Programming Assignment 3

Machine Learning

CS5011

Clustering

Ashutosh Kumar Jha
ME14B148

15 November, 2016

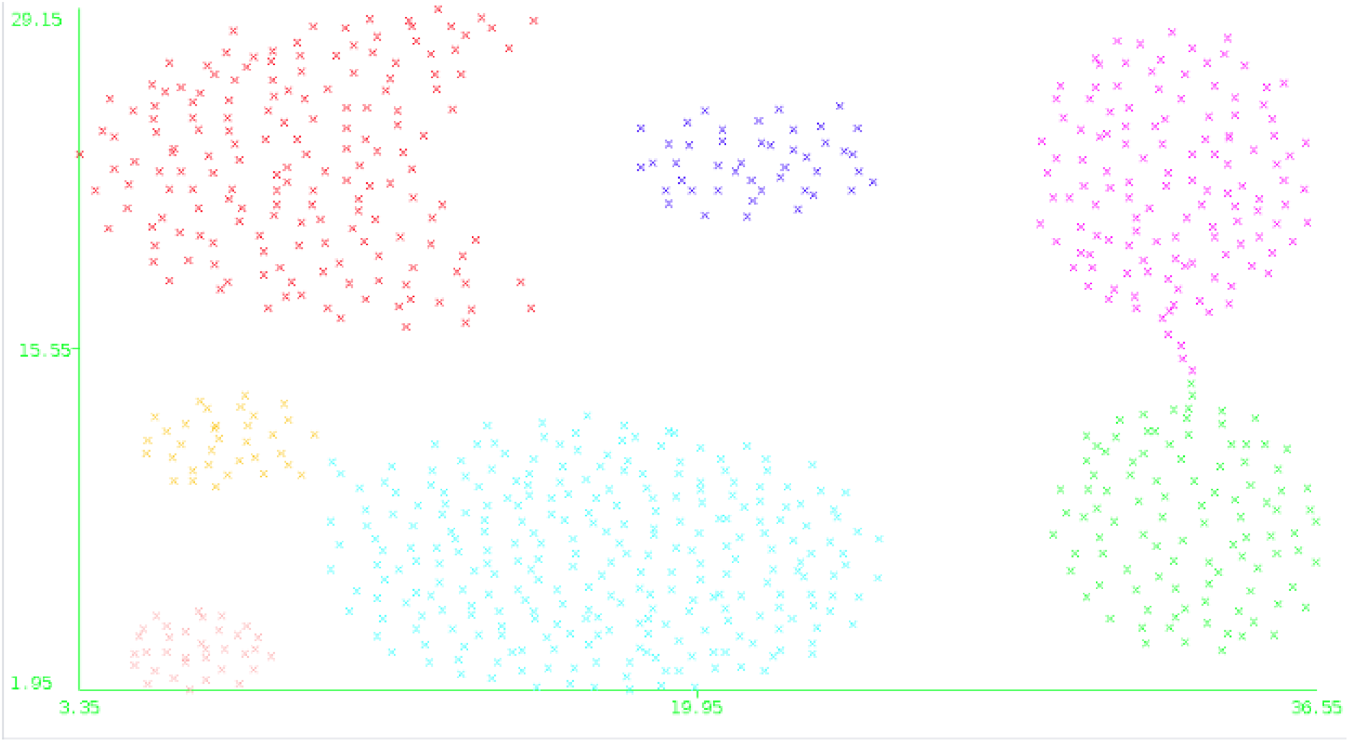

Figure 1: Aggregation Dataset Visualization

1. Converting Datasets to ARFF Format

The datasets were provided in *.txt* format. They were converted to ARFF format by adding the required headers. For the spiral dataset, the name of the columns were removed before adding the headers.

2. Visualization and Analysis of Datasets

2.1. Aggregation

k-Means: Kmeans should work well since the clusters are circular in shape. It might make some errors but it'll make sense on an overall
DBSCAN: DBSCAN will not work here since there are two cases where two clusters are very close to each other i.e. there exists a density connected path.
Hierarchical clustering with single link: Single Link will not work well because it will merge the ones which are path-connected into a single cluster at the base of the dendogram.
Hierarchical clustering with complete link: This will work well because it looks at the farthest points in the cluster.

2.2. Compound

k-Means:k-means will fail due to the annular shape of the clusters.
DBSCAN: This will also fail as the different classes have different densities.
Hierarchical clustering with single link: This will work reasonably well. It might merge some clusters as we go up the dendogram.
Hierarchical clustering with complete link: This will fail as it looks at the distance between the farthest points and will merge the annular clusters.

2.3. Path-Based

k-Means: Will fail due to the outer cluster.
DBSCAN: This will do fail due to a path between the outer class and inner cluster on the right.Furthermore, the density is low at the bottom of the outer class.
Hierarchical clustering with single link: It won't work because of the outliers at the bottom.
Hierarchical clustering with complete link: This will fail as the outer class has extreme points which encompass the full space. As a result, a part of the outer class will be merged low in the dendogram.

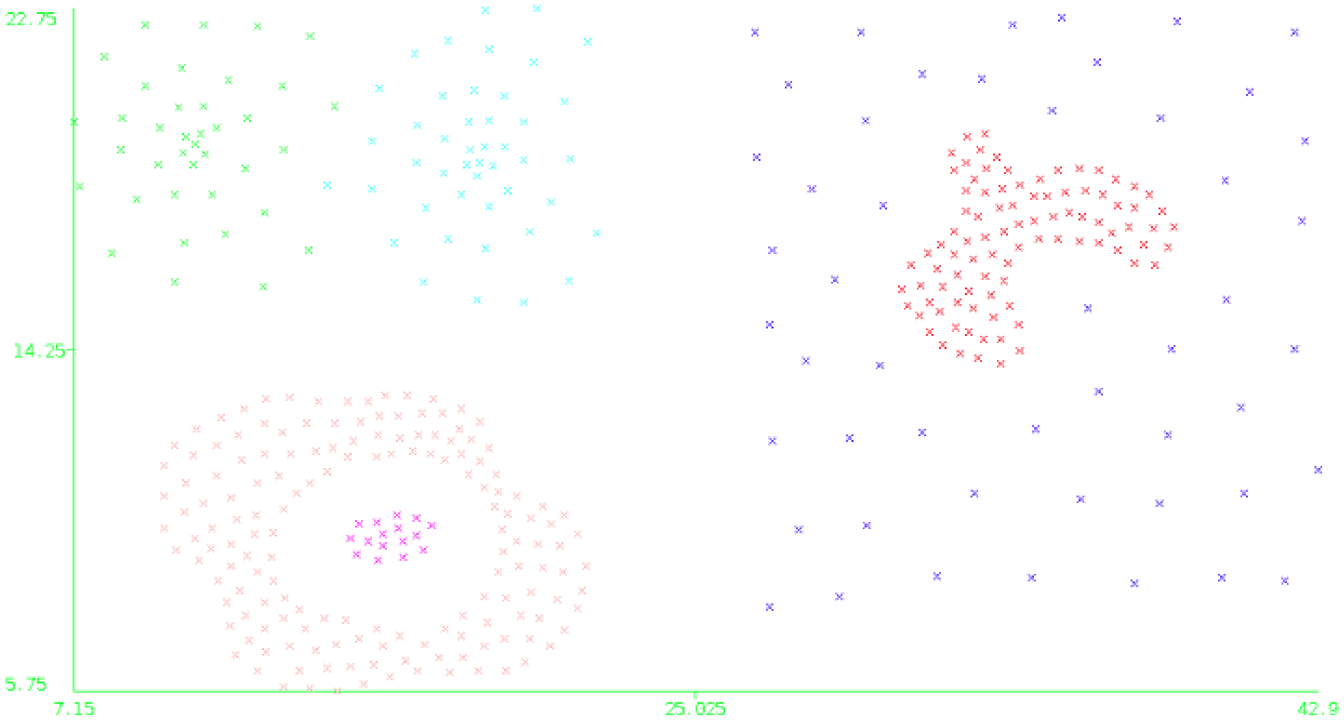

Figure 2: Compound Dataset Visualization

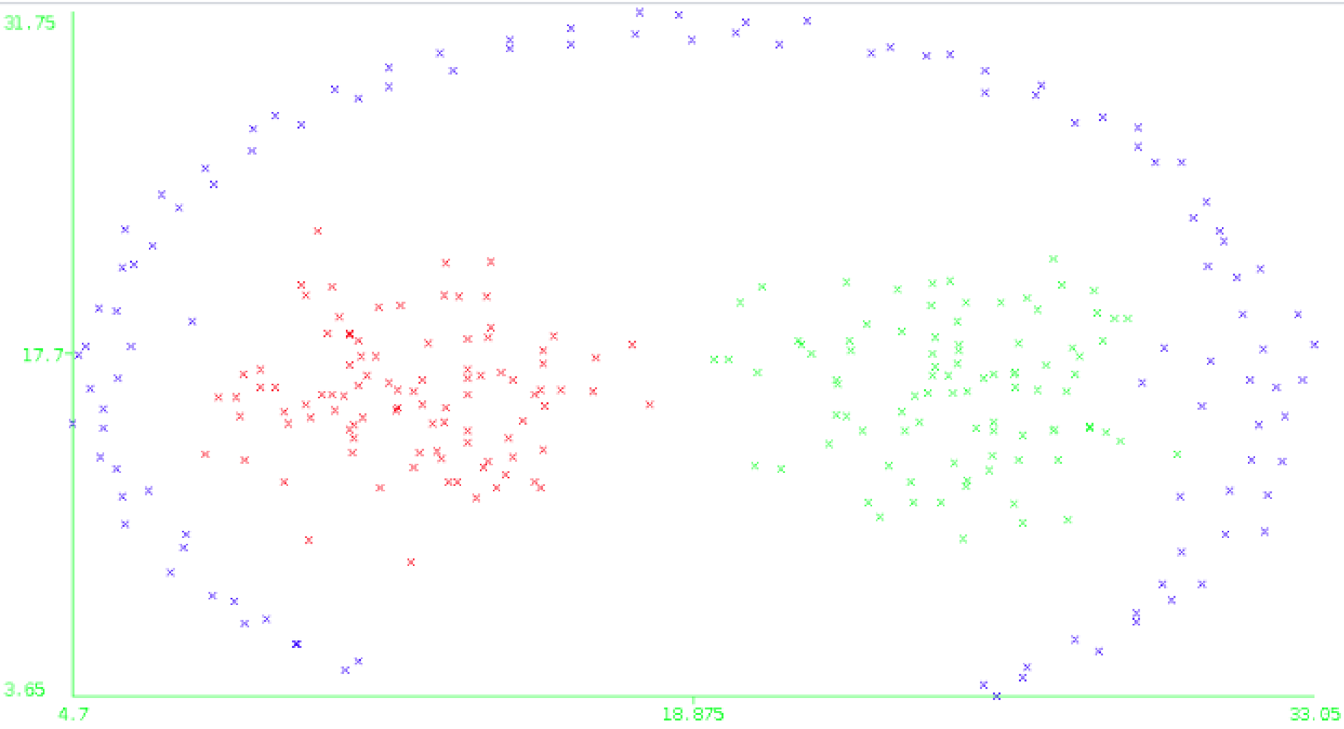

Figure 3: Path-Based Dataset Visualization

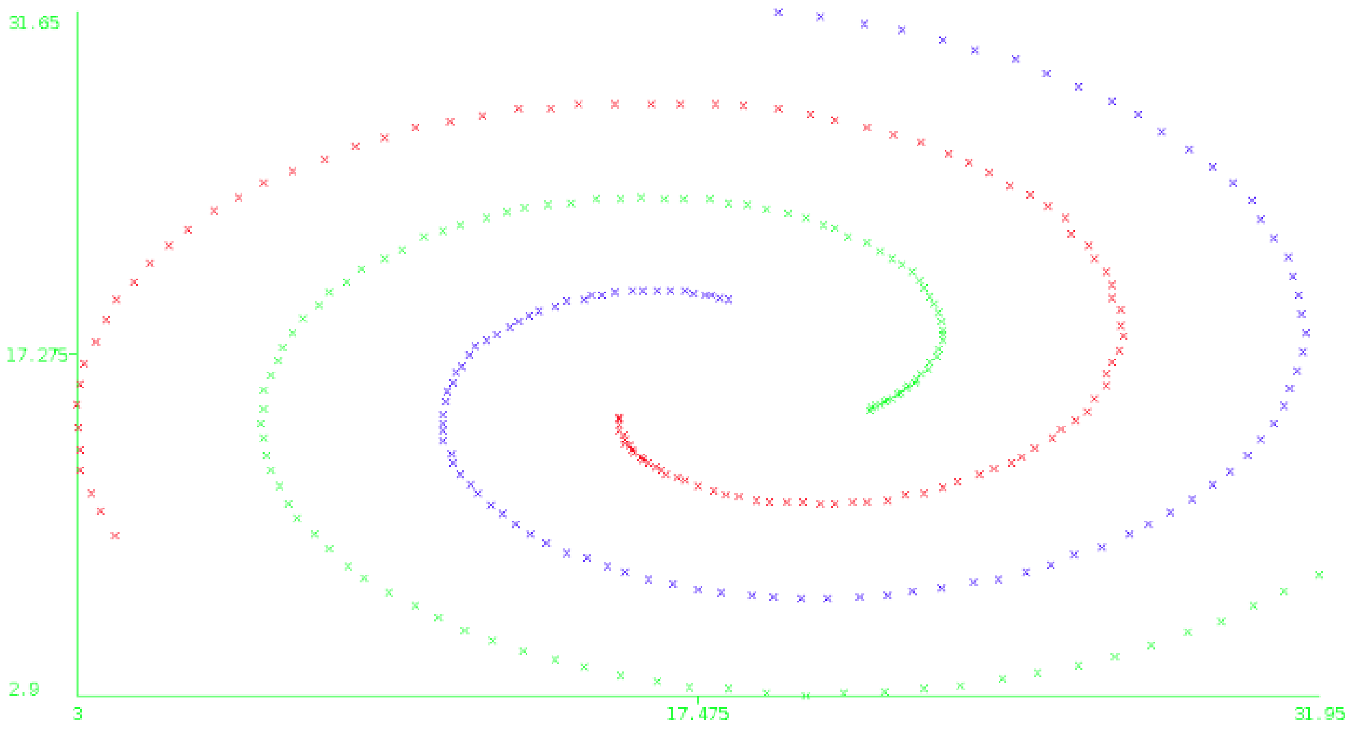

Figure 4: Spiral Dataset Visualization

2.4. Spiral

k-Means: Will fail as the classes are not spherically distributed.
DBSCAN: Will perform reasonably well as the classes are well seperated.
Hierarchical clustering with single link: This will fail as it makes errors because the distance between points from different classes are smaller than the ones in the same class.
Hierarchical clustering with complete link: This will fail as the classes are not tight.

2.5. D31

k-Means: Clusters can be partially recovered but some clusters are too merged. As a result it might make some errors.
DBSCAN: Will fail as the classes are merged at the left part of the space.
Hierarchical clustering with single link: Will fail as it will merge the classes which have overlap at the base of the dendogram.
Hierarchical clustering with complete link: This will work as the clusters are tight and spherical in nature.

2.6. R15

k-Means:k-means will perform well on this cluster for an appropriate value of k as the cluster has circularly distributed classes.
DBSCAN: This will work well as the density is almost same over the space.
Hierarchical clustering with single link: This will fail and merge the classes in the middle of the space as they are close to each other.
Hierarchical clustering with complete link: This will work well as the clusters will form complete links at the base of the dendogram.

2.7. Jain

k-Means: This will fail as the classes are not spherically distributed.
DBSCAN: This will work well as the classes are well seperated. This might make 3 clusters as the upper class has a sparse region in between.
Hierarchical clustering with single link: This will work well but give out more clusters as the upper class has a sparse region.
Hierarchical clustering with complete link: This will not work well as the clusters are not tight in nature.

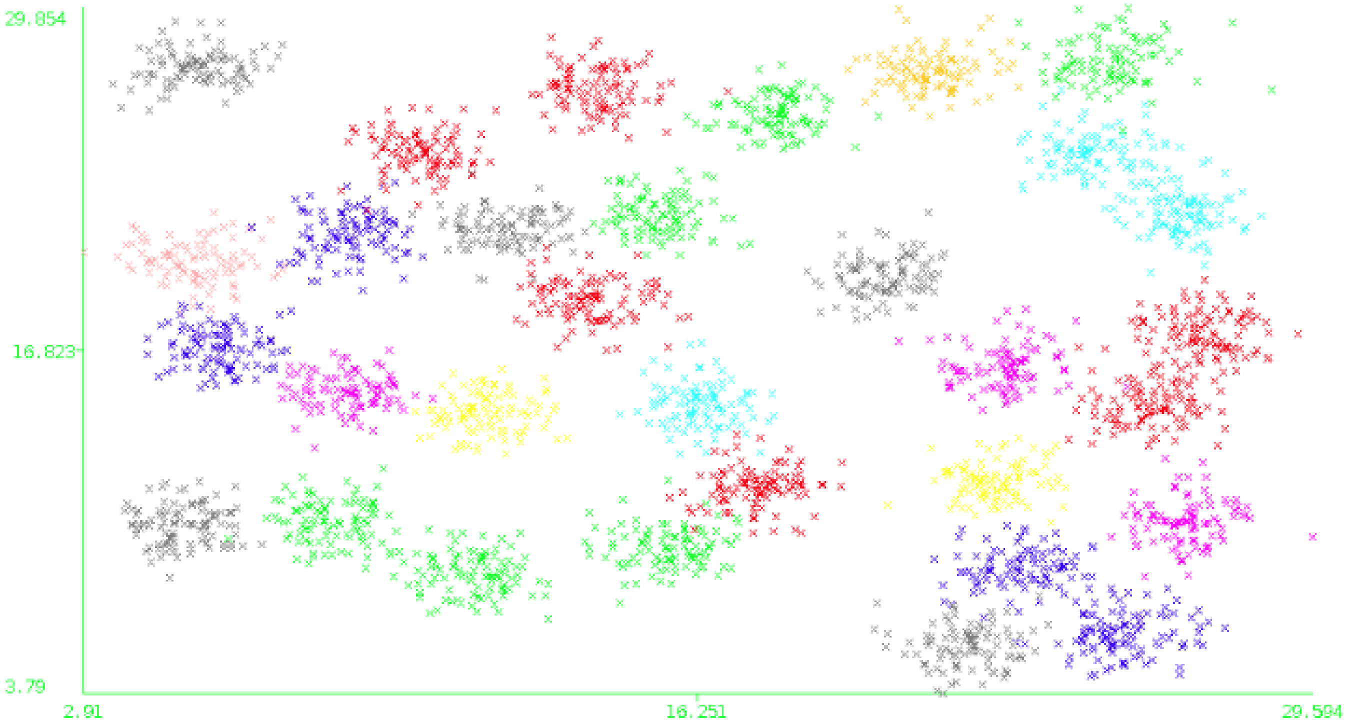

Figure 5: D31 Dataset Visualization

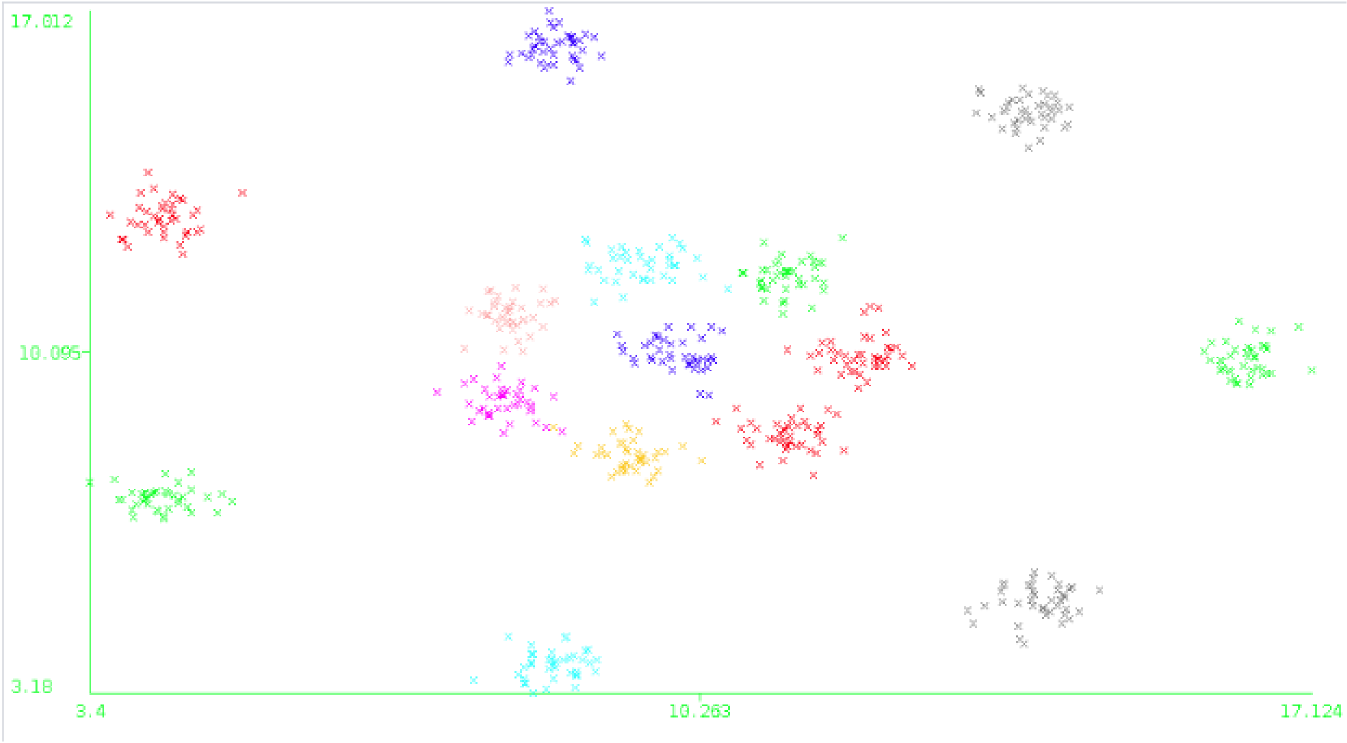

Figure 6: R15 Dataset Visualization

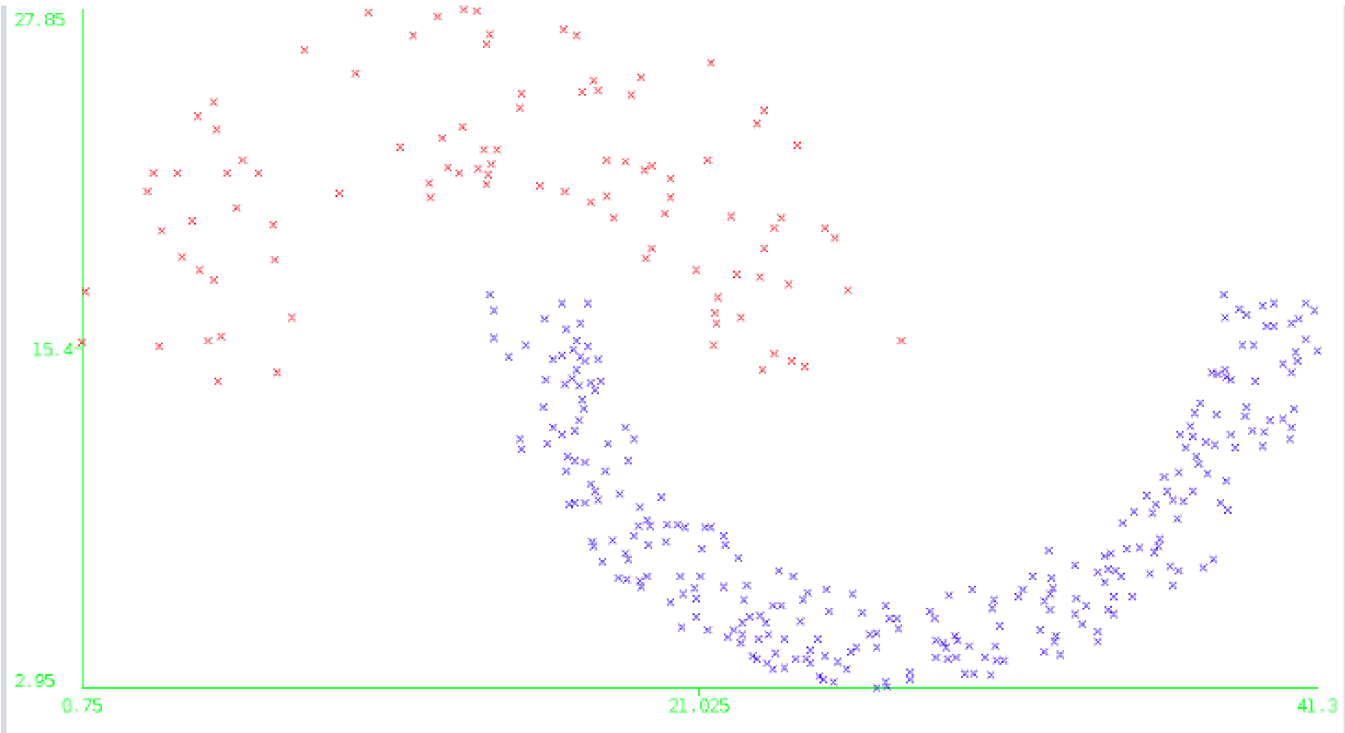

Figure 7: Jain Dataset Visualization

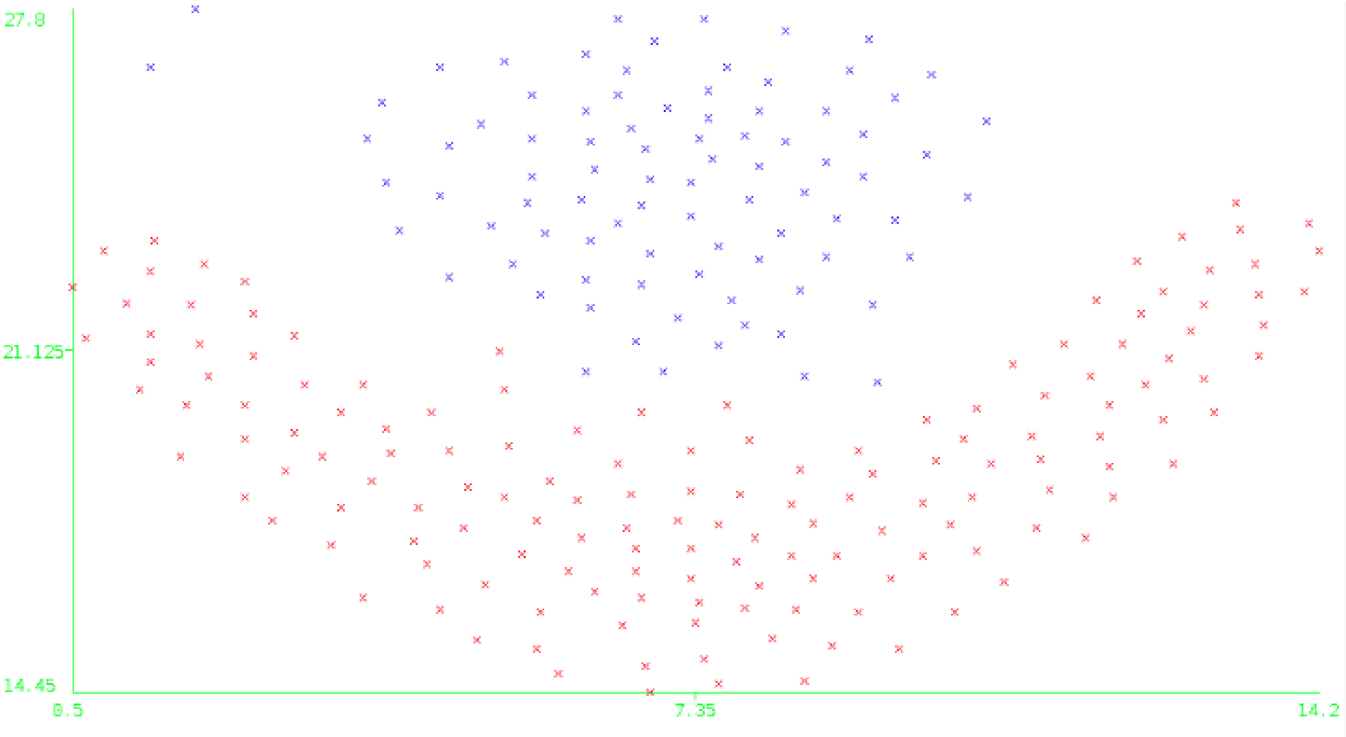

Figure 8: Flames Dataset Visualization

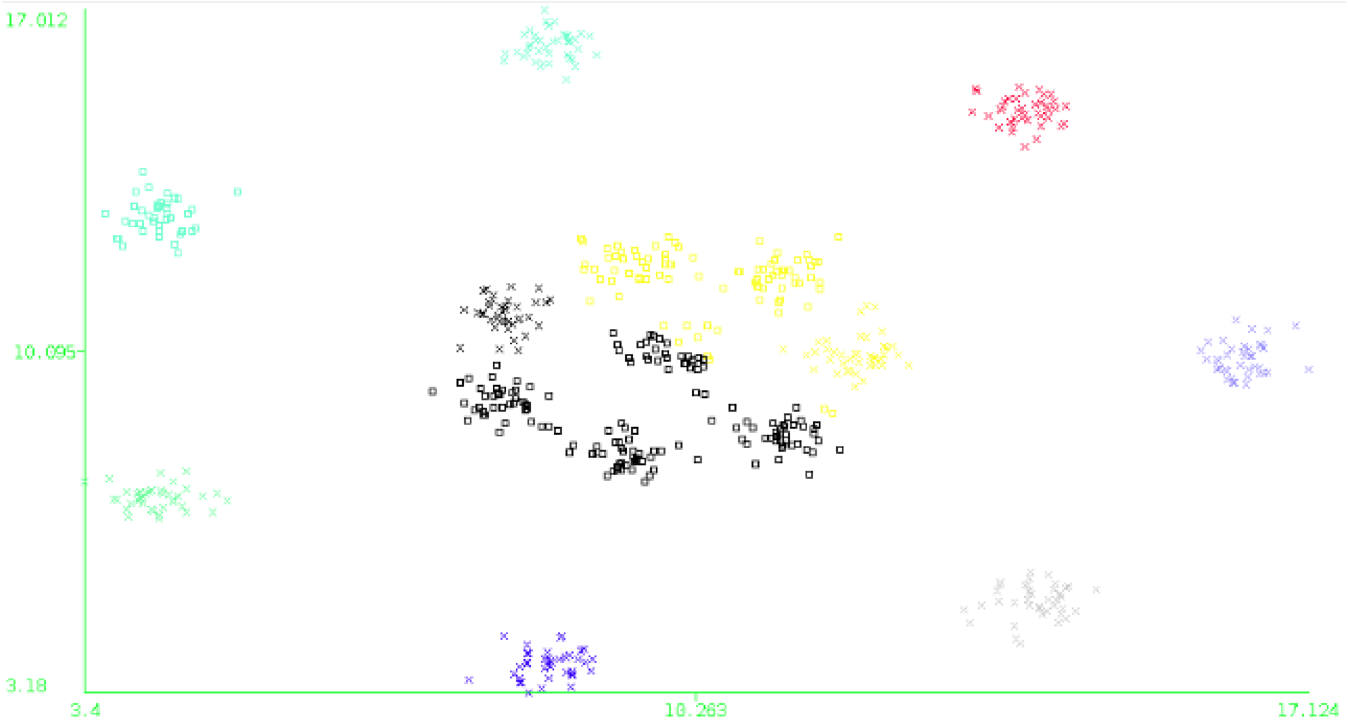

Figure 9: k=8 R15 Clustering Visualization

2.8. Flames

k-Means:k-Means will fail as the classes are not spherically distributed.
DBSCAN:DBSCAN will work well as the classes are well seperated but will leave out the outliers at the top left corner.
Hierarchical clustering with single link: This will not work well if no. of clusters is only 2 as there are outliers in the data. Though for higher number of clusters it does work well.
Hierarchical clustering with complete link: This won't work well as the classes are not tight in the space and complete-link returns only tight classes.

3. k-Means with R15 Dataset

First, k-means was run on the R15 dataset with $k = 8$. The cluster purity obtained was as follows:

Cluster	1	2	3	4	5	6	7	8
Purity	1	0.211	1	1	1	1	0.5	0.305

Table 1: k-means on R15 with k=8

A visualization of the clustering has been provided in figure 9. The value of k was then varied from 1 to 20. A plot of k vs average cluster purity was obtained as shown in Figure 10.
From the plot, we see that the cluster purity keeps increasing till $k = 19$ after which it starts dropping. From the visualization, it is intutive that $k = 15$ would perform best but it is not the case as the clusters are not very seperated and thus, the purity is disturbed and $k = 19$ performs the best with the dataset.

4. DBSCAN with Jain Dataset

DBSCAN was applied to the Jain Dataset. The values of ϵ and $minPoints$ were varied to get the best clustering possible.

A value of $\epsilon = 0.08$ with $minPoints = 5$ gave a good clustering. The visualization of the clustering has been shown in Figure 11.

The cluster purity obtained were:

Cluster	1	2	3
Purity	1	1	1

Table 2: Cluster Purity with DBSCAN on Jain set

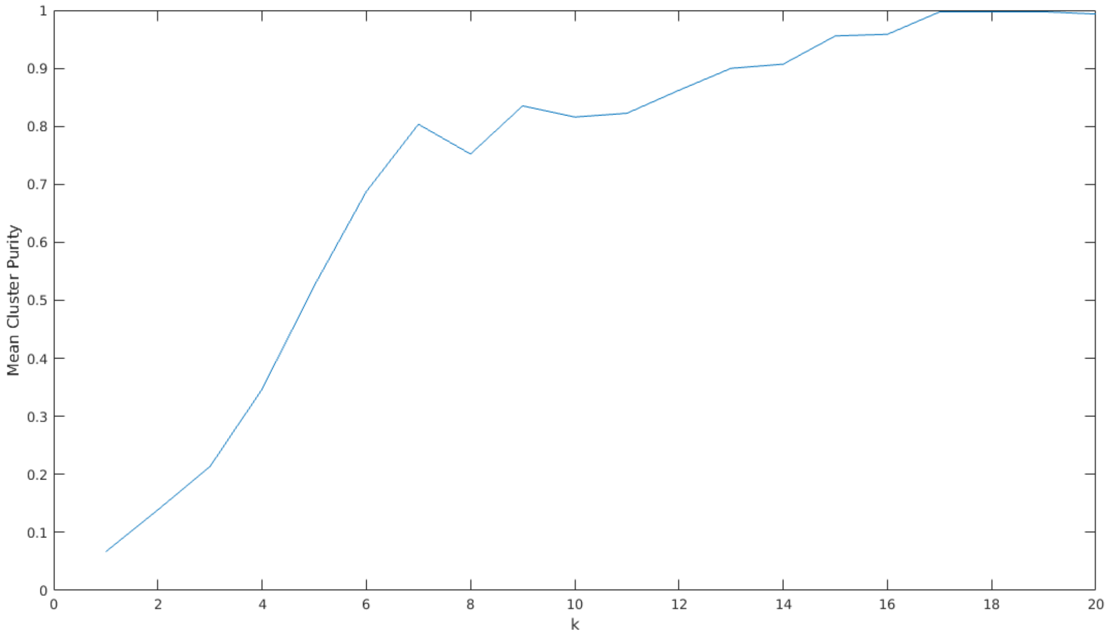

Figure 10: Plot of k vs Average Cluster Purity

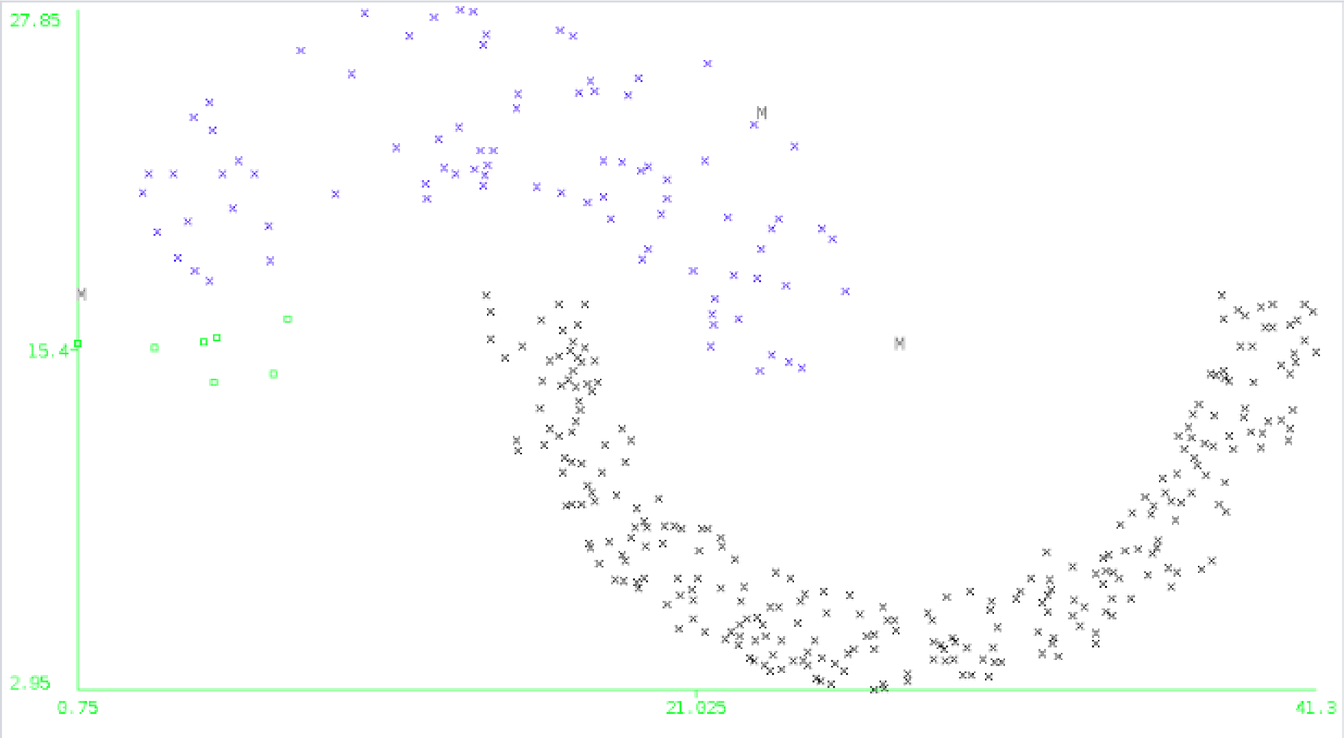

Figure 11: Best DBSCAN Clustering on Jain Dataset

4.1. Effect of Increasing ϵ

The value of *minPoints* was kept at 5 and the value of ϵ was increased in steps.
It was observed that as ϵ increased, all the points came into a single cluster and remained to be so at higher ϵ values. For lower values of ϵ, the no. of clusters went up and the number of unclustered instances went up too. This shows that at lower ϵ, the clusters are tighter.

Epsilon	**No. of Clusters**	**Avg Cluster Purity**	**Unclustered Instances**
0.085	3	1	3
0.1	1	0.74	0
0.2	1	0.74	14
0.05	5	1	45
0.025	21	1	142

Table 3: Effect of Changing ϵ

4.2. Effect of Increasing minPoints

MinPts	**No. of Clusters**	**Avg Cluster Purity**	**Unclustered Instances**
5	3	1	3
6	4	1	5
7	3	1	14
8	4	1	15
9	4	1	37

Table 4: Effect of Increasing *minPoints*

The value of ϵ was kept at 0.85 and the value of *minPoints* was increased in steps.
It was observed that as *minPoints* increases though the cluster purity doesn't change, the no. of unclustered instances keeps going up. Thus, the clusters formed are tighter in nature.

5. Tests on Path-Based, Spiral and Flames datasets

5.1. Tests using DBSCAN

Various values of ϵ and *minPoints* were tried for each of the datasets and the best are reported in the table below:

Dataset	**Number of Clusters**	**Incorrectly Classified Points**
Path Based	3	23
Spiral	3	0
Flames	2	0

Table 5: Tests using DBSCAN

As we see above, spiral and flames are classified well by DBSCAN but Path based isn't as intution stated in section 2. A table of number of incorrectly classified points is given in Table 6 below:

5.1.1. Using Hierarchical Clustering:

Linkage	**Path-Based**	**Spiral**	**Flames**
Single	189	0	85
Complete	88	193	116
Average	81	199	40
Centroid	80	186	85
Ward	74	187	0
Adj-Complete	108	201	86

Table 6: Tests using Hierarchical Clustering

From the table, it's evident that Spiral is clustered well by Single Link, and Flames by Ward Linkages.

Path Based fails in all cases of Hierarchical Clustering and is better neither with DBSCAN.

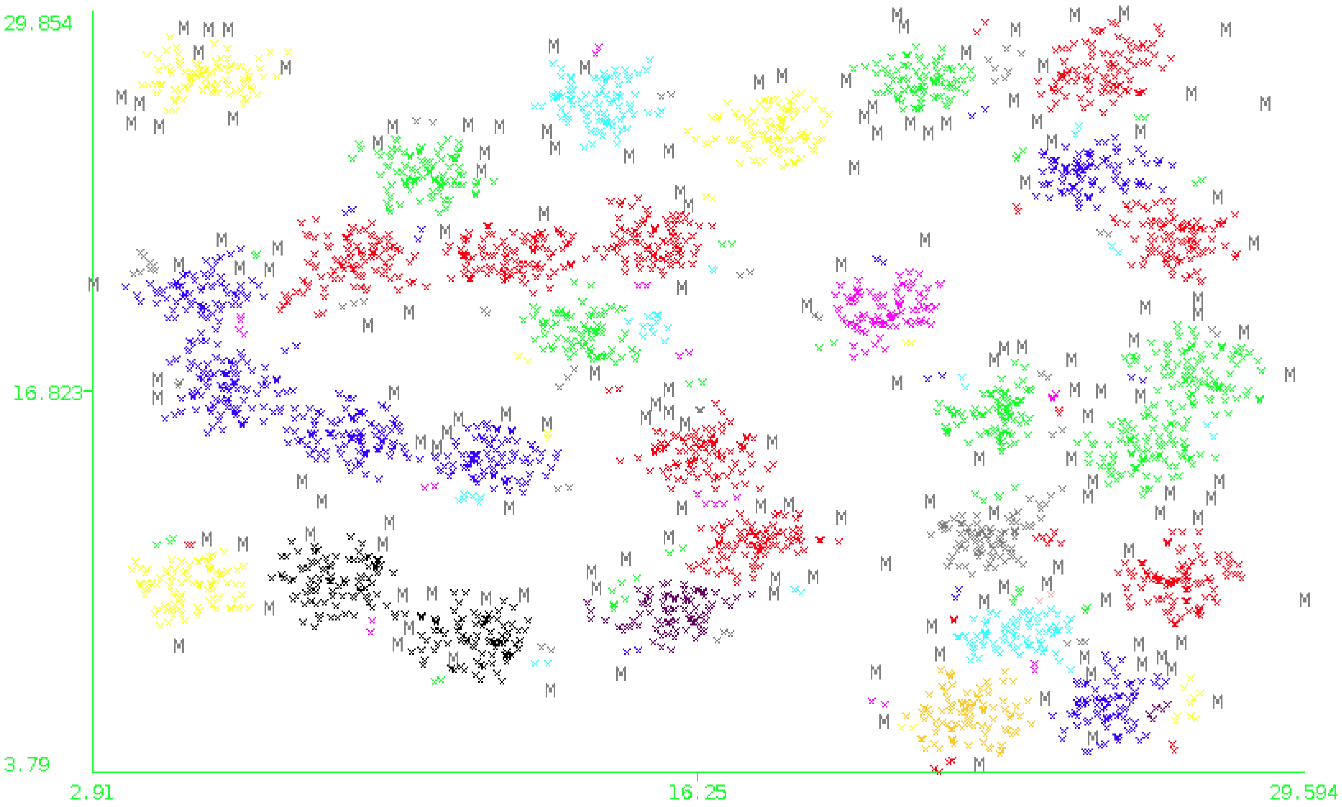

Figure 12: DBSCAN on D31

6. Tests on D31 Dataset

The k-means algorithm doesn't terminate on the full dataset in Weka. Thus, results are not presented in this report.

The DBSCAN algorithm terminates after a long time in Weka. Different values of ϵ and $minPoints$ were tried and the best were obtained at $\epsilon = 0.014$ and $minPoints = 2$. The clustering is a bit bad as some clusters are merged due to density connected paths. The visualization of the clusters is shown in Figure 12.

Hierarchical Clustering with Ward's Linkage also terminates after long time and is able to recover all clusters. The visualization is shown in Figure 13.

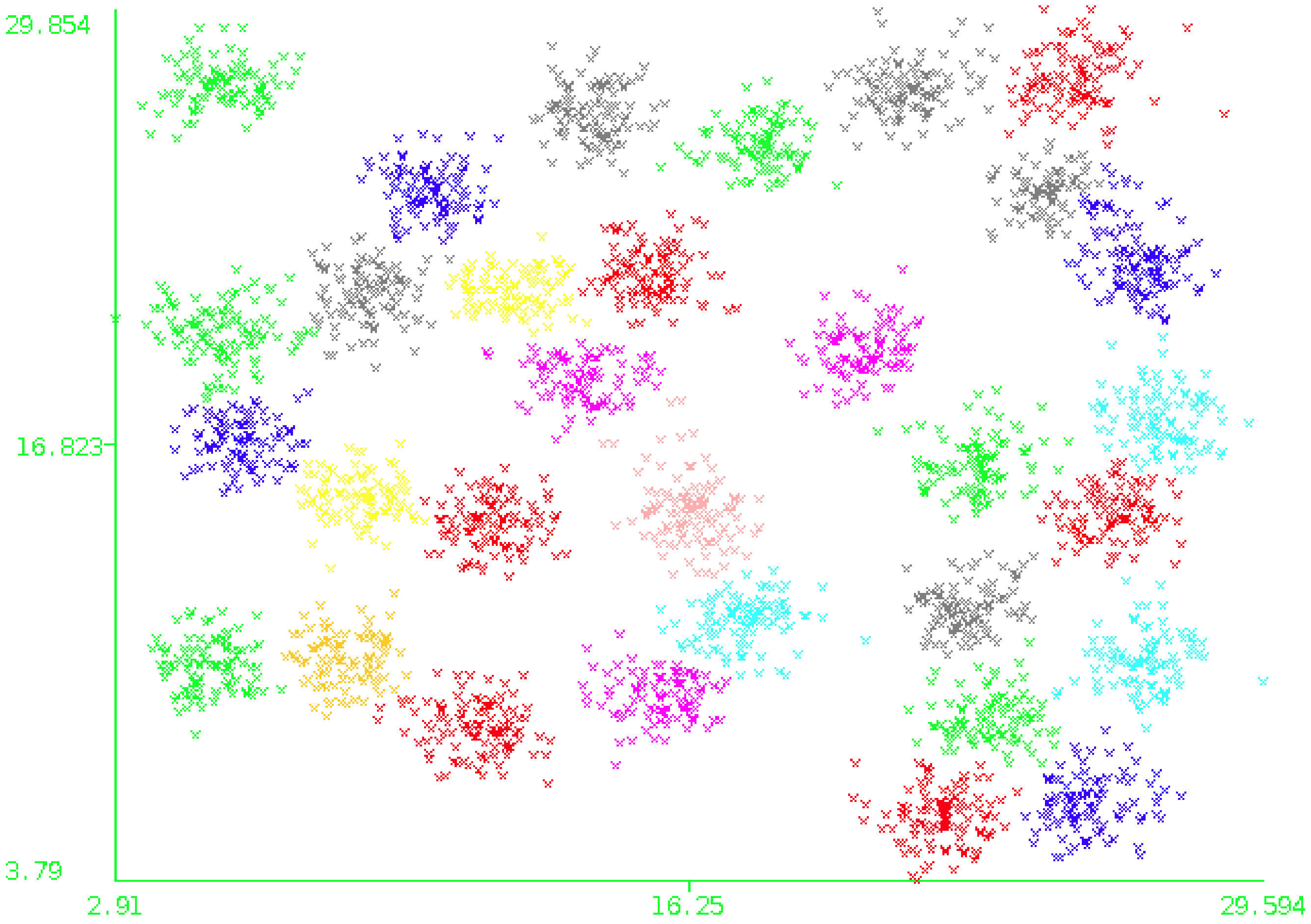

Figure 13: Ward Linkage Hierarchical Clustering on D31