**Introduction**

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by British Statistician and biologist Ronal Fisher in 1936, and is a classic dataset that is referenced in cluster analysis and classification.

The data set consists of 50 samples of each of three species of Iris, namely Iris Setosa, Iris Virginica and Iris Versicolog. Four features were measured from each sample: the length and width of the sepals and petals, in centimeters. Based on the combination of these four features, fisher developed a linear discriminant model to distinguish the species from each other. To have an understanding of the subjects that we are studying, let's have a look at the flowers.



Fig. Iris Setosa Fig. Iris Versicolor Fig. Iris virginica

For this lab, we are going to perform cluster analysis on this dataset, and wish to best classify the elements into their respective cluster, which is Iris Setosa, Versicolor or Virginica.

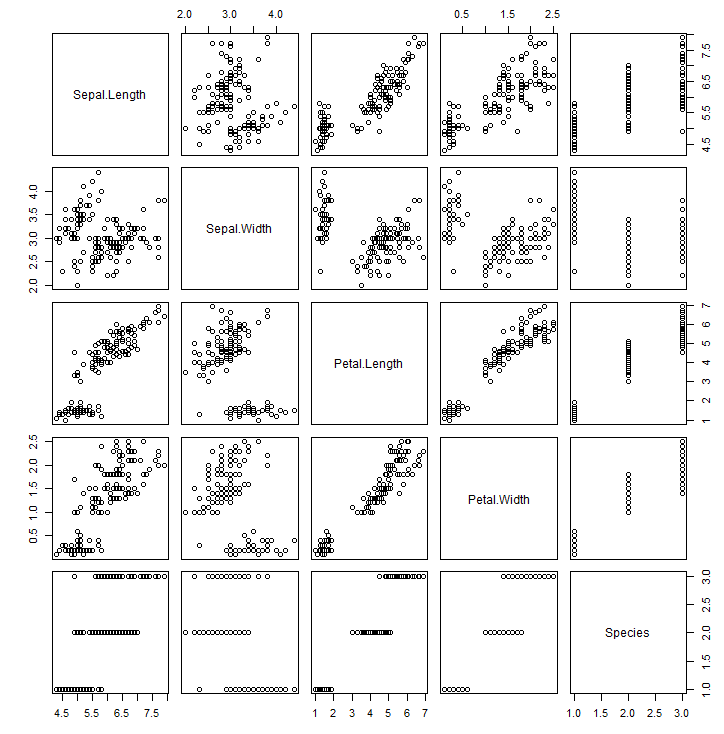


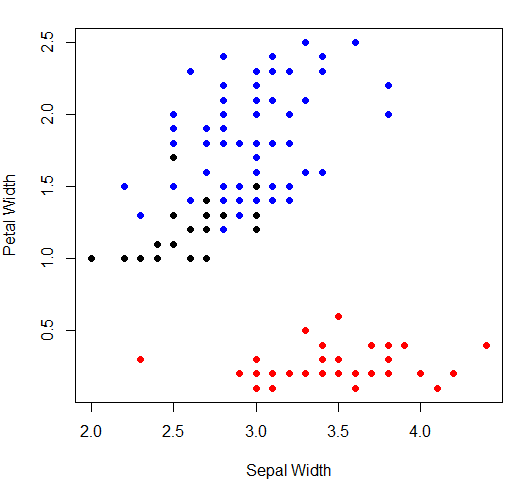
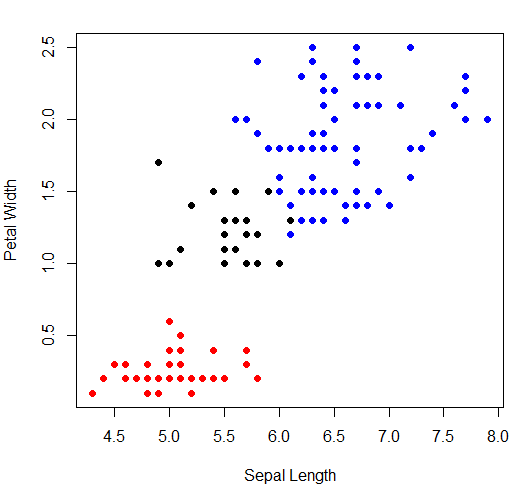
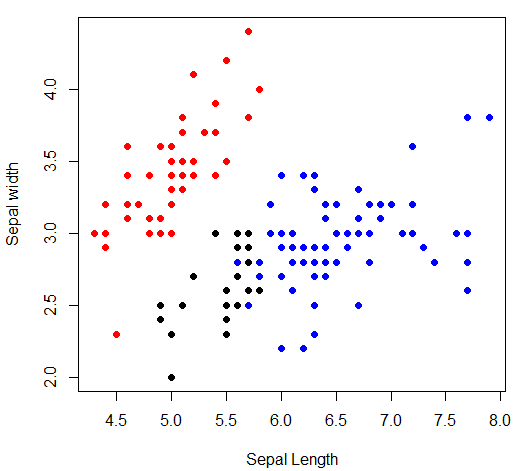
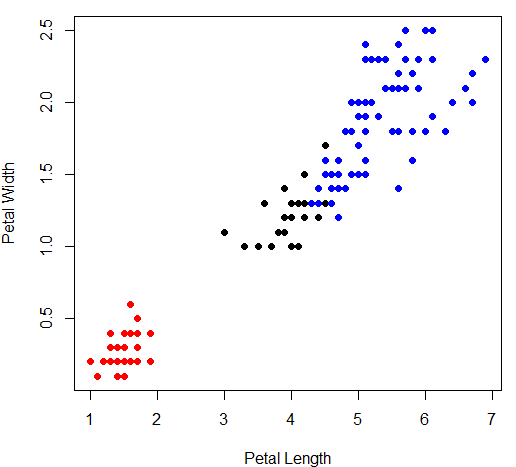
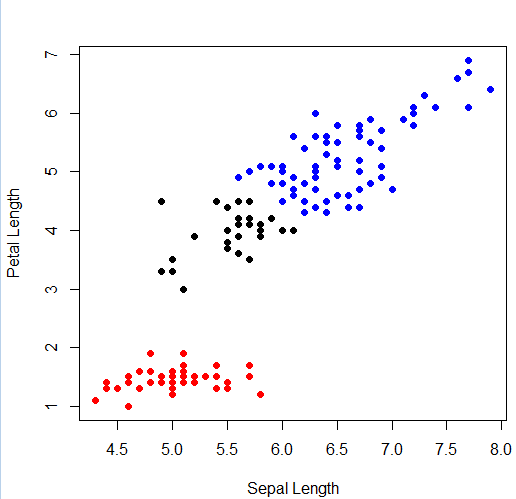
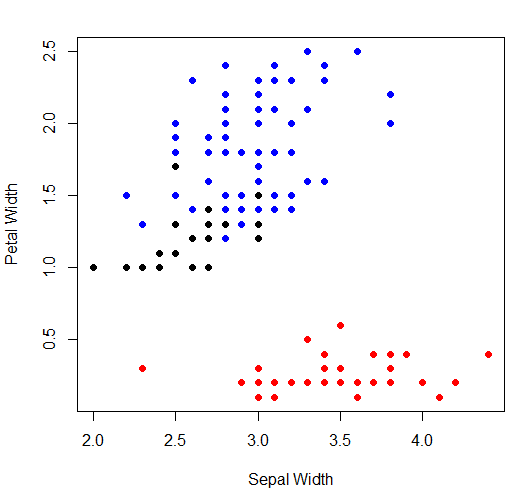
Fig. Scatter Plot of the data

If we plot the all of our data, we can obtain the above picture. In most of the boxes, we can see that the samples can be divided into two distinct clusters; i.e there is linear seperation between the samples while the two of the third classes are not seperated well. This is where our analytical skills should work.

2) Use hierarchical cluster analysis to identify the examined species (iris Setosa, Versicolor, and Virginica) using clustering of their sepal length and width and petal length and width.

1. Plot data in different dimensions, using different colors for different species

In these figure, red represent Setosa, blue represent virginica and black represent versicolor.



It is clear from our plots of various dimensions, that the red samples(Setosa) are always distinctly separated from the rest of the samples. The real problem lies in being able to separate samples into other two classes correctly; Versicolor or Virginica.

When we look at the diagram for Petal Width vs Sepal Length, we can notice that almost all subjects are well separated. Similar intuition can be made about the samples Petal Width and petal Length; they are almost separated by two linear lines.

We cannot observe very distinct separation when we plot data across other dimesions.

In summary, I think we will be able to classify the dataset with high accuracy using the attributes given. For most accuracy, petal width and petal width should be taken as pivot decision making element. If we use another method, then we should be using Petal width and Petal length. If we use other combination of attributes, then we will still end up with comparative classification accuracy.

c. Perform HCA using different distances and cluster methods.

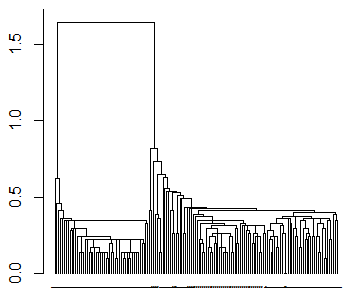
Solution:

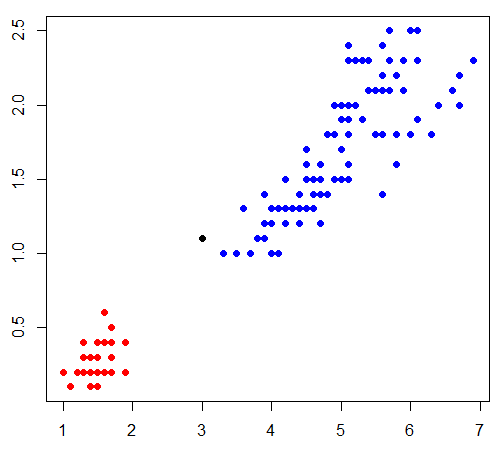
Hierarchical cluster analysis was performed on the dataset using different distances and cluster methods. Cluster 1 is represented by red color, cluster 2 by blue color and cluster 3 by black color.

We used Euclidean, Manhattan and Minkowski as a measure of distance.

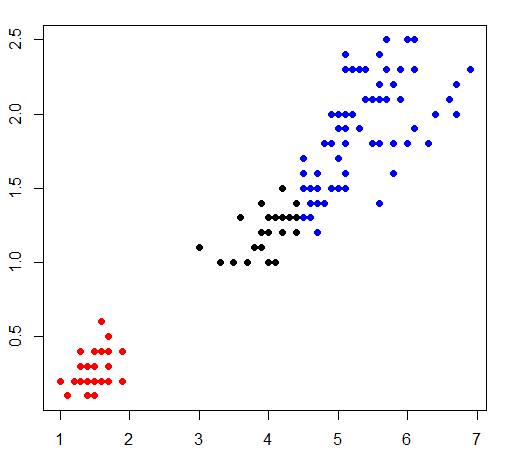
For each of these distances, we used single, complete, average and median approach to calculate distance.

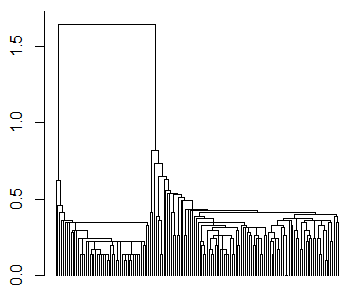
We can see how the dataset is classified into different cluster when we use different parameters.

1) Euclidean Distance, Cluster Method= Single



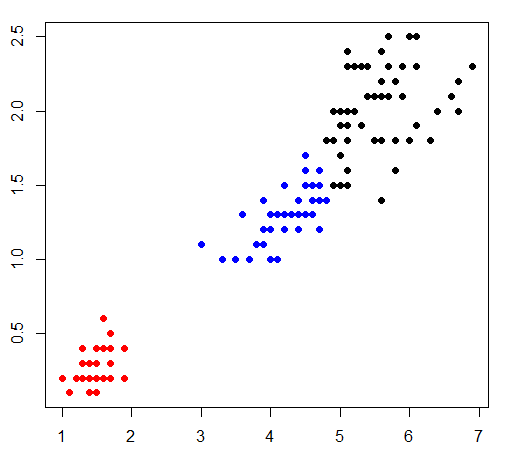
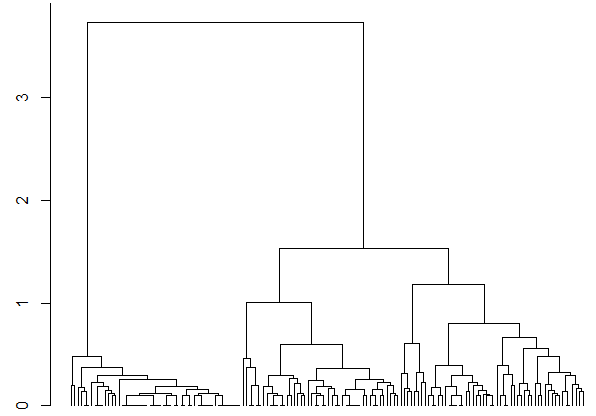
|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 49 | 50 |
| Cluster 3 | 0 | 1 | 0 |

2) Euclidean distance, Complete Linkage



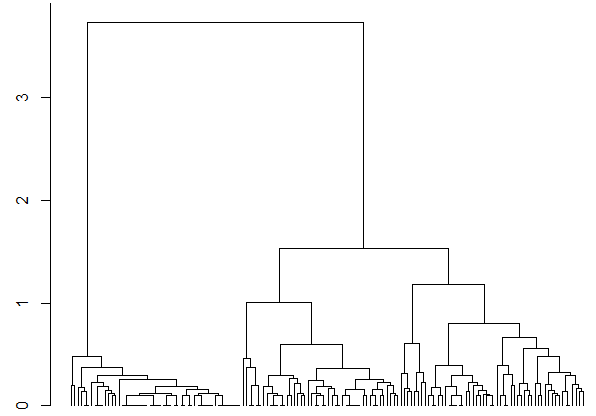
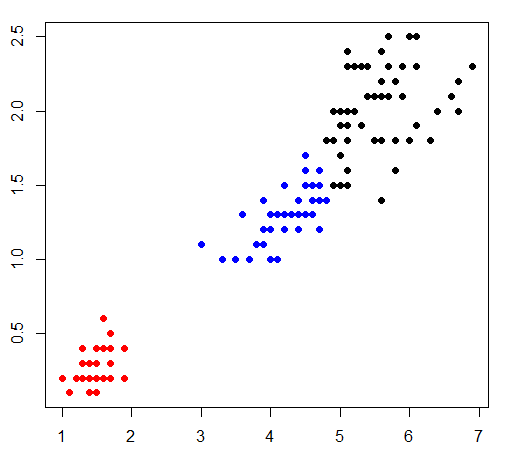
|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 21 | 50 |
| Cluster 3 | 0 | 29 | 0 |

3) Euclidean Distance, Average linkage

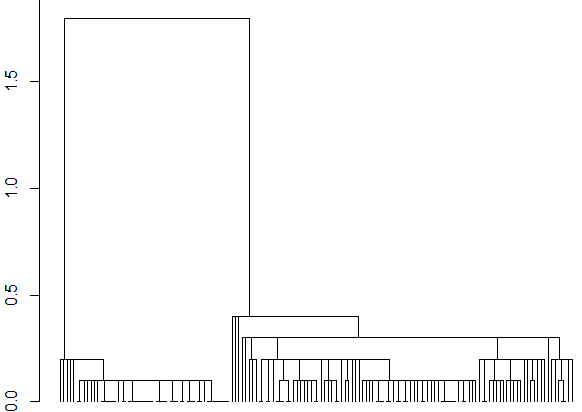


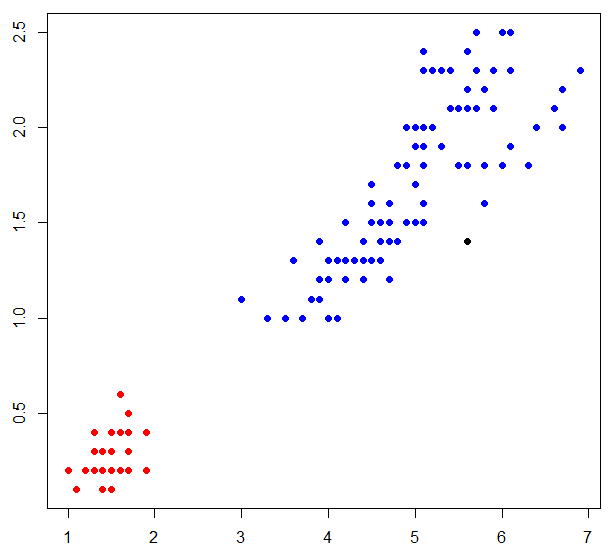
|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 45 | 1 |
| Cluster 3 | 0 | 5 | 49 |

4) Euclidean Distance, Median Linkage

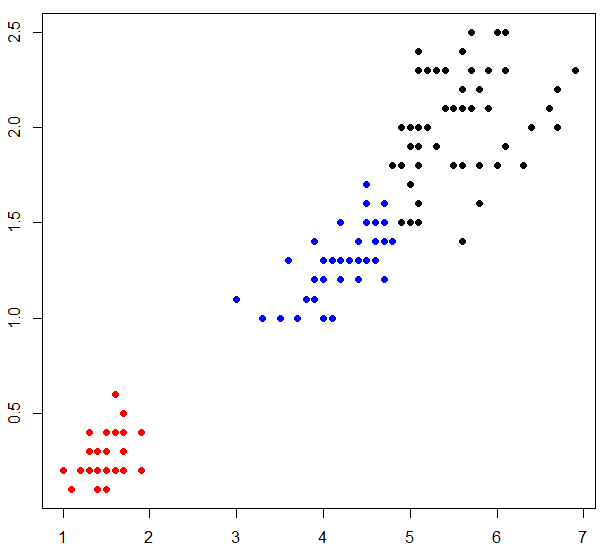
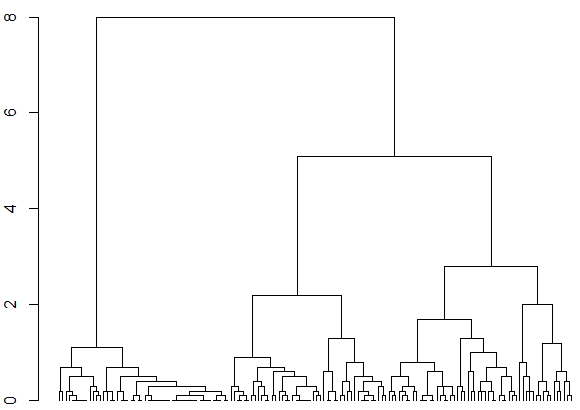


|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 45 | 1 |
| Cluster 3 | 0 | 5 | 49 |

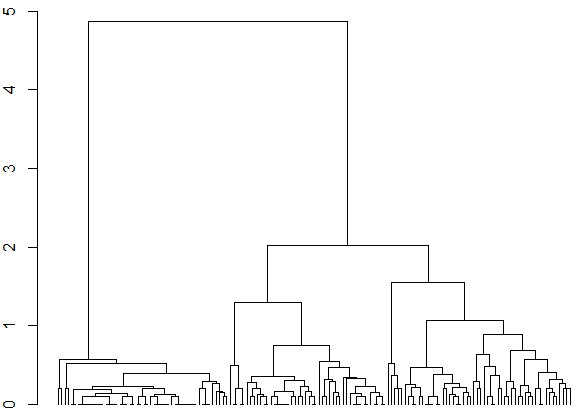
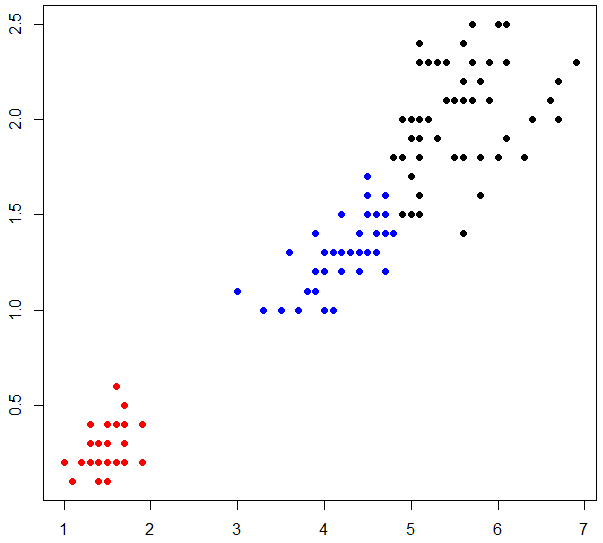
5) Manhattan Distance, Single Linkage



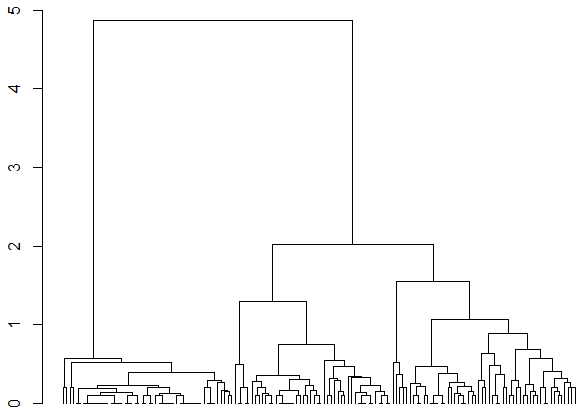
|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 50 | 49 |
| Cluster 3 | 0 | 0 | 1 |

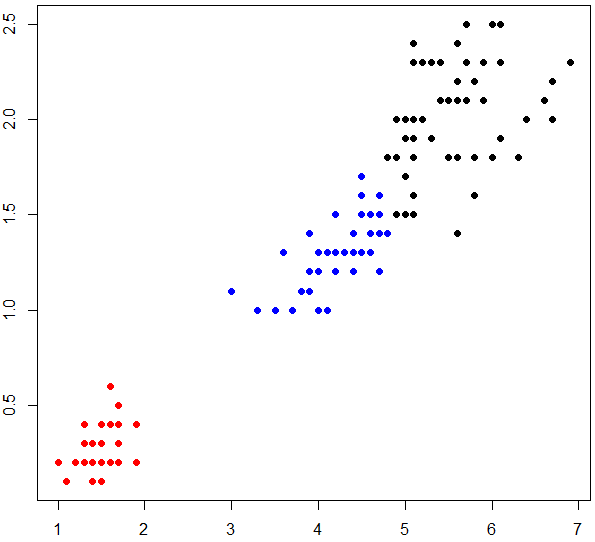
6) Manhattan Distance, Complete Linkage

|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 45 | 1 |
| Cluster 3 | 0 | 5 | 49 |
|  |  |  |  |

7) Manhattan Distance, Median Linkage

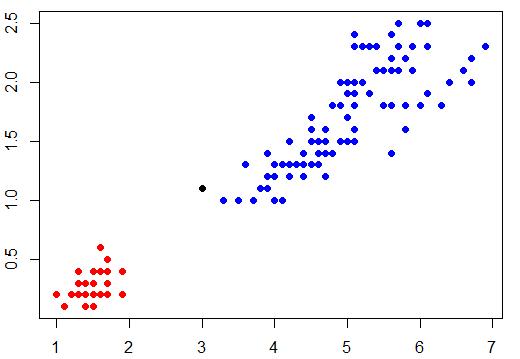
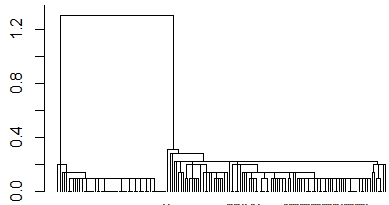
|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 45 | 1 |
| Cluster 3 | 0 | 5 | 49 |

8) Manhattan Distance, Average Linkage



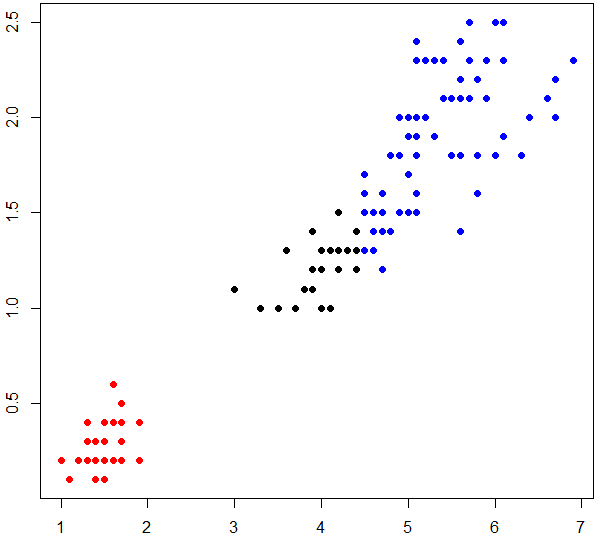
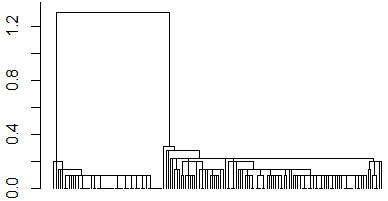
|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 45 | 1 |
| Cluster 3 | 0 | 5 | 49 |

9)Minkowski Distance, Single Link

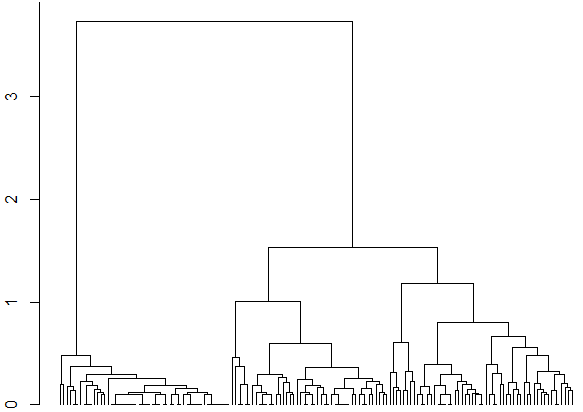


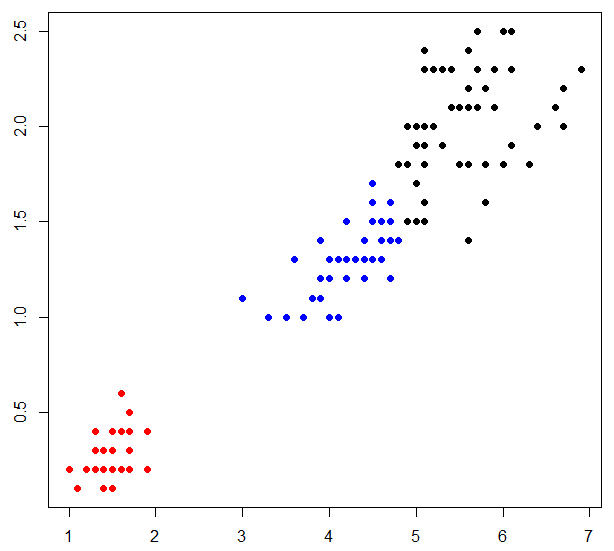
|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 49 | 50 |
| Cluster 3 | 0 | 1 | 0 |

10) Minkowski Distance, Complete Link

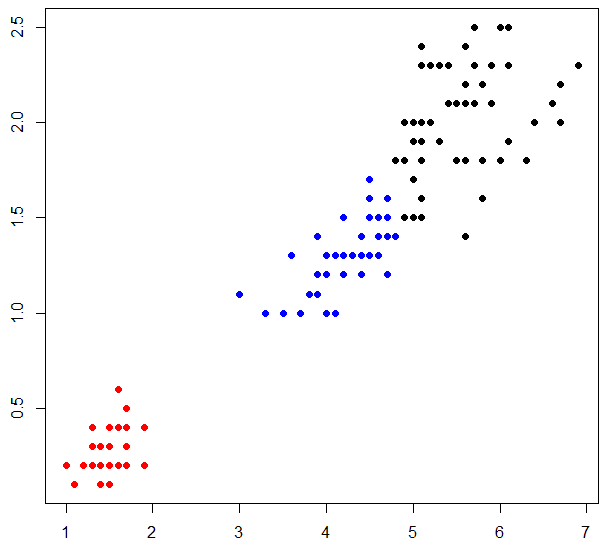


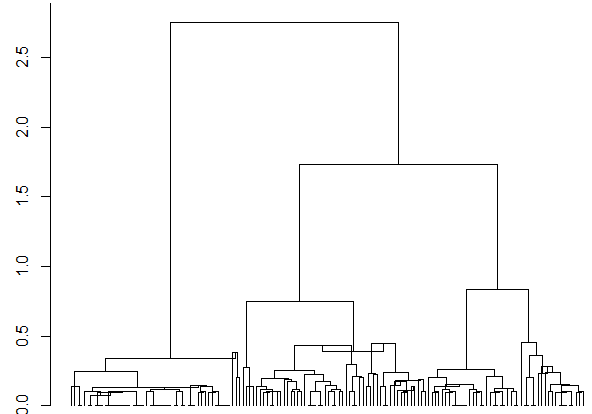
|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 21 | 50 |
| Cluster 3 | 0 | 29 | 0 |

11) Minkowski Distance, Average Link



|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 45 | 1 |
| Cluster 3 | 0 | 25 | 49 |

12) Minkowski Distance, Median Link



|  |  |  |  |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Cluster 1 | 50 | 0 | 0 |
| Cluster 2 | 0 | 45 | 1 |
| Cluster 3 | 0 | 25 | 49 |

Observations:

As our preliminary observation, all of these clustering techniques cluster setosa as a different cluster.

For all the distances taken, we notice that single linkage always groups most of versicolor and virginica together, ie the result is very poor.

Complete linkage also doesn’t generate good results for Euclidean and Minkowski Distances.

Result are good for all other clustering techniques. For (Euclidean, Average), (Euclidean, Median), (Manhatan Complete),(Manhattan, Average), (Manhattan, Median)(Minkowski,Average), (Minkowski,Median) the results are good, ie. only 6 of the samples were misclassified.

If classification accuracy is defined as the number of correctly classified samples among all samples, then

Percentage of classification accuracy = 144/150 =96%.

Conclusion

We studied one of the most classic problem in clustering by performing various combinations of parameter among Distances and linkages, for two sets of attributes.