CO544 – Machine Learning and Data Mining

Lab06 – Part 3

E/17/407

WIJESOORIYA H.D

(01) Attributes and their data types

Attribute	Data Type
sepallength	Numeric
sepalwidth	Numeric
petallength	Numeric
petalwidth	Numeric
class	Nominal

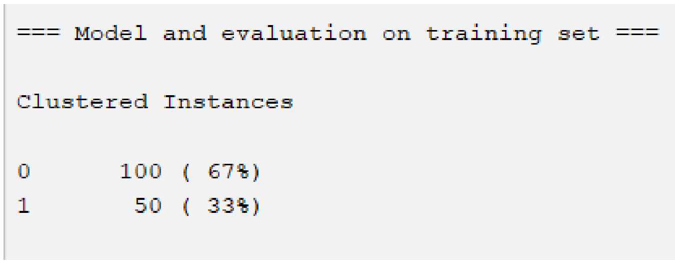
(02) parameters in 'Generic Object Editor'

- seed - a random number
- displayStdDevs - Display standard deviations of numeric attributes and counts of nominal attributes.
- numExecutionSlots - The number of execution slots (threads) to use. Set equal to the number of available cpu/cores
- numClusters - to set number of clusters
- maxIterations - to set maximum number of iterations
- preserveInstancesOrder - to preserve order of instances.
- initializationMethod - an initialization method to use. Random, k-means++, Canopy or farthest first
- distanceFunction - use for instances comparison (default: weka.core.EuclideanDistance)
- fastDistanceCalc - use cut-off values to speed up distance calculation

(03) ‘seed’ in ‘Generic Object Editor’ and the use of seed in KMeans algorithm

Seed is a random number (any integer). In KMeans algorithm seed is used as an initial K point. K point represents the number of clusters. Since the algorithm is sensitive to initial points, we have to try experimentation on the stability of your clusters with different seeds.

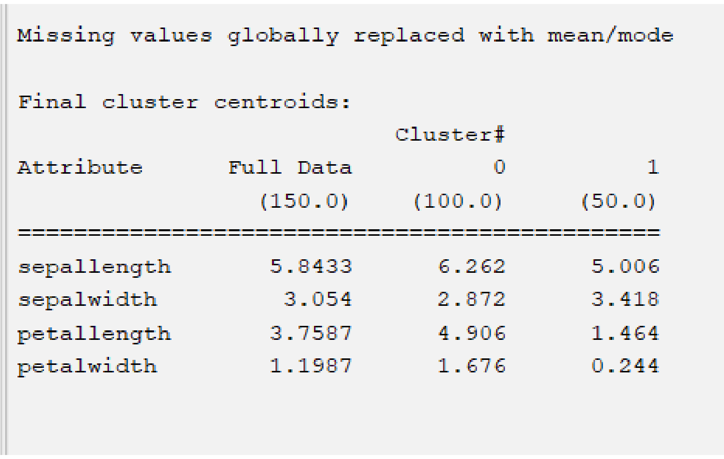
(04) observe the cluster assignments and values in each clusters



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

Clustered Instances

0       100 ( 67%)
1        50 ( 33%)
```

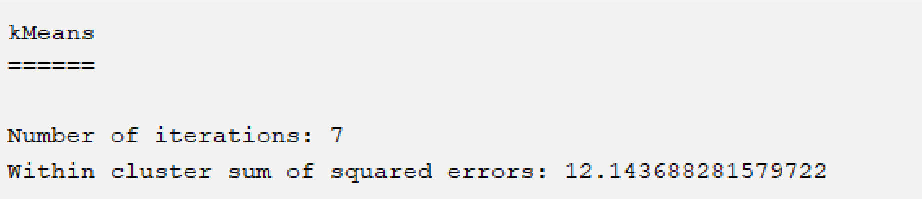
In this data set there are two clusters called ‘0’ and ‘1’.

100 instances are clustered as ‘0’ class and 50 instances are clustered as class ‘1’ . That means 67% of the instances belong to class ‘0’ and the rest (33%) belong to class ‘1’



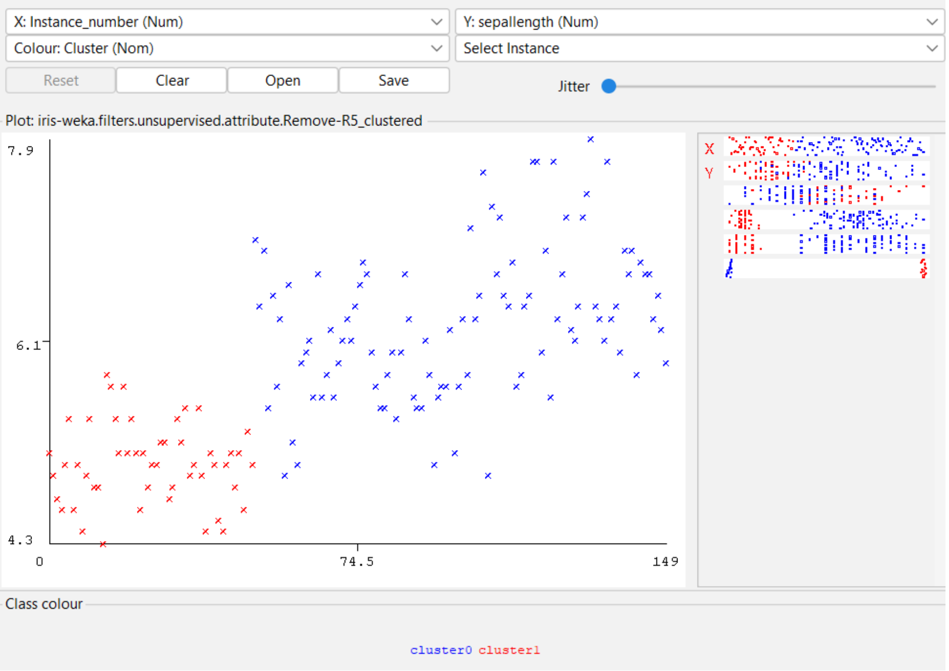
```
Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1
                 (150.0)    (100.0)     (50.0)
==============================================
sepallength       5.8433      6.262      5.006
sepalwidth         3.054      2.872      3.418
petallength       3.7587      4.906      1.464
petalwidth        1.1987      1.676      0.244
```

Sum of squared error



```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 12.143688281579722
```

(05) Visualize cluster assignments



X: Instance_number (Num)

Y: sepallength (Num)

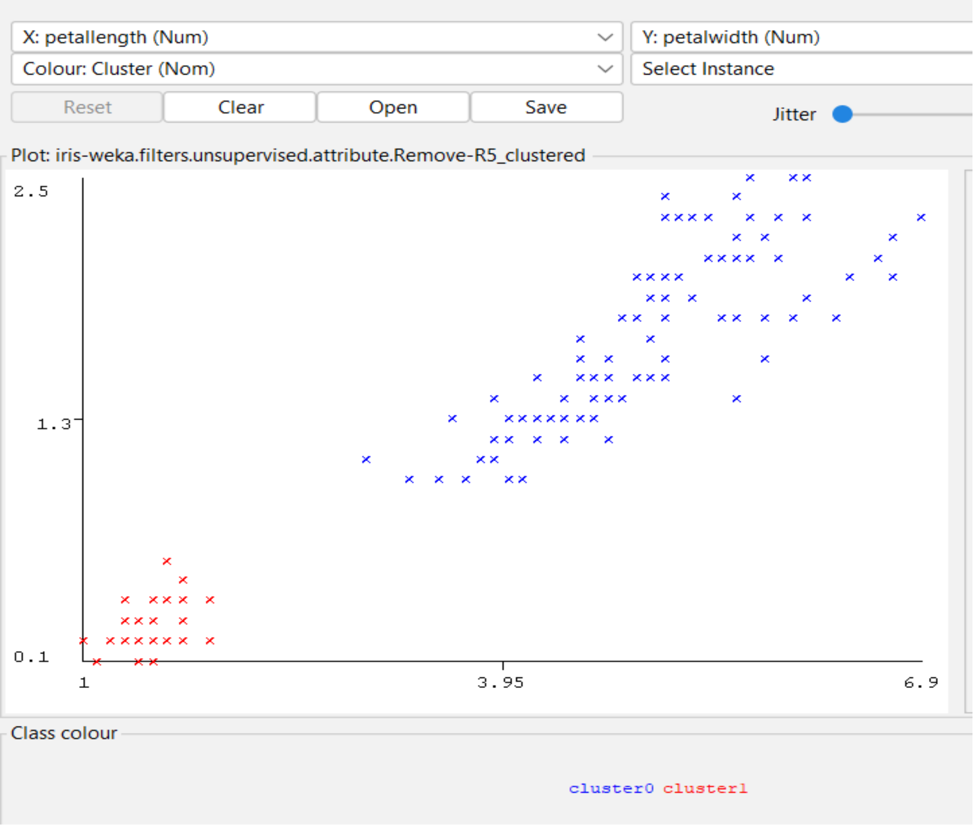
Colour: Cluster (Nom)

Select Instance

Plot: iris-weka.filters.unsupervised.attribute.Remove-R5_clustered

Class colour

Suitable labels for X and Y



X: petallength (Num)

Y: petalwidth (Num)

Colour: Cluster (Nom)

Select Instance

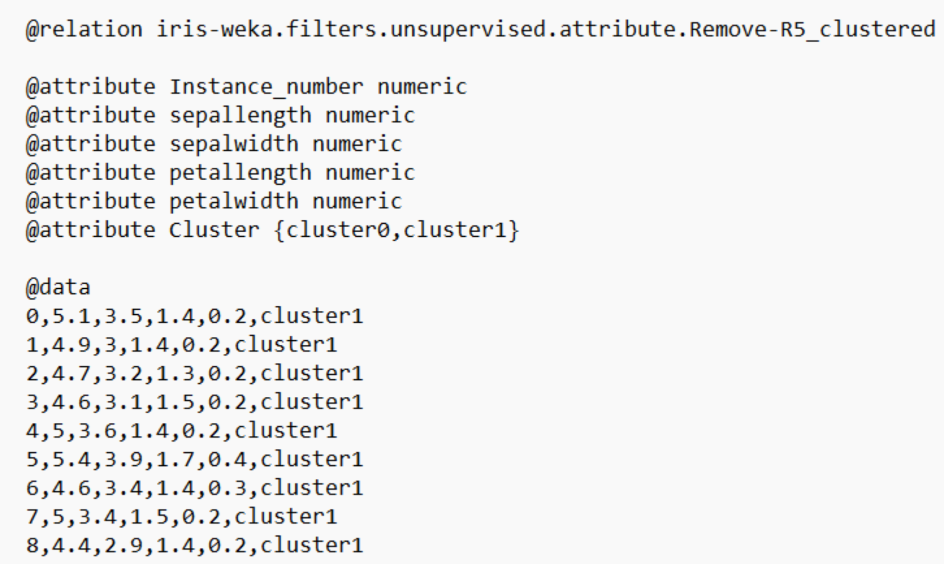
Plot: iris-weka.filters.unsupervised.attribute.Remove-R5_clustered

Class colour

Description for each cluster and observations

Here I have chosen petal length as variable X and petal width as variable Y. I have chosen it because it separated the classes clearly as shown in the above figure.(Red and blue colour instances are clearly separated, no mix)

(06) content of the ARFF file

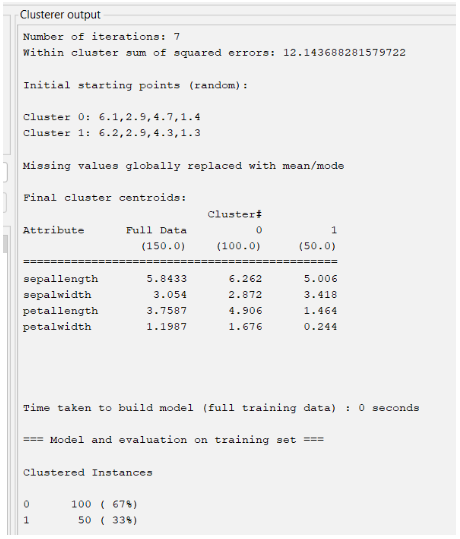


```
@relation iris-weka.filters.unsupervised.attribute.Remove-R5_clustered

@attribute Instance_number numeric
@attribute sepallength numeric
@attribute sepalwidth numeric
@attribute petallength numeric
@attribute petalwidth numeric
@attribute Cluster {cluster0,cluster1}

@data
0,5.1,3.5,1.4,0.2,cluster1
1,4.9,3,1.4,0.2,cluster1
2,4.7,3.2,1.3,0.2,cluster1
3,4.6,3.1,1.5,0.2,cluster1
4,5,3.6,1.4,0.2,cluster1
5,5.4,3.9,1.7,0.4,cluster1
6,4.6,3.4,1.4,0.3,cluster1
7,5,3.4,1.5,0.2,cluster1
8,4.4,2.9,1.4,0.2,cluster1
```

In the ARFF file we can see 150 data records (rows) . Each column (attribute) of data is separated by a ‘,’ and they are correspond to the attributes Instance_number , sepallength , sepalwidth , petallength , petalwidth and cluster . Here the column ‘cluster’ is the target and the rest of the columns are features. There are two clusters , cluster1 and cluster0.

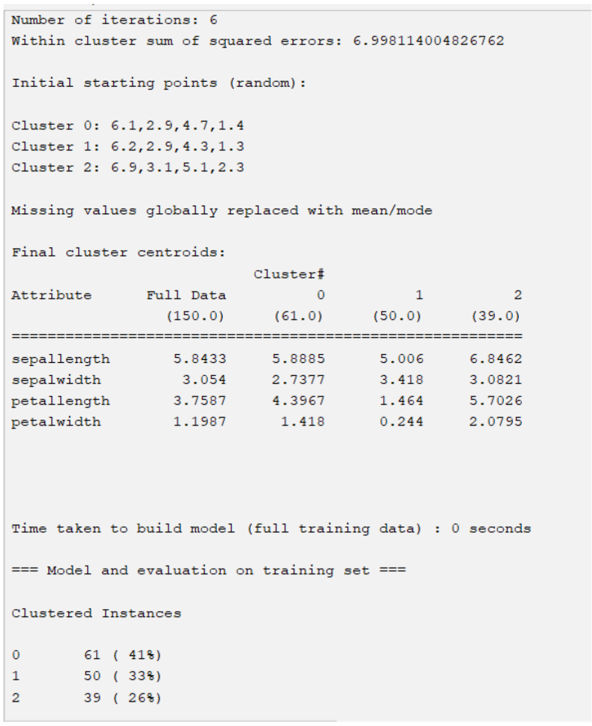
(07) suggest suitable value for k (2<=k<=5) – optimal number of clusters

K=2



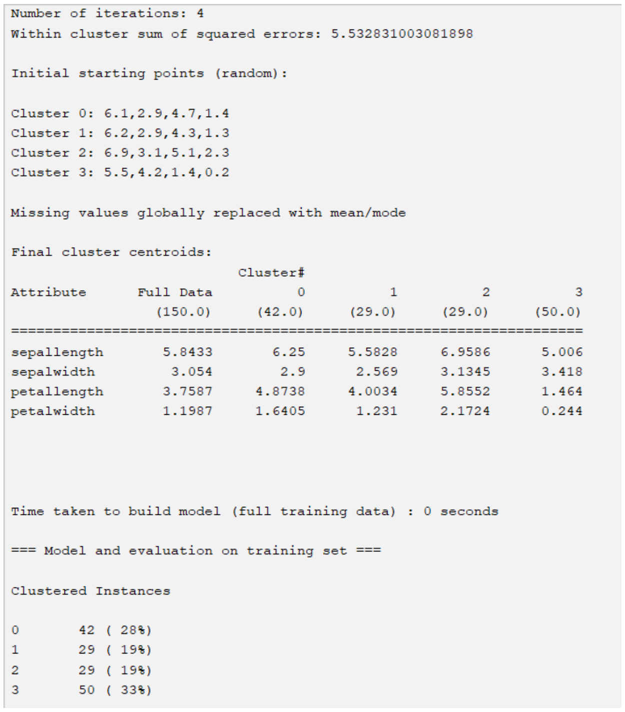
```
Number of iterations: 7
Within cluster sum of squared errors: 12.143688281579722

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1
                 (150.0)    (100.0)     (50.0)
==============================================
sepallength       5.8433      6.262      5.006
sepalwidth         3.054      2.872      3.418
petallength       3.7587      4.906      1.464
petalwidth        1.1987      1.676      0.244




Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0      100 ( 67%)
1       50 ( 33%)
```

K=3



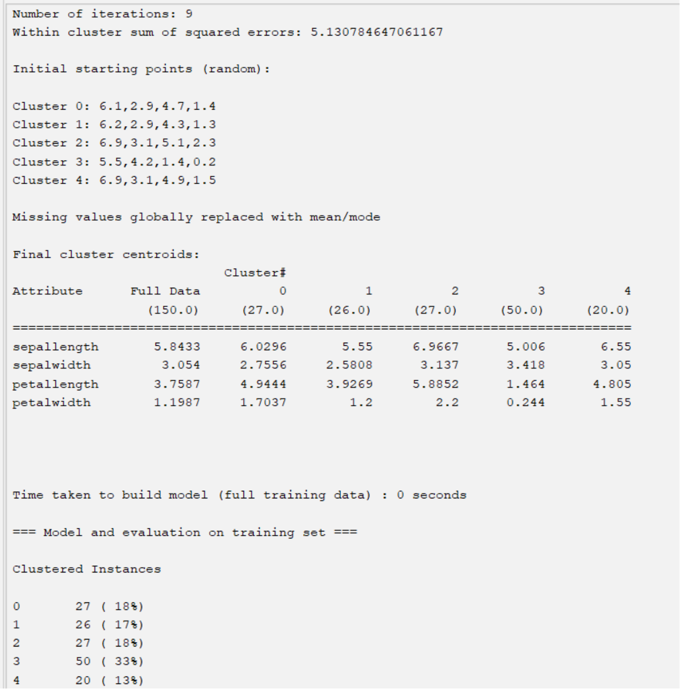
```
Number of iterations: 6
Within cluster sum of squared errors: 6.998114004826762

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3
Cluster 2: 6.9,3.1,5.1,2.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2
                 (150.0)     (61.0)     (50.0)     (39.0)
==========================================================
sepallength       5.8433     5.8885      5.006     6.8462
sepalwidth         3.054     2.7377      3.418     3.0821
petallength       3.7587     4.3967      1.464     5.7026
petalwidth        1.1987      1.418      0.244     2.0795




Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```

K=4



```
Number of iterations: 4
Within cluster sum of squared errors: 5.532831003081898

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3
Cluster 2: 6.9,3.1,5.1,2.3
Cluster 3: 5.5,4.2,1.4,0.2

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2          3
                 (150.0)     (42.0)     (29.0)     (29.0)     (50.0)
=====================================================================
sepallength       5.8433       6.25     5.5828     6.9586      5.006
sepalwidth         3.054        2.9      2.569     3.1345      3.418
petallength       3.7587     4.8738     4.0034     5.8552      1.464
petalwidth        1.1987     1.6405      1.231     2.1724      0.244




Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       42 ( 28%)
1       29 ( 19%)
2       29 ( 19%)
3       50 ( 33%)
```

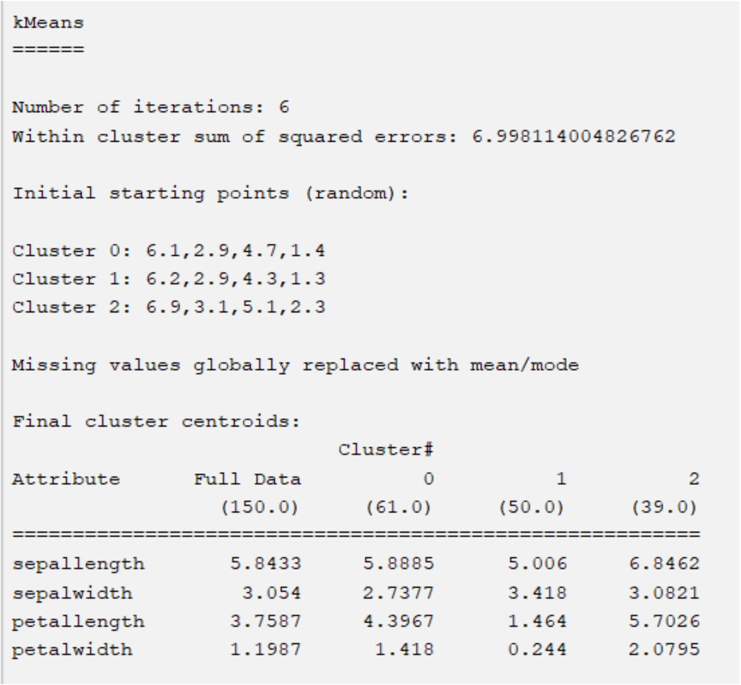
K=5



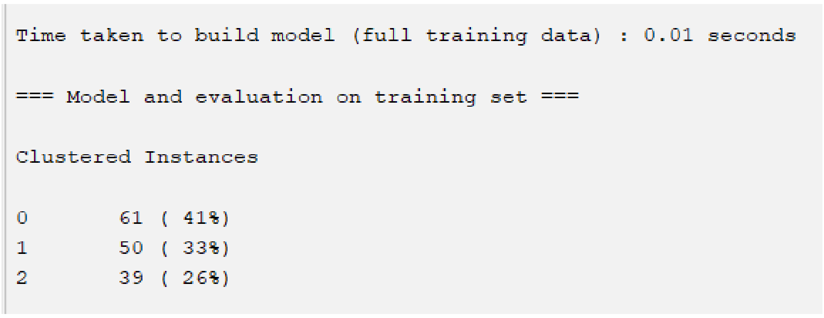
```
Number of iterations: 9
Within cluster sum of squared errors: 5.130784647061167

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3
Cluster 2: 6.9,3.1,5.1,2.3
Cluster 3: 5.5,4.2,1.4,0.2
Cluster 4: 6.9,3.1,4.9,1.5

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2          3          4
                 (150.0)     (27.0)     (26.0)     (27.0)     (50.0)     (20.0)
================================================================================
sepallength       5.8433     6.0296       5.55     6.9667      5.006       6.55
sepalwidth         3.054     2.7556     2.5808      3.137      3.418       3.05
petallength       3.7587     4.9444     3.9269     5.8852      1.464      4.805
petalwidth        1.1987     1.7037        1.2        2.2      0.244       1.55




Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       27 ( 18%)
1       26 ( 17%)
2       27 ( 18%)
3       50 ( 33%)
4       20 ( 13%)
```

To find the optimal number of clusters we can use the ‘elbow’ method. According to the above figures it is clear that the elbow point is at k=3.Therefore the optimal number of clusters is 3.

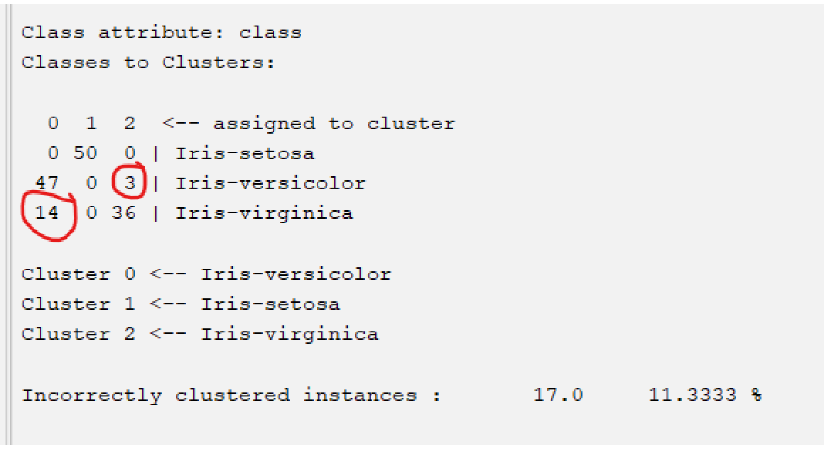
(08) Class to cluster evaluation option (when k=3)



```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 6.998114004826762

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3
Cluster 2: 6.9,3.1,5.1,2.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2
                 (150.0)     (61.0)     (50.0)     (39.0)
=========================================================
sepallength       5.8433     5.8885      5.006     6.8462
sepalwidth         3.054     2.7377      3.418     3.0821
petallength       3.7587     4.3967      1.464     5.7026
petalwidth        1.1987      1.418      0.244     2.0795
```



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

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

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```



```
Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 47  0  3 | Iris-versicolor
 14  0 36 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :       17.0     11.3333 %
```

According to the above figures we can see that there are 3 main clusters. The assigned class values for each cluster are shown bellow.

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

There were 150 instances and out of them 17 instances were misclassified. 3 instances which belong to Iris-versicolor were misclassified as Iris-virginica . 14 actual Iris-virginica instances were incorrectly classified as Iris-versicolor.