Solution to Laboratory Week 4

1. You will need to have access to the WEKA package.

2. The data files for this lab can be found at /KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data

3. Load the file arff/mystery-data5.arff.

(a) Go to the Visualize screen. How many clusters are in the data?

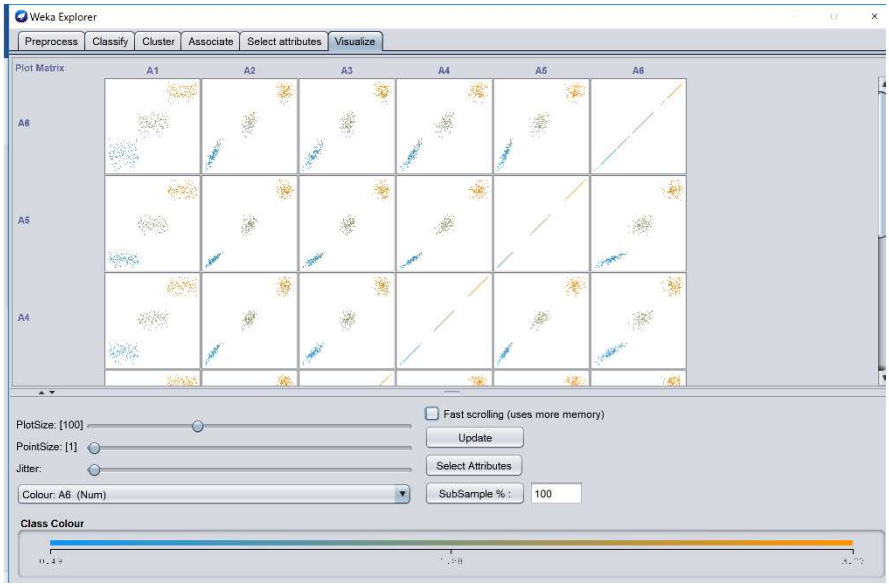


Fig. 1. Screenshot of visualized data.

Figure 1 is the output of visualizing data before running any clustering algorithms on top of the data set. As it can be seen, there are 3 clusters in the data.

(b) Go to the Cluster screen and select SimpleKMeans.

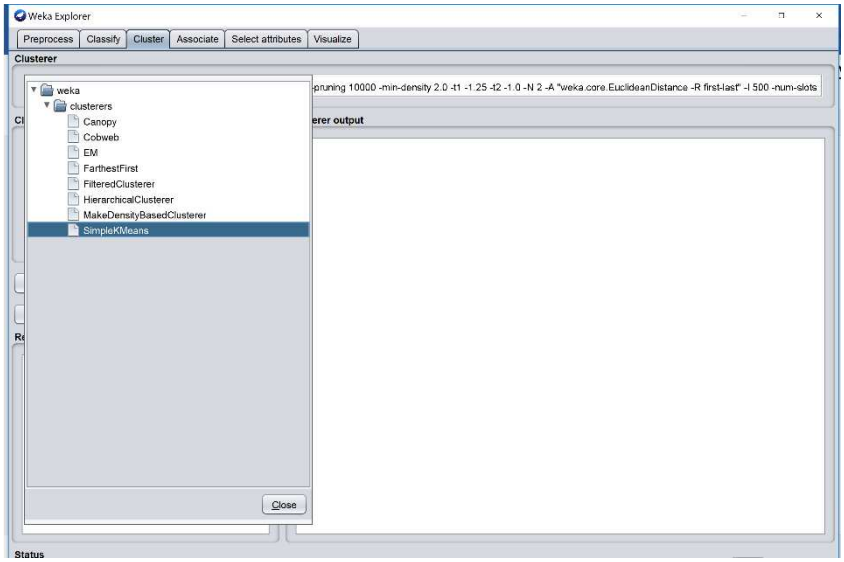


Figure 2.

(c) Run the algorithm for K=2 and analyse the output.

To run k-means clustering, we need to set a list of parameters for the algorithm. In this question, we set K=2, seed= 10. As expected after running K-means, we got two clusters, 0 and 1, where within cluster sum of squared errors is 49.24617244366606. Out of 300 data objects in the data set, 100 data objects are assigned to cluster 0 and the remaining 200 are assigned to cluster 1. Cluster centroids (average) for each attribute on both clusters are as follows.

	Full Data	Cluster0	Cluster 1
A1	150.5	250.5	100.5
A2	2.0045	3.0045	1.5045
A3	2.0038	3.0012	1.5051
A4	2.0141	3.0147	1.5137
A5	2.0099	3.0083	1.5107
A6	1.9951	2.9856	1.4998

(d) Visualise the clusters by a left click on the file in the result list.

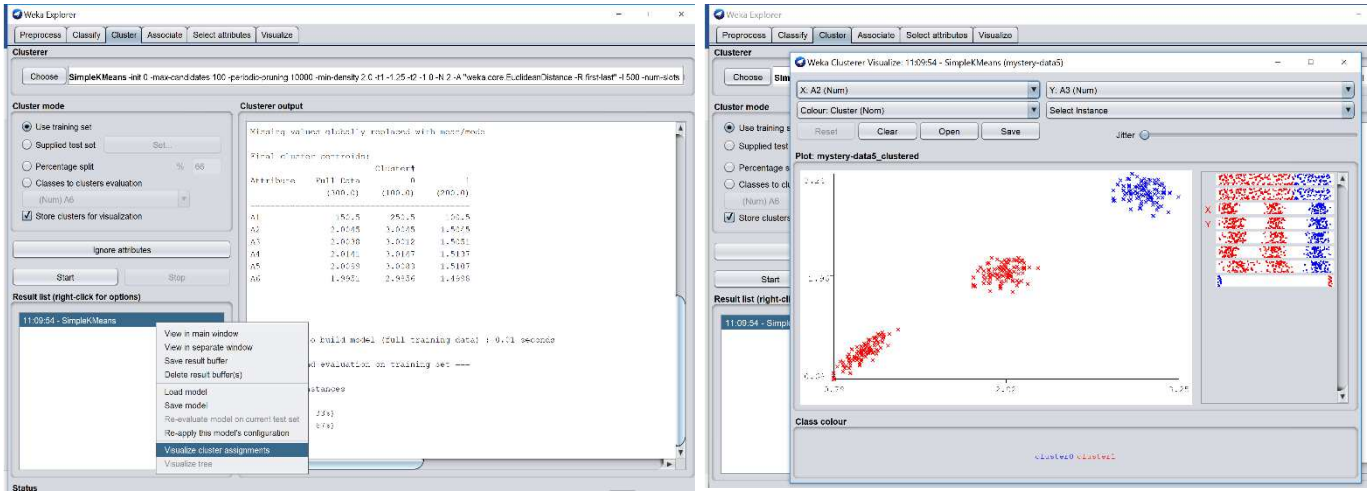


Figure 3. Visualisation of clustering.

(e) Does this appear to be a good clustering result?

No. As shown in figure 3, cluster 1 as represented by red colour is not a proper cluster since cluster 1 includes two well-separated clusters. But forcing number of clusters to k=2 and considering initial random centroids result to have the above output that is not a good clustering result.

(f) Repeat the runs for 5 different initial centres. This done by changing the seed (try 10 and 11) and describe the effect of changing the initial centres.

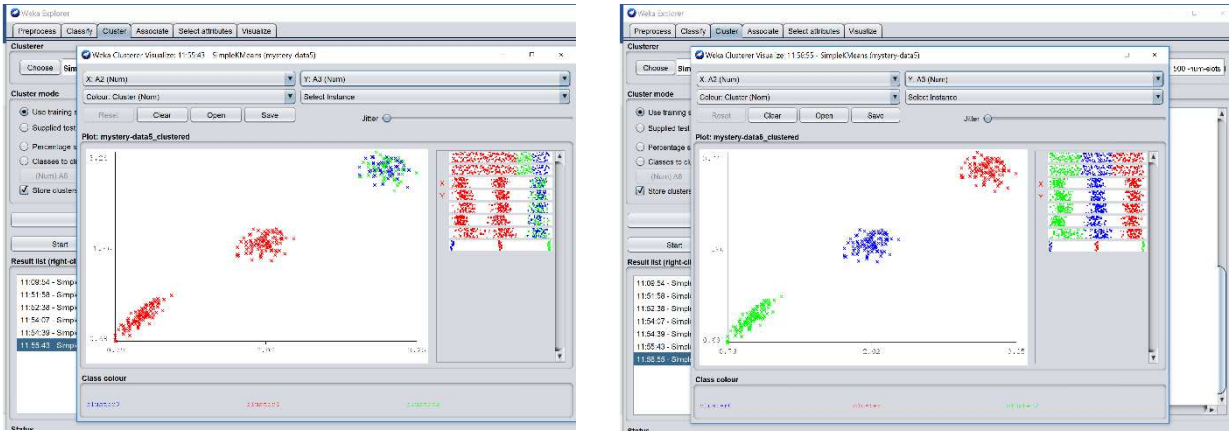
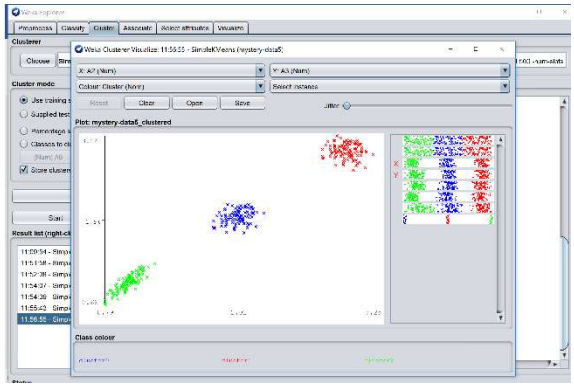


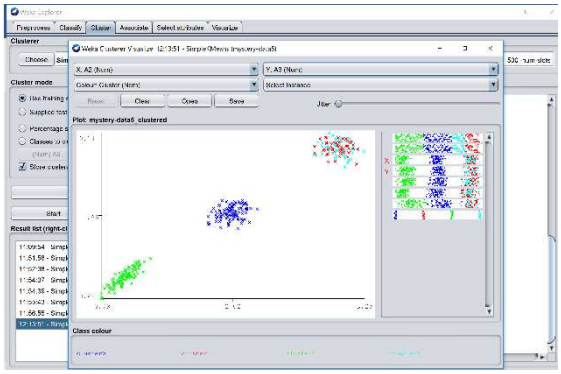
Figure 4. screenshots of run of k-means, a) k=3, seed= 10, b) k=3, seed=11.

Figure 4 represents the run of k-means algorithm where k=3 and seed is 10 in figure 4.a and seed=11 in 4.b. Three clusters were seen by visualizing data before running any clustering on the data set, so we expected to have 3 well separated clusters by setting parameter k to 3. However, as it can be seen from figure 4, initialized centroids may affect on the final output as it can be seen in figure a and b by changing seed to 10 and 11. In figure a, most probably one centre was chosen in between two red clusters in cluster 1, that results to assign those points that are closer this point while in figure b changing seed to 11 results to get 3 well-separated clusters which is a very good clustering result.

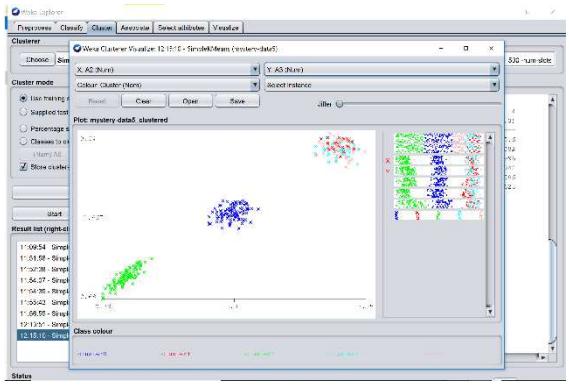
(g) Repeat the runs and visualizations with K = 3,4,5,10,20.



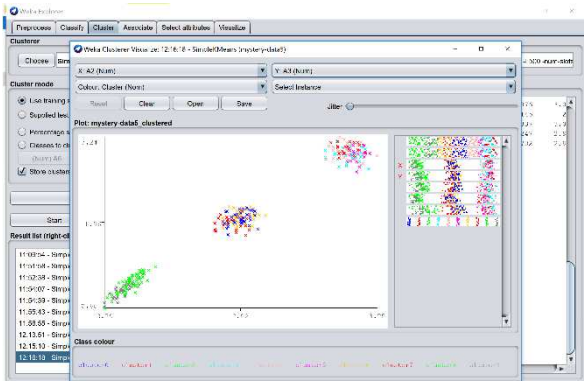
a) k= 3, seed=11



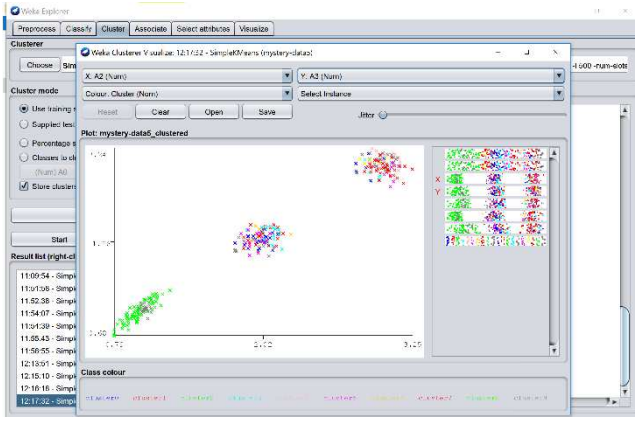
b) k=4, seed= 11



c) K = 5, seed = 11



d) k=10, seed =11



d)k=20, seed= 11.

Figure 5. Different runs of k-means for k= 3,4,5,10 and 20.

(h) How can you tell when you have right number of clusters?

The best approach is by visualization and compare the results of different runs of clustering for varying number of k. The good clustering results in having separated clusters. Also, inspecting the cluster means and standard deviations can give an indication. Two generated clusters with means that are close might be one true cluster.

(i) Does the "Within cluster sum of squared errors" give any indication about the number of clusters?

Yes, as increasing the number of clusters leads to minimise the within cluster sum of squared errors. Since this error shows the average error in each cluster not between clusters. So increasing number of k translates to have more fine clusters. If we set number of k into total number of data points in the data set, you would expect to receive within sum squared errors of zero that means every data points are clustered as one cluster.

4. Restart Weka and reload the file arff/mystery-data5.arff.

(a) Go to the Cluster screen and select EM.

(b) Run the algorithm with K=2,3,4,5 and compare the clusters to the ones found by Kmeans. Are they the same? Would you expect them to be the same?

(c) Run the EM with N (numClusters) set to -1. Visualise the clusters.

(d) Does this appear to be a good clustering result?

(e) How does it compare with the K-Means results?

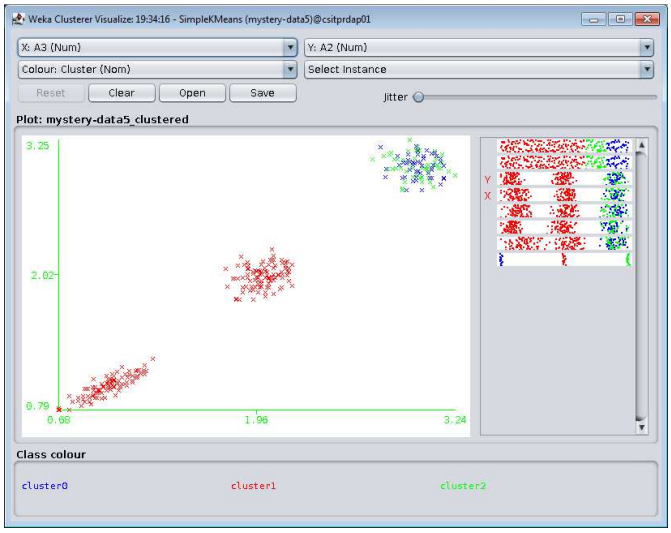
(f) The clusters generated by EM can be quite sensitive to the values of minStdDev. Explore the effect of different values for this parameters [To start with, try order of magnitude changes, ie minStdDev 1.0E-6 --> 1.0E-5 --> .... --> .01 --> .1 -->1 -->10] Summarize your observations.

(g) The clusters generated by EM can also be quite sensitive to the values of minLogLikelihoodImprovementIterating, minLogLikelihoodImprovementCV. Explore the effect of different values for these parameters. Summarise your observations.

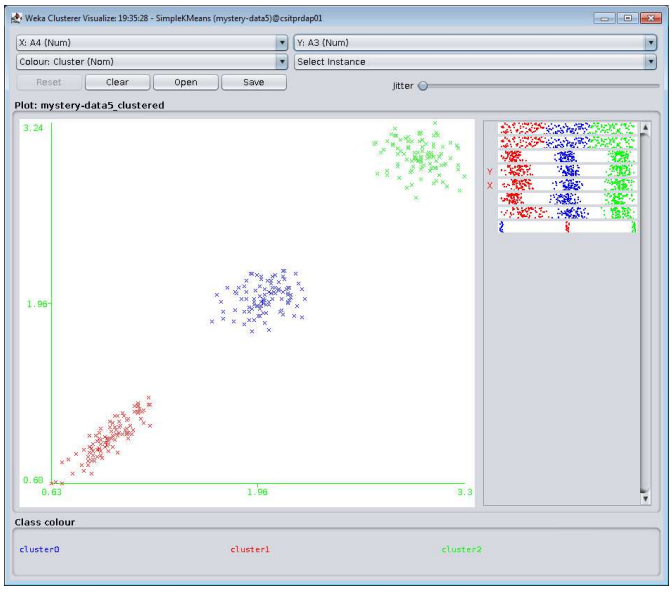
Kmeans can give different clusters based on the initial randomly chosen centres.

The data in these examples is mystery-data5.

K=3 seed = 10 [Unexpected clusters given the visualisation]

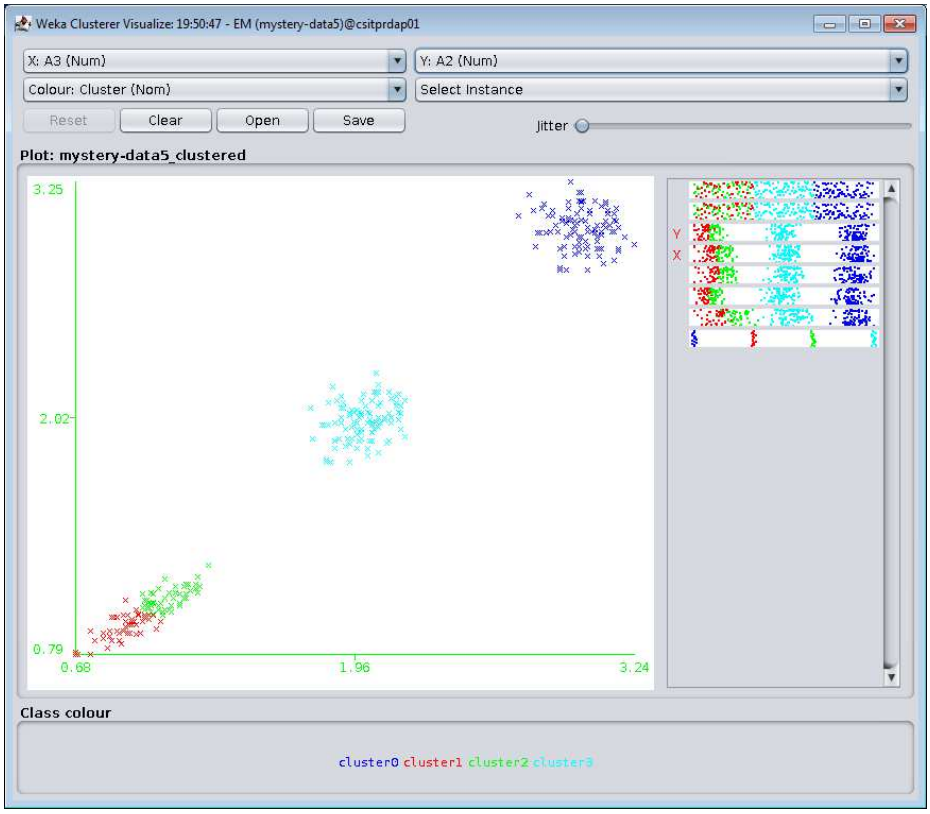


K=3 seed = 100 [Clusters as expected]

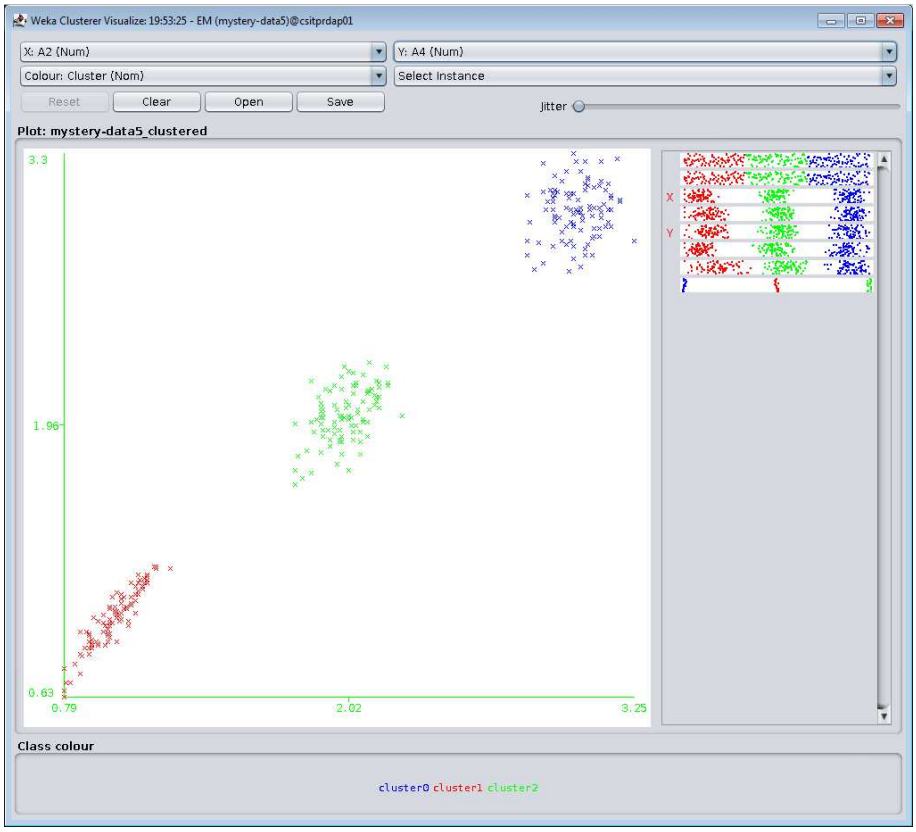


The number of EM clusters depends on the value of MinStdDev

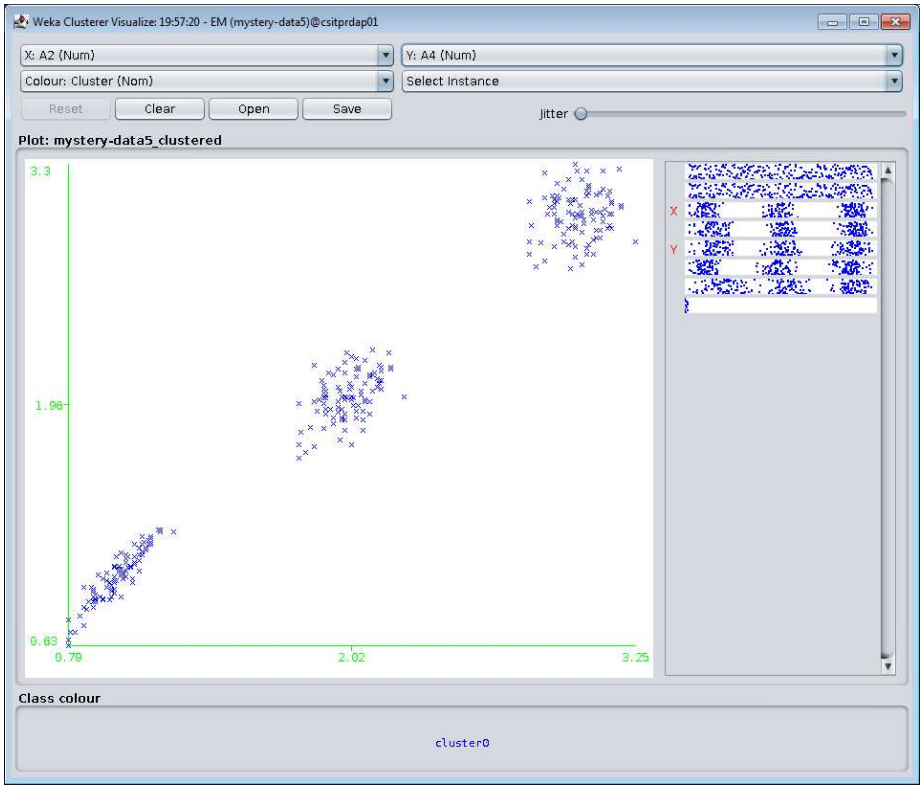
Min stddev is 1.0E-6 [Too many clusters, 4]



minStdDev =0.1 [The expected number of clusters, 3]

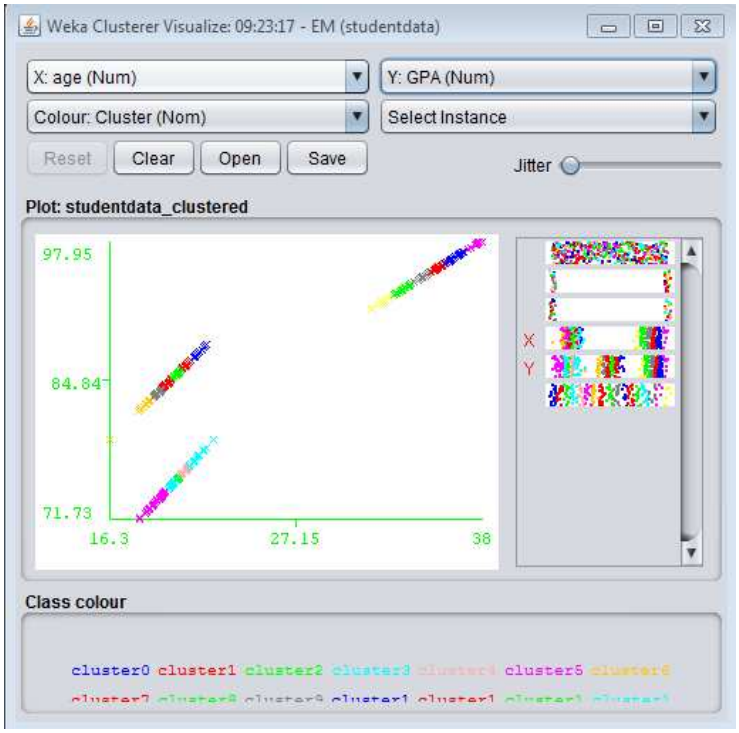


MinStdDev= 100 [Not enough clusters, 1]



For EM the value of minStdDev should correspond roughly to the number of decimal places in the data. This data has 2 decimal places.

5. Load the file `arff/student-data1-small.arff`.
(a) Run the EM algorithm with default parameters on this file. Visualize the output. What do you find?



There are clear clusters, but EM has generated 10.

(b) Adjust the parameters to get right number of clusters.

Setting the standard deviation to 0.95 give 3 clusters:

(c) Give English language descriptions of the clusters.

	Cluster		
Attribute	0	1	2
	(0.36)	(0.33)	(0.3)
sex			
m	183	167	1
f	1	1	153
[total]	184	168	154
course			
MBC	183	1	1
BAPPSCI	1	167	153
[total]	184	168	154
age			
mean	35.0434	20.0663	19.9421
std. dev.	1.403	7.3272	7.3272
GPA			
mean	95.0379	75.1117	84.9145
std. dev.	1.4036	1.5214	1.4926

Cluster 0 is mostly male, mostly studying MBC with age ranging from about 30 to 40 with a GPA over 90.