ML Programming Assignment 3 - Report

Ishu Dharmendra Garg- CS13B060

November 2015

1 Problems on Clustering

Problem 1: Generate Arff

The datasets were converted into ARFF format manually using sublime text.

Problem 2: Visualize data

The given ARFF datasets were plotted (2-D plots) using visualize option of the weka explorer and the plots of all the datasets are as follows:

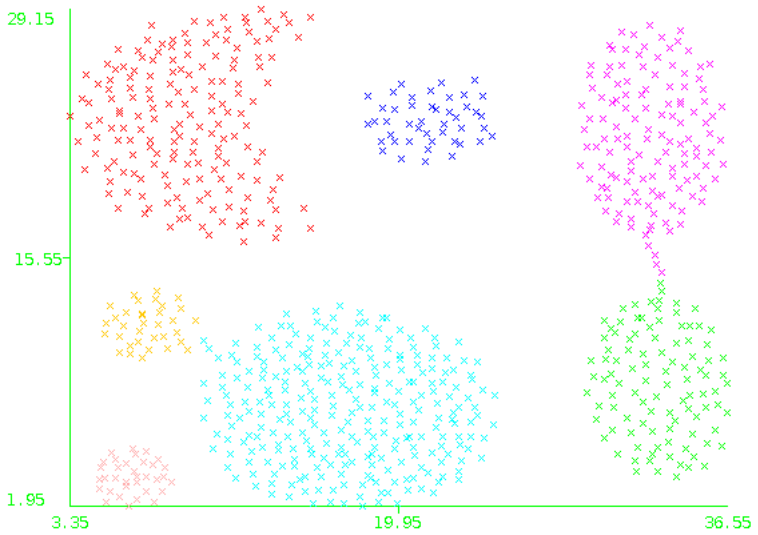

Figure 1: Aggregation

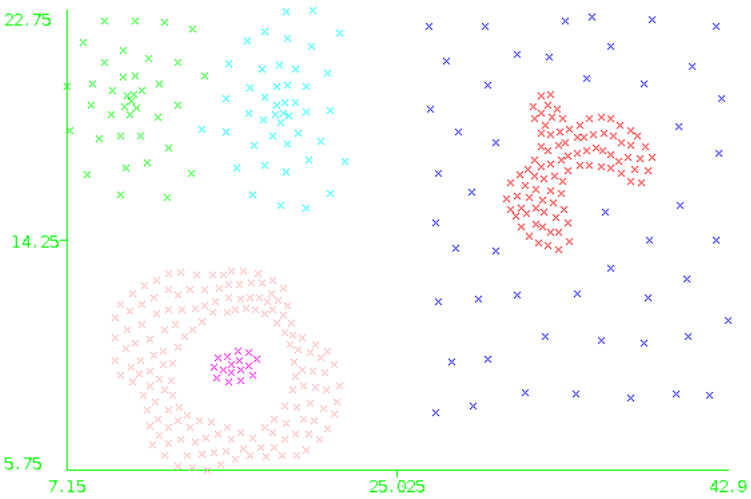

Figure 2: Compound

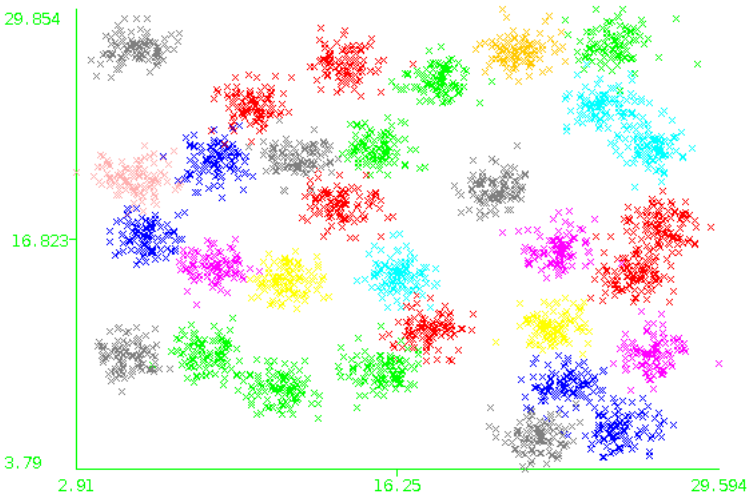

Figure 3: D31

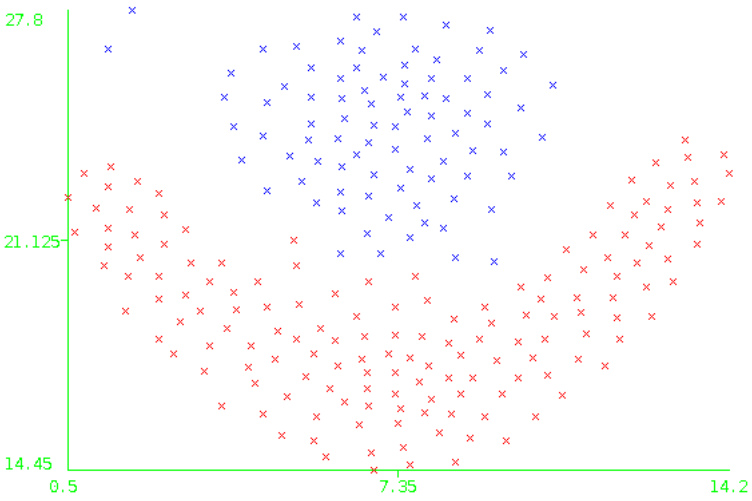

Figure 4: Flames

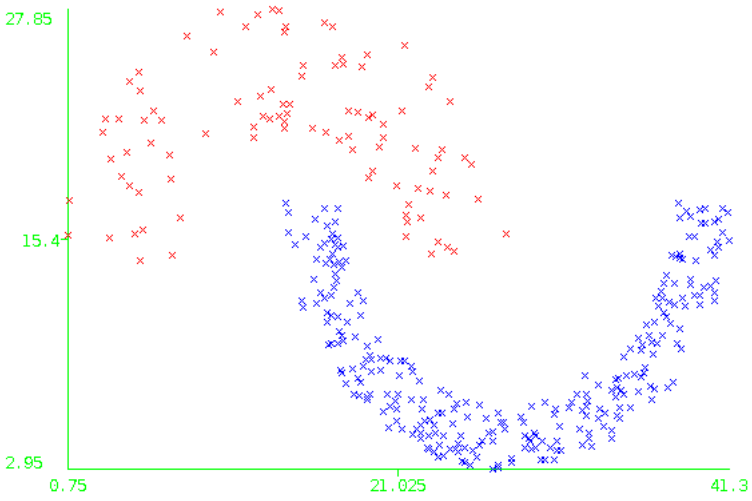

Figure 5: Jain

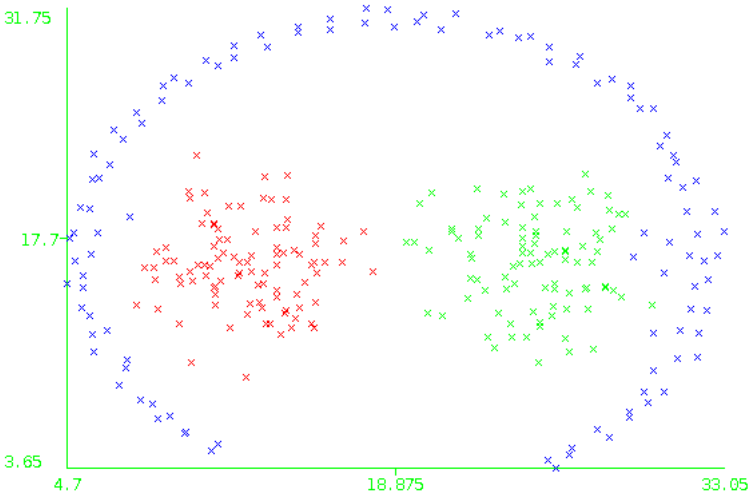

Figure 6: Path Based

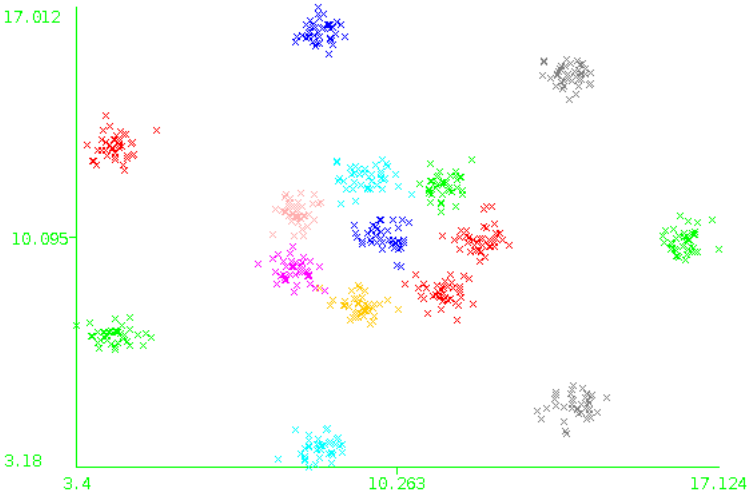

Figure 7: R15

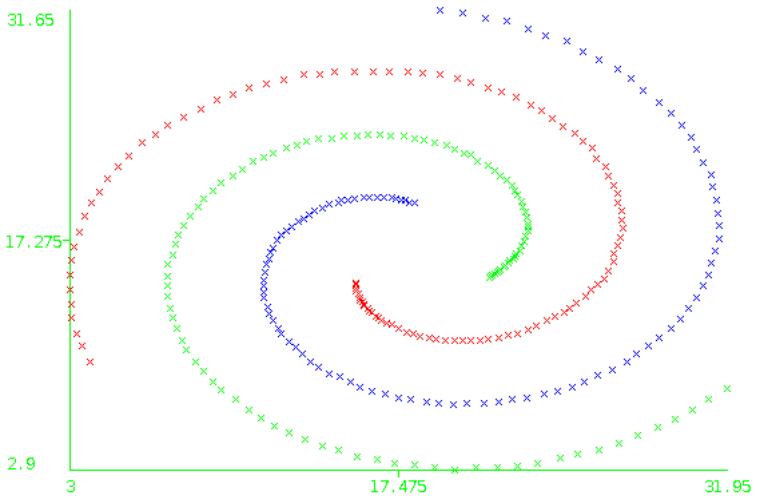

Figure 8: Spiral

Looking at the plots we can infer the following things:

- Aggregation:
  - K-means: We can see that some of the classes are really close . There is a chance the k means will wrongly cluster some portions of these classes. Otherwise this method will work more or less satisfactorily.
  - DBSCAN: Classes that have a link will not be clustered properly due to high density link between them. Well separated classes will be clustered correctly.
  - Hierarchical [single link]: Two connected classes would be treated as a single long cluster. This will be due to particular nature of the single link clustering where it tends to produce long clusters.
  - Hierarchical [complete link]: The max within class distance is more that the max distance between the classes. Hence the results will not be good.
- Compound:
  - K-means: The classes are too much spreaded and interspersed. We also have nested circularly spreaded classes which will cause k-means to cluster them together. And hence the results would be poor for this method.
  - DBSCAN: The class marked with dark blue color is too sparse compared to other classes. For small epsilon and considerable minPoints we can classify the some classes pretty well but at the same time lot of points will be unclassified. If we try to increase the epsilon or

decrease the minPoints some classes would start merging up. Hence due to all these factors this method will not be appropriate for this dataset.

- Hierarchical [single link]: Dark blue class is highly spreaded out. Hence that will be clustered in several other cluster. Rest of the classes are decently densed and hence those will be clustered properly.
- Hierarchical [complete link]: Dense classes will be clustered in the same cluster. But the sparse class will be fragmented among different clusters.

- D31 dataset:
  - K-means: Classes are compact and decently separated. This method will perform decently on the given data set.
  - DBSCAN: Almost all the classes are touching to each other (which sometimes are very dense). Hence It will not be able to separate these classes. The results would be poor.
  - Hierarchical [single link]: No chance! Since the classes are touchy :). There boundary point between two classes will be clustered together and hence we will not see good results.
  - Hierarchical [complete link]: Due to compactness and the spherical nature the farthest points within class are closer than the farthest points between classes. Hence we get some decent results.

- Flames dataset:
  - K-means: The points in the middle of the classes will come in the same cluster. But there would be some mis-clustering when we go towards the boundary side where the two classes touch.
  - DBSCAN: For appropriate parameters the results will be good. The classes are denser than the region where the two class touch.
  - Hierarchical [single link]: There will be problems where the two classes meet. Since between class points here would be more close they will be clustered together.
  - Hierarchical [complete link]: Do you see the two outliers? Those are much away from both the classes. And hence they will cause the two classes to be clustered in the same cluster. Also for the U class the farthest points are much away form each other and at the same time they are relatively closer to the points of the other class. Hence this method will not give us a proper clustering.

- Jain dataset:
  - K-means: It depends. If the centroids are close to the centers of the U shapes the clustering will be good.

  - DBSCAN: Would be good (in-fact very good). Since the two classes are well separated by low density regions and also the classes themselves very dense.
  - Hierarchical [single link]: For 2-3 clusters it will not be good. but for this method will be great if we take 5-6 clusters as the outliers will be be classified together and the U classes will be clustered separately.
  - Hierarchical [complete link]: Some points near the ends of the U classes will be mis-clustered as they are closer to the farthest points of the opposite class. But apart form that it will be decent clustering.

- Path Based dataset:
  - K-means: The big circled class will be broken down into several clusters. While the classes within will be clustered well.
  - DBSCAN: Since the parts of the inner clusters are in touch with the outer ring, some part of the inner clusters will also come in the outer cluster. Other that that the ring will be clustered together.
  - Hierarchical [single link]: Will be satisfactory. Since it is a single link, the long circle will be detected as well as the two inner classes.
  - Hierarchical [complete link]: No way! Points on the ring are more distant than the some point on the ring and the inner cluster.

- R15 dataset:
  - K-means: Clusters are compact, spherical and almost separated hence this will perform well.
  - DBSCAN: For small epsilon and moderate minPoints the performance will be good. This is because the classes are quite dense and separated by low density regions.
  - Hierarchical [single link]: The inner classes would be wrongly clustered. Will identify the outer classes in different clusters.
  - Hierarchical [complete link]: Classes are compact and well separated and hence the performance will be good.

- Spiral dataset:
  - K-means: The classes are not compact and are circularly interleaved and hence it will not be suitable for this dataset.
  - DBSCAN: Since the classes are dense are are well separated by the non dense region this method will perform well.
  - Hierarchical [single link]: Since the closest point within class are much more closer to the closest points between classes hence this method will be good for this dataset.
  - hierarchical [complete link]: There are cross class points that are separated by distance less than the distance between some of the within class points. So this will not be a right choice for the given dataset.

Question 3: K-means on R15

The cluster purity when I ran K-means on R15 for K = 8 was 0.533. Since the classes are compact and are well separated we see that increasing the k increase the cluster purity.

The Plot of cluster purity for various k is as follows:

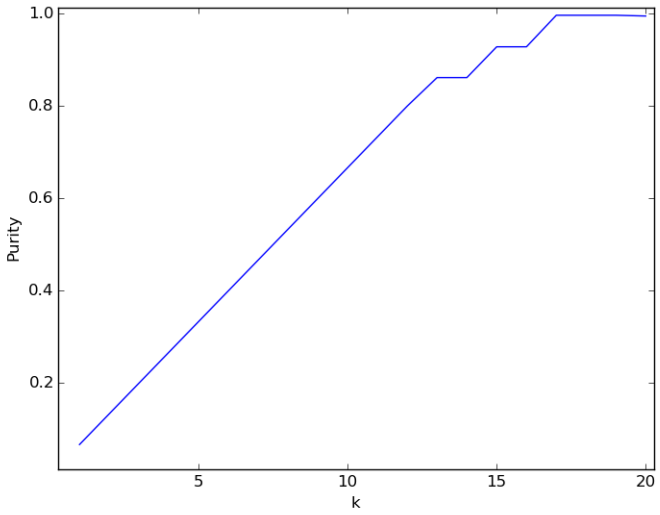

Figure 9: Cluster purity vs K for R15

Question 4: DBSCAN on Jain

If we keep epsilon small the two classes would be separated. But if we increase the epsilon for given minPoints the clusters starts to have points from both the classes and hence the cluster purity starts degrading. For a given epsilon if we change the minPoints we see that red class stars splitting up but the blue class remains clustered. The purity does not change much. The cluster purity for the various values of epsilon and minPoints are as follows:

ϵ	minPoints	Purity
0.01	2	0.7453
0.03	3	0.9946
0.04	5	0.9946
0.05	2	1
0.05	5	0.9973
0.05	10	0.9946
0.08	10	1
0.10	2	0.7399

Table 1: Cluster Purity for various values of ϵ and minPoints

Question 5: DBSCAN and Hierarchical

The comparisons between DBSCAN and Hierarchical clustering for the different data sets are as follows:

- Path Based:
  - Hierarchical: Out of all the best results were given by the Complete linkage. The ring was broken down in many clusters. Also some parts of the inner clusters were included in the clusters of the ring.
  - DBSCAN: The outer ring due to varying density is broken in many clusters. But the inner clusters are almost always recovered.
- Spiral:
  - Hierarchical: As expected the Single linkage gave the optimal results. All the spiral paths were distinctly clustered.
  - DBSCAN: Since the paths are highly dense and are well separated by the non dense region this methods also performs well.
- Flames:
  - Hierarchical: Very good results were obtained by this taking ward linkage. The two classes were clustered separately.
  - DBSCAN: The two classes are dense in the interior but relatively sparse at the boundary. After some parameter search classes were clustered separately and only a few points remained unclustered.

Question 6: K-means, DBSCAN, Hierarchial on D31

- K-means: I did can not recover all 31 clusters with k = 32. There were some clusters which were merged. If we increase k we see no significant difference till k = 42 when there were no cluster with the two different classes. But some of the classes were divided into two or three clusters.

- DBSCAN : For good results I had to fine tune the parameters. At minPoints = 20 and epsilon = 0.02905, I got all the classes in separate cluster with with 227 unclassified points.
- Hierarchical : With the ward linkage we get all the clusters. This is due to the property of the ward linkage where it tries to minimize the variance of the merged clusters. Since the classes are compact we get the result as expected.

2 Decision Trees

Question 1: Convert to ARFF format

Using the information provided in the names and the data file, and by using the sublime text two ARFF files were made, one for training and other for testing.

Question 2: Run J48 Decision Tree

Looking at the results one can say that the results are much better if we have low number of MinNumObj and we do not do the pruning. In general we see that for low MinNumObj the results are similar with or without pruning. But for high MinNumObj the results are not good when we do the pruning.

The following tables shows the observations:

Precision	Recall	F-Measure	Class
1	1	1	e
1	1	1	p

Table 2: MinNumObj = [2, 10, 20]; Pruning = False

Precision	Recall	F-Measure	Class
0.985	1	0.993	e
1	0.989	0.995	p

Table 3: MinNumObj = [20, 40, 100, 500, 1000]; Pruning = True — False

Precision	Recall	F-Measure	Class
1	0.06	0.114	e
0.602	1	0.752	p

Table 4: MinNumObj = 3000, Pruning = False

Precision	Recall	F-Measure	Class
0.413	1	0.584	e
0	0	0	p

Table 5: MinNumObj = 3000, Pruning = True

Question 3: Important Features

By looking at split points of the decision tree we can say that the important features are :-

1. odor
2. stalk-shape
3. spore-print-color
4. gill-size
5. gill-spacing
6. population