ANALYSING SEED STRUCTURE USING DATA CLUSTERING

PROJECT PRESENTATION

Dominic Essuman

Dr. J . Pailden (Advisor)

Southern Illinois University Edwardsville University (SIUE)

dessuma@siue.edu

December 5, 2017

Outline

INTRODUCTION:CLUSTERING

Definition

Clustering is a major technique used 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

- Two groups of clustering methods Hierarchical methods and Nonhierarchical (partitioning) methods
- The kernels of three varieties of wheat were examined under soft X-ray technology and a number of geometric properties were recorded.

DATA DESCRIPTION

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.

Variables

- area A
- perimeter P
- Compactness $C = 4 * pi * A/P^2$
- length of kernel
- width of kernel
- asymmetry coefficient
- length of kernel groove
- Variety of Wheat (target attribute)

OBJECTIVE OF STUDY

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

AVERAGE LINKAGE

PROCEDURE

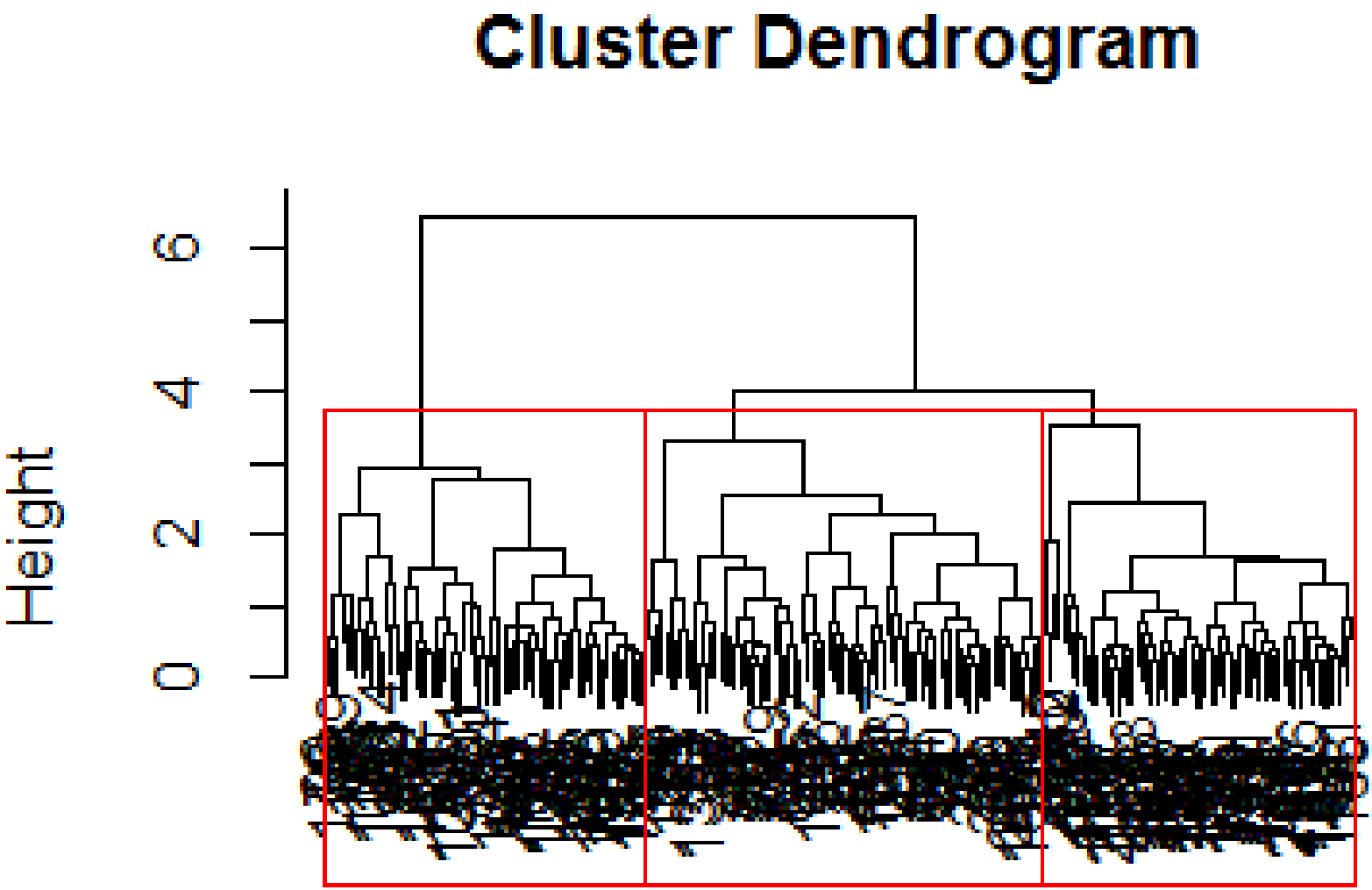
- Treats the distance between two clusters as the average distance between all pairs of items where one member of a pair belongs to each cluster.
- Start by finding the minimum entry in $D = dik$ and merging the corresponding objects to get cluster (UV).
- For Step 3, the distances between (UV) and the 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

AVERAGE LINKAGE

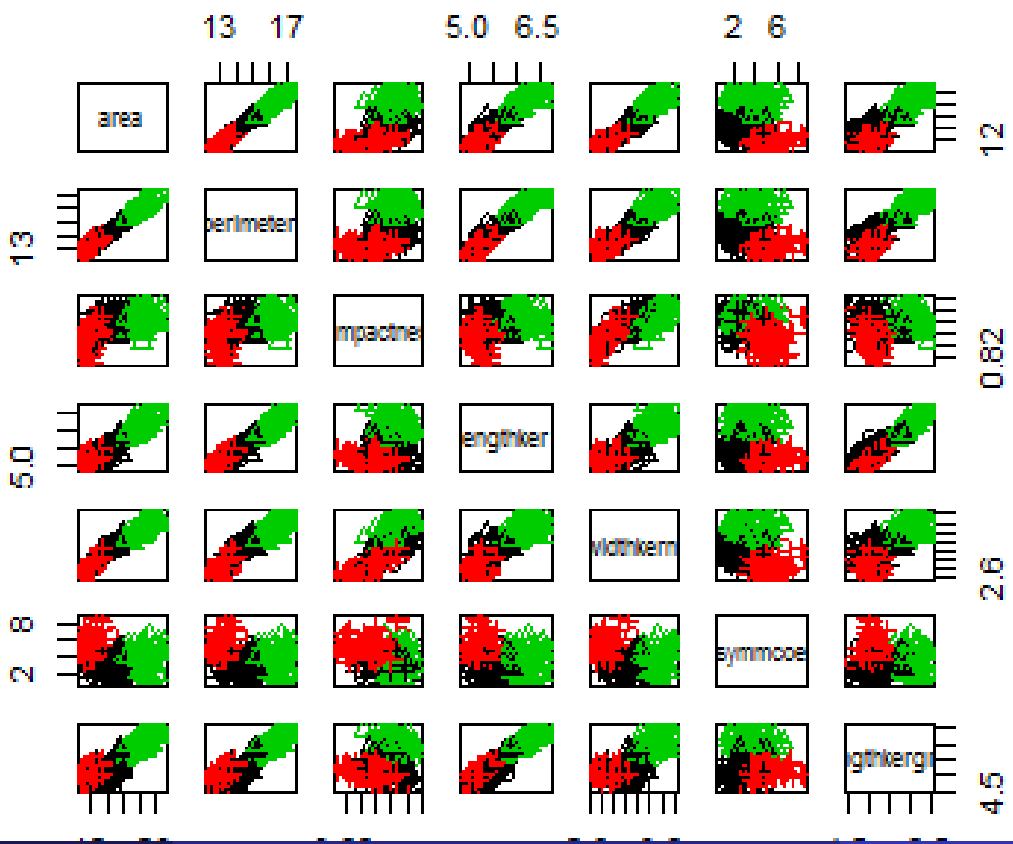
AVERAGE LINKAGE FOR SEED DATA



AVERAGE LINKAGE

AVERAGE LINKAGE FOR SEED DATA

- . The visualisations show that the data does fall into a number of distinct clusters.



AVERAGE LINKAGE

SUMMARY OF RESULTS

Table: **SUMMARY**

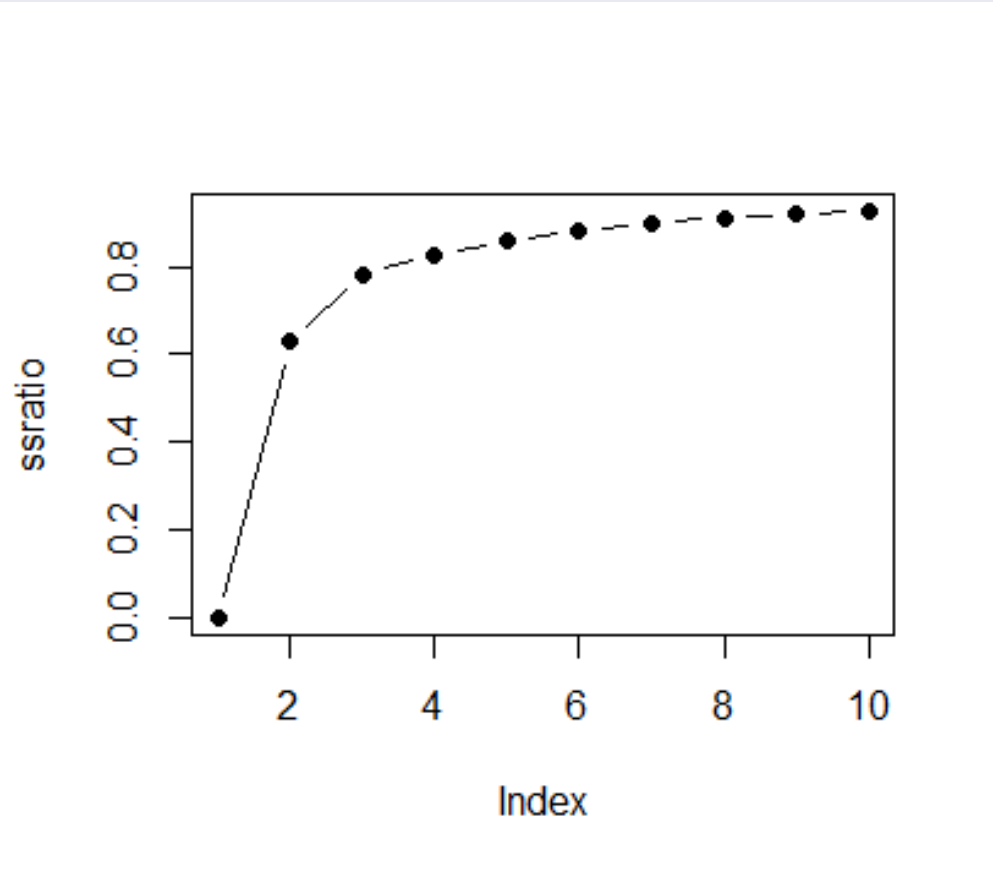
	ACTUAL		
PREDICTED	Kama	Rosa	Canadian
1	66	6	9
2	3	0	61
3	1	64	0

K MEANS

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 data was prepared by scaling so that all of 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.



SUMMARY OF RESULTS K-MEANS

- The K-Means clustering algorithm was run again with the number of clusters set at three. The results were then analysed.

Table: **SUMMARY**

	CLUSTER		
VARIETY	Kama	Rosa	Canadian
1	6	62	2
2	0	5	65
3	66	4	0

DISCUSSION AND 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.
- The K-Means was able to successfully identify three naturally occurring classes
- The results of both algorithms returned an accuracy.
- This experiment also shows that it would be possible to use the K-Means algorithm to generate a set of classes from data that doesnt have a classification label attribute.