DATA MINING ASSIGNMENT

Implementing Clustering Algorithms –

K-Means Clustering

Hierarchical Clustering

GROUP MEMBERS

ABHILASH K MIRJI – 2011AAPS049H

ASHWIN RAGHAVAN – 2011B5A7550H

PRAVEEN VENKATESWARAN – 2011B4A7640H

Introduction: K-Means Clustering

It is a partitioning approach where observations are divided into K groups and reshuffled to form the most cohesive clusters possible according to a given criterion. Conceptually, the K-Means algorithm:

1) Selects K centroids (K rows chosen at random).

2) Assigns each data point to its closest centroid.

3) Recalculates the centroids as the average of all data points in a cluster.

4) Assigns data points to their closest centroids.

5) Continues steps 3 and 4 until either the observations are not reassigned or the maximum number of iterations is reached.

K-means clustering can handle larger datasets than hierarchical cluster approaches. Additionally, observations are not permanently committed to a cluster. They are moved when doing so improves the overall solution. However, the use of means implies that all variables must be continuous and the approach can be severely affected by outliers.

The value of K must be specified in advance. There has been extensive research done on determining the value of K and several heuristics have been identified that can be used to assist in choosing the value of K. We have used the 'elbow-curve' which is a plot of the total within-groups sums of squares against the number of clusters in a K-means solution. A bend in the graph can suggest the appropriate number of clusters.

Hierarchical Clustering

It is a method that seeks to build a hierarchy of clusters. There are two basic strategies- Agglomerative (Bottom-up) and Divisive (Top-down) which differ in whether the clusters are either merged together into bigger clusters or if the clusters keep getting divided into smaller clusters.

The hierarchical clustering method we have chosen defines the cluster distance between two clusters to be the maximum distance between their individual components. At every stage of the clustering process, the two nearest clusters are merged into a new cluster. This process is repeated until the whole data set is agglomerated into a single cluster. The complexity is of the order $O(n^3)$ which makes it too slow for large data sets.

The Dataset: User Knowledge Modelling Data Set

We have used the User Knowledge Modelling dataset available on the UCI dataset repository. It was generated by Hamdi Tolga Kahraman during his PhD thesis. It provides details about students' knowledge on the subject of Electrical DC Machines. The users' knowledge class were classified by the authors using intuitive knowledge classifier (a hybrid ML technique of k-NN and meta-heuristic exploring methods).

Number of Instances: 403

Number of Attributes: 5

Attribute Values:

STG	Degree of study time for main subject materials
SCG	Degree of repetitive study for main subject materials
STR	Degree of study time for related subject materials
LPR	Exam performance for related subject materials
PEG	Exam performance for main subject materials

Missing Attribute Values: none

Class Distribution (number of instances per class)

Class	N
Very_low	50
Low	129
Middle	122
High	130

Reason for choosing:

- It is a numerical dataset that allows easy clustering especially since k-means requires a numeric input.
- The dataset gives a chance to physically interpret and make sense of the clusters generated of the different knowledge levels amongst students.
- The dataset is large enough to ensure the formation of distinct and meaningful clusters and not so large that hierarchical clustering becomes a problem.

Objective

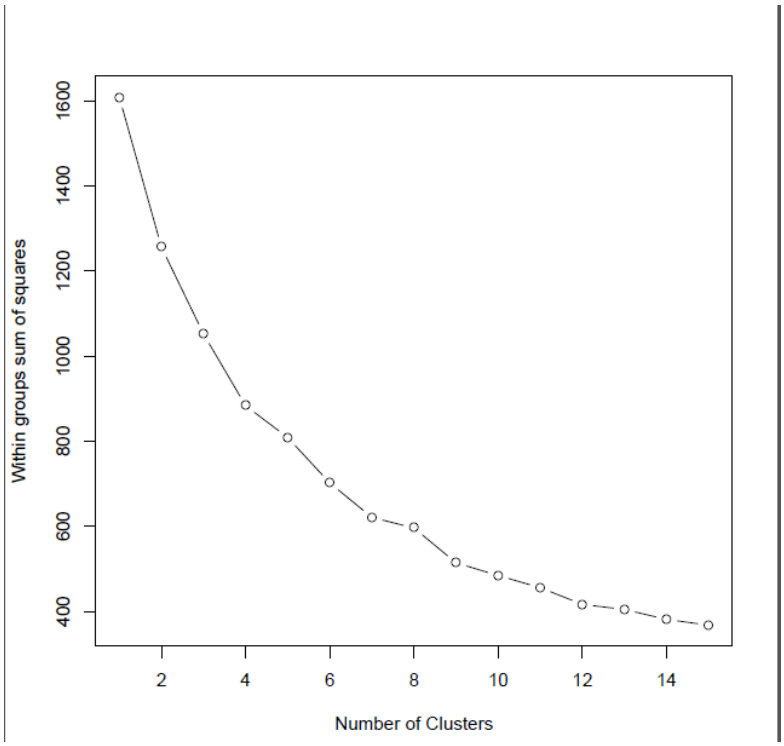
We aim at clustering the given data into distinct and meaningful clusters that can capture different levels of student knowledge in the subject.

Pre-processing

We extract the columns containing the attribute values and create a data frame that we use as input for the clustering algorithms. Since the data has already been normalized there was no need to scale it for ensuring that the elbow graph is plotted correctly.

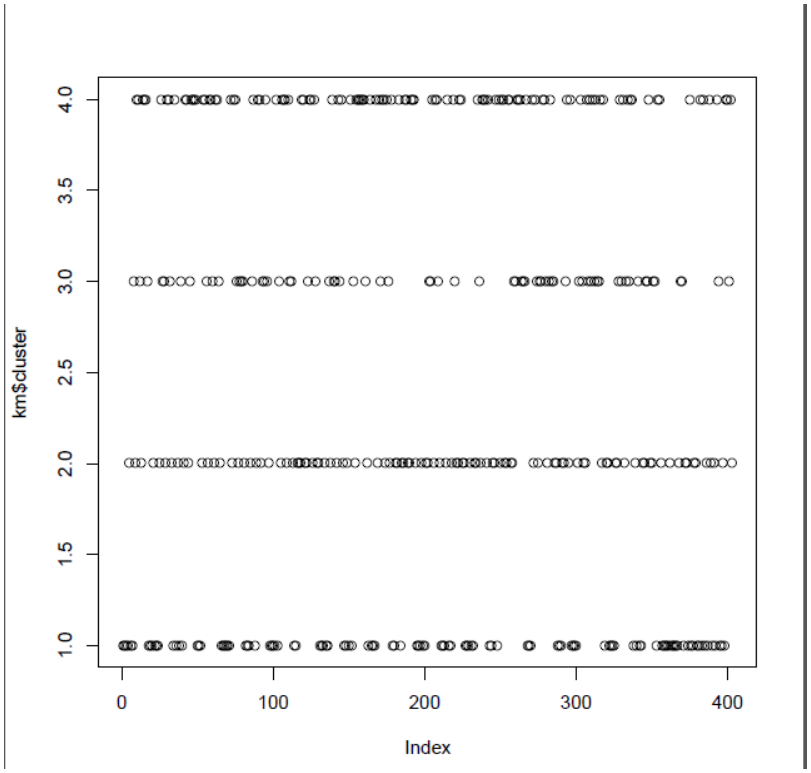
Results

We generated the elbow graph of our data as follows:



We observe that the first low deviation in the sum of squares (first 'elbow') occurs at K=4. This also agrees with the findings of the dataset.

We then ran the K-means algorithm and plotted the clusters into which each individual student has been placed.



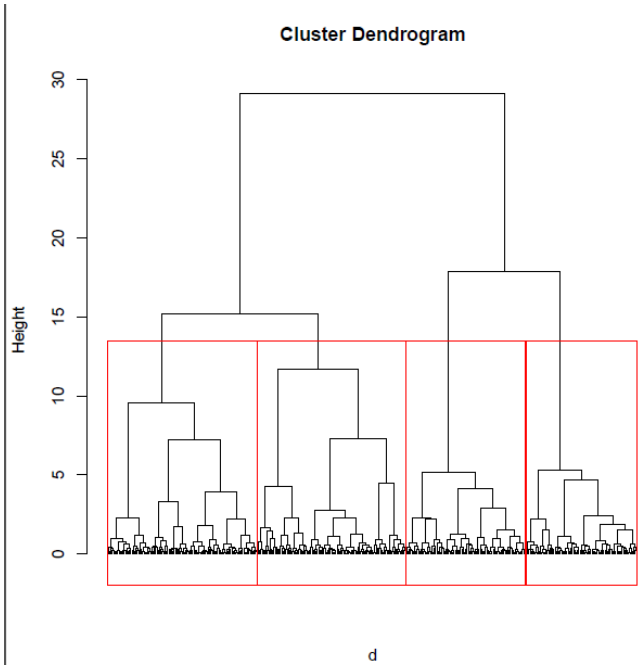
We then extracted the mean values for each cluster with respect to each attribute.

	STG	SCG	STR	LPR	PEG
1	0.3659561	0.2878509	0.2070614	0.2817632	0.4205263
2	0.3887736	0.3956321	0.4492453	0.7666038	0.3705660
3	0.2423429	0.2939857	0.6077143	0.3144286	0.1977571
4	0.3754248	0.4257788	0.6253982	0.3401770	0.7331858

We also obtained the sum of squares within each cluster as:

1) 18.195053 2) 22.358804 3) 9.040071 4) 21.753380

We then used the hierarchical clustering method in order to agglomerate the data into different clusters. The following is the plot of the results-



Where the red rectangles represent the 4 different clusters that were obtained.

Concluding Remarks

Using the mean values in each cluster that was obtained through k-means, we see that the 5th attribute has more significant variance in the values. This makes sense as a student's performance in the main examination has the most importance in deciding on the knowledge levels. In all the clusters we see significant difference in the means with respect to different attributes thus showing that distinct clusters were indeed formed. The Cluster Dendogram further serves to validate our results by dividing data into four clusters with the leaf nodes containing the different students' information. Hence we obtain a measure for clustering students into 4 different knowledge levels.