**ASSIGNMENT# 5**

**Problem 1 (25 points):**

1. (15 points) Perform k-means clustering using all the attributes with the except of the class label, vary the number of clusters from 3 to 4 to 5 to 6 and report:
   1. How the cluster centers were calculated

The cluster centers can be determined by mean.

* 1. What similarity measure was used

Euclidean distance is the similarity measure used to determine the distance of the data points from the cluster center.

* 1. For each k, report the following:

For K=3



For K=4



For K=5



For K=6



* + 1. In your opinion, which k should be selected? Explain your selection.

Based on the observation of the misclassification matrix, I jumped into a conclusion that K=3 cluster to be chosen. From the misclassification matrix it can be noted that Cluster 1 has 1 outlier, Cluster 2 has 9 outliers and cluster 3 has 12 outliers, which is considered to be minimal when compared to rest of the clusters. (K= 4,5,6).

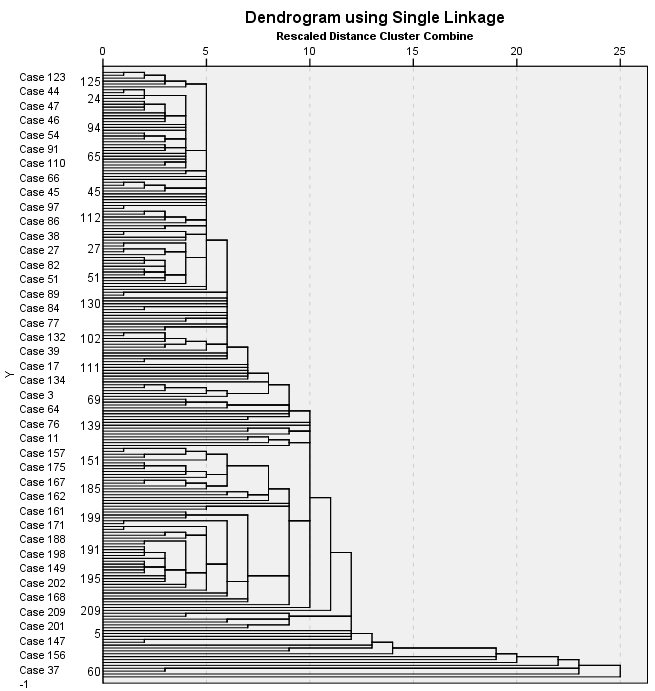
* + 1. For the selected k in iv, analyze and report if the normalization of the attributes will influence the clustering results.

The misclassification matrix for the normalized data is reported below, it can be noted that the misclassified elements is pretty much similar to the once above. The transformed data normalizes the data between the range [-3, 3]. Hence there is no much influence on the clustering results.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class \* Cluster Number of Case Crosstabulation** | | | | | |
| Count | | | | | |
|  | | Cluster Number of Case | | | Total |
| 1 | 2 | 3 |
| Class | 1 | 5 | 63 | 2 | 70 |
| 2 | 0 | 3 | 67 | 70 |
| 3 | 66 | 4 | 0 | 70 |
| Total | | 71 | 70 | 69 | 210 |

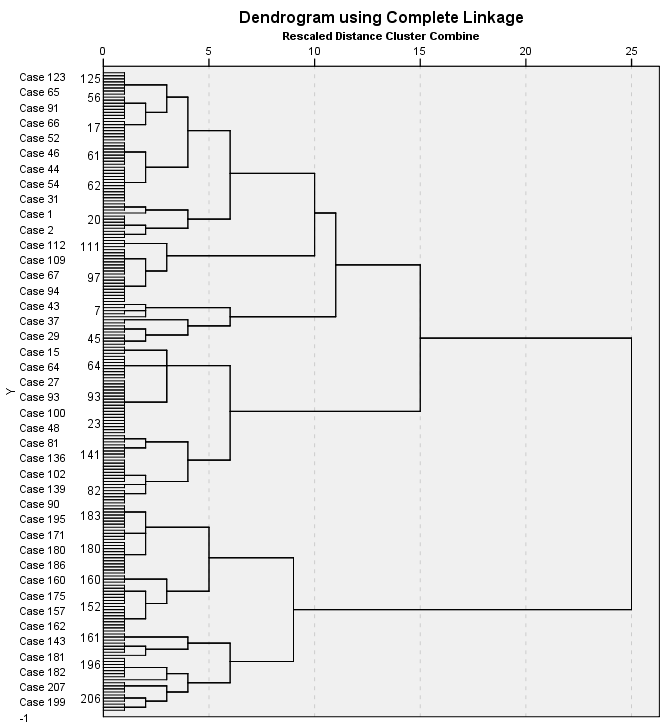
1. (10 points) Perform hierarchical clustering using all attributes except the class label as follows:

Apply single linkage algorithm and report



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class \* Single Linkage Crosstabulation** | | | | | |
| Count | | | | | |
|  | | Single Linkage | | | Total |
| 1 | 2 | 3 |
| Class | 1.000 | 68 | 1 | 1 | 70 |
| 2.000 | 70 | 0 | 0 | 70 |
| 3.000 | 68 | 2 | 0 | 70 |
| Total | | 206 | 3 | 1 | 210 |

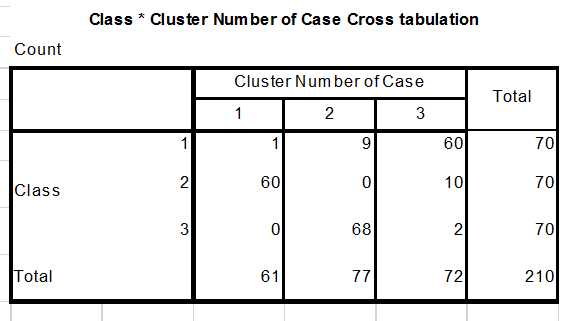
* Apply complete linkage and report



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class \* Complete Linkage Crosstabulation** | | | | | |
| Count | | | | | |
|  | | Complete Linkage | | | Total |
| 1 | 2 | 3 |
| Class | 1.000 | 48 | 2 | 20 | 70 |
| 2.000 | 4 | 66 | 0 | 70 |
| 3.000 | 0 | 0 | 70 | 70 |
| Total | | 52 | 68 | 90 | 210 |

1. (2.5 points) Compare the results with hierarchical clustering and k-means algorithm.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class \* Complete Linkage Crosstabulation** | | | | | |
| Count | | | | | |
|  | | Complete Linkage | | | Total |
| 1 | 2 | 3 |
| Class | 1.000 | 48 | 2 | 20 | 70 |
| 2.000 | 4 | 66 | 0 | 70 |
| 3.000 | 0 | 0 | 70 | 70 |
| Total | | 52 | 68 | 90 | 210 |



From the reported misclassification matrix for the hierarchical clustering and k-means clustering, the misclassified elements in case of k-mean clustering is minimal when compare to hierarchical clustering. The number of outliers in hierarchical clustering is more than the k-means. Hence k-means algorithm is considered to be the best over hierarchical clustering.

1. (2.5 points) Create an executive summary (~half a page) that outlines the problem, summarizes the data, describes the methodology, summarizes the results, and makes recommendations. When creating it, imagine that you will give this summary to someone who is not an expert in data mining.

The problem outlines the three different varieties of wheat: Kama, Rosa and Canadian. The survey is conducted on the following geometric parameters such as- area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove. The data model is performed on 3 classes with 70 elements each.

The methodology chosen for data preprocessing are: K-means clustering and Hierarchical clustering. In K-means 🡺 the cases are divided into clusters, hence select the number of clusters, then perform the similarity measures on it like Euclidean to calculate the distance of each data point in the cluster to the cluster center. Choose the appropriate seed point i.e., the cluster center, then assign the points to the clusters, calculate the new center based on the smallest distance to the new cluster center. Repeat the above mentioned step until there is no much major difference in the change of the cluster elements to the cluster centers. In Hierarchical methodology, the clusters are built in hierarchy. It follows ‘bottom up approach’ – each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy. Other approach ‘top down’ – all observations starts in one cluster, and splits are performed recursively when moved down the hierarchy.

From the above step perform in modeling the data into k-means and hierarchical clustering the k-means is chosen to be the best means of clustering the data since the misclassified elements in this case is minimal when compared to the hierarchical clustering.

Recommendation🡺 K- means algorithm is considered to be the best when the data set is huge and Hierarchical clustering is considered when the data set is minimal.

**Problem 2 (25 points):**

On the same data used in Problem 1, create a decision tree classification model for the three different varieties of wheat: Kama, Rosa and Canadian.

1. Use 10-fold cross validation and at least five different configurations to produce a decision tree classifier. Report the results obtained for the different configurations and choose one as being the best among the configurations you tried. Explain your answer.



I would choose decision tree#1 as the best model, with number of parent as 2, number of child as 1 and also the depth being set to 5. It results in giving the accuracy as 99.5%. Hence it is considered to be the best classification model.

1. For the best tree configuration, report the misclassification matrix and interpret it. In your opinion, is accuracy a good way to interpret the performance of the model? If not, suggest other measures.



In this matrix, of the 70 actual class1, the system predicted that 1 was in Class2, 1 was in Class3, so on. All correct guesses are located in the diagonal of the table, so it’s easy to visually inspect the table for errors, as they will be represented by values outside the diagonal.

Accuracy is not only the option to determine the performance of the classification model. There are different performance metrics to calculate the performance of the model they are: Sensitivity, Specificity and Precision



We got Sensitivity 99.3%, Specificity 99.3% and Precision 99.3% for the best model.

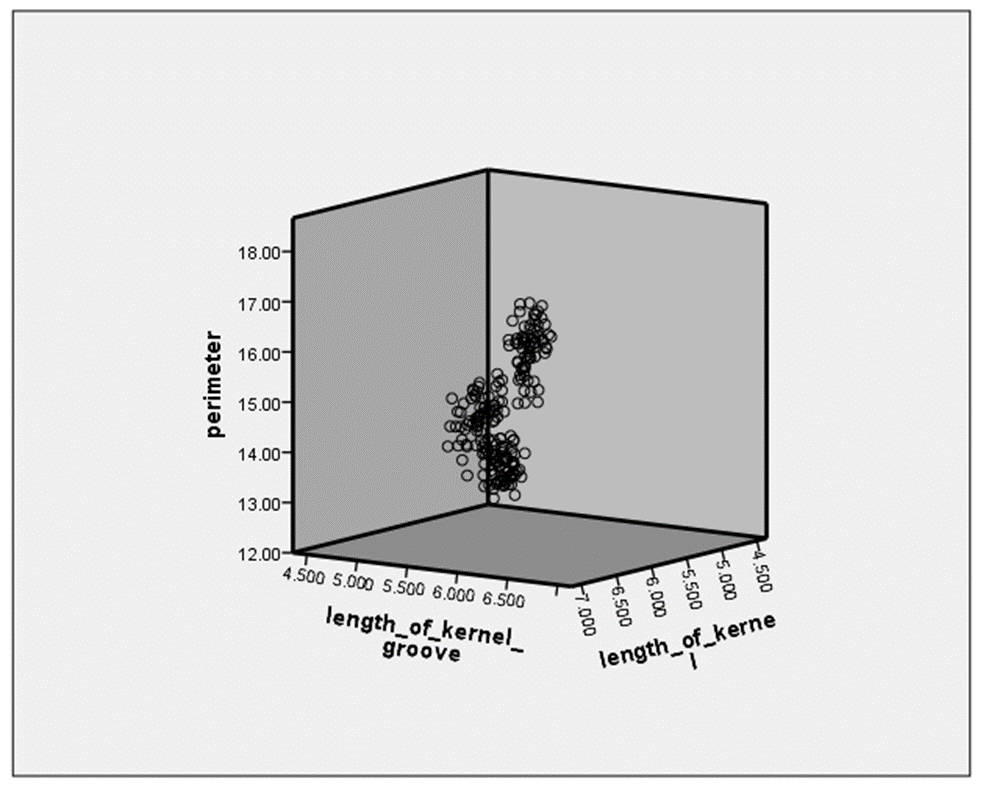
Hence these performance metrics are taken into consideration to decide the performance of the model apart from accuracy.

1. What are the most important three attributes for classifying the wheat data?

Length kernel groove, Perimeter and Length of kernel

1. Create a graph that will allow you to visualize the data in the 3-dimensional space of the most important attributes. Interpret the graph.

The 3D scatterplot – taking 3 most important measures and identify outliers or cluster groupings in the 3 dimensions. It displays Length\_kernel\_groove along X-axis, Perimeter increasing along Y-axis, and Length\_kernel increasing along the Z-axis. From the graph it can clearly observed that the cases are tightly coupled. Also by rotating the scatterplot graph we can observe that the points are clustered and no outliers are detected.



1. Are there any other techniques that can help identify variables for data visualization? Explain your answer and include any analysis you will perform to answer this question.

There are many techniques that help identify variables for data visualizing apart from decision tree.

1. **KNN classifier** provides an option of feature selection that helps visualize the scatterplot based on the most important features.
2. **K-means clustering,** wherein the table that denotes the final cluster centers help us determine the important features and this may thus be used for data visualization.
3. **Forward step-wise regression**

Initially consider an empty set, S= {∅} #Set of features selected,   
1. Perform k-means on each of the features individually for some k.   
2. For each cluster measure some clustering performance metric like Dunn’s Index measure.  
3. Take the feature which gives you the best performance and add it to S  
4. Perform k-means on S and each of the remaining features individually  
5. Take the feature which gives you the best performance and add it to S  
6. If you have reached the desired number of features stop, else go back to 4   
  
once you complete it, you will have a curve of the performance metric chosen for each of the cluster-features. Then you will be able to see if adding features helps getting a better clustering and which subset of your features is the best.