Sanjivani Rural Education Society's

College of Engineering, Kopargaon-423603

DEPARTMENT OF COMPUTER ENGINEERING

Instruction No. 0 CL-III/EL-IIIC/ Sr. No. B-C4 Rev 00 Date: 01/03/09	**Title: EL-I(DMW) :Visualize the clusters using suitable tool (WEKA)**

Aim:

Visualize the clusters using suitable tool (WEKA)

Problem Definition/Objective:

Consider a suitable dataset. For clustering of data instances in different groups, apply different clustering techniques (minimum 2). Visualize the clusters using suitable tool (WEKA)

Input:CSV Dataset

Output: Visualization of the clusters.

Prerequisites:

1. Need java installed in the system.
2. Weka tool in system.

Relevant Theory:

CLUSTERING: - Clustering is a task of assigning a set of objects into groups called as clusters. Clustering is also referred as cluster analysis where the objects in the same cluster are more similar to each other than to those objects in other clusters.
Clustering is the main task of Explorative Data mining and is a common technique for statistical data analysis used in many fields like machine learning, pattern recognition, image analysis, bio informatics etc…
Cluster analysis is not an algorithm but is a general task to be solved.
Clustering is of different types like hierarchical clustering which creates a hierarchy of clusters, partial clustering, and spectral clustering.

SimpleK-Means: -

It is a method of cluster analysis called as partial cluster analysis or partial clustering.
K-Means clustering partition or divides **n** observations into K clusters.
Each observation belongs to the cluster with the nearest mean.
K-means clustering is an algorithm to group the objects based on attributes/features into K number of groups where K is positive integer.
K-Means clustering is used in different types of applications like pattern recognition, artificial intelligent, image processing, etc…
Now Open the WEKA GUI Chooser from start menu◊all programs and click on the EXPLORER button.

Procedure:

Now click on the **Open File** button and choose the file named as "cluster.csv" where the content of cluster.csv is as shown in the figure 1.

customer ID	age	income	student	credit rating	class By computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	Yes
12	middle	medium	no	excellent	Yes
13	middle	high	yes	fair	Yes
14	senior	medium	no	excellent	No

FIGURE 1: INPUT FILE (CLUSTER.CSV)

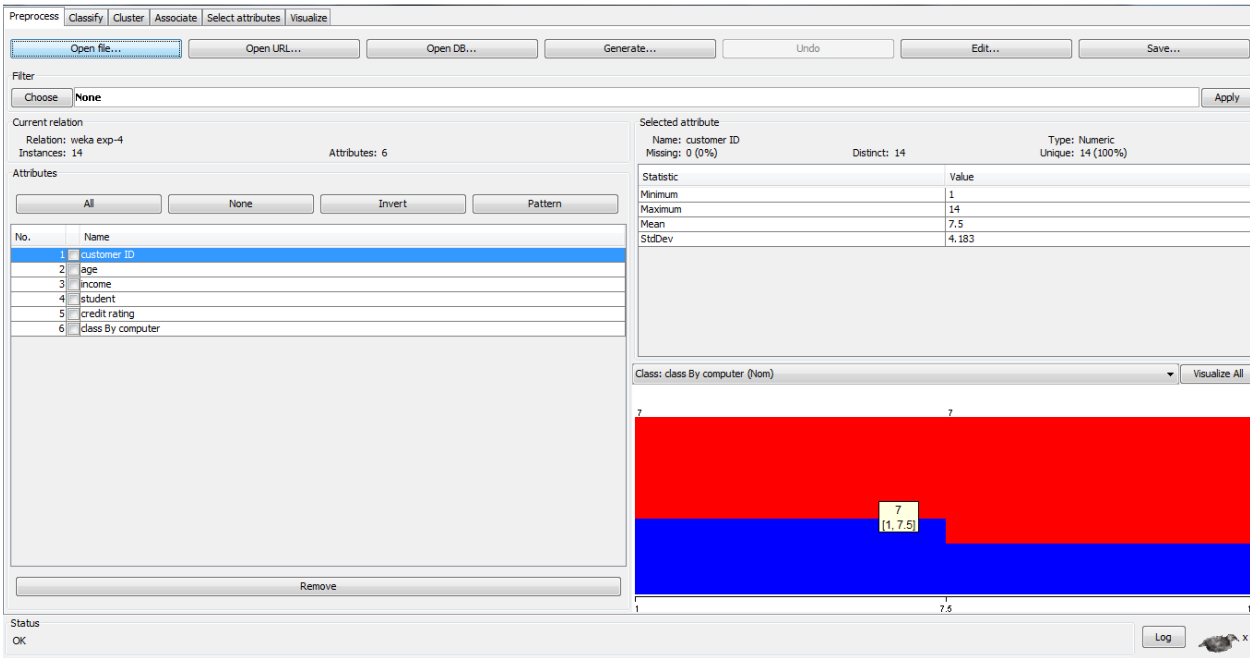


FIGURE 2: LOAD THE FILE CLUSTER.CSV

After loading the input file named cluster.csv as shown in figure 2, choose the cluster tab in the WEKA explorer window.

Under the cluster tab click on choose button and select the **SimpleKMeans** under **clusterers** as shown in the following figure 3.

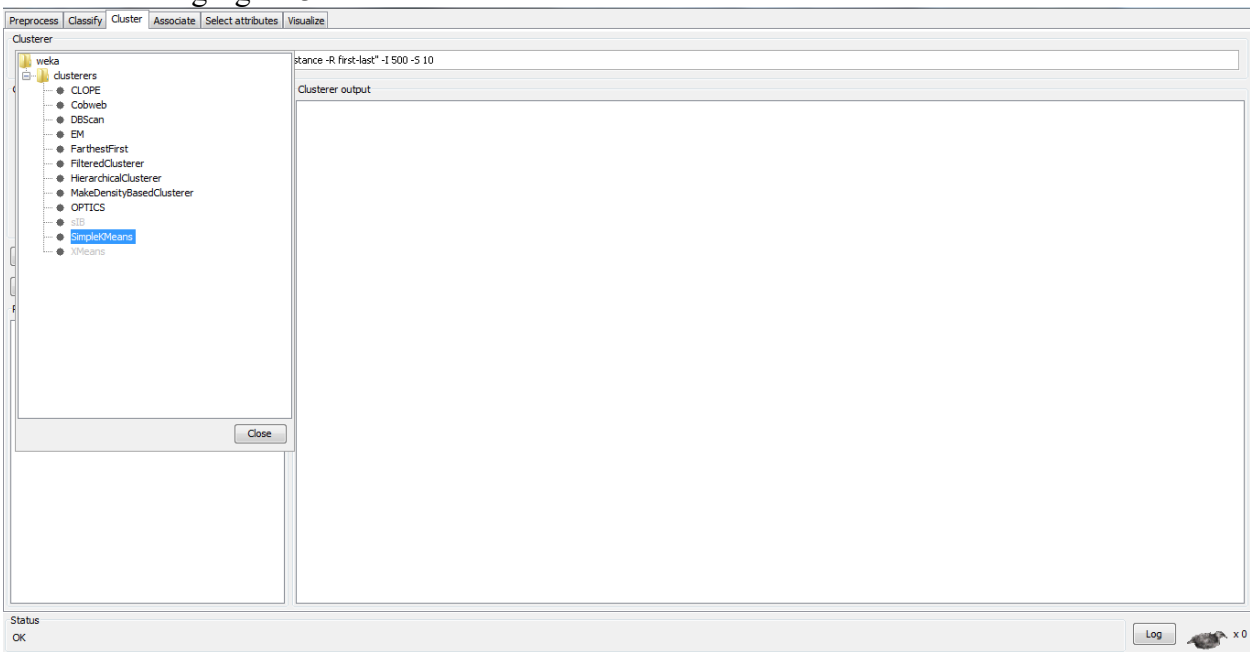


FIGURE 3: SELECTING SIMPLEKMEANS

Now select the "**use training set** "under the **Test Options** located at the left of the WEKA explorer window and click on **start** button.

The output is represented in the **Clusterer Output** window in weka explorer window, which is as shown in figure 4 below:

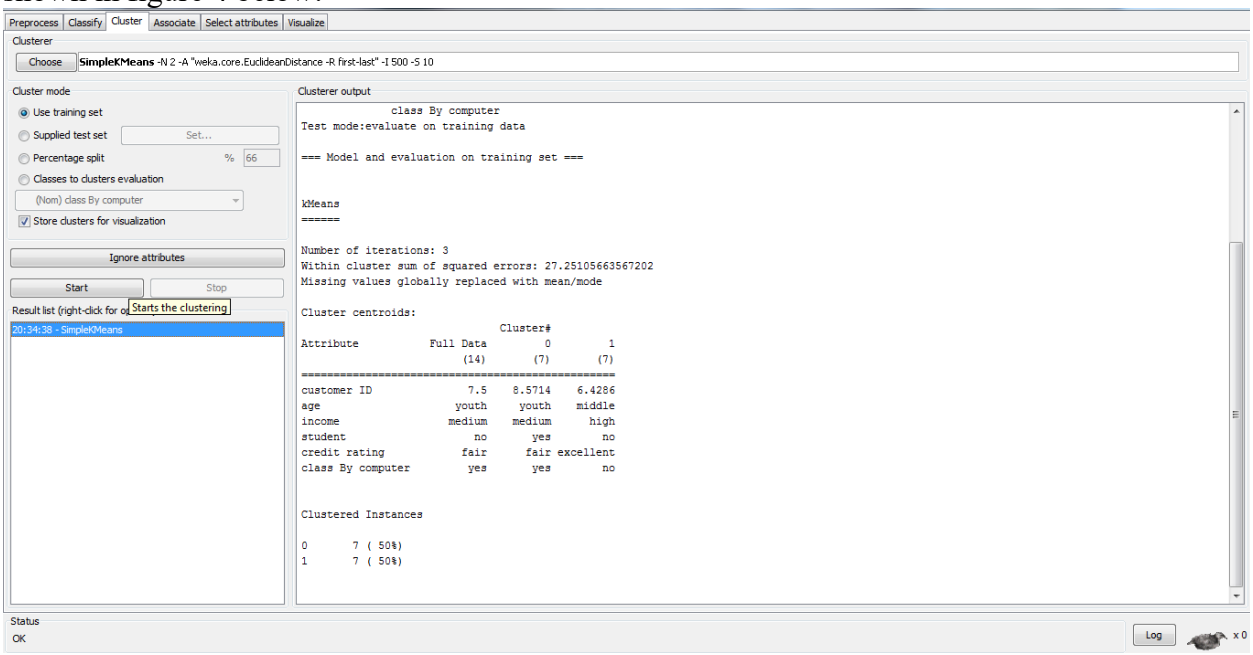


FIGURE 4: SHOWING OUTPUT IN CLUSTERER OUTPUT WINDOW IN WEKA EXPLORER WINDOW

We can also view the output in a separate window by right clicking on the option in **Result list** and clicking on "**view in separate window"** as shown in figure 5**.**

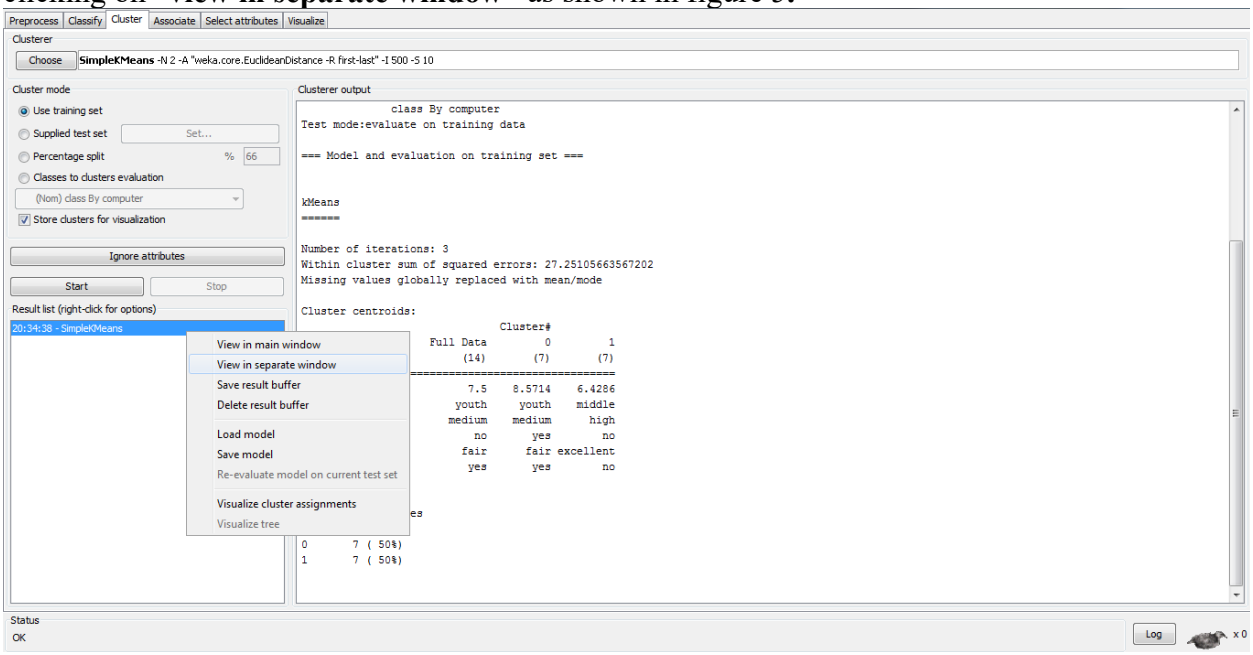


FIGURE 5: SELECTING THE OPTION "VIEW IN SEPARATE WINDOW"

The output is viewed in a separate window is as follows:

=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Relation: weka exp-4

Instances:14

Attributes:6

customer ID

age

income

student

credit rating

class By computer

Test mode:evaluate on training data

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

kMeans

======

Number of iterations: 3

Within cluster sum of squared errors: 27.25105663567202

Missing values globally replaced with mean/mode

Cluster centroids:

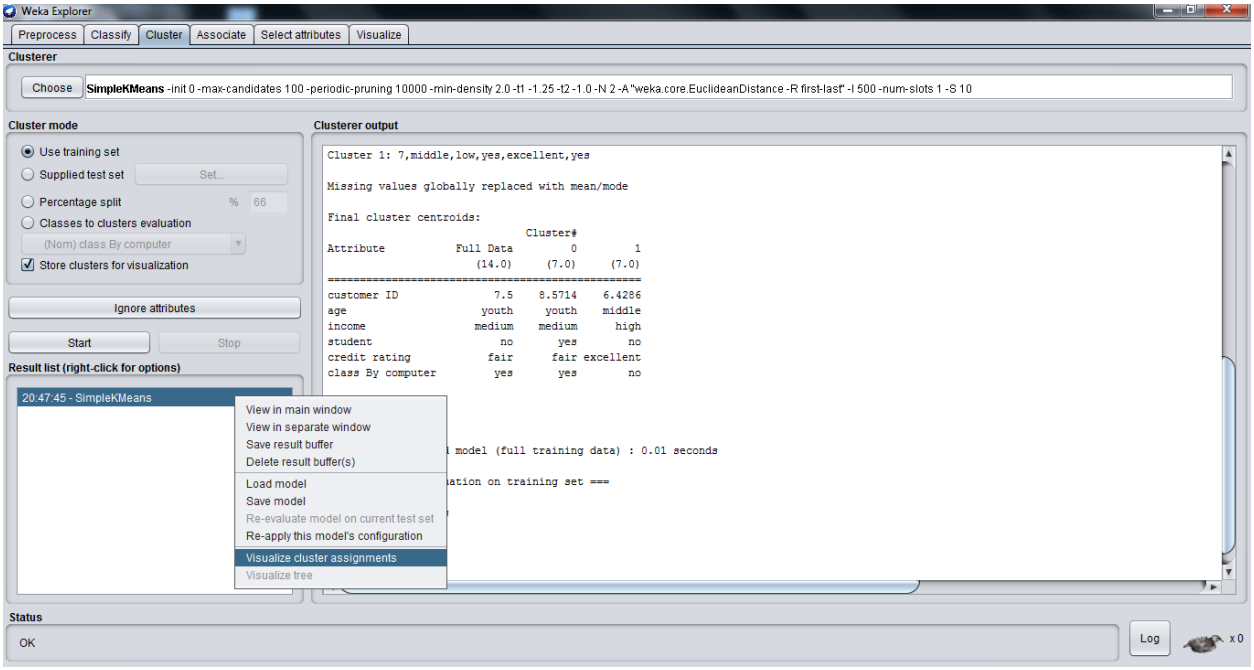
Cluster#

Attribute Full Data 0 1

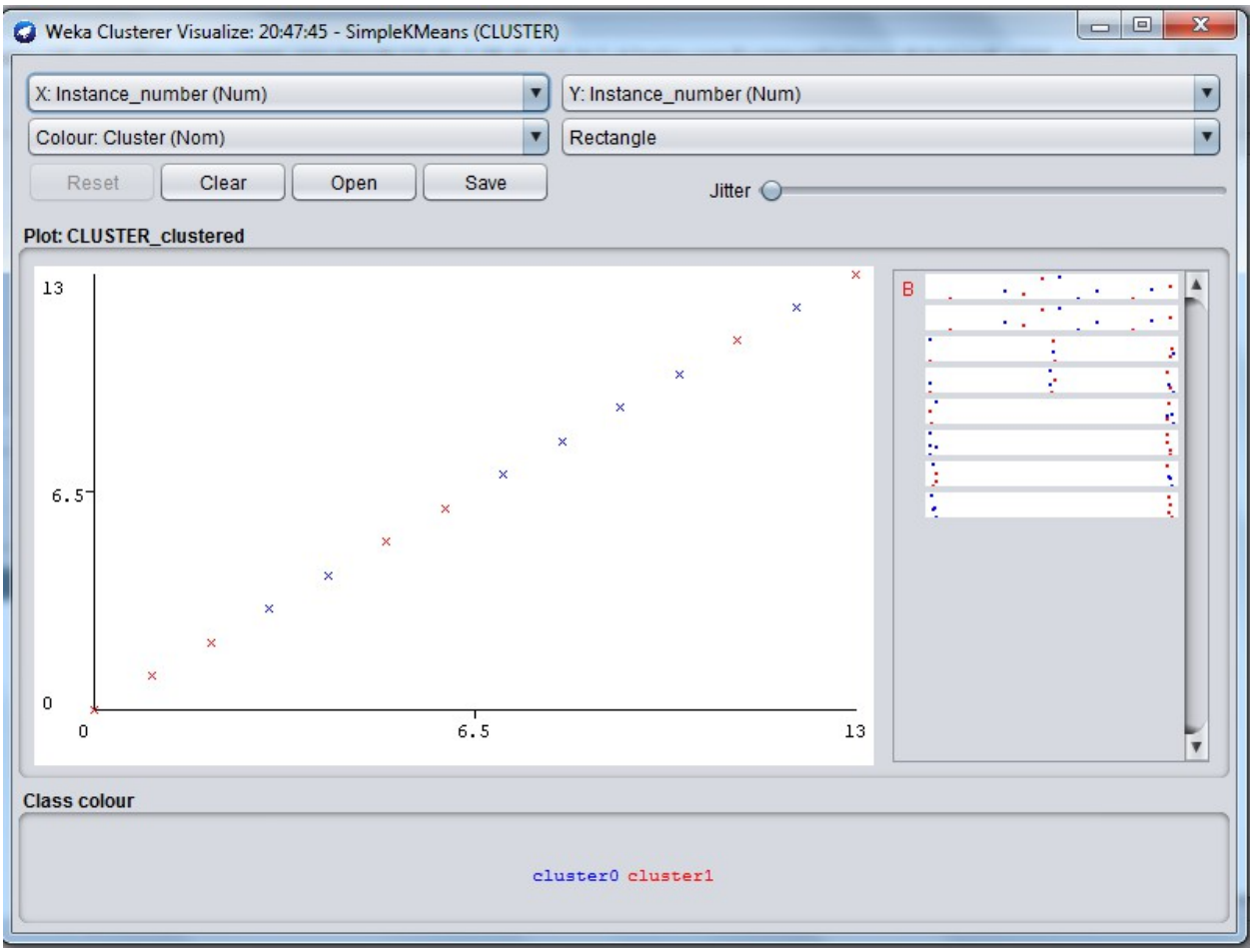
(14) (7) (7)

==========

customer ID 7.5 8.5714 6.4286
age youth youth middle
income medium medium high
student no yes no
credit rating fair fair excellent
class By computer yes yes no
Clustered Instances
0 7 (50%)
1 7 (50%)



Visualize the Cluster



Cluster data using the FarthestFirst algorithm.

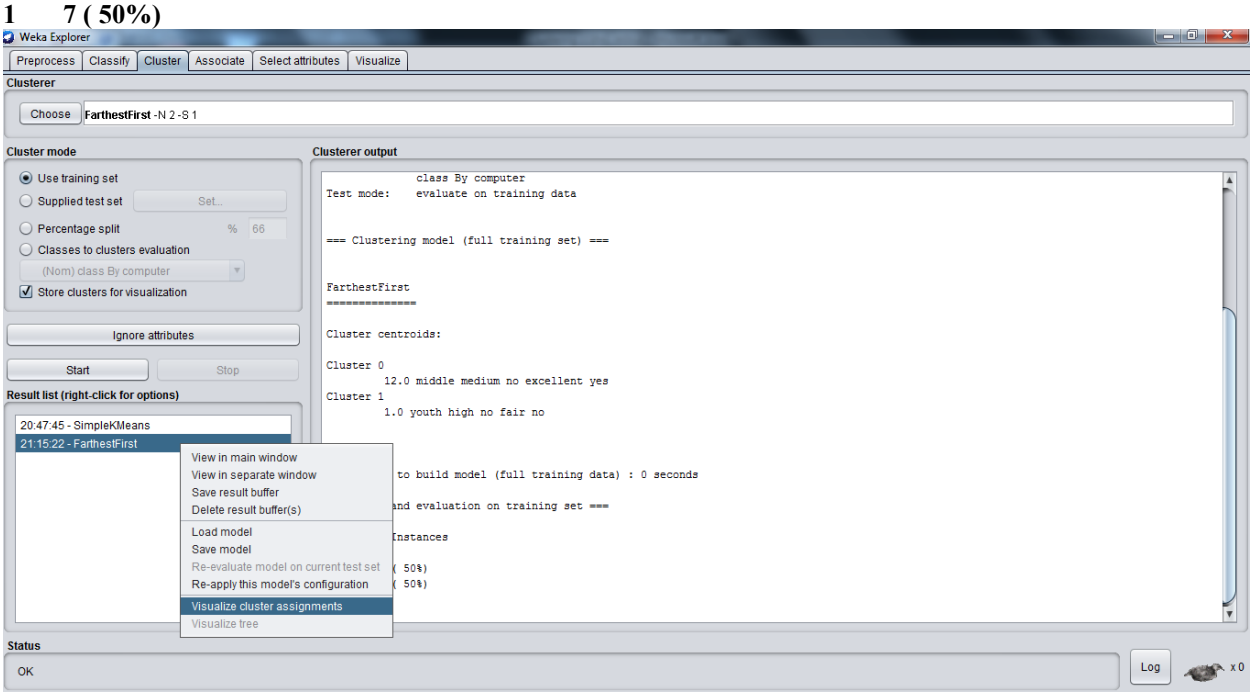
=== Run information ===
Scheme: weka.clusterers.FarthestFirst -N 2 -S 1
Relation: CLUSTER
Instances: 14
Attributes: 6
customer ID
age
income
student
credit rating
class By computer
Test mode: evaluate on training data
=== Clustering model (full training set) ===
FarthestFirst
==========
Cluster centroids:

Cluster 0
12.0 middle medium no excellent yes
Cluster 1
1.0 youth high no fair no
Time taken to build model (full training data) : 0 seconds

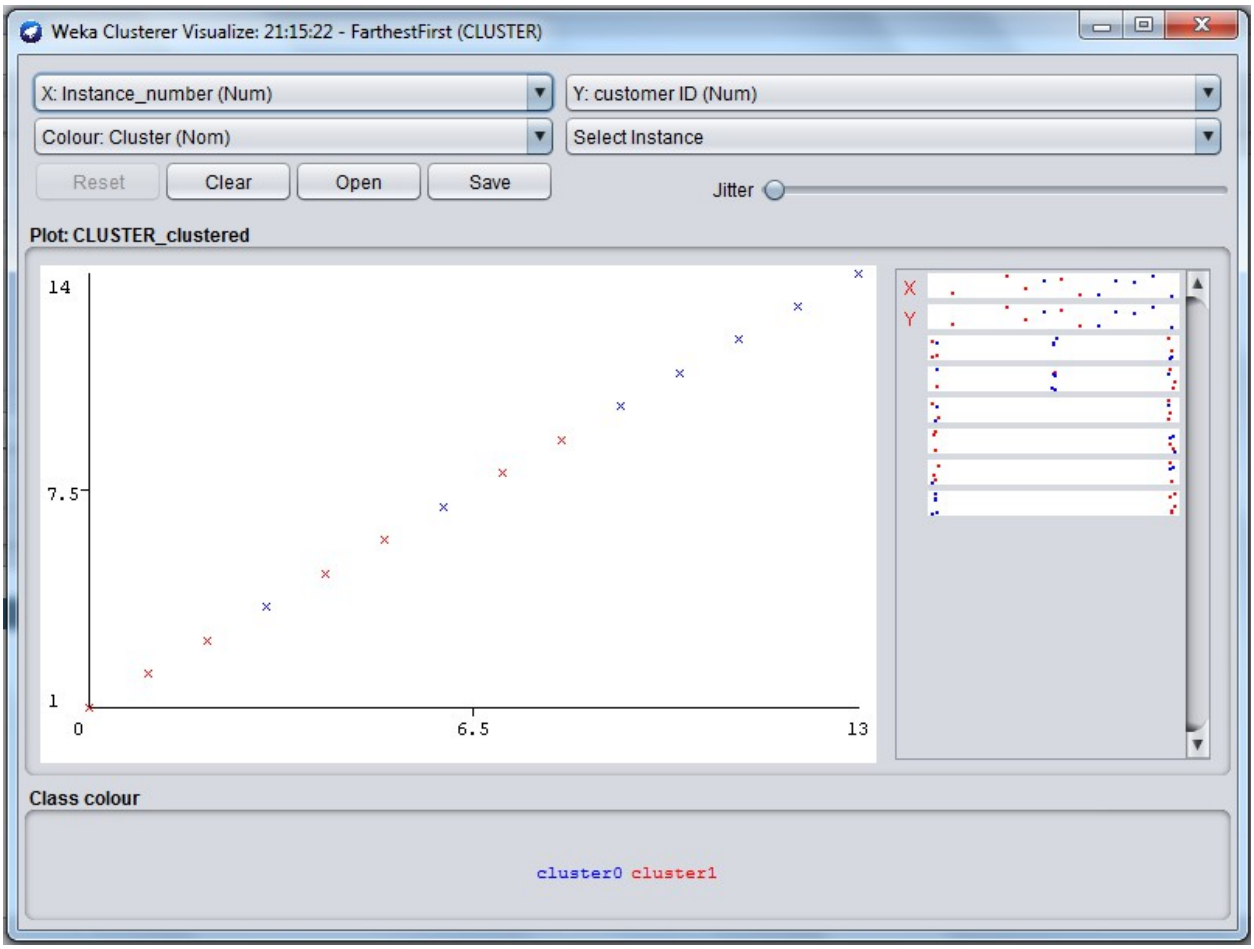
=== Model and evaluation on training set ===

Clustered Instances

0 7 (50%)
1 7 (50%)



Visualize the Cluster



Conclusion:

Created the clusters visualization.

References Used:

Data Mining: Concepts and Techniques	Han, Jiawei Kamber, Micheline Pei and Jia	Elsevier Publishers

Prepared by:
Prof.T.Bhaskar
Subject Teacher

Approved by:
Prof. D.B. Kshirsagar
Head, Computer Engineering