Lab 2
732A61 Data Mining - Clustering and Association Analysis

Carles Sans Fuentes

March 2, 2017

Assignment 1

The main goals of this assignment is to cluster a given dataset and use association analysis to describe the clusters obtained.

For this, the data set iris.arff data (a data set providing with 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample, they are the length and the width of sepal and petal.

Before doing the analysis, some treatment on the data set is needed such that the four first variables are discretized to three bins using some filtering from the program.

Choice

It will be also assessed results for (1) 4 clusters and (2) 4 bins

Exercise 1 Clustering

3 Clusters

The algorithm "SimpleKMeans" has been applied to your data. In Weka euclidian distance is implemented in SimpleKmeans and the seed chosen for this case (ignoring the class attribute) is 10. The results are provided here below for three clusters:

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000
    -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I
    500 -num-slots 1 -S 10
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
Ignored:
              class
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 96.0

```

```
Initial staring points (random):

Cluster 0: '\'(5.5-6.7]\'','\'(2.8-3.6]\'','\'(2.966667-4.933333]\'','\'(0.9-1.7]\''
Cluster 1: '\'(6.7-inf)\'','\'(2.8-3.6]\'','\'(4.933333-inf)\'','\'(1.7-inf)\''
Cluster 2: '\'(-inf-5.5]\'','\'(3.6-inf)\'','\'(-inf-2.966667]\'','\'(-inf-0.9]\''

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                            Cluster#
Attribute                             Full Data                    0                      1                      2
                                          (150)                 (55)                   (45)                   (50)
=====================================================================================================================
sepallength                          '(5.5-6.7]'          '(5.5-6.7]'            '(5.5-6.7]'            '(-inf-5.5]'
sepalwidth                           '(2.8-3.6]'         '(-inf-2.8]'            '(2.8-3.6]'             '(2.8-3.6]'
petallength                '(2.966667-4.933333]' '(2.966667-4.933333]'      '(4.933333-inf)'      '(-inf-2.966667]'
petalwidth                           '(0.9-1.7]'          '(0.9-1.7]'            '(1.7-inf)'            '(-inf-0.9]'




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0        55 ( 37%)
1        45 ( 30%)
2        50 ( 33%)
```

It can be seen that the three clusters show big differences across petal characteristics, whereas in the case of sepal characteristics, clusters 1 and 2 share sepal width characteristics as well as clusters 0 and 1 for sepal length.

4 Clusters

The algorithm "SimpleKMeans" has been applied to your data for 4 clusters. It is already known from the class the best clustering for this data is 3, but it will be tried to see what happens with 3 even though it is know that there are three types of Iris flowers. The seed chosen for this case (ignoring the class attribute) is 10. The results are provided here below for four clusters:

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000
    -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 4 -A "weka.core.EuclideanDistance -R first-last" -I
    500 -num-slots 1 -S 10
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
Ignored:
              class
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 90.0

Initial staring points (random):

Cluster 0: '\'(5.5-6.7]\'','\'(2.8-3.6]\'','\'(2.966667-4.933333]\'','\'(0.9-1.7]\''
```

```
Cluster 1: '\'(6.7-inf)\'','\'(2.8-3.6]\'','\'(4.933333-inf)\'','\'(1.7-inf)\''
Cluster 2: '\'(-inf-5.5]\'','\'(3.6-inf)\'','\'(-inf-2.966667]\'','\'(-inf-0.9]\''
Cluster 3: '\'(6.7-inf)\'','\'(2.8-3.6]\'','\'(2.966667-4.933333]\'','\'(0.9-1.7]\''

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                            Cluster#
Attribute                            Full Data                     0                      1
                        2                        3
                                         (150)                  (52)                   (44)
                                                              (50)                    (4)
==================================================================================================================

sepallength                        '(5.5-6.7]'           '(5.5-6.7]'            '(5.5-6.7]'
             '(-inf-5.5]'            '(6.7-inf)'
sepalwidth                         '(2.8-3.6]'          '(-inf-2.8]'            '(2.8-3.6]'
              '(2.8-3.6]'            '(2.8-3.6]'
petallength              '(2.966667-4.933333]' '(2.966667-4.933333]'      '(4.933333-inf)'
    '(-inf-2.966667]' '(2.966667-4.933333]'
petalwidth                         '(0.9-1.7]'           '(0.9-1.7]'            '(1.7-inf)'
             '(-inf-0.9]'            '(0.9-1.7]'




Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0        52 ( 35%)
1        44 ( 29%)
2        50 ( 33%)
3         4 (  3%)
```

Trying to associate rules to all clusters

In order to find rules for all clusters, it has been necessary to get 50 rules with minimum confidence equals to 0.6 and minimum support equals to 0.2. Also, to ease the finding these I have set the "car" parameter to true and set the class index to my cluster index which is 6. The following result has been got:

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 50 -T 0 -C 0.6 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -A -c 6
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-weka.
    filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-
    candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.
    core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10-I5-weka.filters.
    unsupervised.attribute.Remove-R6-weka.filters.unsupervised.attribute.AddCluster-Wweka.
    clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density
    2.0 -t1 -1.25 -t2 -1.0 -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-
    slots 1 -S 10-I5
Instances:    150
Attributes:   6
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
              cluster
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.2 (30 instances)
Minimum metric <confidence>: 0.6
Number of cycles performed: 16

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 20

Size of set of large itemsets L(3): 15
```

```

Size of set of large itemsets L(4): 6

Size of set of large itemsets L(5): 1

Best rules found:

 1. petallength='(-inf-2.966667]' 50 ==> cluster=cluster3 50    conf:(1)
 2. petalwidth='(-inf-0.9]' 50 ==> cluster=cluster3 50    conf:(1)
 3. class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 4. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 50 ==> cluster=cluster3 50    conf:(1)
 5. petallength='(-inf-2.966667]' class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 6. petalwidth='(-inf-0.9]' class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 7. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 8. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' 47 ==> cluster=cluster3 47    conf:(1)
 9. sepallength='(-inf-5.5]' petalwidth='(-inf-0.9]' 47 ==> cluster=cluster3 47    conf:(1)
10. sepallength='(-inf-5.5]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
11. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 47 ==> cluster=cluster3 47    conf:(1)
12. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
13. sepallength='(-inf-5.5]' petalwidth='(-inf-0.9]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
14. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
15. petallength='(4.933333-inf)' petalwidth='(1.7-inf)' 40 ==> cluster=cluster2 40    conf:(1)
16. petallength='(4.933333-inf)' petalwidth='(1.7-inf)' class=Iris-virginica 40 ==> cluster=cluster2 40    conf:(1)
17. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' 36 ==> cluster=cluster3 36    conf:(1)
18. sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' 36 ==> cluster=cluster3 36    conf:(1)
19. sepalwidth='(2.8-3.6]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
20. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' 36 ==> cluster=cluster3 36    conf:(1)
21. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' 36 ==> cluster=cluster3 36    conf:(1)
22. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
23. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 36 ==> cluster=cluster3 36    conf:(1)
24. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
25. sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
26. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 36 ==> cluster=cluster3 36    conf:(1)
27. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
28. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
29. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
30. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
31. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' 33 ==> cluster=cluster1 33    conf:(1)
32. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' class=Iris-versicolor 33 ==> cluster=cluster1 33    conf:(1)
33. sepalwidth='(-inf-2.8]' petalwidth='(0.9-1.7]' 31 ==> cluster=cluster1 31    conf:(1)
34. sepallength='(5.5-6.7]' petalwidth='(0.9-1.7]' 38 ==> cluster=cluster1 37    conf:(0.97)
35. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' 37 ==> cluster=cluster3 36    conf:(0.97)
36. sepallength='(5.5-6.7]' petalwidth='(0.9-1.7]' class=Iris-versicolor 35 ==> cluster=cluster1 34    conf:(0.97)
37. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' class=Iris-versicolor 34 ==> cluster=cluster1 33    conf:(0.97)
38. sepallength='(5.5-6.7]' class=Iris-versicolor 36 ==> cluster=cluster1 34    conf:(0.94)
39. petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' 48 ==> cluster=cluster1 45    conf:(0.94)
40. petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' class=Iris-versicolor 47 ==> cluster=cluster1 44    conf:(0.94)
41. petalwidth='(1.7-inf)' 46 ==> cluster=cluster2 43    conf:(0.93)
42. petalwidth='(1.7-inf)' class=Iris-virginica 45 ==> cluster=cluster2 42    conf:(0.93)
43. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' 39 ==> cluster=cluster1 36    conf:(0.92)
44. petalwidth='(0.9-1.7]' class=Iris-versicolor 49 ==> cluster=cluster1 45    conf:(0.92)
45. petallength='(2.966667-4.933333]' class=Iris-versicolor 48 ==> cluster=cluster1 44    conf:(0.92)
46. petallength='(4.933333-inf)' class=Iris-virginica 44 ==> cluster=cluster2 40    conf:(0.91)
47. petalwidth='(0.9-1.7]' 54 ==> cluster=cluster1 49    conf:(0.91)
48. class=Iris-versicolor 50 ==> cluster=cluster1 45    conf:(0.9)
49. petallength='(4.933333-inf)' 46 ==> cluster=cluster2 41    conf:(0.89)
50. petallength='(2.966667-4.933333]' 54 ==> cluster=cluster1 48    conf:(0.89)
```

I am going to comment a rule for each cluster: Rule 1.When petal length is in between '(-inf-2.966667]', which happens to be in 50 cases (1/3 of the data) in all 50 cases (confidence 1) they belong to cluster 3.

Rule 41.When petal width is in between '(1.7-inf)', which happens to be in 46 cases (almost 1/3 of the data) in almost all cases (43 out of 46 being confidence 0.93) they belong to cluster 2.

Rule 34.When sepal length is in between '(5.5-6.7]' and petal width between (0.9-1.7], which happens to be in 38 cases (almost 1/4 of the data) in almost all cases (37 out of 38 being confidence 0.97) they belong to cluster 1. No rules for the fourth cluster has been found. The minimum support should be lowered more, but then we run into the problem that these rules will not be general enough, so that I will not dig more into it.

Explanation of the 4 clusters and comparison with 3 clusters

It can be seen that one of the clusters is really small compared with the others (cluster 3 with just four observations). Clusters from 0 to 2 have exactly the same characteristics as before. For that reason, the new cluster (number 3), being a little bit more different than the previous ones accounts for sepal.Length from (6.7-inf), sepal.width(2.8-3.6]', petal.Length (2.966667-4.933333] and petal.width(0.9-1.7].

4 bins 3 clusters

The algorithm "SimpleKMeans" has been applied to your data. In Weka euclidian distance is implemented in SimpleKmeans and the seed chosen for this case (ignoring the class attribute) is 10. The results are provided here below for three clusters:

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000
    -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I
    500 -num-slots 1 -S 10
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B4-M-1.0-Rfirst-last
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
Ignored:
              class
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 161.0

Initial staring points (random):

Cluster 0: '\'(5.2-6.1]\'','\'(2.6-3.2]\'','\'(3.95-5.425]\'','\'(1.3-1.9]\''
Cluster 1: '\'(6.1-7]\'','\'(2.6-3.2]\'','\'(3.95-5.425]\'','\'(0.7-1.3]\''
Cluster 2: '\'(6.1-7]\'','\'(2.6-3.2]\'','\'(3.95-5.425]\'','\'(1.9-inf)\''

Missing values globally replaced with mean/mode

Final cluster centroids:
                                         Cluster#
Attribute               Full Data              0              1              2
                            (150)           (54)           (66)           (30)
==========================================================================
sepallength           '(5.2-6.1]'   '(-inf-5.2]'    '(5.2-6.1]'      '(6.1-7]'
sepalwidth            '(2.6-3.2]'    '(3.2-3.8]'    '(2.6-3.2]'    '(2.6-3.2]'
petallength        '(3.95-5.425]' '(-inf-2.475]' '(3.95-5.425]'  '(5.425-inf)'
petalwidth            '(-inf-0.7]'   '(-inf-0.7]'    '(1.3-1.9]'    '(1.9-inf)'
```

```
Time taken to build model (full training data) : 0.07 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       54 ( 36%)
1       66 ( 44%)
2       30 ( 20%)
```

Explanation of the 4 bins and comparison with 3 bins

It can be seen that the three clusters show big differences across almost all clusters, given the fact that now we have 4 bins. There is one cluster now that it is much bigger than the previous ones compared to 3 bins, where all clusters had more or less the same amount of data. This is the cluster 1, with 44% of the data.

Visualization

Here below a visualization of the simplekmeans algorithm is represented in figure 1 and 2 for the 3 clusters with 3 and 4 bins respectively. It is basically a representation of what I have just explained above in a matrix way (symmetrical). Later on, with the association rules, these graphics will be better explained as well the clustering association in terms of size and length of the petal and sepal regarding classes of Iris flowers. Just notice that on Figure2 there are sometimes four points in the graph from the same color accounting for the different number of bins and its classification.

Figure 1: Simplekmeans algorithm for 3 clusters from the iris data set using seed 10- 3 bins

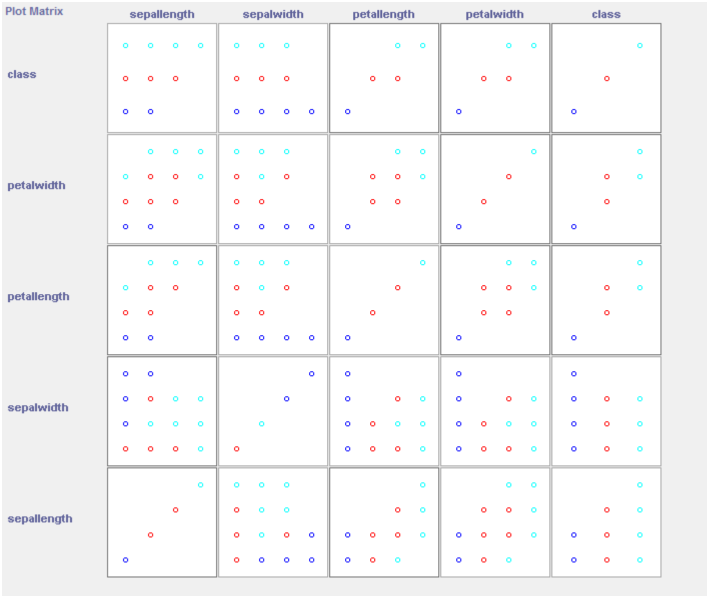

Figure 2: Simplekmeans algorithm for 3 clusters from the iris data set using seed 10- 4 bins

Exercise 2 Association analysis

3 and 4 Clusters- 3 bins

Now I have been asked to use the apriori algorithm on the data set to perform some evaluation on rules base on the characteristics I prefer. The characteristics chosen are the following ones: (1) Support (% of cases where the rule appears out of all cases) between 0.1 and 1 with changes of 0.05 to set 15 rules and (2) usage of a confidence interval to set these rules and (3) minimum confidence for a rule of 0.9 (rule must happen in 9 out of 10 cases where the first part of the rule happens).

Output Association analysis

The minimum support got for our 15 rules is 0.3, having 45 instances with a minimum confidence of 0.9 (90%) and the number of cycles performed has been 14 in order to achieve the 15 rules.

Out of the non-ranked four item sets, 15 ranked rules were given from which I will explain the first 3: (1)The first rule says that 33% of the cases (50 instances), when the petal width is in between (-inf,0.9), the petal length is between (-inf, 2.96667) for all the cases (confidence of 1). (2) The second rule says that 33% of the cases (50 instances), when the petal length is in between (-inf, 2.96667), the petal width is between (-inf,0.9) for all the cases (confidence of 1). This is the same rule as before but on the other way round. It makes sense that if on all cases we get something, on the reverse it should be the same. (3) The third rule says that for all iris setosa class (1/3 of the instances, which is 50), petal length is between (-inf, 2.96667) for all cases (confidence of 1).

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 15 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.3 (45 instances)
```

```
Minimum metric <confidence>: 0.9
Number of cycles performed: 14

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 10

Size of set of large itemsets L(3): 5

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

Best rules found:

 1. petalwidth='(-inf-0.9]' 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 2. petallength='(-inf-2.966667]' 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 3. class=Iris-setosa 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 4. petallength='(-inf-2.966667]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 5. class=Iris-setosa 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 6. petalwidth='(-inf-0.9]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 7. petalwidth='(-inf-0.9]' class=Iris-setosa 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 8. petallength='(-inf-2.966667]' class=Iris-setosa 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 9. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
10. class=Iris-setosa 50 ==> petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
11. petalwidth='(-inf-0.9]' 50 ==> petallength='(-inf-2.966667]' class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
12. petallength='(-inf-2.966667]' 50 ==> petalwidth='(-inf-0.9]' class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
13. sepallength='(-inf-5.5]' petalwidth='(-inf-0.9]' 47 ==> petallength='(-inf-2.966667]' 47    <conf:(1)> lift:(3) lev:(0.21) [31] conv:(31.33)
14. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' 47 ==> petalwidth='(-inf-0.9]' 47    <conf:(1)> lift:(3) lev:(0.21) [31] conv:(31.33)
15. sepallength='(-inf-5.5]' class=Iris-setosa 47 ==> petallength='(-inf-2.966667]' 47    <conf:(1)> lift:(3) lev:(0.21) [31] conv:(31.33)
```

3 Clusters- 4 bins explanation

Now I have been asked to use the apriori algorithm on the data set to perform some evaluation on rules base on the characteristics I prefer. The characteristics chosen are the following ones: (1) Support (% of cases where the rule appears out of all cases) between 0.1 and 1 with changes of 0.05 to set 15 rules and (2) usage of a confidence interval to set these rules and (3) minimum confidence for a rule of 0.9 (rule must happen in 9 out of 10 cases where the first part of the rule happens).

Output Association analysis

The minimum support got for our 15 rules is 0.25, having 37 instances with a minimum confidence of 0.9 (90%) and the number of cycles performed has been 15 in order to achieve the 15 rules.

Out of the non-ranked four item sets, 15 ranked rules were given from which I will explain the first 3: (1)The first rule says that 33% of the cases (50 instances), when the petal width is in between (-inf,-0.7), the petal length is between (-inf, -2.475) for all the cases (confidence of 1). (2) The second rule says that 33% of the cases (50 instances), when the petal length is in between (-inf, -2.475), the petal width is between (-inf,-0.7) for all the cases (confidence of 1). This is the same rule as before but on the other way round. It makes sense that if on all cases we get something, on the reverse it should be the same. (3) The third rule says that for all iris setosa class (1/3 of the instances, which is 50), petal length is between (-inf, -2.475) for all cases (confidence of 1).

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 15 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B4-M-1.0-Rfirst-last
```

```
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.25 (37 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 8

Size of set of large itemsets L(3): 4

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

Best rules found:

 1. petalwidth='(-inf-0.7]' 50 ==> petallength='(-inf-2.475]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 2. petallength='(-inf-2.475]' 50 ==> petalwidth='(-inf-0.7]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 3. class=Iris-setosa 50 ==> petallength='(-inf-2.475]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 4. petallength='(-inf-2.475]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 5. class=Iris-setosa 50 ==> petalwidth='(-inf-0.7]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 6. petalwidth='(-inf-0.7]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 7. petalwidth='(-inf-0.7]' class=Iris-setosa 50 ==> petallength='(-inf-2.475]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 8. petallength='(-inf-2.475]' class=Iris-setosa 50 ==> petalwidth='(-inf-0.7]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 9. petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
10. class=Iris-setosa 50 ==> petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
11. petalwidth='(-inf-0.7]' 50 ==> petallength='(-inf-2.475]' class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
12. petallength='(-inf-2.475]' 50 ==> petalwidth='(-inf-0.7]' class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
13. sepallength='(-inf-5.2]' petalwidth='(-inf-0.7]' 39 ==> petallength='(-inf-2.475]' 39    <conf:(1)> lift:(3) lev:(0.17) [26] conv:(26)
14. sepallength='(-inf-5.2]' petallength='(-inf-2.475]' 39 ==> petalwidth='(-inf-0.7]' 39    <conf:(1)> lift:(3) lev:(0.17) [26] conv:(26)
15. sepallength='(-inf-5.2]' class=Iris-setosa 39 ==> petallength='(-inf-2.475]' 39    <conf:(1)> lift:(3) lev:(0.17) [26] conv:(26)
```

3 Clusters- 4 bins comparison

Rules obtained are really similar. The only thing that changes is the interval being smaller. This makes sense, since when using the apriori function the sets always work for any subsets so we get the same rules but being the intervals smaller (e.g. the subsets being smaller than the set). All in all, the rules for 3 bins are better since they are more global and general.

Exercise 3 Describing clustering through association analysis

3 Clusters-3 bins

In this section a new variable called "cluster" has been created using the process filter "addClusterfilter" with the information previously used for clustering (with the algorithm simplekmeans) to run again and the apriori algorithm which outputs the data below:

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 15 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10-I5
Instances:    150
Attributes:   6
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
              cluster
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.3 (45 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16

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

Size of set of large itemsets L(3): 14

Size of set of large itemsets L(4): 6

Size of set of large itemsets L(5): 1

Best rules found:

 1. petalwidth='(-inf-0.9]' 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 2. petallength='(-inf-2.966667]' 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 3. class=Iris-setosa 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 4. petallength='(-inf-2.966667]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 5. cluster=cluster3 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 6. petallength='(-inf-2.966667]' 50 ==> cluster=cluster3 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 7. class=Iris-setosa 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 8. petalwidth='(-inf-0.9]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 9. cluster=cluster3 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
10. petalwidth='(-inf-0.9]' 50 ==> cluster=cluster3 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
11. cluster=cluster3 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
12. class=Iris-setosa 50 ==> cluster=cluster3 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
13. petalwidth='(-inf-0.9]' class=Iris-setosa 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
14. petallength='(-inf-2.966667]' class=Iris-setosa 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
15. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
```

Explanation

By creating one extra variable from our cluster, it is intended to find new rules with our cluster, to see whether our clustering is useful or not. Nevertheless, the output rules are still really similar comparing it to ones shown before. An example of this is that the three first rules shown before haven't changed. This might mean that these rules are really reliable. On the other hand, new rules have arisen which takes into account our variable "cluster" (i.e. 5,6,9,10,11,12). Rules 5 and 6 are going to be explained: (Rule 5) From those classified as cluster 3 (which are 50 instances, so 1/3 of the dataset) , all of them have each petallength='(-inf-2.966667] meaning that the Confidence is 1. (Rule 6) From those who have petallength='(-inf-2.966667] (which are 50 instances, so 1/3 of the dataset) , all of them have been classified as cluster 3 meaning that the Confidence is 1. (Exactly the same rule as before but the other way round). The visualization of the classification with the cluster can be also viewed below:

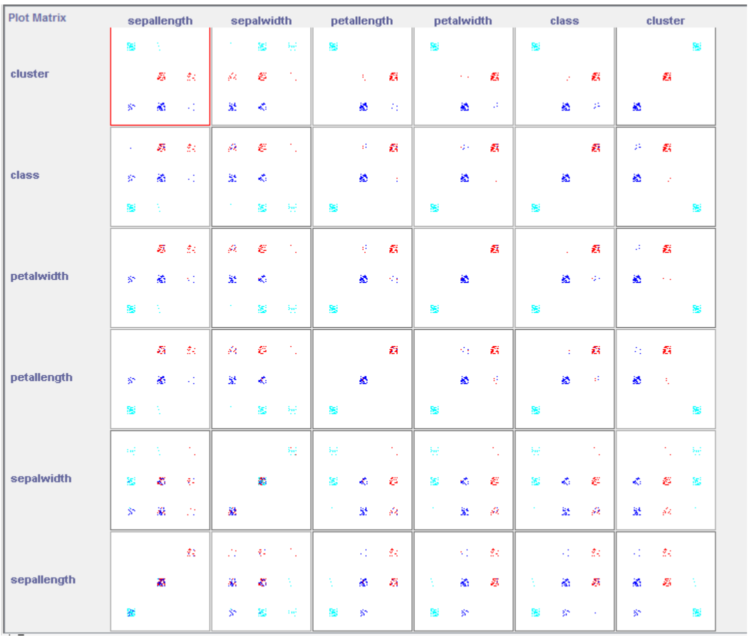

Figure 3: Simplekmeans algorithm for 3 clusters from the iris data set using seed 10 with the variable cluster-3 bins

Trying to associate rules to all clusters

In order to find rules for all clusters, it has been necessary to get 50 rules with minimum confidence equals to 0.6. Also, to ease the finding these I have set the "car" parameter to true and set the class index to my cluster index which is 6. The following result has been got:

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 50 -T 0 -C 0.6 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -A -c 6
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-weka.
    filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-
    candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.
    core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10-I5
Instances:    150
Attributes:   6
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
              cluster
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.2 (30 instances)
Minimum metric <confidence>: 0.6
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14

Size of set of large itemsets L(2): 20

Size of set of large itemsets L(3): 15

Size of set of large itemsets L(4): 6
```

```

Size of set of large itemsets L(5): 1

Best rules found:

 1. petallength='(-inf-2.966667]' 50 ==> cluster=cluster3 50    conf:(1)
 2. petalwidth='(-inf-0.9]' 50 ==> cluster=cluster3 50    conf:(1)
 3. class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 4. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 50 ==> cluster=cluster3 50    conf:(1)
 5. petallength='(-inf-2.966667]' class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 6. petalwidth='(-inf-0.9]' class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 7. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 8. petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' 48 ==> cluster=cluster1 48    conf:(1)
 9. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' 47 ==> cluster=cluster3 47    conf:(1)
10. sepallength='(-inf-5.5]' petalwidth='(-inf-0.9]' 47 ==> cluster=cluster3 47    conf:(1)
11. sepallength='(-inf-5.5]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
12. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 47 ==> cluster=cluster3 47    conf:(1)
13. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
14. sepallength='(-inf-5.5]' petalwidth='(-inf-0.9]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
15. petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' class=Iris-versicolor 47 ==> cluster=cluster1 47    conf:(1)
16. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
17. petallength='(4.933333-inf)' petalwidth='(1.7-inf)' 40 ==> cluster=cluster2 40    conf:(1)
18. petallength='(4.933333-inf)' petalwidth='(1.7-inf)' class=Iris-virginica 40 ==> cluster=cluster2 40    conf:(1)
19. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' 36 ==> cluster=cluster3 36    conf:(1)
20. sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' 36 ==> cluster=cluster3 36    conf:(1)
21. sepalwidth='(2.8-3.6]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
22. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' 36 ==> cluster=cluster3 36    conf:(1)
23. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' 36 ==> cluster=cluster3 36    conf:(1)
24. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
25. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 36 ==> cluster=cluster3 36    conf:(1)
26. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
27. sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
28. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 36 ==> cluster=cluster3 36    conf:(1)
29. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
30. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
31. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
32. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
33. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' 33 ==> cluster=cluster1 33    conf:(1)
34. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' class=Iris-versicolor 33 ==> cluster=cluster1 33    conf:(1)
35. sepalwidth='(-inf-2.8]' petalwidth='(0.9-1.7]' 31 ==> cluster=cluster1 31    conf:(1)
36. petalwidth='(0.9-1.7]' class=Iris-versicolor 49 ==> cluster=cluster1 48    conf:(0.98)
37. petallength='(2.966667-4.933333]' class=Iris-versicolor 48 ==> cluster=cluster1 47    conf:(0.98)
38. sepallength='(5.5-6.7]' petalwidth='(0.9-1.7]' 38 ==> cluster=cluster1 37    conf:(0.97)
39. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' 37 ==> cluster=cluster3 36    conf:(0.97)
40. sepallength='(5.5-6.7]' petalwidth='(0.9-1.7]' class=Iris-versicolor 35 ==> cluster=cluster1 34    conf:(0.97)
41. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' class=Iris-versicolor 34 ==> cluster=cluster1 33    conf:(0.97)
42. petalwidth='(0.9-1.7]' 54 ==> cluster=cluster1 52    conf:(0.96)
43. class=Iris-versicolor 50 ==> cluster=cluster1 48    conf:(0.96)
44. petallength='(2.966667-4.933333]' 54 ==> cluster=cluster1 51    conf:(0.94)
45. sepallength='(5.5-6.7]' class=Iris-versicolor 36 ==> cluster=cluster1 34    conf:(0.94)
46. petalwidth='(1.7-inf)' 46 ==> cluster=cluster2 43    conf:(0.93)
47. petalwidth='(1.7-inf)' class=Iris-virginica 45 ==> cluster=cluster2 42    conf:(0.93)
48. petallength='(4.933333-inf)' class=Iris-virginica 44 ==> cluster=cluster2 41    conf:(0.93)
49. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' 39 ==> cluster=cluster1 36    conf:(0.92)
50. petallength='(4.933333-inf)' 46 ==> cluster=cluster2 42    conf:(0.91)

t
```

I am going to comment a rule for each cluster: Rule 1.When petal length is in between '(-inf-2.966667]', which happens to be in 50 cases (1/3 of the data) in all 50 cases (confidence 1) they belong to cluster 3.

Rule 46.When petal width is in between '(1.7-inf)', which happens to be in 46 cases (almost 1/3 of the data) in almost all cases (43 out of 46 being confidence 0.93) they belong to cluster 2. Rule 38.When sepal length is in between '(5.5-6.7]' and petal width between (0.9-1.7], which happens to be in 38 cases (almost 1/4 of the data) in almost all cases (37 out of 38 being confidence 0.97) they belong to cluster 1.

4 Clusters-3 bins

The same has been done as above but with the variable of 4 clusters, leading to the following output:

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 15 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10-I5
Instances:    150
Attributes:   6
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
              cluster
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.3 (45 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 15

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

Size of set of large itemsets L(3): 13

Size of set of large itemsets L(4): 5

Size of set of large itemsets L(5): 1

Best rules found:

 1. petalwidth='(-inf-0.9]' 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 2. petallength='(-inf-2.966667]' 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 3. class=Iris-setosa 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 4. petallength='(-inf-2.966667]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 5. cluster=cluster3 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 6. petallength='(-inf-2.966667]' 50 ==> cluster=cluster3 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 7. class=Iris-setosa 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 8. petalwidth='(-inf-0.9]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 9. cluster=cluster3 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
10. petalwidth='(-inf-0.9]' 50 ==> cluster=cluster3 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
11. cluster=cluster3 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
12. class=Iris-setosa 50 ==> cluster=cluster3 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
13. petalwidth='(-inf-0.9]' class=Iris-setosa 50 ==> petallength='(-inf-2.966667]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
```

```
14. petallength='(-inf-2.966667]' class=Iris-setosa 50 ==> petalwidth='(-inf-0.9]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
15. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
```

Trying to associate rules to all clusters

In order to find rules for all clusters, it has been necessary to get 50 rules with minimum confidence equals to 0.9 and minimum support equals to 0.1. Also, to ease the finding these I have set the "car" parameter to true and set the class index to my cluster index which is 6. The following result has been got:

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 50 -T 0 -C 0.6 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -A -c 6
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10-I5-weka.filters.unsupervised.attribute.Remove-R6-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10-I5
Instances:    150
Attributes:   6
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
              cluster
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.2 (30 instances)
Minimum metric <confidence>: 0.6
Number of cycles performed: 16

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 20

Size of set of large itemsets L(3): 15

Size of set of large itemsets L(4): 6

Size of set of large itemsets L(5): 1

Best rules found:

 1. petallength='(-inf-2.966667]' 50 ==> cluster=cluster3 50    conf:(1)
 2. petalwidth='(-inf-0.9]' 50 ==> cluster=cluster3 50    conf:(1)
 3. class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 4. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 50 ==> cluster=cluster3 50    conf:(1)
 5. petallength='(-inf-2.966667]' class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 6. petalwidth='(-inf-0.9]' class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 7. petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-setosa 50 ==> cluster=cluster3 50    conf:(1)
 8. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' 47 ==> cluster=cluster3 47    conf:(1)
 9. sepallength='(-inf-5.5]' petalwidth='(-inf-0.9]' 47 ==> cluster=cluster3 47    conf:(1)
10. sepallength='(-inf-5.5]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
11. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 47 ==> cluster=cluster3 47    conf:(1)
12. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
13. sepallength='(-inf-5.5]' petalwidth='(-inf-0.9]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
14. sepallength='(-inf-5.5]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-setosa 47 ==> cluster=cluster3 47    conf:(1)
15. petallength='(4.933333-inf)' petalwidth='(1.7-inf)' 40 ==> cluster=cluster2 40    conf:(1)
16. petallength='(4.933333-inf)' petalwidth='(1.7-inf)' class=Iris-virginica 40 ==> cluster=cluster2 40    conf:(1)
17. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' 36 ==> cluster=cluster3 36    conf:(1)
```

```
18. sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' 36 ==> cluster=cluster3 36    conf:(1)
19. sepalwidth='(2.8-3.6]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
20. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' 36 ==>
     cluster=cluster3 36    conf:(1)
21. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' 36 ==> cluster=
     cluster3 36    conf:(1)
22. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' class=Iris-setosa 36 ==> cluster=cluster3
     36    conf:(1)
23. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' 36 ==> cluster
     =cluster3 36    conf:(1)
24. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' class=Iris-setosa 36 ==> cluster=
     cluster3 36    conf:(1)
25. sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' class=Iris-setosa 36 ==> cluster=cluster3 36
         conf:(1)
26. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='
     (-inf-0.9]' 36 ==> cluster=cluster3 36    conf:(1)
27. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' class=Iris-
     setosa 36 ==> cluster=cluster3 36    conf:(1)
28. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petalwidth='(-inf-0.9]' class=Iris-setosa
     36 ==> cluster=cluster3 36    conf:(1)
29. sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' class=Iris-
     setosa 36 ==> cluster=cluster3 36    conf:(1)
30. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' petallength='(-inf-2.966667]' petalwidth='
     (-inf-0.9]' class=Iris-setosa 36 ==> cluster=cluster3 36    conf:(1)
31. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' 33 ==>
     cluster=cluster1 33    conf:(1)
32. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' class=Iris
     -versicolor 33 ==> cluster=cluster1 33    conf:(1)
33. sepalwidth='(-inf-2.8]' petalwidth='(0.9-1.7]' 31 ==> cluster=cluster1 31    conf:(1)
34. sepallength='(5.5-6.7]' petalwidth='(0.9-1.7]' 38 ==> cluster=cluster1 37    conf:(0.97)
35. sepallength='(-inf-5.5]' sepalwidth='(2.8-3.6]' 37 ==> cluster=cluster3 36    conf:(0.97)
36. sepallength='(5.5-6.7]' petalwidth='(0.9-1.7]' class=Iris-versicolor 35 ==> cluster=
     cluster1 34    conf:(0.97)
37. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' class=Iris-versicolor 34 ==>
     cluster=cluster1 33    conf:(0.97)
38. sepallength='(5.5-6.7]' class=Iris-versicolor 36 ==> cluster=cluster1 34    conf:(0.94)
39. petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' 48 ==> cluster=cluster1 45    conf
     :(0.94)
40. petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' class=Iris-versicolor 47 ==>
     cluster=cluster1 44    conf:(0.94)
41. petalwidth='(1.7-inf)' 46 ==> cluster=cluster2 43    conf:(0.93)
42. petalwidth='(1.7-inf)' class=Iris-virginica 45 ==> cluster=cluster2 42    conf:(0.93)
43. sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' 39 ==> cluster=cluster1 36
     conf:(0.92)
44. petalwidth='(0.9-1.7]' class=Iris-versicolor 49 ==> cluster=cluster1 45    conf:(0.92)
45. petallength='(2.966667-4.933333]' class=Iris-versicolor 48 ==> cluster=cluster1 44    conf
     :(0.92)
46. petallength='(4.933333-inf)' class=Iris-virginica 44 ==> cluster=cluster2 40    conf:(0.91)
47. petalwidth='(0.9-1.7]' 54 ==> cluster=cluster1 49    conf:(0.91)
48. class=Iris-versicolor 50 ==> cluster=cluster1 45    conf:(0.9)
49. petallength='(4.933333-inf)' 46 ==> cluster=cluster2 41    conf:(0.89)
50. petallength='(2.966667-4.933333]' 54 ==> cluster=cluster1 48    conf:(0.89)
```

I am going to comment a rule for each cluster: Rule 1.When petal length is in between '(-inf-2.966667]', which happens to be in 50 cases (1/3 of the data) in all 50 cases (confidence 1) they belong to cluster 3.

Rule 41.When petal width is in between '(1.7-inf)', which happens to be in 46 cases (almost 1/3 of the data) in almost all cases (43 out of 46 being confidence 0.93) they belong to cluster 2.

Rule 34.When sepal length is in between '(5.5-6.7]' and petal width between (0.9-1.7], which happens to be in 38 cases (almost 1/4 of the data) in almost all cases (37 out of 38 being confidence 0.97) they belong to cluster 1. No rules for the fourth cluster has been found. The minimum support should be lowered more, but then we run into the problem that these rules will not be general enough, so that I will not dig more into it.

Explanation and comparison of the 4 clusters with the 3 cluster variables

Giving the fact that the clusters are almost the same and that the new cluster does not have instances enough to overcome the minimum support which is 0.3 and 0.2 respectively, the rules are exactly as the same above, as expected. For the new extra cluster, no rules are got given the fact that its support is really really low (less than 0.1) and I am not interested in lowering it more because we will find too many rules which are not general enough and will difficult our search. Therefore, the same rules are got since the new cluster got is really small and

3 Clusters-4 bins

The same has been done as above but with the variable of 3 clusters and 4 bins, leading to the following output:

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 15 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B4-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10-I5
Instances:    150
Attributes:   6
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
              cluster
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.3 (45 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 10

Size of set of large itemsets L(3): 4

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

Best rules found:

 1. petalwidth='(-inf-0.7]' 50 ==> petallength='(-inf-2.475]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 2. petallength='(-inf-2.475]' 50 ==> petalwidth='(-inf-0.7]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 3. class=Iris-setosa 50 ==> petallength='(-inf-2.475]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 4. petallength='(-inf-2.475]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 5. petallength='(-inf-2.475]' 50 ==> cluster=cluster1 50    <conf:(1)> lift:(2.78) lev:(0.21) [31] conv:(32)
 6. class=Iris-setosa 50 ==> petalwidth='(-inf-0.7]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 7. petalwidth='(-inf-0.7]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
 8. petalwidth='(-inf-0.7]' 50 ==> cluster=cluster1 50    <conf:(1)> lift:(2.78) lev:(0.21) [31] conv:(32)
 9. class=Iris-setosa 50 ==> cluster=cluster1 50    <conf:(1)> lift:(2.78) lev:(0.21) [31] conv:(32)
10. petalwidth='(-inf-0.7]' class=Iris-setosa 50 ==> petallength='(-inf-2.475]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
11. petallength='(-inf-2.475]' class=Iris-setosa 50 ==> petalwidth='(-inf-0.7]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
12. petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' 50 ==> class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
13. class=Iris-setosa 50 ==> petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
14. petalwidth='(-inf-0.7]' 50 ==> petallength='(-inf-2.475]' class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
15. petallength='(-inf-2.475]' 50 ==> petalwidth='(-inf-0.7]' class=Iris-setosa 50    <conf:(1)> lift:(3) lev:(0.22) [33] conv:(33.33)
```

Trying to associate rules to all clusters

In order to find rules for all clusters, it has been necessary to get 50 rules with minimum confidence equals to 0.9 and minimum support equals to 0.1. Also, to ease the finding these I have set the "car" parameter to true and set the class index to my cluster index which is 6. The following result has been got:

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 50 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -A -c 6
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B4-M-1.0-Rfirst-last-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10-I5
Instances:    150
Attributes:   6
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
              cluster
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (15 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

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

Size of set of large itemsets L(2): 43

Size of set of large itemsets L(3): 34

Size of set of large itemsets L(4): 11

Size of set of large itemsets L(5): 2

Best rules found:

 1. petallength='(-inf-2.475]' 50 ==> cluster=cluster1 50    conf:(1)
 2. petalwidth='(-inf-0.7]' 50 ==> cluster=cluster1 50    conf:(1)
 3. class=Iris-setosa 50 ==> cluster=cluster1 50    conf:(1)
 4. petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' 50 ==> cluster=cluster1 50    conf:(1)
 5. petallength='(-inf-2.475]' class=Iris-setosa 50 ==> cluster=cluster1 50    conf:(1)
 6. petalwidth='(-inf-0.7]' class=Iris-setosa 50 ==> cluster=cluster1 50    conf:(1)
 7. petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' class=Iris-setosa 50 ==> cluster=cluster1 50    conf:(1)
 8. sepallength='(-inf-5.2]' petallength='(-inf-2.475]' 39 ==> cluster=cluster1 39    conf:(1)
 9. sepallength='(-inf-5.2]' petalwidth='(-inf-0.7]' 39 ==> cluster=cluster1 39    conf:(1)
10. sepallength='(-inf-5.2]' class=Iris-setosa 39 ==> cluster=cluster1 39    conf:(1)
11. petallength='(3.95-5.425]' class=Iris-versicolor 39 ==> cluster=cluster2 39    conf:(1)
12. sepallength='(-inf-5.2]' petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' 39 ==> cluster=cluster1 39    conf:(1)
13. sepallength='(-inf-5.2]' petallength='(-inf-2.475]' class=Iris-setosa 39 ==> cluster=cluster1 39    conf:(1)
14. sepallength='(-inf-5.2]' petalwidth='(-inf-0.7]' class=Iris-setosa 39 ==> cluster=cluster1 39    conf:(1)
15. sepallength='(-inf-5.2]' petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' class=Iris-setosa 39 ==> cluster=cluster1 39    conf:(1)
16. petallength='(3.95-5.425]' petalwidth='(1.3-1.9]' 33 ==> cluster=cluster2 33    conf:(1)
17. sepallength='(5.2-6.1]' petallength='(3.95-5.425]' 32 ==> cluster=cluster2 32    conf:(1)
18. sepalwidth='(2.6-3.2]' class=Iris-versicolor 32 ==> cluster=cluster2 32    conf:(1)
19. sepallength='(5.2-6.1]' class=Iris-versicolor 29 ==> cluster=cluster2 29    conf:(1)
20. sepalwidth='(2.6-3.2]' petallength='(3.95-5.425]' class=Iris-versicolor 29 ==> cluster=cluster2 29    conf:(1)
21. sepallength='(5.2-6.1]' sepalwidth='(2.6-3.2]' 26 ==> cluster=cluster2 26    conf:(1)
22. sepalwidth='(3.2-3.8]' petallength='(-inf-2.475]' 26 ==> cluster=cluster1 26    conf:(1)
23. sepalwidth='(3.2-3.8]' petalwidth='(-inf-0.7]' 26 ==> cluster=cluster1 26    conf:(1)
24. sepalwidth='(3.2-3.8]' class=Iris-setosa 26 ==> cluster=cluster1 26    conf:(1)
25. sepalwidth='(2.6-3.2]' petallength='(3.95-5.425]' petalwidth='(1.3-1.9]' 26 ==> cluster=cluster2 26    conf:(1)
26. sepalwidth='(3.2-3.8]' petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' 26 ==> cluster=cluster1 26    conf:(1)
27. sepalwidth='(3.2-3.8]' petallength='(-inf-2.475]' class=Iris-setosa 26 ==> cluster=cluster1 26    conf:(1)
28. sepalwidth='(3.2-3.8]' petalwidth='(-inf-0.7]' class=Iris-setosa 26 ==> cluster=cluster1 26    conf:(1)
29. sepalwidth='(3.2-3.8]' petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' class=Iris-setosa 26 ==> cluster=cluster1 26    conf:(1)
30. sepallength='(5.2-6.1]' sepalwidth='(2.6-3.2]' petallength='(3.95-5.425]' 24 ==> cluster=cluster2 24    conf:(1)
31. sepallength='(5.2-6.1]' petallength='(3.95-5.425]' class=Iris-versicolor 23 ==> cluster=cluster2 23    conf:(1)
32. petalwidth='(1.3-1.9]' class=Iris-versicolor 22 ==> cluster=cluster2 22    conf:(1)
33. petallength='(3.95-5.425]' petalwidth='(1.3-1.9]' class=Iris-versicolor 21 ==> cluster=cluster2 21    conf:(1)
34. sepallength='(-inf-5.2]' sepalwidth='(3.2-3.8]' 20 ==> cluster=cluster1 20    conf:(1)
```

```
35. sepallength='(5.2-6.1]' petalwidth='(0.7-1.3]' 20 ==> cluster=cluster2 20    conf:(1)
36. sepallength='(-inf-5.2]' sepalwidth='(3.2-3.8]' petallength='(-inf-2.475]' 20 ==> cluster=
    cluster1 20    conf:(1)
37. sepallength='(-inf-5.2]' sepalwidth='(3.2-3.8]' petalwidth='(-inf-0.7]' 20 ==> cluster=
    cluster1 20    conf:(1)
38. sepallength='(-inf-5.2]' sepalwidth='(3.2-3.8]' class=Iris-setosa 20 ==> cluster=cluster1
    20    conf:(1)
39. sepallength='(5.2-6.1]' petalwidth='(0.7-1.3]' class=Iris-versicolor 20 ==> cluster=
    cluster2 20    conf:(1)
40. sepallength='(-inf-5.2]' sepalwidth='(3.2-3.8]' petallength='(-inf-2.475]' petalwidth='(-
    inf-0.7]' 20 ==> cluster=cluster1 20    conf:(1)
41. sepallength='(-inf-5.2]' sepalwidth='(3.2-3.8]' petallength='(-inf-2.475]' class=Iris-
    setosa 20 ==> cluster=cluster1 20    conf:(1)
42. sepallength='(-inf-5.2]' sepalwidth='(3.2-3.8]' petalwidth='(-inf-0.7]' class=Iris-setosa
    20 ==> cluster=cluster1 20    conf:(1)
43. sepallength='(-inf-5.2]' sepalwidth='(3.2-3.8]' petallength='(-inf-2.475]' petalwidth='(-
    inf-0.7]' class=Iris-setosa 20 ==> cluster=cluster1 20    conf:(1)
44. petallength='(5.425-inf)' petalwidth='(1.9-inf)' 19 ==> cluster=cluster3 19    conf:(1)
45. sepallength='(5.2-6.1]' sepalwidth='(2.6-3.2]' class=Iris-versicolor 19 ==> cluster=
    cluster2 19    conf:(1)
46. petallength='(5.425-inf)' petalwidth='(1.9-inf)' class=Iris-virginica 19 ==> cluster=
    cluster3 19    conf:(1)
47. sepallength='(6.1-7]' petalwidth='(1.9-inf)' 18 ==> cluster=cluster3 18    conf:(1)
48. petallength='(3.95-5.425]' petalwidth='(0.7-1.3]' 18 ==> cluster=cluster2 18    conf:(1)
49. sepallength='(6.1-7]' petalwidth='(1.9-inf)' class=Iris-virginica 18 ==> cluster=cluster3
    18    conf:(1)
50. sepalwidth='(2.6-3.2]' petalwidth='(1.3-1.9]' class=Iris-versicolor 18 ==> cluster=cluster2
     18    conf:(1)
```

I am going to comment a rule for each cluster: Rule 1.When petal length is in between '(-inf-2.475]', which happens to be in 50 cases (1/3 of the data) in all 50 cases (confidence 1) they belong to cluster 3.

Rule 34.When sepal length is in between '(-inf-5.2]' and petal width between (3.2-3.8], which happens to be in 20 cases in all cases (confidence 1) they belong to cluster 1.

Rule 35.When sepal length is in between '(5.2-6.1]' and petal width between (0.9-1.3], which happens to be in 20 cases in all cases (confidence 1) they belong to cluster 2.

Comparison of the 4 bins with the 3 cluster variables with 3 bins

Giving the fact that the clusters are almost equal but with subsets of the former given the fact that there are more bins, the same rules are outputted leading to similar results but more concrete. For that, association rules for 3 bins are better than those for 4 bins given that they are more general and this ones are just a subset of the previous rules. Nevertheless, in some cases they can differ a bit.