Assignment 4

1. Consider the data set shown in Table 6.1.

Customer ID	**Transaction ID**	**Items Bought**
1	0001	{a,d,e}
1	0024	{a,b,c,e}
2	0012	{a,b,d,e}
2	0031	{a,c,d,e}
3	0015	{b,c,e}
3	0022	{b,d,e}
4	0029	{c,d}
4	0040	{a,b,c}
5	0033	{a,d,e}
5	0038	{a,b,e}

a) Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket.

Solution:

- Support of {e} : Number of occurrence of {e}/ Total number of transaction
  S({e})= 8/10 =**0.8**
- Support of {b,d}
  S({b,d})= 2/10 =**0.2**
- Support of {b,d,e}= 2/10= **0.2**

b) Use the results in part (a) to compute the confidence for the association rules {b, d} −→ {e} and {e} −→ {b, d}. Is confidence a symmetric measure?

Solution:

- Confidence ({b,d}-> {e})= s({b,d,e}) /s({b,d})
  = 0.2 /0.2
  = 1
- Confidence ({b,d}-> {e})=s({b,d,e})/ s({e})

  = 0.2/0.8

  = 0.25

From the above result we can say confidence is not a symmetric measure.

2. Consider the market basket transactions shown in Table 6.2.

Transaction ID	**Item Bought**
1	{Milk,Beer,Daipers}
2	{Bread,Butter,Milk }
3	{Milk,Daipers,Cookies}
4	{Bread,Butter,Cookies }
5	{Beer,Cookies,Daipers}
6	{Milk,Daipers,Bread,Butter}
7	{Bread,Butter,Daipers}
8	{Beer, Daiper}
9	{Milk,Daipers,Bread,Butter}
10	{ Beer, Cookies}

a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

Solution:

The data set contains 6 items namely Milk Beer Diapers Bread Butter Cookies. Hence maximum association rule that can be extracted are 602.

b) What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?

Solution:

The maximum items in a transaction is 4 hence maximum size that can be extracted are 4.

c) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

Solution:

As there are total of 6 items in data set, 3 out of 6 can be selected in ${}^{6}C_{3}$ i.e. 20 ways.

d) Find an itemset (of size 2 or larger) that has the largest support.

Solution:

{bread,butter } has the largest support 0.5.

e) Find a pair of items, a and b, such that the rules {a} −→ {b} and {b} −→ {a} have the same confidence.
Solution: { Beer, Cookies } and {Bread,Butter}

3. The Apriori algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size k+1 are created by joining a pair of frequent itemsets of size k (this is known as the candidate generation step). A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the Apriori algorithm is applied to the data set shown in Table 6.3 with minsup = 30%, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

Transaction ID	**Items Bought**
1	{a,b,d,e}
2	{b,c,d}
3	{a,b,d,e}
4	{a,c,d,e}
5	{b,c,e,d}
6	{b,d,e}
7	{c,d}
8	{a,b,c}
9	{a,d,e}
10	{a,b,e}

a) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?
Solution:
Percentage of frequent itemsets = 16/32
= 0.5

b) What is the pruning ratio of the Apriori algorithm on this data set? (Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.)
Solution:
Pruning ratio = N / total number of itemsets.
Pruning ratio is 11/32 = 34.4%.

c) What is the false alarm rate (i.e, percentage of candidate itemsets that are found to be infrequent after performing support counting)?
Solution:
False alarm rate = I / total number of itemsets.
False alarm rate is 5/32 = 15.6%.

4. The Apriori algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider the hash tree for candidate 3-itemsets shown in Figure 6.2.

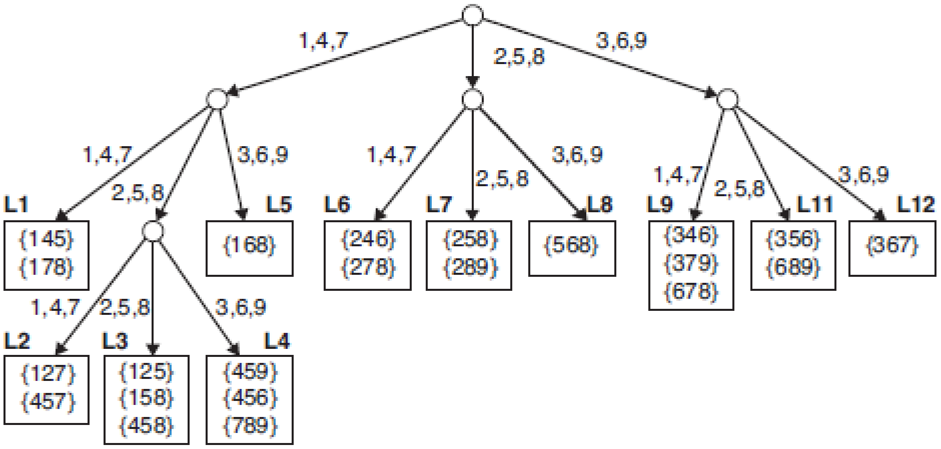


Figure 6.2. An example of a hash tree structure.

a. Given a transaction that contains items {1, 3, 4, 5, 8}, which of the hash tree leaf nodes will be visited when finding the candidates of the transaction?
Solution:
Nodes visited are:
L1 L3 L5 L9 L11

b. Use the visited leaf nodes in part (b) to determine the candidate itemsets that are contained in the transaction {1, 3, 4, 5, 8}.
Solution:
{ 1 4 5} { 1 5 8} { 4 5 8}

Weka Experiments

1. Explain what are the important parameters, what are the default values for them

   Solution:

   - car -If enabled class association rules are mined instead of (general) association rules i.e. it uses class association rules are used for association if car= true. Default value: False.
   - classIndex - Index of the class attribute (used for class association rule). If set to -1, the last attribute is taken as class attribute. Default value: -1.
   - delta – The support is subtracted by this factor until minimum support is reached. Default value: 0.05
   - lowerBoundMinSupport -- Lower bound for minimum support.
   - metricType -- Set the type of metric by which to rank rules. Confidence is the proportion of the examples covered by the premise that are also covered by the consequence(Class association rules can only be mined using confidence). Lift is confidence divided by the proportion of all examples that are covered by the consequence. This is a measure of the importance of the association that is independent of support. Leverage is the proportion of additional examples covered by both the premise and consequence above those expected if the premise and consequence were independent of each other. The total number of examples that this represents is presented in brackets following the leverage. Conviction is another measure of departure from independence. Conviction is given by P(premise)P(!consequence) / P(premise, !consequence).Default value: Confidence
   - minMetric -- Minimum metric score. Consider only rules with scores higher than this value. Default value: 0.9
   - numRules -- Number of rules to find. Default value: 10
   - outputItemSets -- If enabled the itemsets are output as well. Default value:false
   - removeAllMissingCols -- Remove columns with all missing values. Default value: false

- significanceLevel -- Significance level. Significance test (confidence metric only).Default value: -1.0

- pperBoundMinSupport -- Upper bound for minimum support. Start iteratively decreasing minimum support from this value. Default value: 1.0

- verbose -- If enabled the algorithm will be run in verbose mode.Default value: false.

2. Show the top 4 association rules using the default parameters

Solution:

a. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 conf:(0.92)

b. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 conf:(0.92)

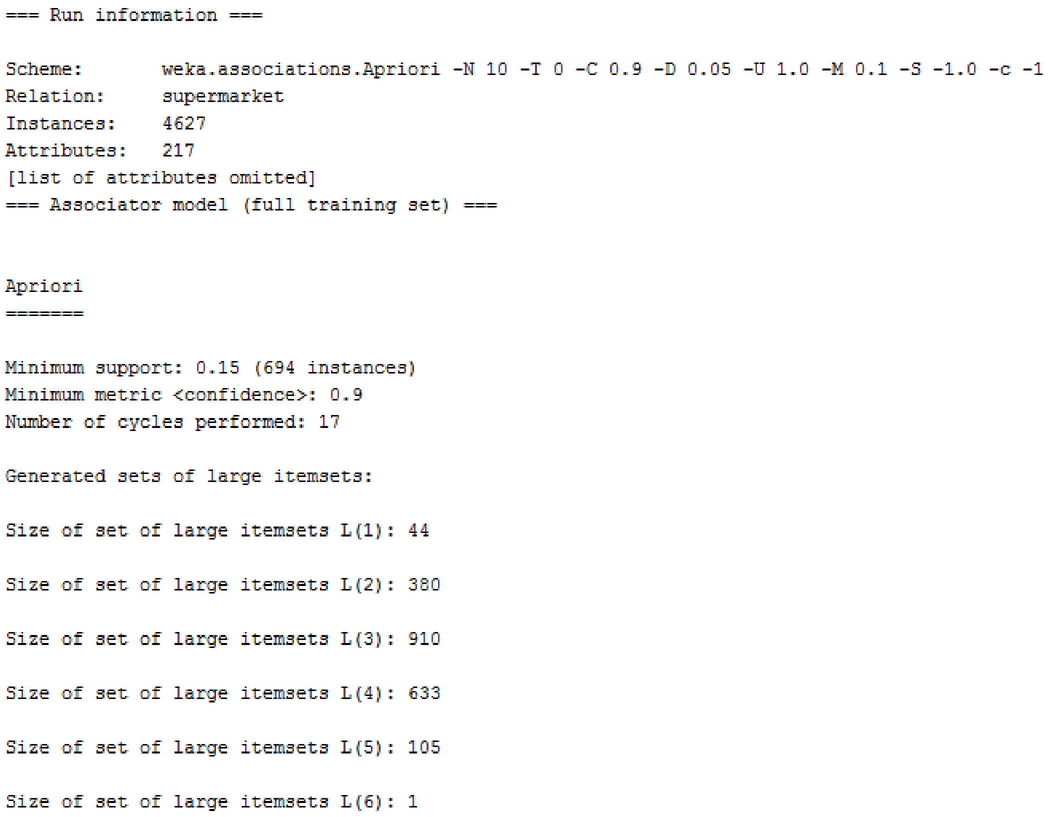
c. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t705 conf:(0.92)

d. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 conf:(0.92)

3. Modify support/confidence 3 times: set them to low, medium and high values. You can decide yourself what exact values to use. Experiment with each of the 9 of combinations of the values for support and confidence and compare the results.

Solution:

a. lowerBoundMinSupport:0.1 minMetric:0.9

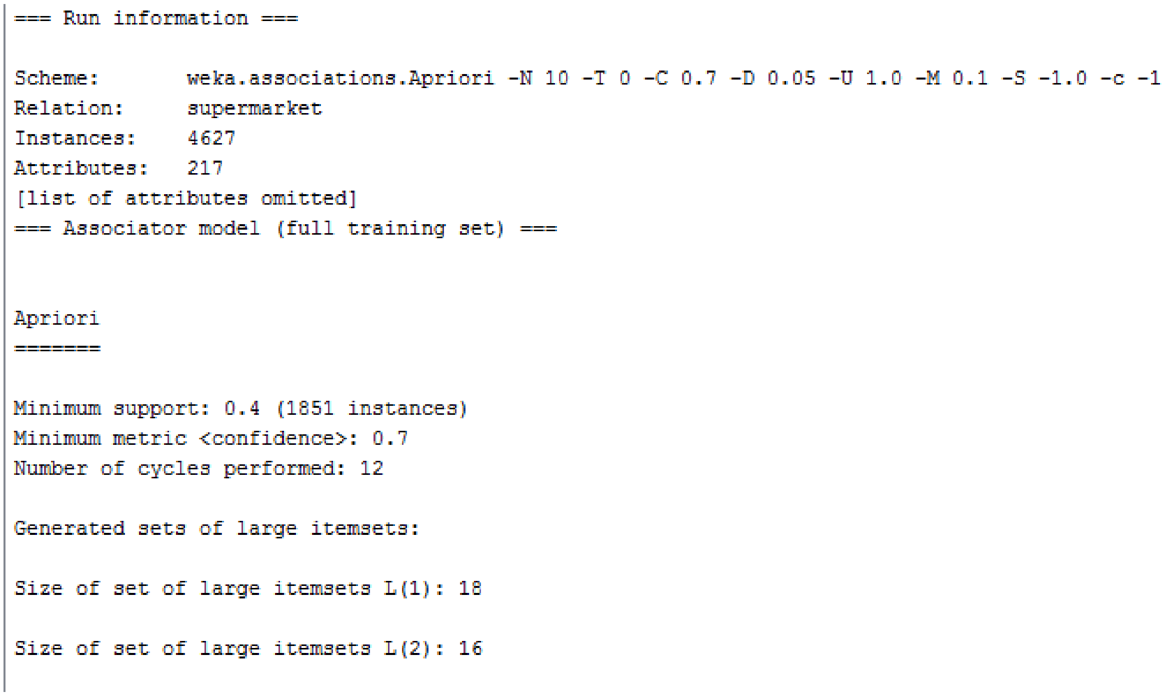


```
=== Run information ===

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:     supermarket
Instances:    4627
Attributes:   217
[list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 44

Size of set of large itemsets L(2): 380

Size of set of large itemsets L(3): 910

Size of set of large itemsets L(4): 633

Size of set of large itemsets L(5): 105

Size of set of large itemsets L(6): 1
```

The above result shows that:

- There are 6 items in the largest itemset and there exist only one item set that satisfy association rule.
- The minimum support is 0.15 and the minimum metric i.e. confidence is 0.9.
- The above result shows that count for each set with 1,2,3,4,5 and 6 itemsets.

b. lowerBoundMinSupport:0.1 minMetric:0.7

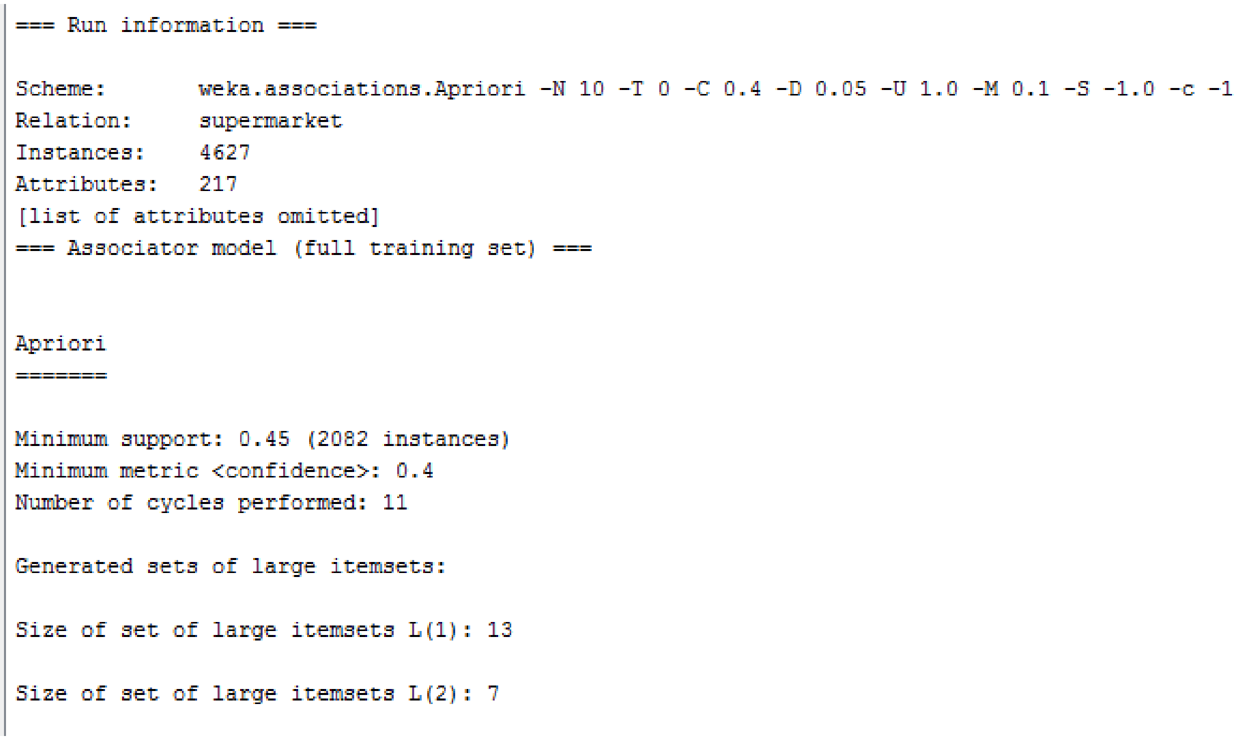


```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.7 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
[list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.4 (1851 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 18

Size of set of large itemsets L(2): 16
```

The above result shows that:

- There are 2 items in the largest itemset and there exist 16 itemsets that satisfy association rule.
- The minimum support increases to 0.4 and the minimum metric i.e. confidence is 0.7.
- The above result shows that count for each set with 1 and 2.

c. lowerBoundMinSupport:0.1 minMetric:0.4

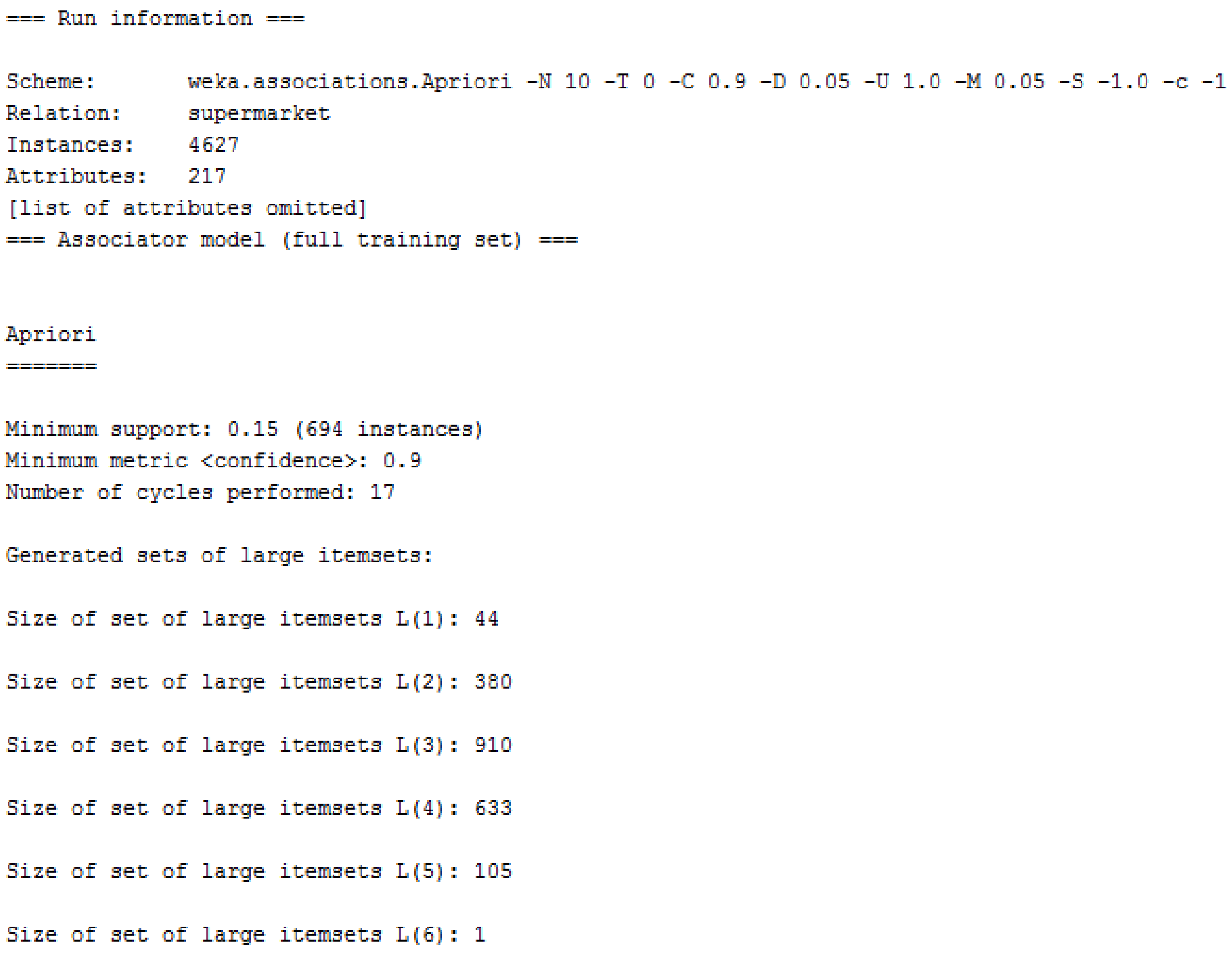


```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.4 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
[list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.45 (2082 instances)
Minimum metric <confidence>: 0.4
Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 13

Size of set of large itemsets L(2): 7
```

The above result shows that:

- There are 2 items in the largest itemset and there exist 7 itemsets that satisfy association rule.
- The minimum support increases to 0.45 and the minimum metric i.e. confidence is 0.4.
- The above result shows that count for each set with 1 and 2.

d. lowerBoundMinSupport:0.05 minMetric:0.9

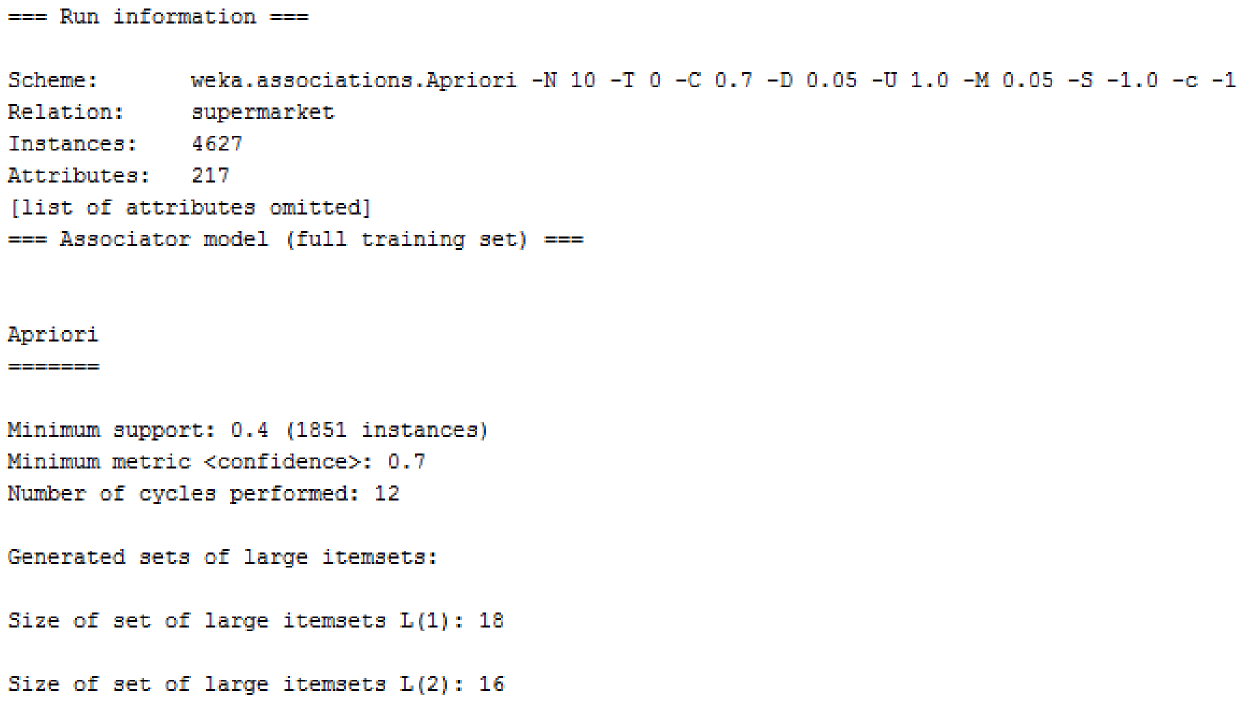


```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.05 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
[list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 44

Size of set of large itemsets L(2): 380

Size of set of large itemsets L(3): 910

Size of set of large itemsets L(4): 633

Size of set of large itemsets L(5): 105

Size of set of large itemsets L(6): 1
```

The above result shows that:

- There are 6 items in the largest itemset and there exist 7 itemsets of size 6 items each that satisfy association rule.
- The minimum support increases to 0.15 and the confidence remains constant at 0.9.
- The above result shows that count for each set with 1,2,3,4,5 and 6.

e. lowerBoundMinSupport:0.05 minMetric:0.7

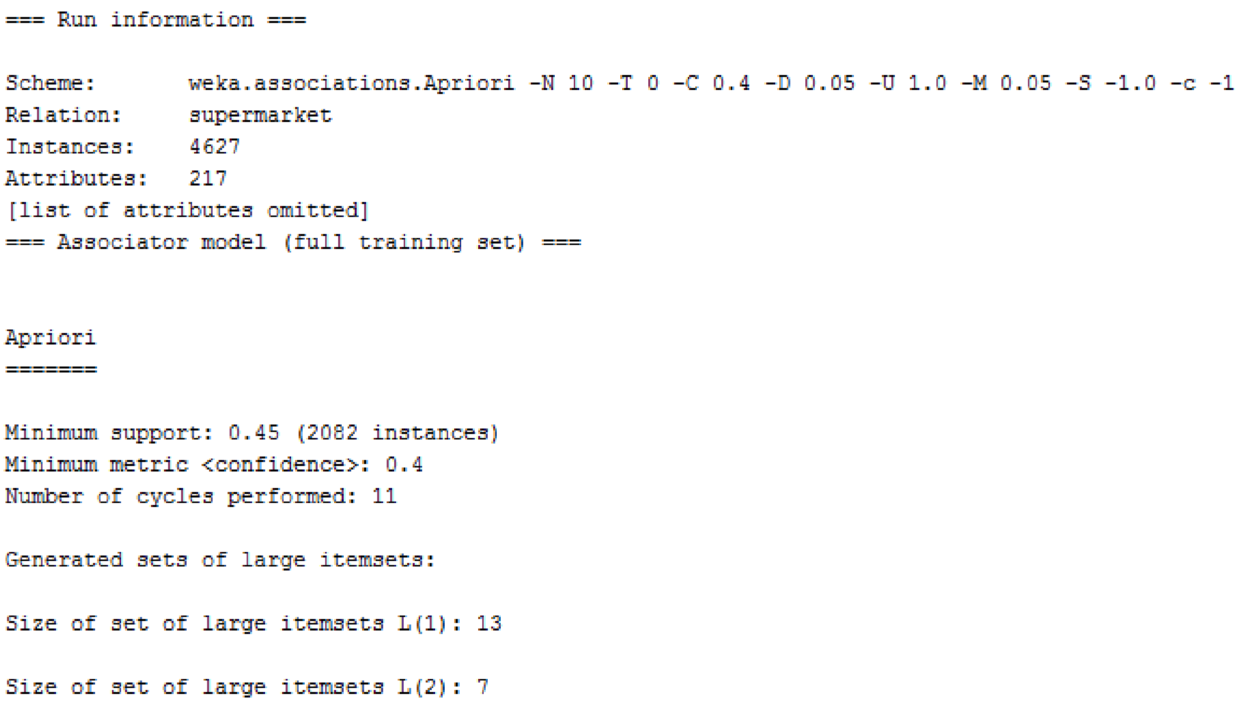


```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.7 -D 0.05 -U 1.0 -M 0.05 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
[list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.4 (1851 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 18

Size of set of large itemsets L(2): 16
```

The above result shows that:

- There are 2 items in the largest itemset and there exist 16 itemsets of size 2 items each that satisfy association rule.
- The minimum support increases to 0.4 and the confidence remains constant at 0.7.
- The above result shows that count for each set with 1 and2.

f. lowerBoundMinSupport:0.05 minMetric:0.4

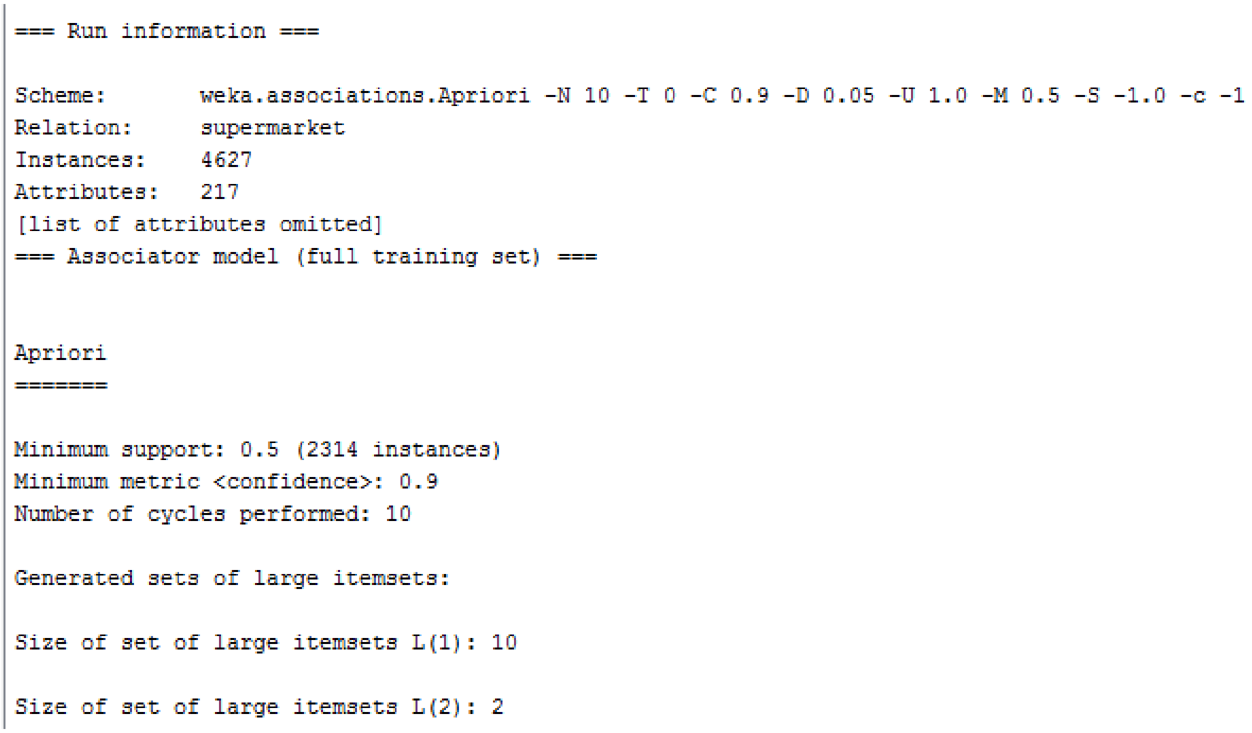


```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.4 -D 0.05 -U 1.0 -M 0.05 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
[list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.45 (2082 instances)
Minimum metric <confidence>: 0.4
Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 13

Size of set of large itemsets L(2): 7
```

The above result shows that:

- There are 2 items in the largest itemset and there exist 7 itemsets of size 2 items each that satisfy association rule.
- The minimum support increases to 0.45 and the confidence remains constant at 0.4.
- The above result shows that count for each set with 1 and2.

g. lowerBoundMinSupport:0.5 minMetric:0.9

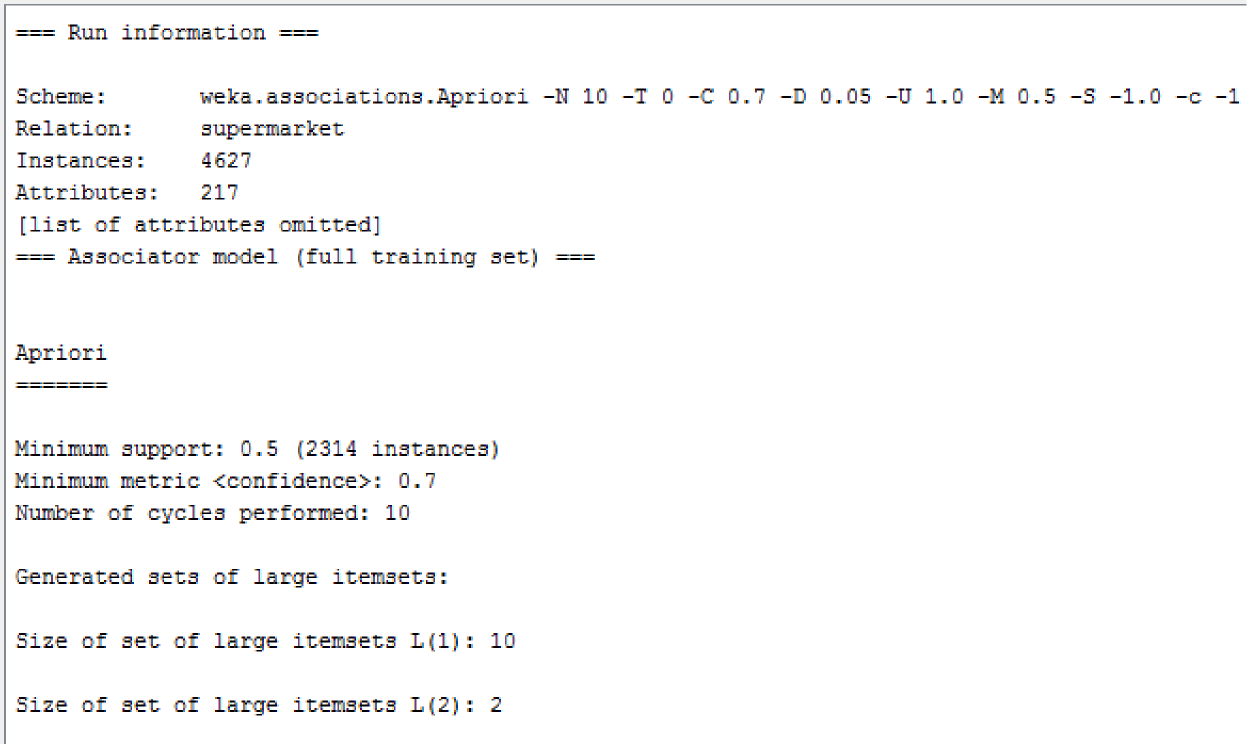


```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.5 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
[list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.5 (2314 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 2
```

The above result shows that:

- There are 2 items in the largest itemset and there exist 2 itemsets of size 2 items each that satisfy association rule.
- The minimum support remains same 0.5 and the confidence remains constant at 0.9.
- The above result shows that count for each set with 1 and2.

h. lowerBoundMinSupport:0.5 minMetric:0.7

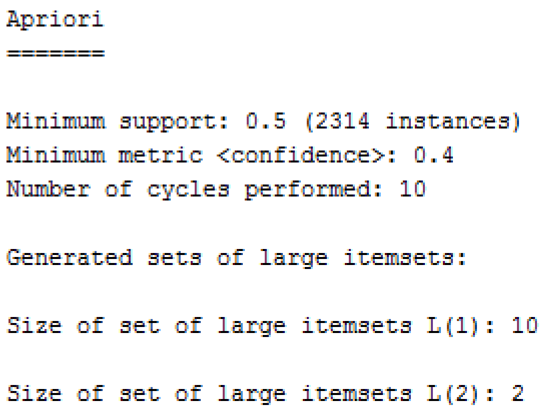


```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.7 -D 0.05 -U 1.0 -M 0.5 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
[list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.5 (2314 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 2
```

The above result shows that:

- There are 2 items in the largest itemset and there exist 2 itemsets of size 2 items each that satisfy association rule.
- The minimum support remains same 0.5 and the confidence remains constant at 0.9.
- The above result shows that count for each set with 1 and2.
- The result remains the same even though the confidence metric is decreases.

i. lowerBoundMinSupport:0.5 minMetric:0.4



```
Apriori
=======

Minimum support: 0.5 (2314 instances)
Minimum metric <confidence>: 0.4
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 2
```

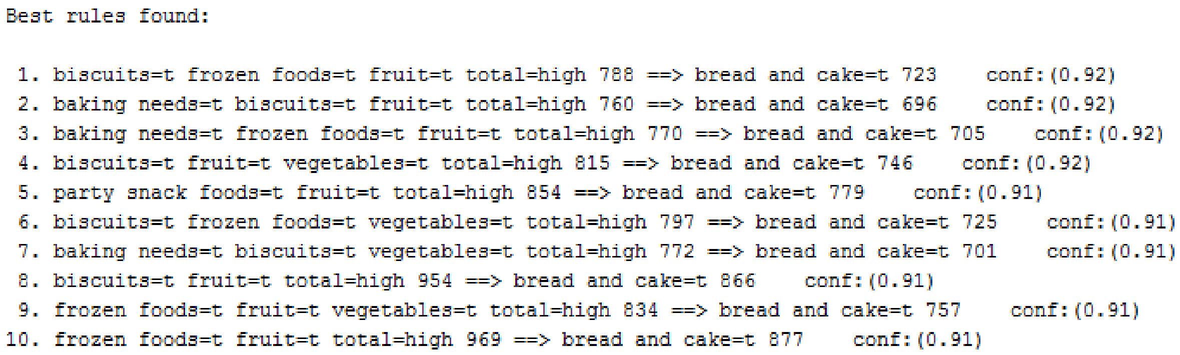
The above result shows that:

- There are 2 items in the largest itemset and there exist 2 itemsets of size 2 items each that satisfy association rule.
- The minimum support remains same 0.5 and the confidence remains constant at 0.4.
- The above result shows that count for each set with 1 and2.
- The result remains the same even though the confidence metric is decreases.

4. Use the attribute selection tab, select some of the available items for your association rule computation. You can remove the most frequent items or select items to remove using other criteria. Can you see any interesting results?

   Solution:

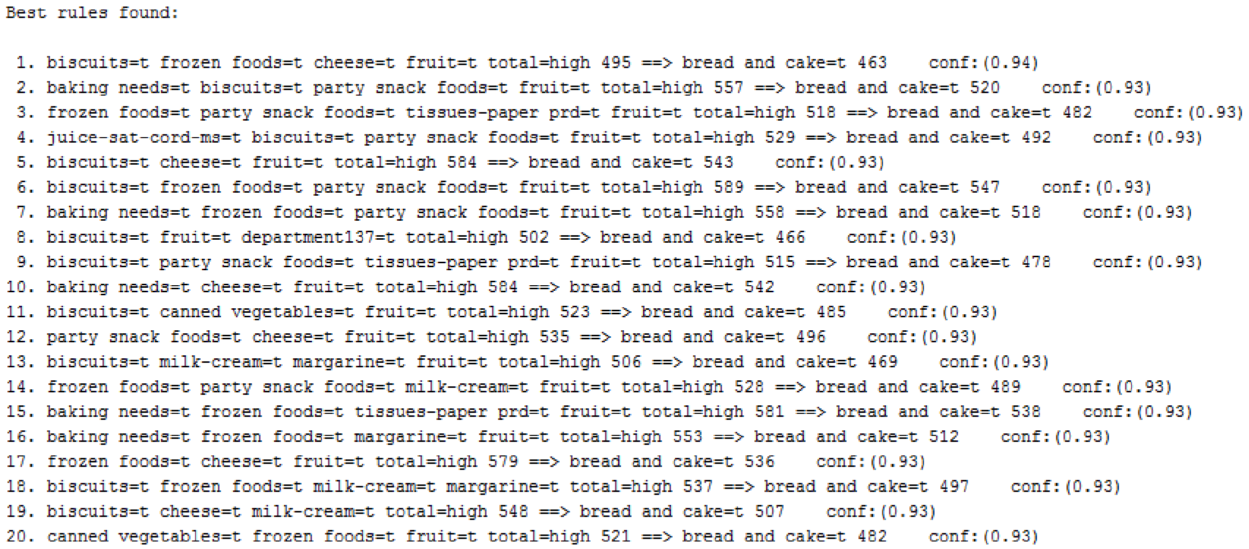
   a. Best association rules for default case:



```
Best rules found:

 1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723    conf:(0.92)
 2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696    conf:(0.92)
 3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705    conf:(0.92)
 4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746    conf:(0.92)
 5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779    conf:(0.91)
 6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725    conf:(0.91)
 7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701    conf:(0.91)
 8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866    conf:(0.91)
 9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757    conf:(0.91)
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877    conf:(0.91)
```

- The above result shows that shows top 10 association rule for which the confidence is as high as 0.92
- The minimum support increased to 0.15 from 0.1 and the value of the
- confidence remained the same.
- The output consist of 6 attribute item sets and took 17 cycles for completion of the run.

b. Removing selected attributes
After removing attributes grocery misc, coupon, vegetable, baby needs we get:



```
Best rules found:

 1. biscuits=t frozen foods=t cheese=t fruit=t total=high 495 ==> bread and cake=t 463    conf:(0.94)
 2. baking needs=t biscuits=t party snack foods=t fruit=t total=high 557 ==> bread and cake=t 520    conf:(0.93)
 3. frozen foods=t party snack foods=t tissues-paper prd=t fruit=t total=high 518 ==> bread and cake=t 482    conf:(0.93)
 4. juice-sat-cord-ms=t biscuits=t party snack foods=t fruit=t total=high 529 ==> bread and cake=t 492    conf:(0.93)
 5. biscuits=t cheese=t fruit=t total=high 584 ==> bread and cake=t 543    conf:(0.93)
 6. biscuits=t frozen foods=t party snack foods=t fruit=t total=high 589 ==> bread and cake=t 547    conf:(0.93)
 7. baking needs=t frozen foods=t party snack foods=t fruit=t total=high 558 ==> bread and cake=t 518    conf:(0.93)
 8. biscuits=t fruit=t department137=t total=high 502 ==> bread and cake=t 466    conf:(0.93)
 9. biscuits=t party snack foods=t tissues-paper prd=t fruit=t total=high 515 ==> bread and cake=t 478    conf:(0.93)
10. baking needs=t cheese=t fruit=t total=high 584 ==> bread and cake=t 542    conf:(0.93)
11. biscuits=t canned vegetables=t fruit=t total=high 523 ==> bread and cake=t 485    conf:(0.93)
12. party snack foods=t cheese=t fruit=t total=high 535 ==> bread and cake=t 496    conf:(0.93)
13. biscuits=t milk-cream=t margarine=t fruit=t total=high 506 ==> bread and cake=t 469    conf:(0.93)
14. frozen foods=t party snack foods=t milk-cream=t fruit=t total=high 528 ==> bread and cake=t 489    conf:(0.93)
15. baking needs=t frozen foods=t tissues-paper prd=t fruit=t total=high 581 ==> bread and cake=t 538    conf:(0.93)
16. baking needs=t frozen foods=t margarine=t fruit=t total=high 553 ==> bread and cake=t 512    conf:(0.93)
17. frozen foods=t cheese=t fruit=t total=high 579 ==> bread and cake=t 536    conf:(0.93)
18. biscuits=t frozen foods=t milk-cream=t margarine=t total=high 537 ==> bread and cake=t 497    conf:(0.93)
19. biscuits=t cheese=t milk-cream=t total=high 548 ==> bread and cake=t 507    conf:(0.93)
20. canned vegetables=t frozen foods=t fruit=t total=high 521 ==> bread and cake=t 482    conf:(0.93)
```

- The deletion of the attribute results in increase in the confidence of associative rules from 0.92 to 0.94
- The minimum support and confidence remained the same.
- The output consist of 6 attribute item sets and took 18 cycles for completion of the run.

Conclusion:

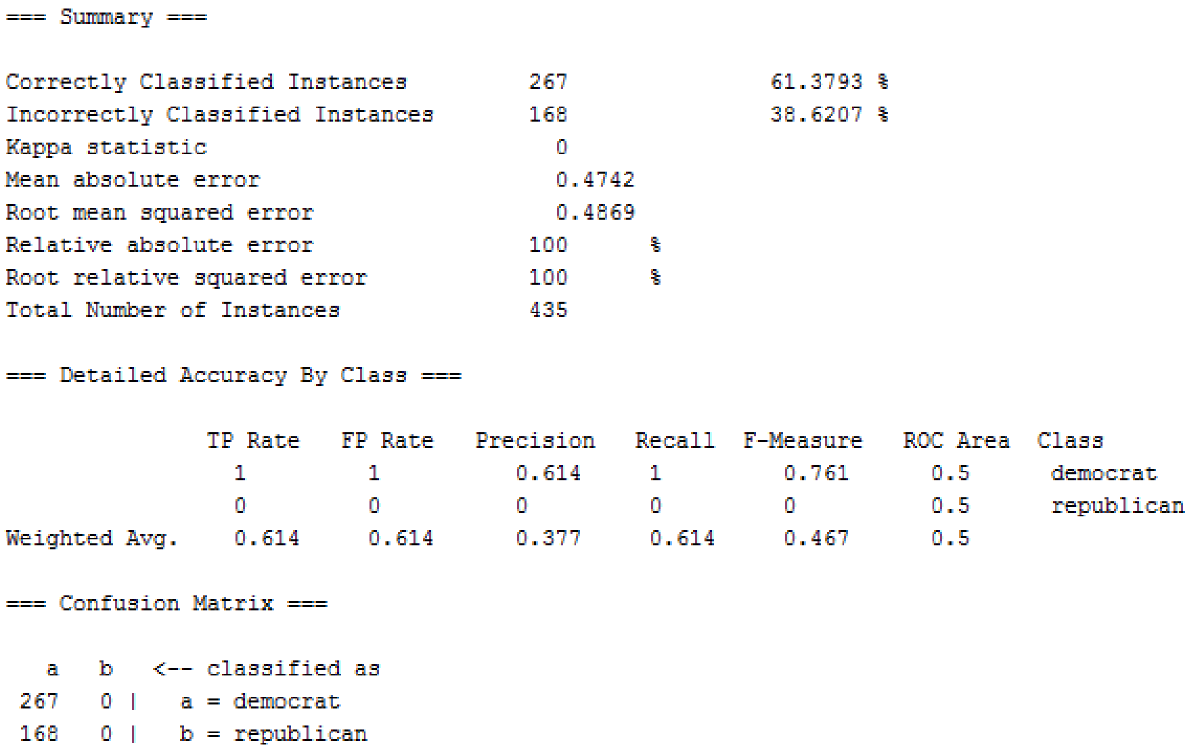
The confidence values increased when some frequent attributes were deleted. The number of 6 attribute itemset also increased from 1 to 214 and other itemsets are also affected.

5. Report the summary of what your learned, explain how your modifications affected the generation of the rules
   - The apriori algorithm takes min support and confidence metric into consideration while applying association rule.
   - These parameter of the algorithm were change and the modification in the result were observed.

- If the min support is reduced to below the default value 0.1 and the confidence is decreased the min support is increased to higher values. The association rule with confidence up to 0.92 and reduces to 0.8 as confidence value are produced by varying these parameters.
- If the min support is equal to the default value and confidence is decreased the minimum support increases. Maximum confidence remain the same as that for the above case.
- If the min support is increased and confidence is decreased the support value remains the same. The confidence is this case reduces to 0.8.
- If some of the frequent items from the data set are deleted the confidence value increases for the best relation from 0.92 to 0.94.

6. For the Vote dataset: Compare the association rules to the SimpleCart decision tree

a) SimpleCart on vote data set.

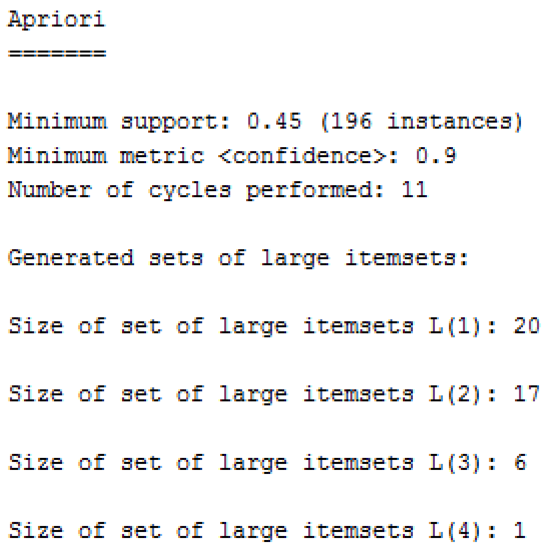


```
=== Summary ===

Correctly Classified Instances         267               61.3793 %
Incorrectly Classified Instances       168               38.6207 %
Kappa statistic                          0
Mean absolute error                      0.4742
Root mean squared error                  0.4869
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1         0.614     1         0.761      0.5      democrat
                 0         0         0         0         0          0.5      republican
Weighted Avg.    0.614     0.614     0.377     0.614     0.467      0.5

=== Confusion Matrix ===

   a   b   <-- classified as
 267   0 |   a = democrat
 168   0 |   b = republican
```

The above results shows that:

- The error rate is 38.62% i.e. 168 instances of the vote data set are misclassified.
- The SimpleCart algorithm classifies the training data to either a democrat or republican.

b) Apriori algorithm on vote data set



```
Apriori
=======

Minimum support: 0.45 (196 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 6

Size of set of large itemsets L(4): 1
```

The above result shows

- The minimum support increases to 0.45 and the confidence remains at 0.9.
- The itemset has maximum 4 attributes L (4) and there is 1 such itemset that satisfies association set.