**Data Mining Assignment 2**

**Rajat Saxena**

**2011C6PS649H**

**Arpit Sharma**

**2011A7PS078H**

**1. List of files**

* Kmeans.r : R script for k means clustering
* Dataset.csv : Integrated Data set file
* Report.docx : Complete description of assignment
* Run1,2,3,4,5,6 : Results of diff executions includes output as well as graphs for diffrent attributes.

**2. Introduction**

Clustering allows a user to make groups of data to determine patterns from the data. Clustering has its advantages when the data set is defined and a general pattern needs to be determined from the data. You can create a specific number of groups, depending on your business needs. One defining benefit of clustering over classification is that every attribute in the data set will be used to analyze the data.

**3. Dataset**

We chose a dataset which are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

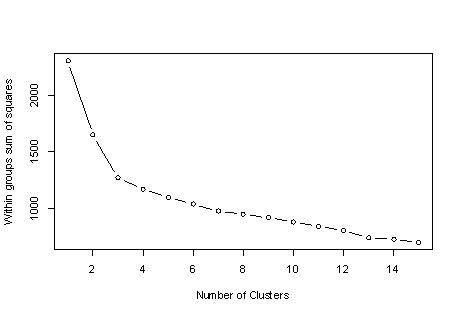
The data originally come from the UCI Machine Learning Repository (<http://www.ics.uci.edu/~mlearn/MLRepository.html>).

**4. Overview of Maths involved**

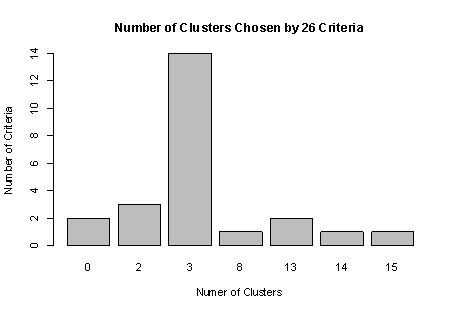
* Every attribute in the data set should be normalized, whereby each value is divided by the difference between the high value and the low value in the data set for that attribute. For example, if the attribute is age, and the highest value is 72, and the lowest value is 16, then an age of 32 would be normalized to 0.5714.
* Given the number of desired clusters, randomly select that number of samples from the data set to serve as our initial test cluster centers. For example, if you want to have three clusters, you would randomly select three rows of data from the data set.
* Compute the distance from each data sample to the cluster center (our randomly selected data row), using the least-squares method of distance calculation.
* Assign each data row into a cluster, based on the minimum distance to each cluster center.
* Compute the centroid, which is the average of each column of data using only the members of each cluster.
* Calculate the distance from each data sample to the centroids you just created. If the clusters and cluster members don't change, you are complete and your clusters are created. If they change, you need to start over by going back to step 3, and continuing again and again until they don't change clusters.

**5. Implementation in R**

* As K-means clustering requires that the number of clusters to extract be specified in advance, we initially used a R script to find out the optimal number of clusters.
* The NbClust package can be used as a guide. Additionally, a plot of the total within-groups sums of squares against the number of clusters in a K-means solution can be helpful. A bend in the graph can suggest the appropriate number of clusters.
* The nummber of clusters is determined using the wwsplot() and NbClust()functions. Figure 1 indicates that there is a distinct drop in within groups sum of squares when moving from 1 to 3 clusters. After three clusters, this decrease drops off, suggesting that a 3-cluster solution may be a good fit to the data.
* The figure is as follows:



* In figure 2, 14 of 24 criteria provided by the NbClust package suggest a 3-cluster solution. Note that not all 30 criteria can be calculated for every dataset. The figure is given below:



* A final cluster solution is obtained with kmeans() function and the cluster centroids are printed. Since the centroids provided by the function are based on standardized data, the aggregate() function is used along with the cluster memberships to determine variable means for each cluster in the original metric.
* To check how well the K-means clustering uncovered the actual data contained in the type variable, a cross-tabulation of type(wine varietal) and cluster membership was done by us.
* We then quantified the agreement between type and cluster, using an adjusted Rank index provided by the flexclust package.
* The adjusted Rand index provides a measure of the agreement between two partitions, adjusted for chance. It ranges from -1 (no agreement) to 1 (perfect agreement). Agreement between the wine varietal type and the cluster solution is 0.897.
* So finally as result of R script analysis we found that the number of custers = 3 and agreement = 0.897.

**6. Implementation in Weka**

* Next we moved onto Weka for better visualization of the dataset.
* Using the data set from the repository loaded the file in the explorer since all the attributes are numeric no preprocessing is required and we can directly jump to clustering.
* As we derived 3 as optimal number of clusters from the above R Script, we will use it as ur default value.

**Approach 1**

* Initially we used **Simple Kmeans** functions using Eucledian distance as our distance function and **Seed= 3**.
* The run information is given in file Run1.txt under Run1 folder.
* Next we compared wine type on the basis of different attributes and plotted the graph for each. The result can be found in Run1 folder.
* In the next run we selected **Seed= 6.**
* The run information is given in file Run2.txt under folder Run 2.
* Again we plotted graph for different attribute and wine type, results of which is stored in Run2 folder.
* In the next run we selected **Seed= 9.**
* The run information is given in file Run3.txt under folder Run 3.
* Again we plotted graph for different attribute and wine type, results of which is stored in Run3 folder.

**Approach 2**

* Next we moved onto another method of initial seed selection which is **X- Means**.
* We ran X Means algorithm for 3 different values of seed which are 3,6 and 9 respectively and the result is stored in Run4,Run5 and Run6 folder.

**7. Description of final clusters discovered**

**Cluster 0**

* These refers to wine having fresh, crisp taste because of high content of malic acid and are derived from grapes.
* Have high color intensity(red wines mostly).
* Have highest amount of non flavanoid phenols.
* Are more healthy.

**Cluster 1**

* Least sweet and viscous wine types.
* Least color intensity(mostly refers to white wines).
* Can be both flavored or non flavored.
* Least amount of alcohol in these types of wines.
* Least haziness.

**Cluster 2**

* Wine of this type have more sweetness and viscosity cause of high amount of alcohol.
* Mostly have high astringency and texture.
* Consists of both red and white wine.
* Have high concentration of flavanoid phenols.
* Have the highest turbidity or haze.
* Highest protein concentration.

**8. Conclusion**

We came upto the conclusion that to decide what is the best number of cluster is a different problem than to decide how to set the values of the seeds.

* The first problem is how to decide the"value of k" in k-means (k= amount of clusters), because any additional cluster improves the quality of the clustering but at a decreasing rate, and having too many clusters may be useless to decision makers, data comprehension, data explanation, etc.
* The number of initial seeds (initial centers of clusters) is the same as number of clusters (at leats in the original k-means). The problem of the values of the seeds is different than problem of number of clusters.. Normally one would use random cluster center. With better seeds, k-means converges faster and the quality of the clusters is good.
* We used two technique, **SimpleKmeans** and **Xmeans**.
* **Xmeans** turned out to be a **better** technique among the above two.