EX : 8 Applying k-means clustering on a given data set

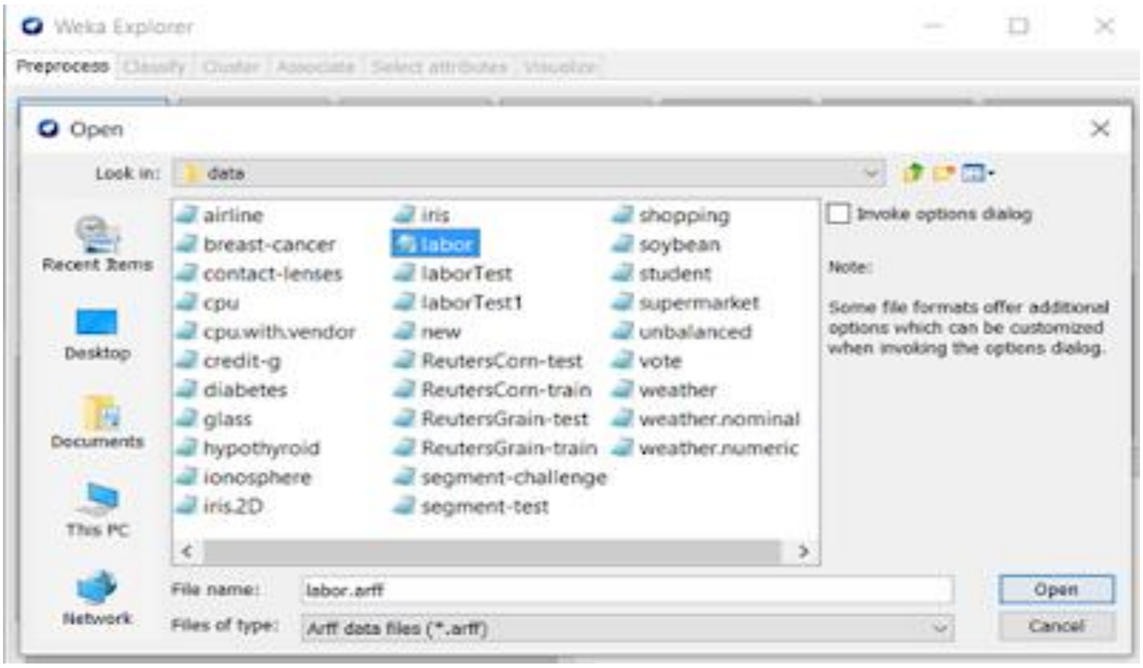
Procedure:

Step1: Open the data file in Weka Explorer. It is presumed that the required data fields have been discretized. In this example it is age attribute.

Step2: Clicking on the associate tab will bring up the interface for association rule algorithm.

Step3: We will use K-means algorithm. This is the default algorithm.

Step4: Inorder to change the parameters for the run (example support, confidence etc) we click on the text box immediately to the right of the choose button.

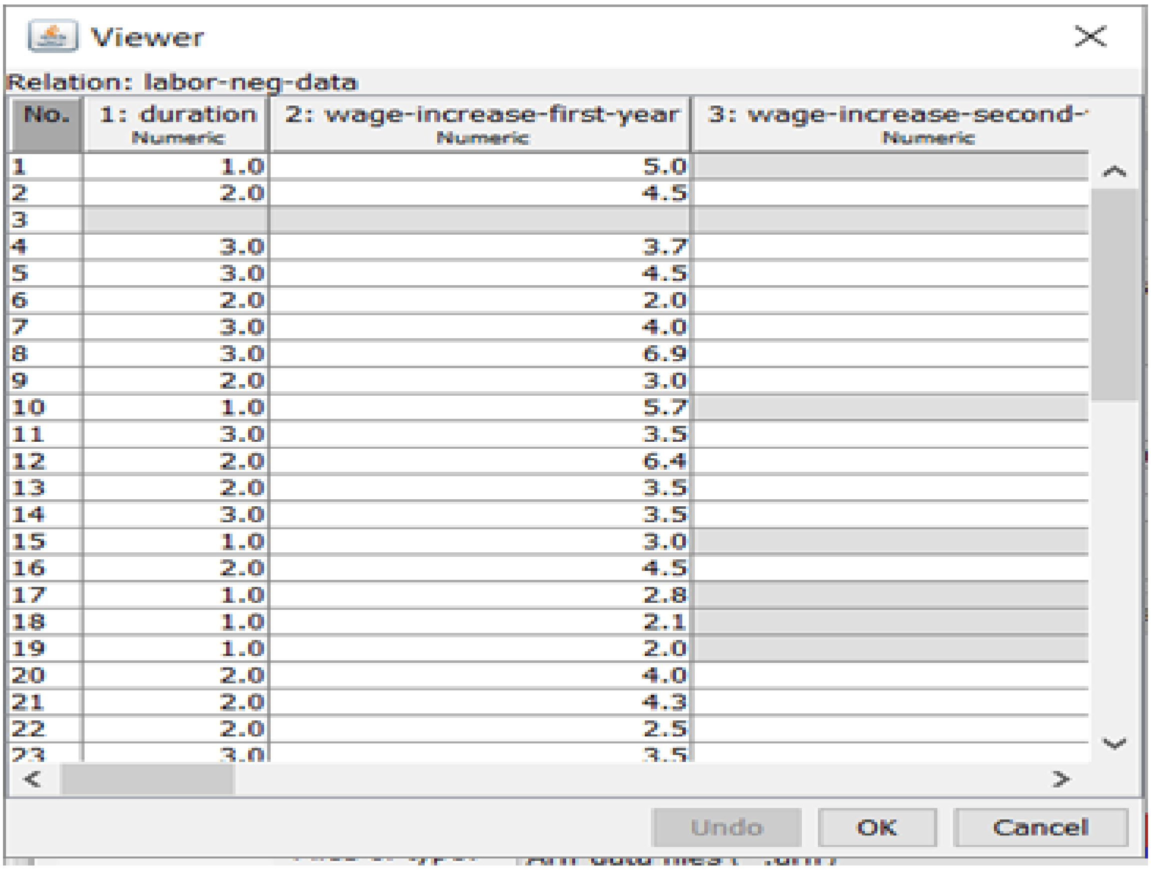


Note:

Some file formats offer additional options which can be customized when invoking the options dialog.

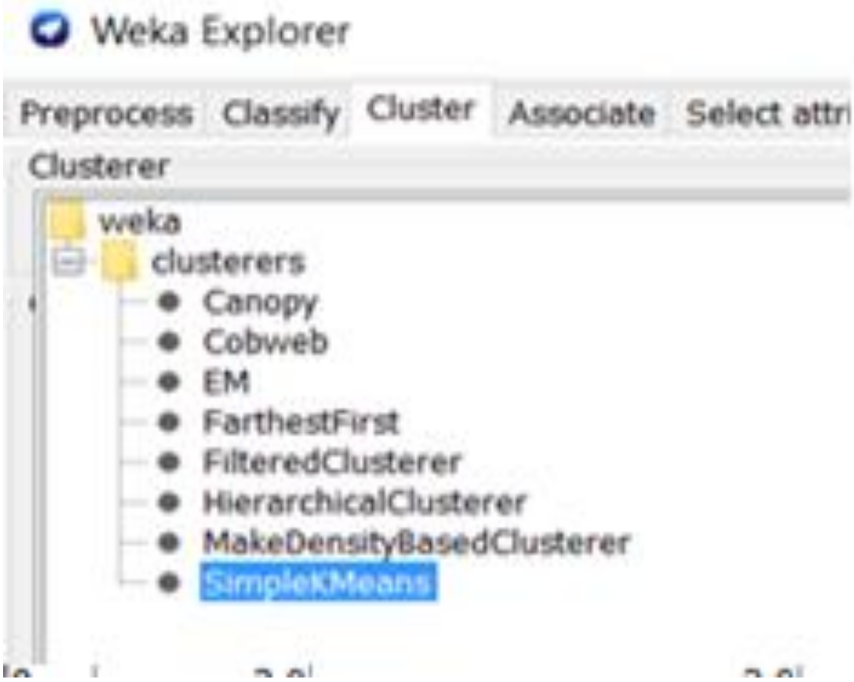
File name: labor.arff

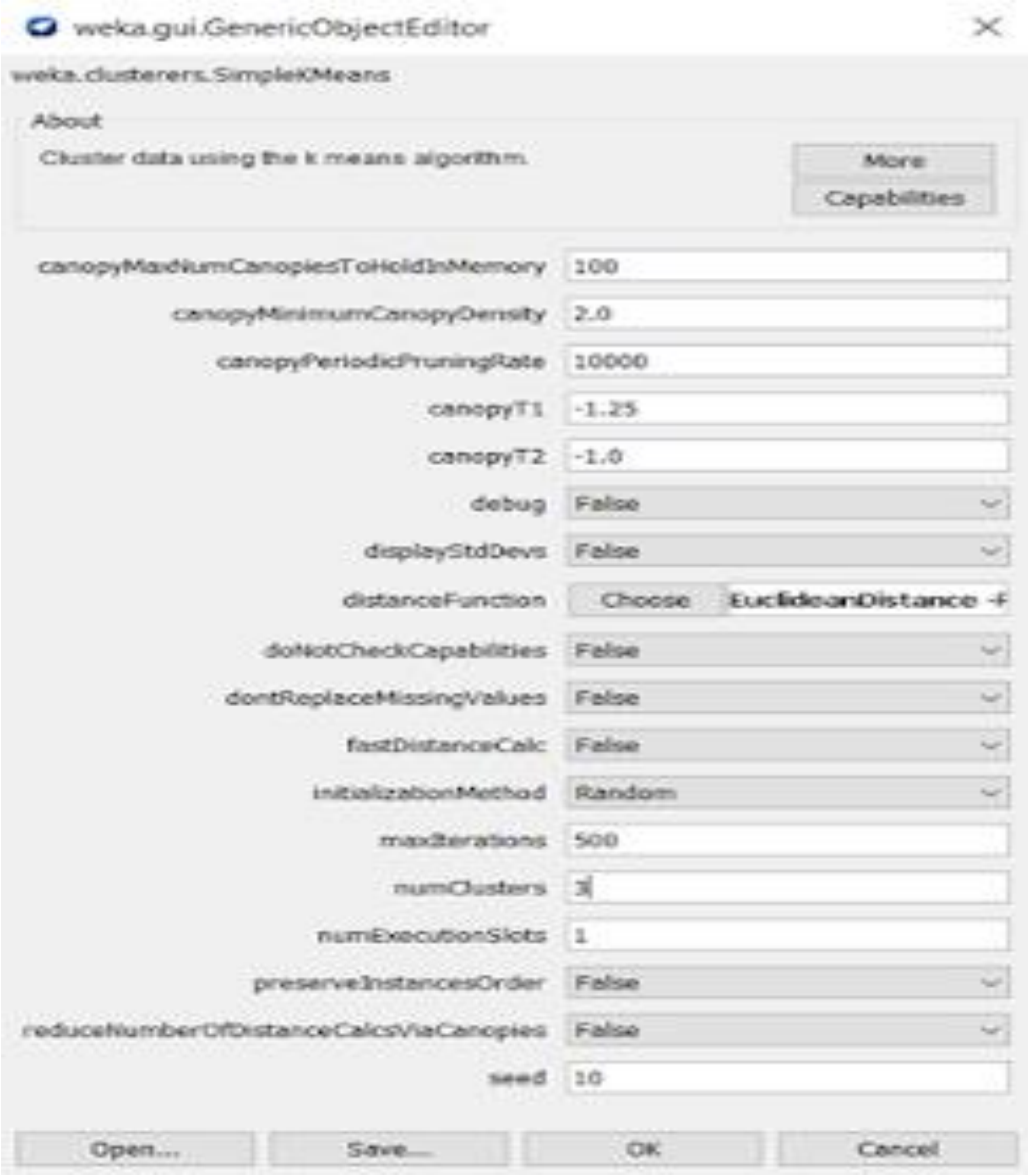
Files of type: Arff data files (*.arff)



Relation: labor-neg-data

No.	1: duration Numeric	2: wage-increase-first-year Numeric	3: wage-increase-second- Numeric
1	1.0	5.0	
2	2.0	4.5	
3			
4	3.0	3.7	
5	3.0	4.5	
6	2.0	2.0	
7	3.0	4.0	
8	3.0	6.9	
9	2.0	3.0	
10	1.0	5.7	
11	3.0	3.5	
12	2.0	6.4	
13	2.0	3.5	
14	3.0	3.5	
15	1.0	3.0	
16	2.0	4.5	
17	1.0	2.8	
18	1.0	2.1	
19	1.0	2.0	
20	2.0	4.0	
21	2.0	4.3	
22	2.0	2.5	
23	3.0	3.5	





weka.clusterers.SimpleKMeans

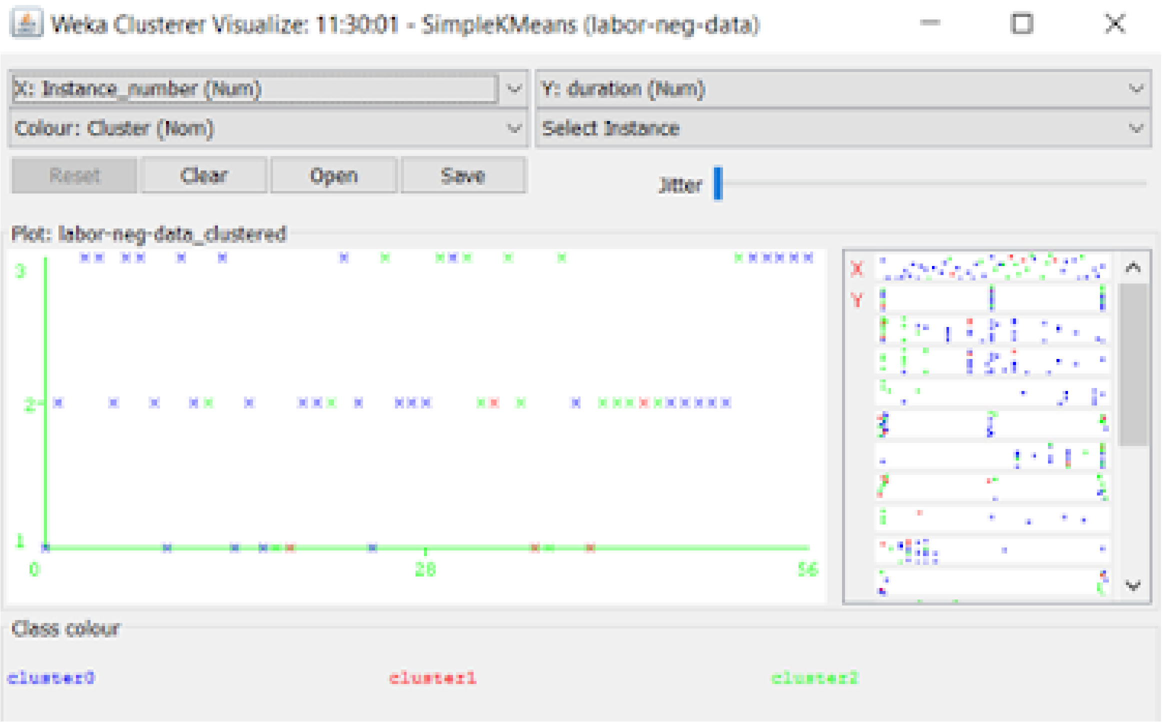
About

Cluster data using the k means algorithm.

canopyMaxNumCanopiesToHoldInMemory	100
canopyMinimumCanopyDensity	2.0
canopyPeriodicPruningRate	10000
canopyT1	-1.25
canopyT2	-1.0
debug	False
displayStdDevs	False
distanceFunction	EuclideanDistance -F
doNotCheckCapabilities	False
dontReplaceMissingValues	False
fastDistanceCalc	False
initializationMethod	Random
maxIterations	500
numClusters	3
numExecutionSlots	1
preserveInstancesOrder	False
reduceNumberOfDistanceCalcsViaCanopies	False
seed	10

```
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:     labor-neg-data
Instances:    57
Attributes:   17
              duration
              wage-increase-first-year
              wage-increase-second-year
              wage-increase-third-year
              cost-of-living-adjustment
              working-hours
              pension
              standby-pay
              shift-differential
              education-allowance
              statutory-holidays
              vacation
              longterm-disability-assistance
              contribution-to-dental-plan
              bereavement-assistance
              contribution-to-health-plan
              class


Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 119.5224194214812

Initial starting points (random):

Cluster 0: 1,5.7,3.971739,3.913333,none,40,empl_contr,7.444444,4,no,11,generous,yes,full,yes,full,good
Cluster 1: 1,2,3.971739,3.913333,tc,40,ret_allw,4,0,no,11,generous,no,none,no,none,bad
Cluster 2: 2,2.5,3,3.913333,tcf,40,none,7.444444,4.870968,no,11,below_average,yes,half,yes,full,bad

Missing values globally replaced with mean/mode
```

```
Final cluster centroids:
                                    Cluster#
Attribute                    Full Data          0          1          2
                             (57.0)      (36.0)       (5.0)      (16.0)
========================================================================
=======
duration                        2.1607       2.2267         1.4        2.25
wage-increase-first-year            3.8036       4.4695         3.2       2.4938
wage-increase-second-year             3.9717       4.4175       4.183       2.9027
wage-increase-third-year             3.9133       4.1093       3.9133       3.4725
cost-of-living-adjustment              none         none         none         none
working-hours                    38.0392      37.4766      39.2078       38.94
pension                      empl_contr   empl_contr        none   empl_contr
standby-pay                      7.4444       7.9938       6.7556       6.4236
shift-differential                4.871        5.4776       3.1484       4.0444
education-allowance                   no           no           no           no
statutory-holidays                11.0943      11.4801        10.6      10.3809
vacation                   below_average     generous below_average below_average
longterm-disability-assistance          yes          yes          no          yes
contribution-to-dental-plan           half         half        none         half
bereavement-assistance               yes          yes          no          yes
contribution-to-health-plan           full         full        none         full
class                            good         good         bad          bad

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

=== Model and evaluation on training set ===
Clustered Instances

0      36 ( 63%)
1       5 (  9%)
2      16 ( 28%)
```



Weka Clusterer Visualize: 11:30:01 - SimpleKMeans (labor-neg-data)

X: Instance_number (Num)

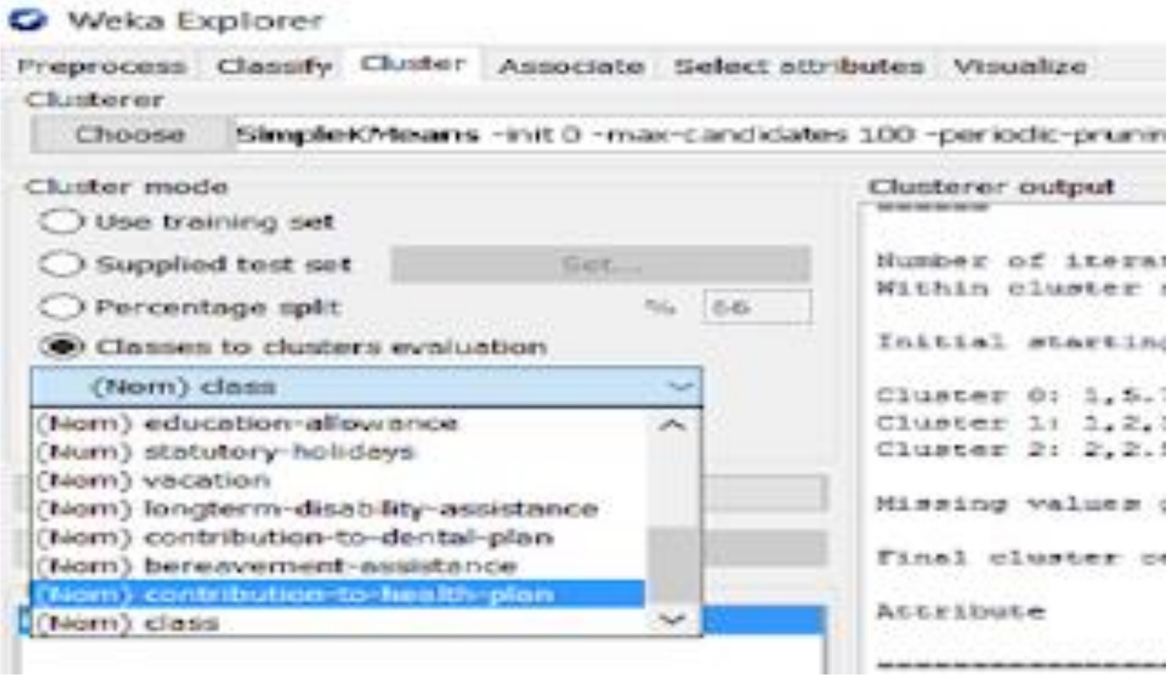
Y: duration (Num)

Colour: Cluster (Nom)

Select Instance

Plot: labor-neg-data_clustered

Class colour



Weka Explorer

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose

SimpleKMeans -init 0 -max-candidates 100 -periodic-prunin

Cluster mode

○ Use training set

○ Supplied test set

○ Percentage split % 66

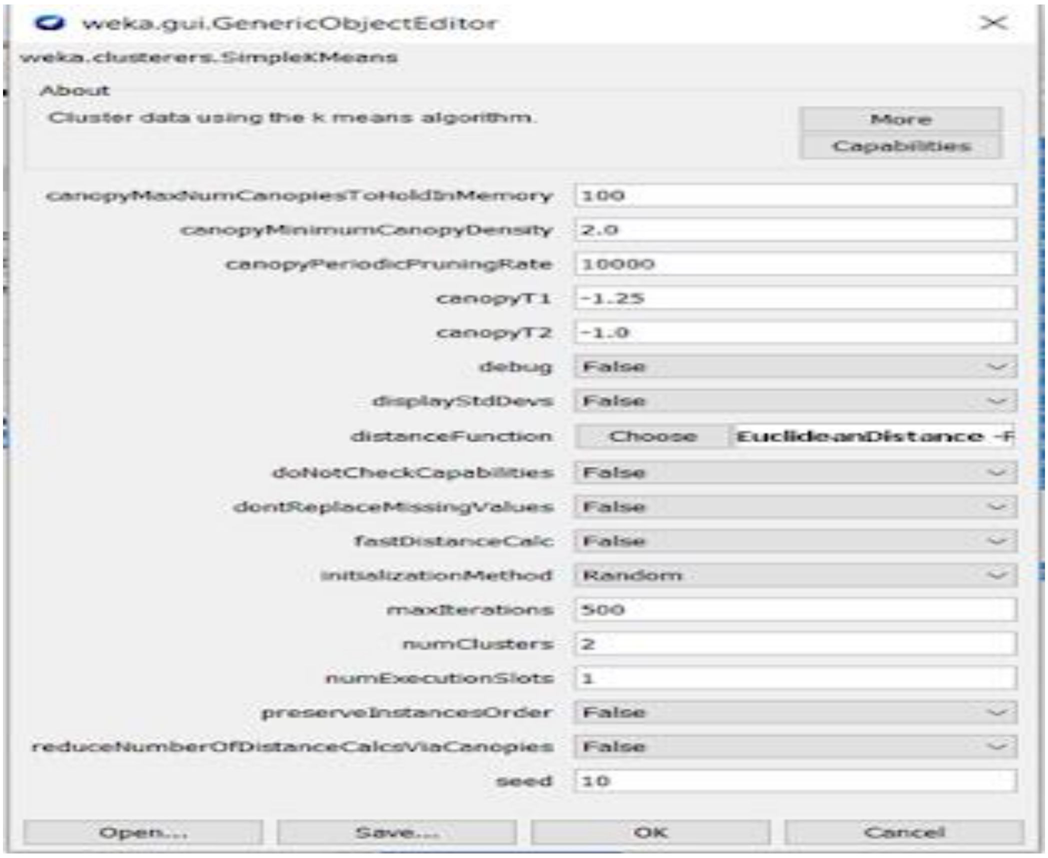
● Classes to clusters evaluation

(Nom) class

(Nom) education-allowance
(Num) statutory-holidays
(Nom) vacation
(Nom) longterm-disability-assistance
(Nom) contribution-to-dental-plan
(Nom) bereavement-assistance
(Nom) contribution-to-health-plan
(Nom) class
Set...

Clusterer output

```
Number of iterat
Within cluster s

Initial startin

Cluster 0: 1,5.
Cluster 1: 1,2,
Cluster 2: 2,2.

Missing values

Final cluster c

Attribute

----------------
```



weka.clusterers.SimpleKMeans

About

Cluster data using the k means algorithm.

canopyMaxNumCanopiesToHoldInMemory	100
canopyMinimumCanopyDensity	2.0
canopyPeriodicPruningRate	10000
canopyT1	-1.25
canopyT2	-1.0
debug	False
displayStdDevs	False
distanceFunction	EuclideanDistance -R
doNotCheckCapabilities	False
dontReplaceMissingValues	False
fastDistanceCalc	False
initializationMethod	Random
maxIterations	500
numClusters	2
numExecutionSlots	1
preserveInstancesOrder	False
reduceNumberOfDistanceCalcsViaCanopies	False
seed	10

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: labor-neg-data

Instances: 57

Attributes: 17

duration

wage-increase-first-year

wage-increase-second-year

wage-increase-third-year

cost-of-living-adjustment

working-hours

pension

standby-pay

shift-differential

education-allowance

statutory-holidays

vacation

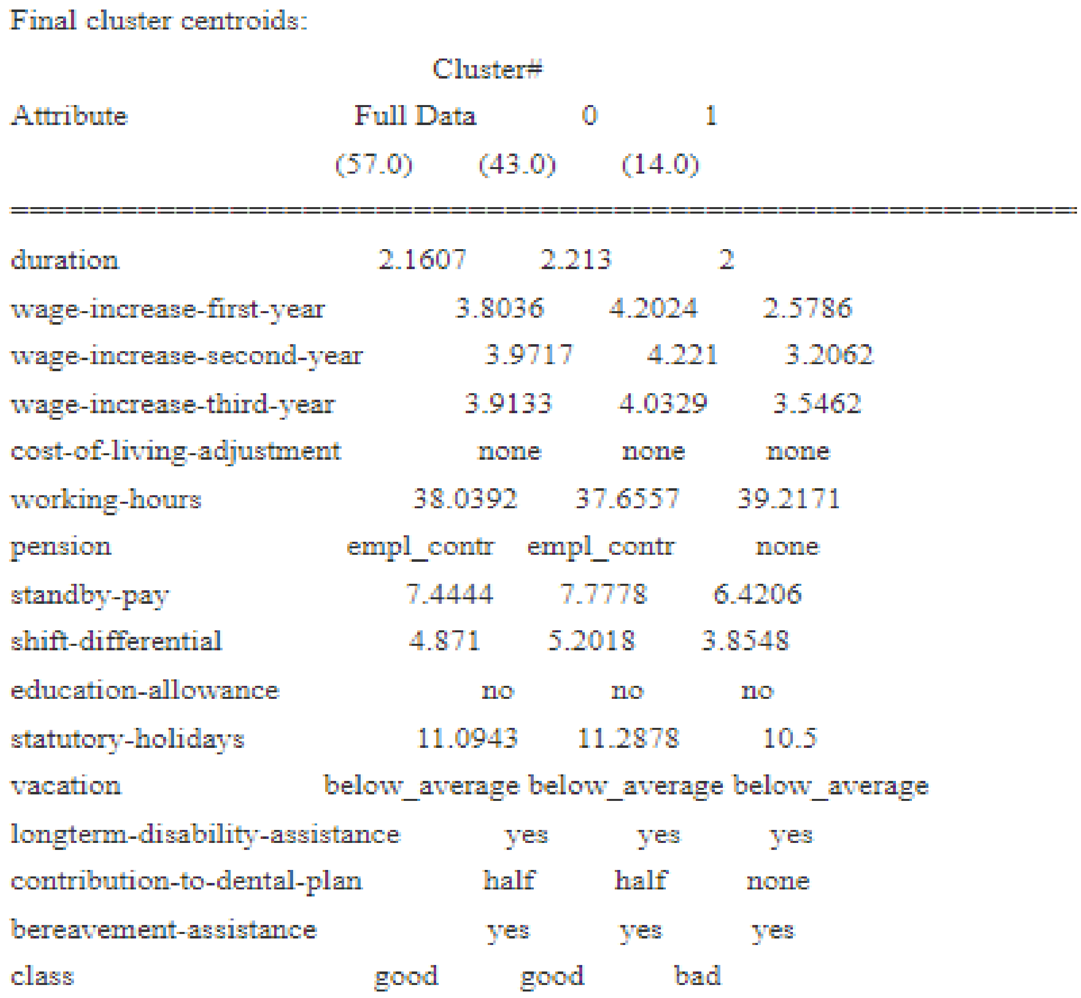
longterm-disability-assistance

contribution-to-dental-plan

bereavement-assistance

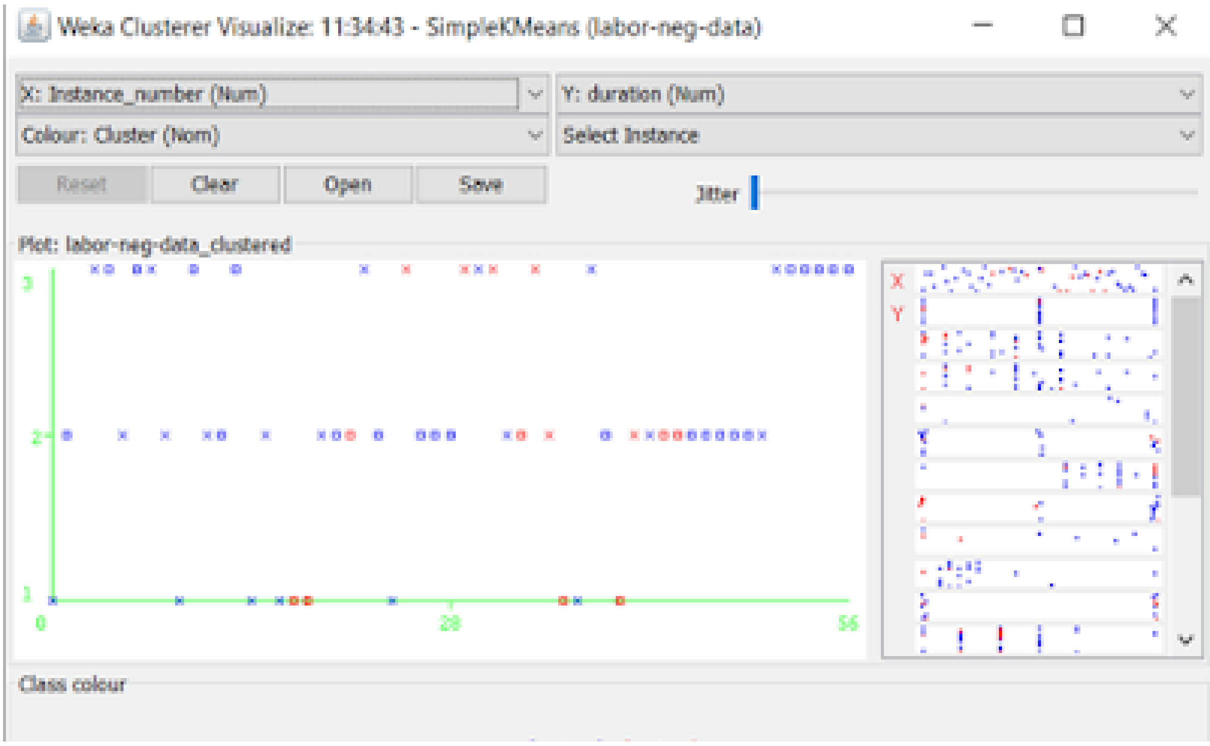
class

```
Ignored:
            contribution-to-health-plan
Test mode:    Classes to clusters evaluation on training data
=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 122.05464734126849

Initial starting points (random):

Cluster 0: 1,5.7,3.971739,3.913333,none,40,empl_contr,7.444444,4,no,11,generous,yes,full,yes,good
Cluster 1: 1,2,3.971739,3.913333,tc,40,ret_allw,4,0,no,11,generous,no,none,no,bad

Missing values globally replaced with mean/mode
```



```
Final cluster centroids:
                                        Cluster#
Attribute                       Full Data          0           1
                              (57.0)       (43.0)       (14.0)
==================================================================
duration                          2.1607        2.213           2
wage-increase-first-year               3.8036       4.2024       2.5786
wage-increase-second-year                3.9717        4.221       3.2062
wage-increase-third-year               3.9133       4.0329       3.5462
cost-of-living-adjustment                none         none         none
working-hours                        38.0392      37.6557      39.2171
pension                         empl_contr    empl_contr         none
standby-pay                          7.4444       7.7778       6.4206
shift-differential                   4.871        5.2018       3.8548
education-allowance                     no           no           no
statutory-holidays                   11.0943      11.2878        10.5
vacation                      below_average below_average below_average
longterm-disability-assistance            yes          yes          yes
contribution-to-dental-plan              half         half         none
bereavement-assistance                    yes          yes          yes
class                               good        good         bad
```

```
Time taken to build model (full training data) : 0 seconds
=== Model and evaluation on training set ===
Clustered Instances

0      43 ( 75%)
1      14 ( 25%)

Class attribute: contribution-to-health-plan
Classes to Clusters:

  0  1  <-- assigned to cluster
 20  8 | none
  9  0 | half
 14  6 | full

Cluster 0 <-- none
Cluster 1 <-- full

Incorrectly clustered instances :          31.0       54.386  %
```



Weka Clusterer Visualize: 11:34:43 - SimpleKMeans (labor-neg-data)

X: Instance_number (Num)

Y: duration (Num)

Colour: Cluster (Nom)

Select Instance

Plot: labor-neg-data_clustered