedent and the consequent of the rule, respectively.

Let's take an easy example from the supermarket sphere. The example that we are considering is quite small and in practical situations, datasets contain millions or billions of transactions. The set of itemsets, I ={Onion, Burger, Potato, Milk, Beer} and a database consisting of six transactions. Each transaction is a tuple of 0's and 1's where 0 represents the absence of an item and 1 the presence.

An example for a rule in this scenario would be {Onion, Potato} => {Burger}, which means that if onion and potato are bought, customers also buy a burger.

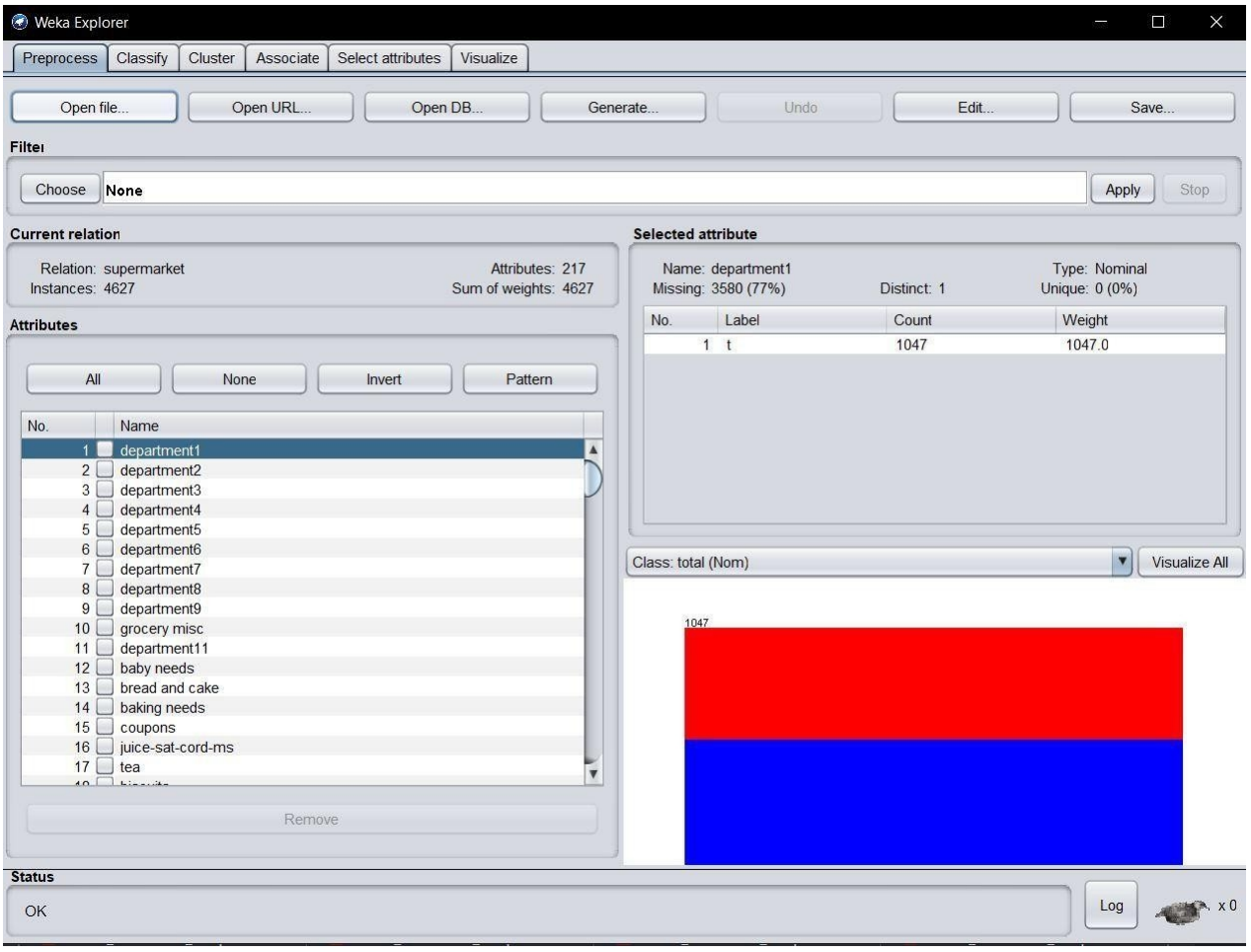
TransactioID	Onion	Potato	Burger	Milk	Beer
t1	1	1	1	0	0
t2	0	1	1	1	0
t3	0	0	0	1	1
t4	1	1	0	1	0
t5	1	1	1	0	1

There are multiple rules possible even from a very small database, so in order to select the interesting ones, we use constraints on various measures of interest and significance. We will look at some of these useful measures such as support, confidence, lift and conviction.

Step1:Load the Supermarket Dataset

Weka comes with a number of real datasets in the "*data*" directory of the Weka installation. This is very handy because you can explore and experiment on these well known problems and learn about the various methods in Weka at your disposal.

Load the Supermarket dataset (*data/supermarket.arff*). This is a dataset of point of sale information. The data is nominal and each instance represents a customer transaction at a supermarket, the products purchased and the departments involved. There is not much information about this dataset online, although you can see this comment from the personal that collected the data.

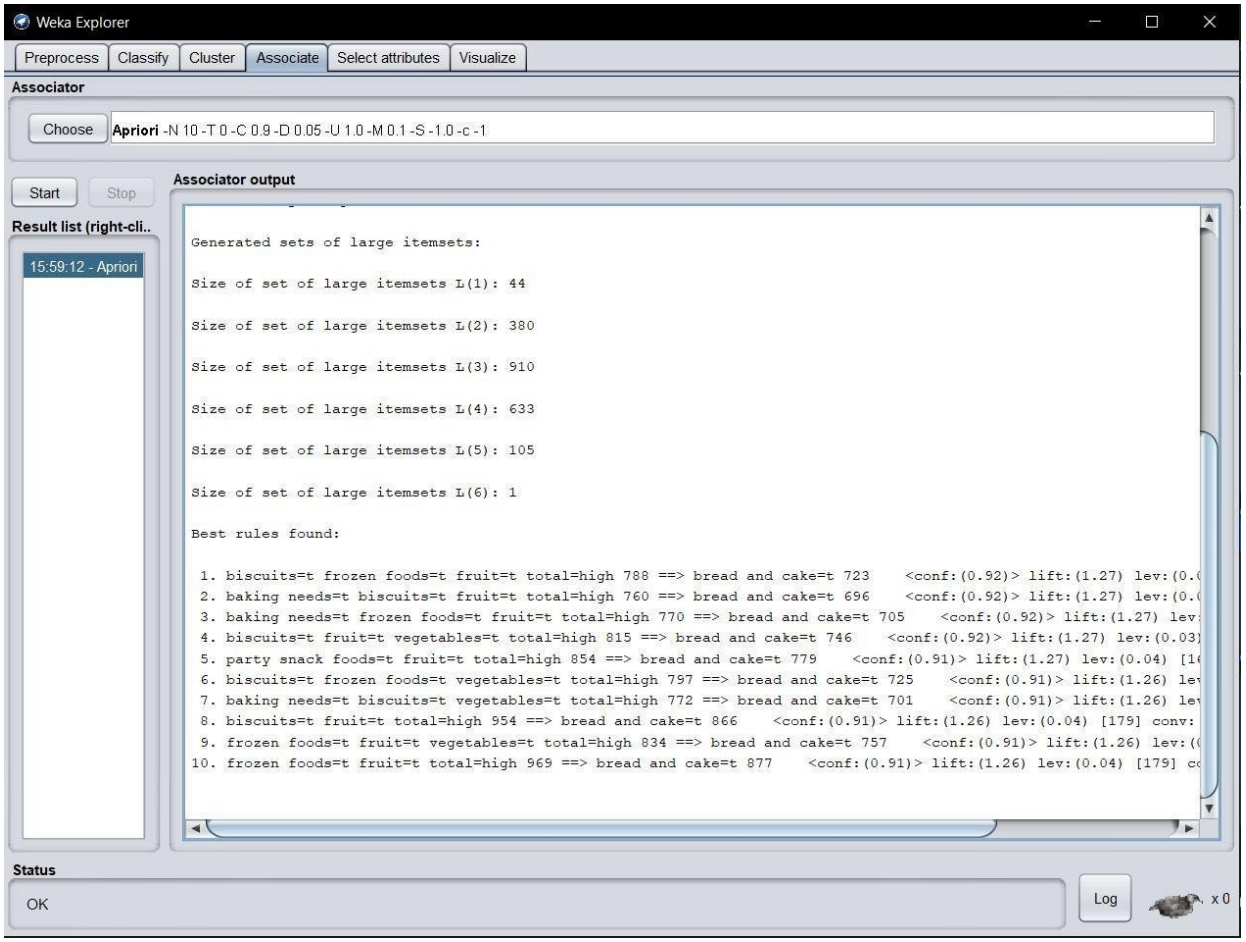


The data contains 4,627 instances and 217 attributes. The data is denormalized. Each attribute is binary and either has a value ("*t*" for true) or no value ("*?*" for missing). There is a nominal class attribute called"total" that indicates whether the transaction was less than $100 (low) or greater than $100 (high).

Step2:Discover Association Rules

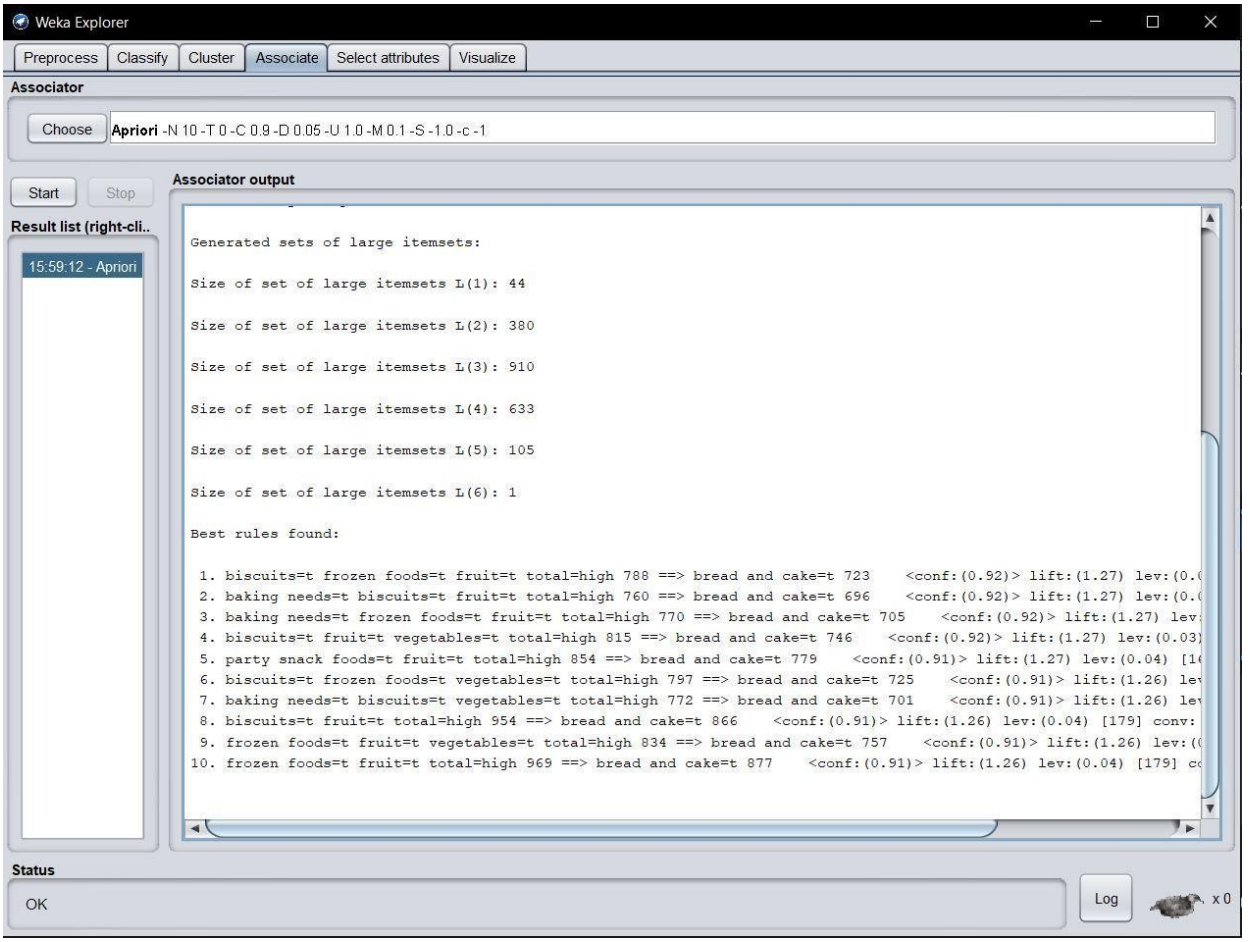
Click the “Associate” tab in the Weka Explorer. The “Apriori” algorithm will already be selected. This is the most well known association rule learning method because it may have been the first (Agrawal and Srikant in 1994) and it is very efficient.

In principle the algorithm is quite simple. It builds up attribute-value (item) sets that maximize the number of instances that can be explained (coverage of the dataset). The search through item space is very much similar to the problem faced with attribute selection and subset search.



Step 3: Analyze Result

The real work for association rule learning is in the interpretation of results.



You can see rules are presented in antecedent => consequent format. The number associated with the antecedent is the absolute coverage in the dataset (in this case a number out of a possible total of 4,627). The number next to the consequent is the absolute number of instances that match the antecedent and the consequent. The number in brackets on the end is the support for the rule (number of antecedent divided by the number of matching consequents). You can see that a cutoff of 91% was usedin selecting rules, mentioned in the "Associator output" window and indicated in that no rule has a coverage less than 0.91.

I don't want to go through all 10 rules, it would be too onerous. Here are few observations:

- We can see that all presented rules have a consequent of "bread and cake".
- All presented rules indicate a high total transaction amount.
- "biscuits" an "frozen foods" appear in many of the presented rules.

The rules discovered where:

1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)
2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.2

3. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746<conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)
4. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779<conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)
5. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725<conf:(0.91)> lift:(1.26) lev:(0.03) [151] conv:(3.06)
6. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701<conf:(0.91)> lift:(1.26) lev:(0.03) [145] conv:(3.01)
7. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(3)
8. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757<conf:(0.91)> lift:(1.26) lev:(0.03) [156] conv:(3)
9. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92)

Conclusion: Hence implementation of Apriori algorithm for association rule mining is studied.

VIVA QUESTIONS.

Q1.Define association rule mining?
Ans: Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is 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.

Q2.Define apriori algorithm?
Ans: The Apriori algorithm is used for mining frequent itemsets and devising association rules from a transactional database. The parameters “support” and “confidence” are used.Support refers to items’ frequency of occurrence; confidence is a conditional probability.

Q3.what is meant by frequent itemset mining?
Ans:Frequent Pattern Mining is a Data Mining subject with the objective of extracting frequent itemsets from a database. Frequent itemsets play an essential role in many Data Mining tasks and are related to interesting patterns in data, such as Association Rules.

Q4.Define Support and Confidence?
Ans: The number of transactions that include items in the {X} and {Y} parts of the rule as apercentage of the total number of transaction.It is a measure of how frequently the collection of items occur together as a percentage of all transactions.