CS5011: INTRODUCTION TO MACHINE LEARNING

PROGRAMMING ASSIGNMENT - 3

Adarsh B
MM14B001

CLUSTERING

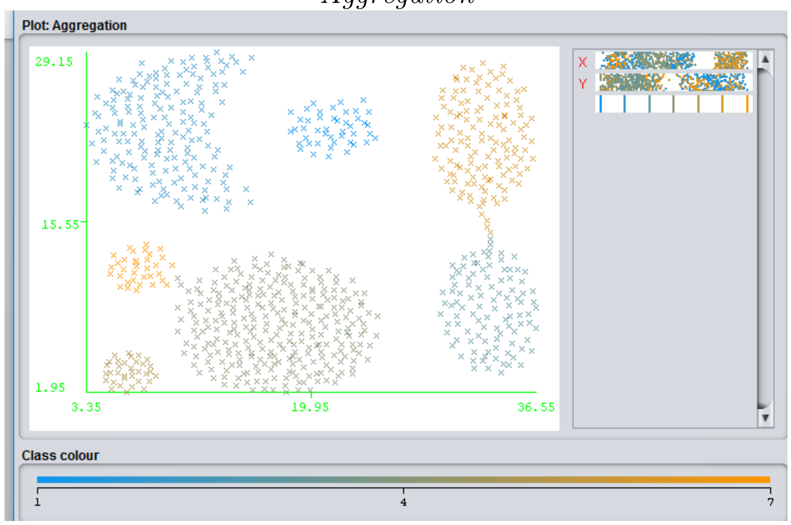
1. Conversion to ARFF format

The eight datasets have been converted from .txt to ARFF format using Excel and Weka.

2. Visualization

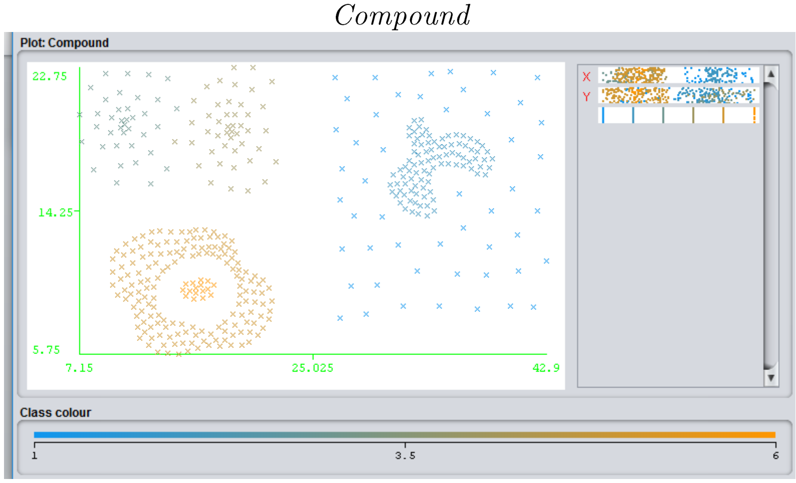
Data visualization outputs are as follows:
(<u>Note:</u> Algorithms marked **green** will work (either normally or under particular conditions), and **red** won't work well)

Aggregation



<u>Usable Clustering Algorithms:</u>
Since the compounds are clearly convex, ***k-means*** and ***Complete Link Hierarchial Clustering*** will work well with the right set of parameters (like k=7). ***Single link Hierarchial Clustering*** will work well except for the region between the clusters, due to which there is a lot of chance for improper cluster reporting. It may classify the blue and brown clusters on the right as a single cluster. **DBSCAN** will work well if minpoints are set to be sufficiently high, so that the points on the boundary of two clusters will get classified as outliers.

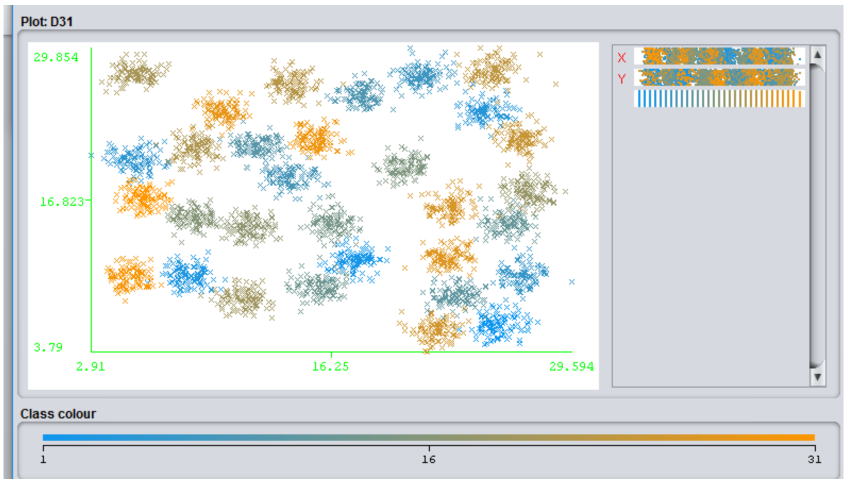
Compound



Usable Clustering Algorithms:

K-means and **Complete link Hierarchial Clustering** will not work well due to the non-convex nature of some of the clusters. **DBSCAN** will work for particular values of minpoints, but won't classify the cluster in blue in the right of the dataset image as the cluster is less dense. **Single link Hierarchial Clustering** will also work well except for the blue cluster on the right half, which will merge as a single cluster.

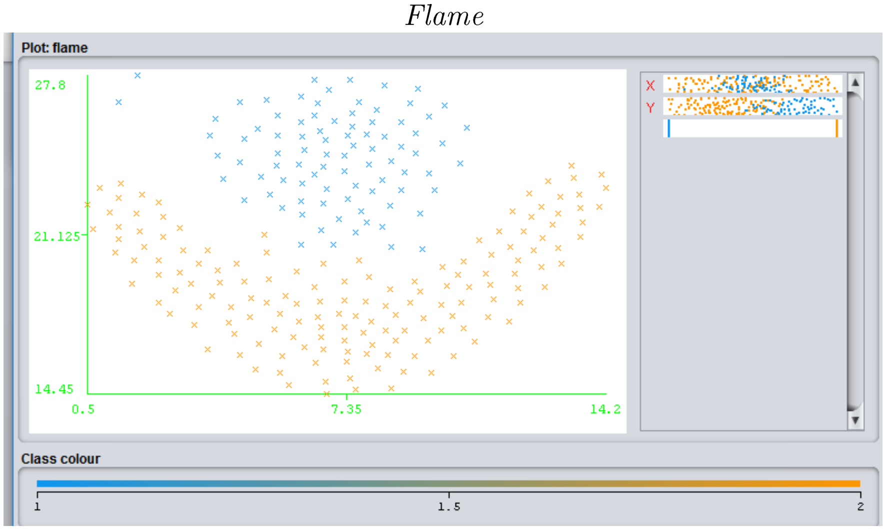
D31



Usable Algorithms:

Single Link Hierarchial Clustering won't give good results here since the clusters are very close to each other. **K-means** will work well, as the clusters are fairly discrete and we can obtain distinct centroid assignments. **Complete Link Hierarchial Clustering**, will be more robust to noise points. **DBSCAN** will work well for a large value of minpts and small eps, but might slightly suffer from noise points or unclassified points.

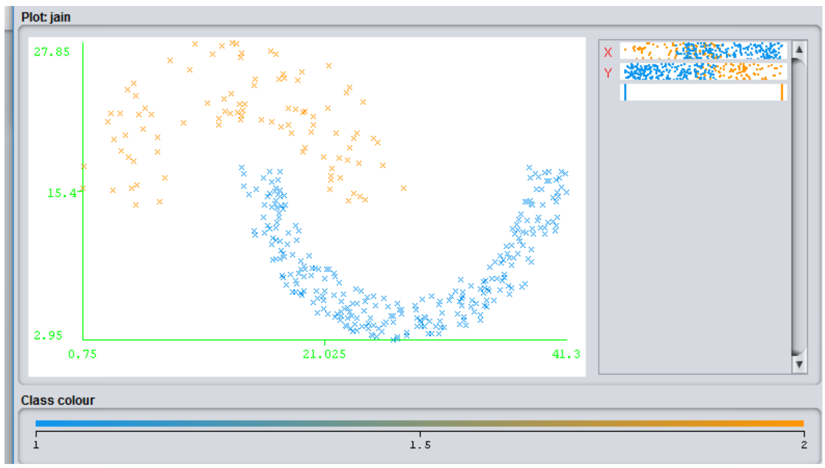
Flame



Usable Algorithms:

K-means won't work well as the left and right end points of class 2 (red) will get classified with class 1 (blue) since they are more closer to the centroid of blue than that of red. **Single link** would fail to distinguish the two clusters as the separation is very less and the shortest distance between the closest neighbours would make them into a single cluster. Whereas in case of **Complete link** we won't have this issue and might be able to get the original clusters if we choose the number of clusters as 2. **DBSCAN** will have no issues for the right choice of parameters.

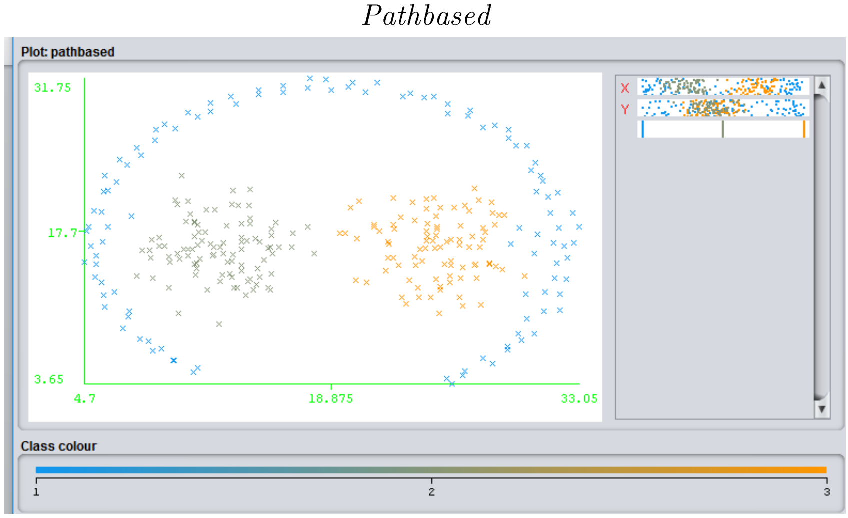
Jain



Usable Algorithms:

As the data set is non-spherical and center points with red and blue classes will get merged together, **K-means** and **Single link Hierarchial Clustering** won't work well. This dataset is ideal for **Complete link Hierarchial Clustering** and **DBSCAN** as the clusters are well separated and densely separate as well.

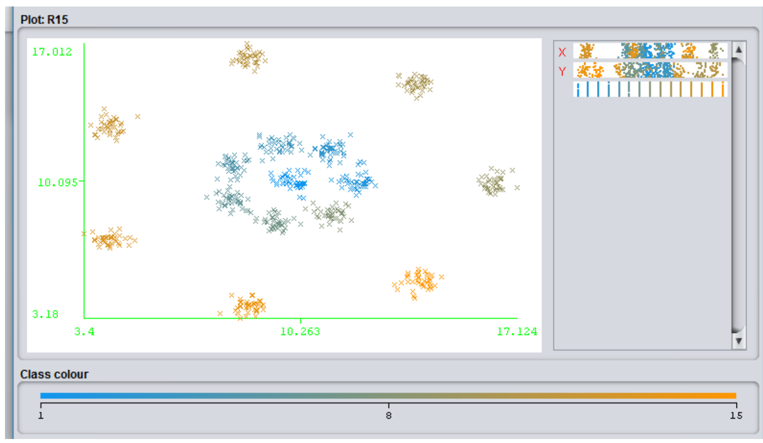
Pathbased



Usable Algorithms:

Due to the non-convex (non-spherical) nature of the clusters, **K-means** and **Complete link Hierarchial Clustering** won't work well due to more closeness of some of the outliers to foreign clusters. **DBSCAN** also wont work properly as the densities are jagged around. We might need to trade off between the number of clusters and the purity levels. **Single Link** will work well, due to well-spaced points in the clusters.

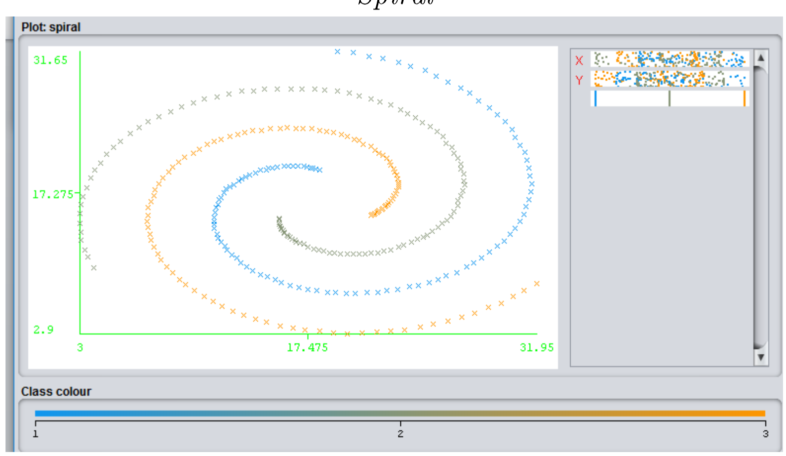
R15



Usable Algorithms:

All the approaches will yield good results for the right values of k / eps and minpts etc. as the clusters are well spaced, density separated and chances of outliers are very minimal.

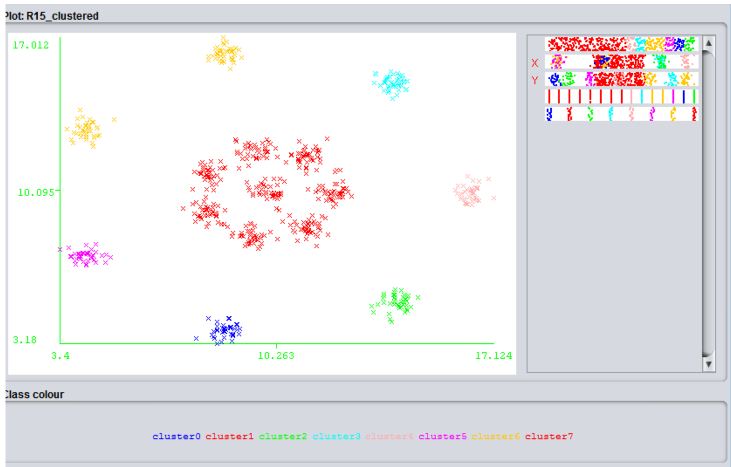
Spiral



Usable Algorithms:

This dataset has clearly separated clusters. Hence, **Single Link** and **DBSCAN** will work very well for this dataset. However, due to non-convexity of the clusters and as some of the end points of class 1 (blue) are closer to class 2 (red) than they are to its own class, **k-means** and **Complete Link** won't work well.

3. K-Means with R15 Dataset



Reported clusters for k = 8

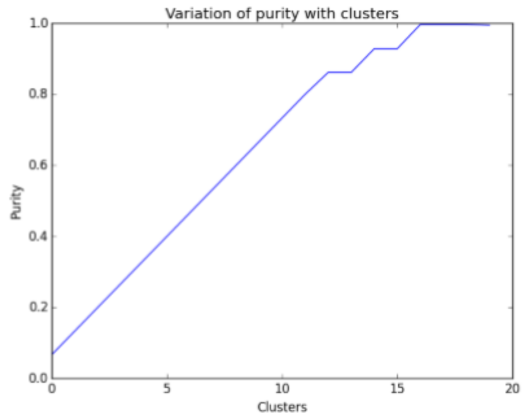
For K=8, the overall cluster purity is **0.533**.
The individual cluster purities are as follows:

Cluster	Purity
0	1.00
1	0.211
2	1.00
3	1.00
4	1.00
5	1.00
6	0.50
7	0.305

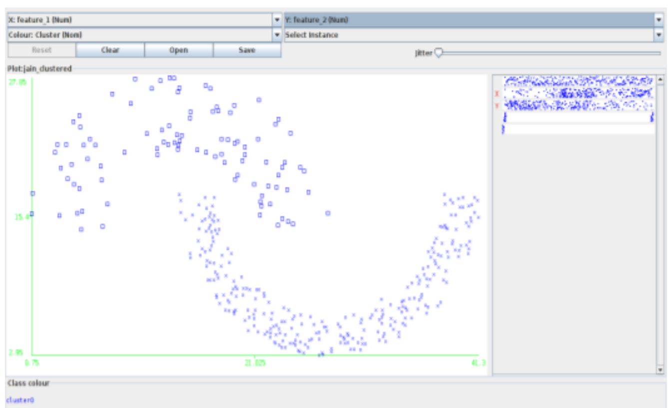
Effect of K on overall purity:

K	**Purity**
2	0.133
4	0.267
6	0.400
8	0.533
10	0.667
12	0.800
14	0.862
16	0.928
18	0.996
20	0.995

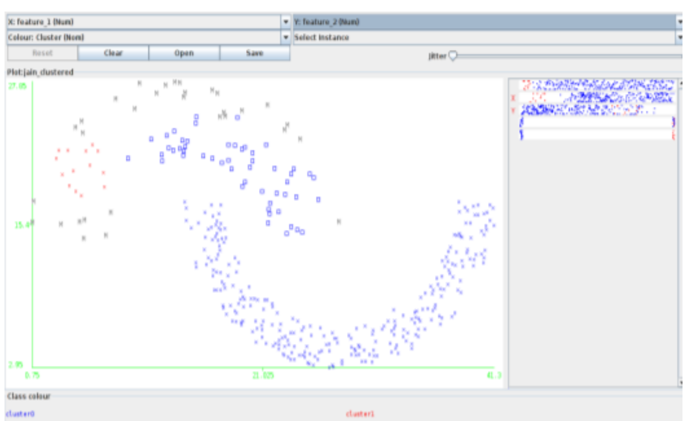
As K increases, purity increases since the clusters become more modular, thereby resulting in more skewed distribution in each cluster.



Question 4 – DBSCAN on Jain Dataset



Eps = 0.9 and minpts = 6



Eps = 0.1 and minpts = 14

DBSCAN doesn't work optimally here because the upper cluster has sparsely distributed points, hence lower density, while the lower cluster has high density. Due to differences in densities, and non-uniform density within the upper cluster itself, it is difficult to trade off and get an optimal value of eps and minpts. If we choose a high value of eps or low value of minpts it is observed that the entire dataset gets classified as one cluster. Tabulated below are the results for various trials, and the highest purities are observed for eps = 0.1 and minpts = 12 or 14.

eps	**Minpoints**	**Purity**	**Misclassified**	**No. of Clusters**
0.1	12	0.785	80	2
0.1	14	0.777	83	2
0.1	16	0.739	97	1
0.1	2	0.739	97	1
0.9	6	0.739	97	1
0.2	100	0.5147	181	1
0.075	2	0.997	1	3

Question 5 – Path based, Spiral and Flames

Path based

DBSCAN performs very poor in this case since the densities are very low and the separation between the three classes is not wide enough for DBSCAN to recognize. If we want to get only three clusters then the purity obtained will be very less for any value of eps and minpts. But we can increase the purity at the cost of having large number of clusters by decreasing the value of minpts.

eps	**Minpoints**	**Purity**	**Misclassified**	**No. of Clusters**
0.9	2	0.366	190	1
0.06	1	0.870	39	11
0.05	1	1.000	0	30
0.05	7	0.533	140	3

The purity values for different types of *hierarchical clustering* with number of clusters = 3 are tabulated below. And as we can observe the purity obtained is maximum when the linkage type is ***Ward***.

Linkage Type	**Purity**	**Misclassified**
Single	0.373	188
Complete	0.706	88
Average	0.73	81
Mean	0.70	90
Centroid	0.733	80
Ward	0.753	74
AdjComplete	0.64	108
NeighborJoin	0.366	190

Spiral

DBSCAN performs really well when we have a small value of eps and a small value of Minpts as well since the points are thinly arranged very close to each other. When we take eps as 0.1 and minpts as 2 we get the exact input classes as our output clusters. But for large value of eps it groups all the datapoints as a single cluster.

eps	Minpoints	Purity	Misclassified	No. of clusters
0.9	6	0.340	206	1
0.1	2	1.000	0	3
0.05	1	1.000	0	3
0.25	4	0.340	206	1

The purity values for different types of *hierarchical clustering* with number of clusters = 3 are tabulated below. And as we can observe clearly the purity obtained is maximum when the linkage type is ***Single***.

Linkage Type	Purity	Misclassified
Single	1.00	0
Complete	0.381	193
Average	0.362	199
Mean	0.0.391	190
Centroid	0.414	186
Ward	0.411	187
AdjComplete	0.356	201
NeighborJoin	0.339	206

Flames

DBSCAN performs well on flames dataset. We even get a purity of 1 for some value of eps and minpts but the number of clusters in that case is 57 while we have only 2 classes in our dataset. But for eps=0.1 and minpts=9 we get a purity of 0.975 with just two clusters formed which is really close to the original dataset.

eps	Minpoints	Purity	Misclassified	No. of clusters
0.9	6	0.637	87	1
0.1	9	0.975	6	2
0.05	1	1.000	0	57
0.1	3	0.637	87	1

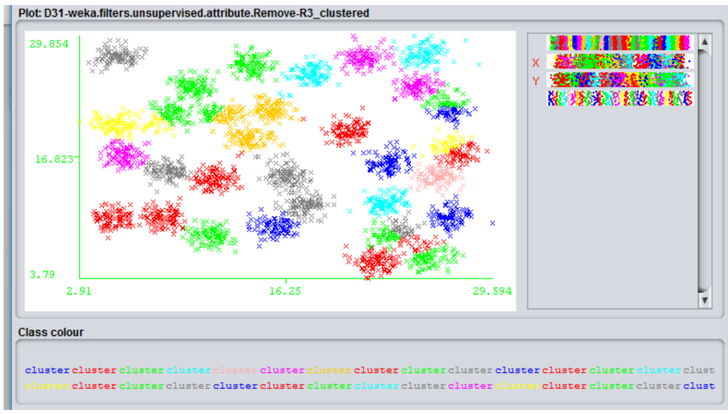
The purity values for different types of *hierarchical clustering* with number of clusters = 2 are tabulated below. And as we can observe the purity obtained is maximum when the linkage type is ***Ward***.

Linkage Type	Purity	Misclassified
Single	0.646	85
Complete	0.516	116
Average	0.833	40
Mean	0.921	19
Centroid	0.646	85
Ward	1.000	0
AdjComplete	0.641	86
NeighborJoin	0.637	87

Question 6 – D31 Dataset

K-Means

For K=32 we get a purity of **0.879**. We are able to recover the 31 clusters but with some amount of error.

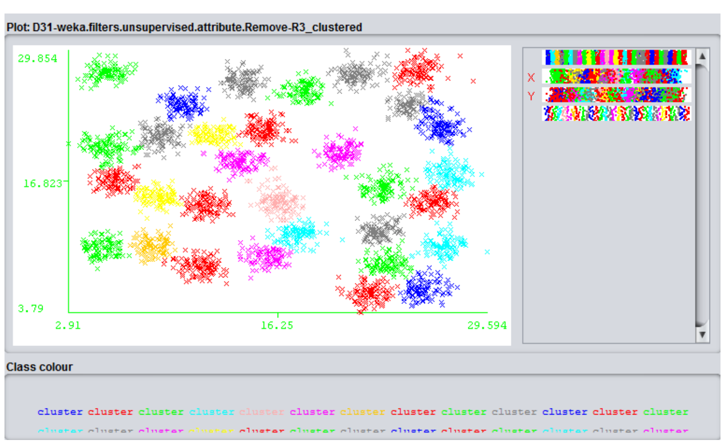


As we keep increasing K the purity also goes up linearly and becomes saturated at a point as we can observe. This is because of increasing granularity and thus the data distribution becomes skewed in each cluster.

K	Purity
32	0.879
40	0.897
48	0.957
56	0.965
64	0.967
128	0.966

Hierarchical Clustering

Using Ward linkage we get a best purity of **0.9632** when the input number of clusters is 32.



DBSCAN

DBSCAN doesn't work well because the clusters are not properly density-separated. Many of the times, nearby clusters are merged together as a single cluster. Varying minpts between 1 and 20, eps between 0.05 and 1, we never get more than 5 clusters in the output. The best purity obtained is **0.1613** when eps = 0.05 and Minpts = 14.

eps	Minpoints	Purity	Misclassified	No. of clusters
0.9	6	0.032	3000	1
0.05	17	0.161	2596	5
0.05	14	0.161	2597	5
0.06	7	0.065	2899	2