MULTIVARIATE ANALYSIS PROJECT FOR STAT 589

ANALYSING SEED STRUCTURE USING CLUSTERING

BY DOMINIC ESSUMAN

SOUTHERN ILLINOIS UNIVERSITY EDWARDSVILLE

DECEMBER 2017

1 ABSTRACT

In this paper clustering is applied to the seed dataset in which different varieties of seeds are categorized into different clusters on the basis of their morphological features. In the present work We performed seed clustering using both hierarchical method and the nonhierarchical method. The data was collected from UCI website's database. The features of seed used are area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove. Both method clustererd the dataset into three and the result provided a greater accuracy of more than 90 percent.

2 INTRODUCTION

Clustering is the process of finding naturally occurring groups in data. Clustering is one of the most widely studied techniques in the context of data mining and has many applications,including disease classification, image processing, pattern recognition, and document retrieval. Clustering is a major technique unsupervised learning method. The main aim of cluster analysis is to partition a given population into groups or clusters with common characteristics, since similar objects are grouped together, while dissimilar objects belong to different cluster. As a result, a new set of categories of interest, characterizing the population, is discovered. The clustering methods are generally divided into two groups; hierarchical and Non hierarchical. Hierarchical Clusters are formed sequentially, with the number of clusters decreasing as clusters merged with other similar clusters agglomerative hierarchical methods or split into less homogeneous groups divisive methods.Non hierarchical is designed to group items into a collection of K clusters where the number of clusters may either be specified in advance or determined as part of the procedure (method). These numerous concepts of clustering are implied by different techniques of determination of the similarity and dissimilarity between objects. A classical partitioning k-means algorithm is concentrated on measuring and comparing the distances among them. It is computationally attractive and easy to interpret and implement in comparison to other methods. On the other hand, the number of clusters is assumed here by user in advance and therefore the nature of the obtained groups may be unreliable for the nature of the data, usually unknown before processing. The rigidity of arbitrary assumptions concerning the number or shape of clusters among data can be overcome by density-based methods that let the data detect inherent data structures. The main idea of this algorithm assumes that each cluster is identified by local maxima of the kernel density estimator of the data distribution. The procedure does not need any assumptions concerning the data and may be applied to a wide range of topics and areas of cluster analysis. The main purpose of this work is to propose an effective technique for forming proper categories of wheat, so the Hierarchical Clustering method (average linkage) and the Non Hierarchical method(K means) was used for the analysis of the data and the result were compared. In the earliest attempts to classify wheat grains a geometry and set of parameters were defined. The size, shape and colour of grain because of their heritable characters, can be used for wheat variety recognition. Accomplished studies showed that digital image processing techniques commonly used in multivariate analysis give reliable results in classification process. In this paper, the algorithm proposed will be used to identify wheat varieties, using their main geometric features. Visualisation of the data usig scatter plot matrix suggested the data was subject to clustering thus the average linkage and the kmeans method was used for the analysis and the result suggest that both methods were a good and appropriate method to be used clustering this data

2.1 OBJECTIVE OF STUDY

The objective of this investigation is to compare the Hierachical(Average linkage) and the Non Hierachical(K-Means) machine learning algorithms and to identify which method is the best at identifying the variety of wheat from its geometric properties.

3 DATA SET

The data of wheat seeds is gathered from UCI website which is a great dataset repository. The numbers of samples of wheat seeds are 210 from three wheat classes Kama, Rosa and Canadian are collected for clustering process. Seven geometrical or morphological features of seeds are considered on the basis of which seeds are classified into three.The features of seed used are area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. The data set can be used for the tasks of classification and cluster analysis

4 DATA ANALYSIS

Here the Average linkage method for the Hierarchical and the KMeans method for the NonHierarchical were used to analyze the data and the results were compared.

4.1 HEIRARCHICAL CLUSTERING-AVERAGE LINKAGE

4.1.1 PROCEDURE

- Treats the distance between two clusters as the average distance between all pairs of items where one member of a pair belongs to each cluster.

- Start by finding the minimum entry in $D = dik$ and merging the corresponding objects to get cluster (UV).

- For Step 3, the distances between (UV) and the other cluster W are determined by

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \tag{1}$$

where d_{ik} is the distance between object i in the cluster (UV) and object k in the cluster W, and $N_{(UV)}$ and N_W are the number of items in clusters (UV) and W

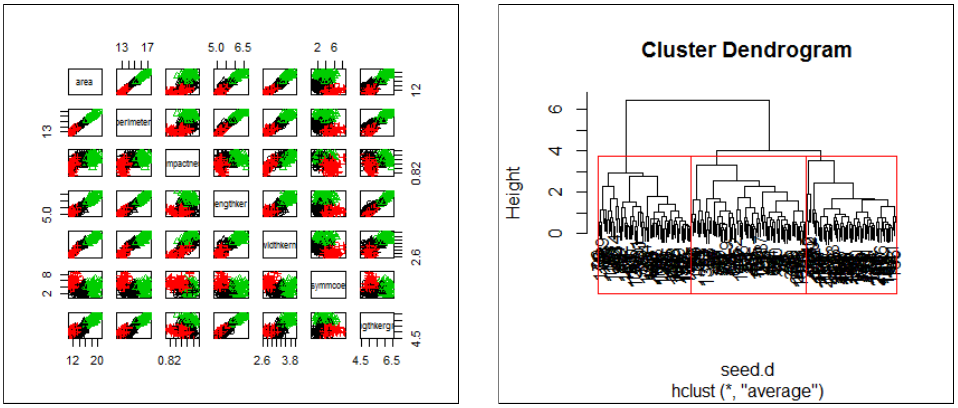


Figure 1: scatter matrix and the Dendogram

Looking at the scatter plot it can be seen that the data has some relationships between the variables thus, clustering can be applied. Different clusters were used for the clustering and it was found that three clusters was the best using the average linkage.This is shown in the Dendogram diagram

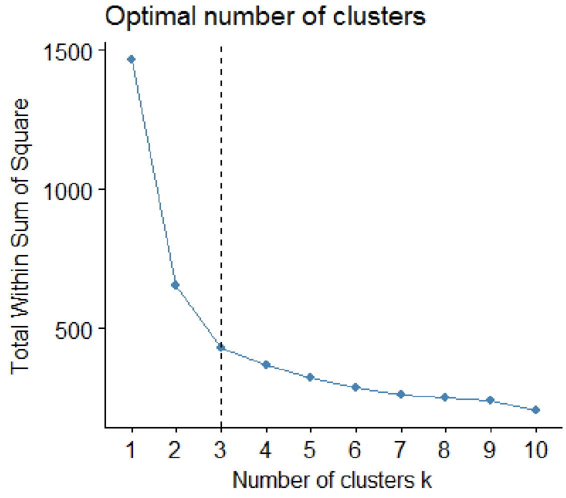
4.2 NON HIERACHICAL K MEANS

4.2.1 PROCEDURE

- Partition the items into K initial clusters.
- Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest (usually used Euclidean distance).
  - Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.
- Repeat Step 2 until no more reassignments take place.
  - Rather than starting with a parition of all items into K preliminary groups in Step 1, we could specify K initial centroids (seed points) and then proceed to Step 2.
  - Final assignment is dependent on the initial partition.

The seed data was prepared by scaling so that all the attributes fall within a similar range of values. The scaled data is then passed through the K-Means clustering algorithm to find the optimum number of clusters. A plot of the within groups sum of squares was produced suggesting that the optimum number of clusters was three, as shown in figure 2 Using the function Nbclust in R 26 different algoritms was run to find the optimum of clusters and the output in figure 3 confirmed it to be three. The bar chart in figure 4 shows the number of statitics of the various algorithm found and that shows that the optimum number of clusters foud is three. The K mean clustering was then run again with the number of clusters set to three.

Figure 2: Within group Sum of Squares



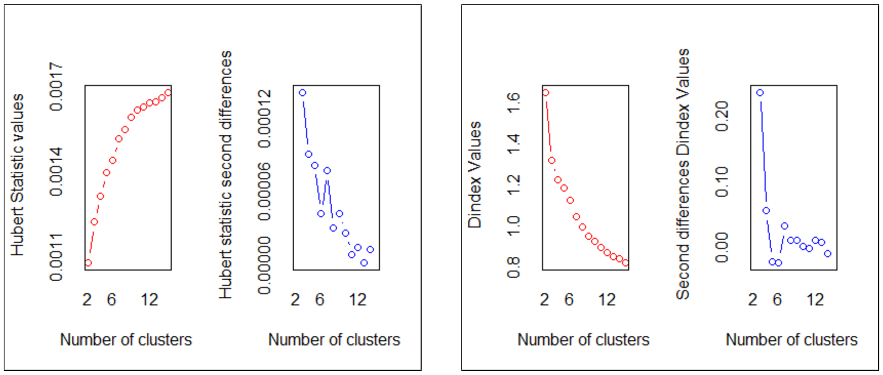


Figure 3: Hubert Statistics and the Dindex Values

5 RESULTS AND DISCUSSION

From te analysis for the Average linkage it can be seen that cluster 1 successfully picked 66 out of 70, cluster 2 picked 61 out of 70 and cluster 3 picked 64 out of 70 which is a good clustering comparing the predictive to the actual attribute. Also for the K Means, cluster 1 successfully picked up 62 out of 70, cluster 2 successfully picked up 65 out of 70 and cluster three successfully picked up 66 out of 70. Comparing the target to the predictive attribute it is seen that the method provide a good clusttering.

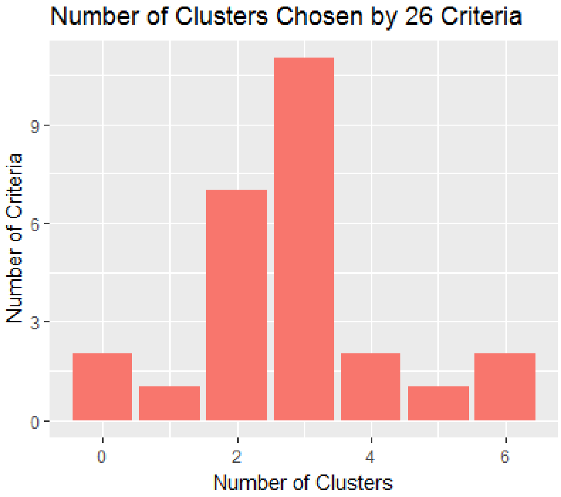


Figure 4: Number of Clusters the found by the various algorithms

	ACTUAL			
PRED	K	R	C	TOT
1	66	6	9	81
2	3	0	61	64
3	1	64	0	65
	70	70	70	

Table 1: Avg Link

	CLUSTER			
VAR	1	2	3	TOT
K	6	2	62	70
R	0	65	5	70
C	66	0	4	70
	72	67	71	

Table 2: K means

6 CONCLUSION

Both the K-Means and Average linkage algorithms performed very well on this data set and were able to identify the variety of wheat by the geometric properties of the seed kernels.Both methods returned a good accuracy of more than 90 percent comparitively to the target attribute.The K-Means was able to successfully identify three naturally occurring classes without using the label attribute.

References

[1] Applied Multivariate Statistical Analysis (6th Edition), Pearson Prentice Hall, 2007, by Richard Johnson, Dean W. Wichern.

[2] ANDERSON T.W. An introduction to Multivariate Statistical Analysis (3rd edition). New york : John Wiley 2003.

[3] Bacon Shone ,J, and W.K Fung." A New Graphical Method for Detecting Single and Multiple Outliers in Univariate nad Multivariate data", Applied Statistics,36no. 2 (1987), 153-162

[4] Barlett, M.S "Multivariate Analysis" Journal of the Royal Statistical Society Supplement(b),9 (1947),176-197

[5] Johnson R.A and G.K Bhattacharyya Statistics Principles and methods (5th ed). New York: John Wiley, 2005