IST565 Data Mining
K-Means clustering algorithm
Work Examples
Professor Bei Yu

This document describes a case study that uses k-Means clustering to explore a large data collection.

Use a portion of the US census data, “UScensus_exercise_arff.arff” to generate clusters using Weka’s Simple K-Means algorithm.

Run the Simple K-Means clustering. Choose a number of clusters, not too large but greater than 2. Choose two attributes to report findings. For example, YEARSCH (Years in school) and INCOME1 (wages and salaries) are two interesting attributes. What kind of people does each cluster characterize in terms of their education and income? Hint: look at the cluster centroids.

Sample solution:

See below my analysis of the clustering result. My interpretation could have been much shorter with more editing on the tabular outputs. -Bei

K-Means: set k=4, other parameters as default values.

```
Cluster centroids:

                 Cluster#

Attribute    Full Data        0        1        2        3

           (8100)    (1222)    (1274)    (3668)    (1936)

=================================================================

dIncome1          0        0        0        1        0
dIncome2          0        0        0        0        0
dIncome3          0        0        0        0        0
dIncome4          0        0        0        0        0
dIncome5          0        0        0        0        0
dIncome6          0        0        0        0        0
dIncome7          0        0        0        0        0
dIncome8          0        0        0        0        0

iYearsch         10       10       10       10        4

Clustered Instances
```

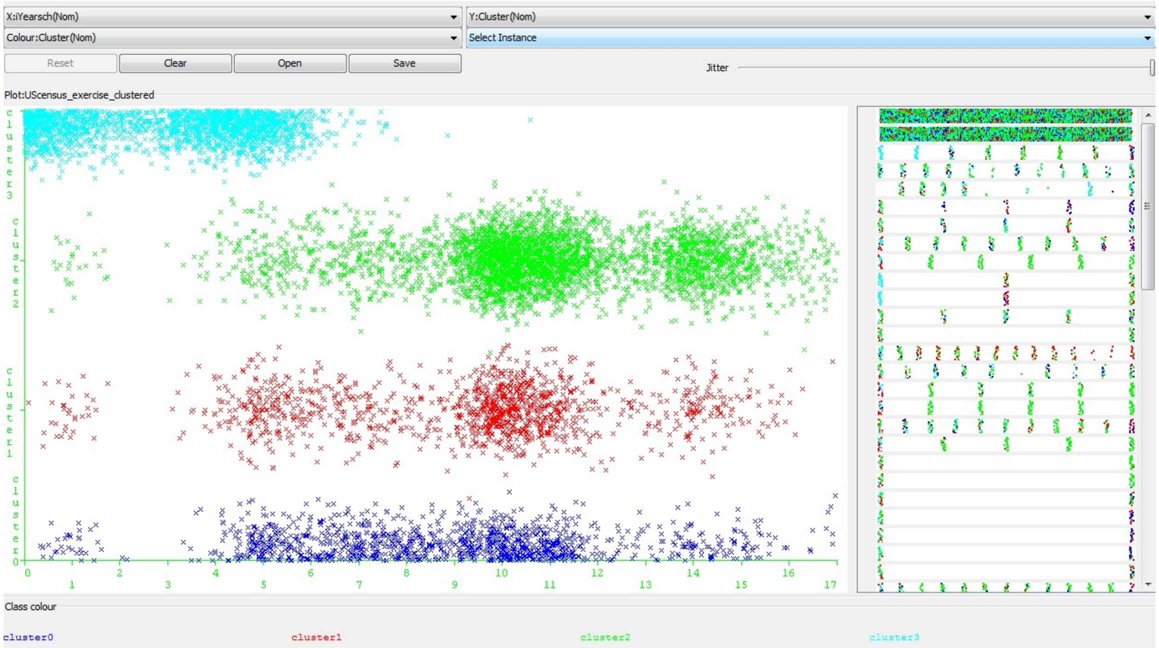
0	1222 (15%)
1	1274 (16%)
2	3668 (45%)
3	1936 (24%)

Observations:

(1) The YearsInSchool attribute is a categorical attribute taking values from 0 to 17. In the entire data set of 8100 people, the 10- and 11-year groups are the largest: 1864 persons finish 10 years of school and 1184 persons finish 11 years.

In the above result table, the centroids of clusters#0,1,2 all take value “10” on this attribute, meaning that people with 10-year school education are the largest group in these clusters. The centroid of cluster #3 takes value “4”, meaning the 4-th graders are the largest group in this cluster.

A scatter plot of the YearsInSchool attribute and the cluster membership in the following figure visualizes the above finding. The figure shows that the members of cluster #3 are mostly elementary school students with fewer than six years of school education. The other three clusters do not differ very much from each other in terms of members’ years of education.



X:iYearsch(Nom)

Y:Cluster(Nom)

Colour:Cluster(Nom)

Select Instance

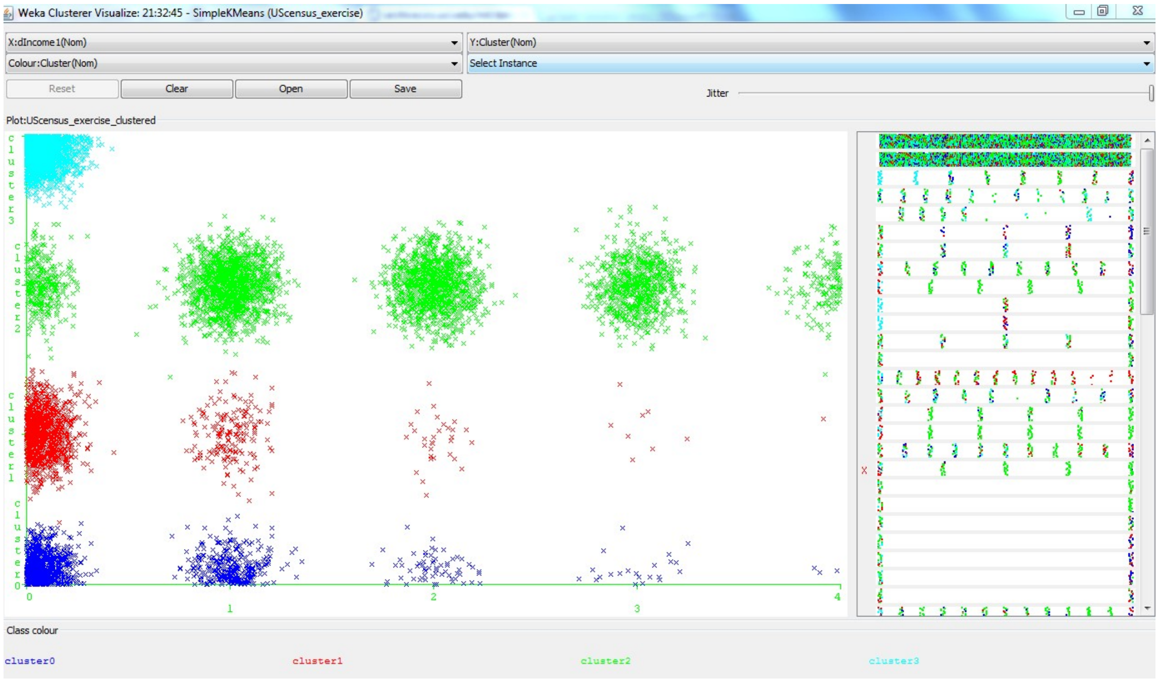
Plot:UScensus_exercise_clustered

(2) The Income1 attribute means wages and salary. Originally it is a numeric attribute, but it is discretized to 5 bins 0 (=0),1(<15000),2(<30000),3(<60000),4(else). The

Income2-8 attributes are all binary attributes, meaning a person has or does not have a certain type of income, such as social security. The majority of people have value “0” on these attributes.

In the above result table, the centroids of all four clusters take same value “0” on all income attributes except for cluster#2, which takes value “1” on dIncome1, meaning the largest group in this cluster have salary but the amount is below $15,000.

The following scatter plot of dIncome1 and cluster membership first explains the composition of people along the income dimension. Both cluster#0 and 1 include mostly people with no income, and some with income lower than $15,000, and few people at other income levels. Cluster#2 is actually the largest cluster (see the table result, 3668 members). The largest group in this cluster is people at income level 1, and also higher income levels. This cluster does have a small number of people with no income.



Weka Clusterer Visualize: 21:32:45 - SimpleKMeans (UScensus_exercise)

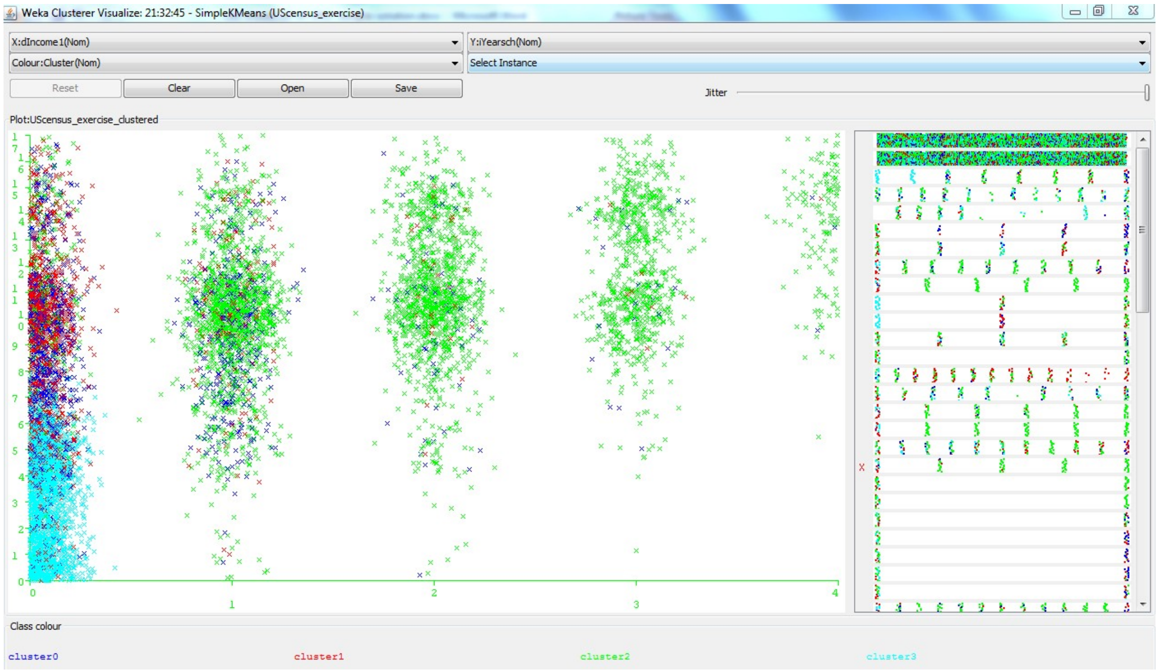
X:dIncome1(Nom)

Y:Cluster(Nom)

Colour:Cluster(Nom)

Plot:UScensus_exercise_clustered

(3) To combine my observation on both attributes, I scatter plot them in the following figure, dIncome1 as the x-axis, YearsInSchool as the y-axis. The following scatter plot shows that cluster#3 (light blue) includes mostly elementary school students with no income; cluster#0 and #1 (blue and red) includes mostly people with more years of school education but has no salary income. These two clusters are hardly distinguishable based on these two attributes. Cluster#2 includes mostly people with beyond-elementary level education and also with low to high income levels.



Weka Clusterer Visualize: 21:32:45 - SimpleKMeans (UScensus_exercise)

X:dIncome1(Nom)

Y:iYearsch(Nom)

Colour:Cluster(Nom)

Plot:UScensus_exercise_clustered