Clustering Methods

Aditya N. Govardhan

Abstract: In this report, the problem of clustering a given data set is approached using clustering methods: Hierarchical Agglomerative Clustering (HAC) and K-Means Clustering. A five dimensional dataset is provided, which is reduced to its two principal components. It has been found that the cluster formation is highly influenced by the distance metric and algorithm chosen for computing distance between clusters. This report aims to put light on the working of these algorithms and factors influencing its clustering methods. It also discusses about the preliminary design aspects to be taken into consideration while implementing clustering methods.

***Index Terms*—hierarchical agglomerative clustering, k-means clustering, principal component analysis, gap statistics**

I. Introduction

Clustering is one of the important classes of unsupervised learning in which data is divided into separate groups based on defined methods for similarity and dissimilarity. This can be done by various attributes of data: 1) location based 2) shape based and 3) density based.

Hierarchical Agglomerative Clustering (HAC) and K-means clustering are two of the various methods to cluster data. This paper analyses the aspects of implementing clustering methods for given five dimensional data with four thousand data points. In order to simplify the analysis and clustering of data, principal component analysis (PCA) is carried out and the first two principal components are chosen.

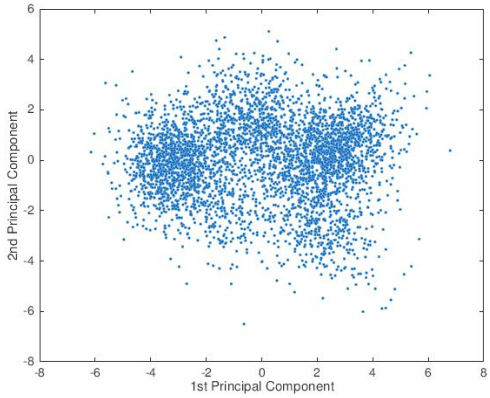

Fig. 1. Dataset after PCA

II. Method

A. Principal Component Analysis (PCA)

The given data is a five dimensional data with four thousand data points. Data lying in a dimension higher than three dimensions is usually difficult to visualize and can lead to misjudgement. However on the other hand, neglecting features usually leads to loss of information and incorrect clustering. Hence PCA is carried out on such data where it is projected on a new basis vectors with minimum variance to the original data. These basis vectors are orthogonal to each other and thus we can choose the basis vectors which carry the maximum desirable variance of the original data.

PCA on given data results in the following vector of variances along each new basis vector: [49.8978 19.7701 13.5004 9.4361 7.3956]^{T}. It can be seen that the first two components carry a total of 69.6679 percent of the original variance which is approximately 70 percent and a good value to carry out clustering. Thus the new data can be represented using two dimensional scatter plots (Fig. 1.). Note that choosing first three components is a viable option too but only two are chosen for visual convenience.

B. Hierarchical Agglomerative Clustering (HAC)

HAC is a data driven clustering method. For initialization, each data point is considered as a cluster. Then the distance between each pair of clusters is calculated using a distance metric and these distances are sorted in ascending order. The closest two clusters are merged together to form a new cluster. This process is repeated eventually leading to one cluster containing all data points.

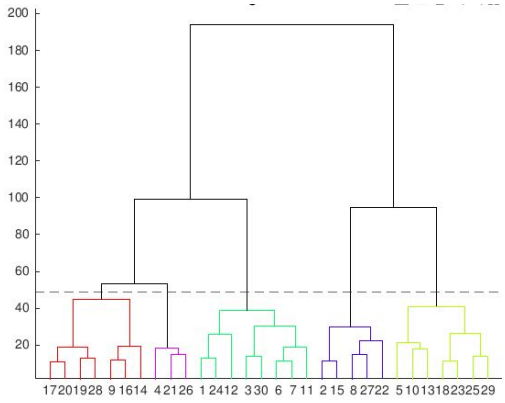

Fig. 2. Dendrogram for HAC

This process of merging closest clusters iteratively can be visualized using a “dendrogram”. HAC algorithm is carried out for a given dataset using “linkage” function in

Aditya N. Govardhan is with the Arizona State University, Tempe, AZ 85281 USA (e-mail: Aditya.Govardhan@asu.edu).

MATLAB and the corresponding simplified dendrogram is shown in Fig. 2. The height of each link between two clusters is equal to the distance between the clusters.

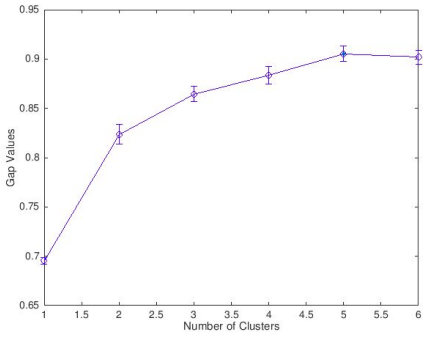

Fig. 3. Gap Statistics for HAC

The parameters that influence clustering is the distance metric chosen and how the distance between two clusters is calculated. The distance metric chosen for this dataset is simple euclidean distance since the data is two dimensional. To calculate the distance between two clusters, ward method is chosen which calculates the distance between two clusters, r and s, according to the following formula:

$$d(r,s) = \sqrt{\frac{n_r \cdot n_s}{n_r + n_s}} \, ||\overline{x_r} - \overline{x_s}||$$

where n_r and n_s are number of datapoints in the clusters and x_r and x_s are the centroids of clusters. These metrics leads to a more intuitive clustering as compared to other metrics. To decide the number of clusters to be formed, the dendrogram is cut off at a threshold and the leaves on each tree belongs to the same cluster. Gap statistics is performed for various cluster numbers and the elbow point is chosen, which in our case is five. Fig. 3. shows the gap statistic and the obtained clusters are shown in Fig. 4.

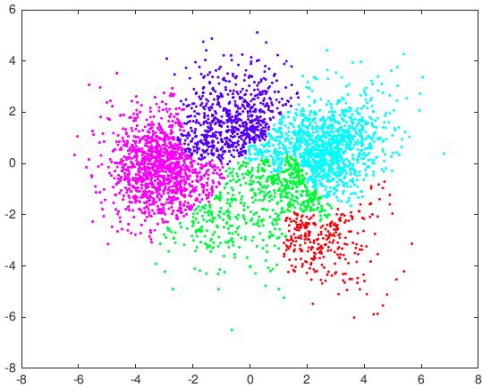

Fig. 4. Clusters using HAC

C. K-Means Clustering

K-Means clustering is a data driven model as well, where K is the number of clusters desired. In this method, K points are chosen which are initial centroids of the clusters. For each data point, its distance is measured from these K chosen points according to a distance metric and the data point is assigned to the cluster with minimum distance. The centroid of the corresponding cluster is updated.

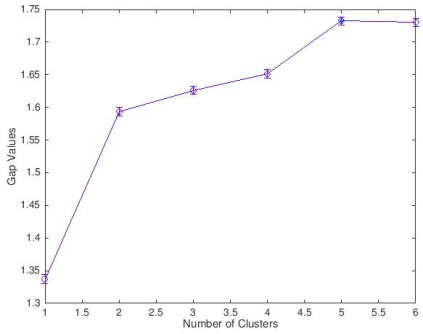

Fig. 5. Gap Statistics for K-Means Clustering

The parameters that influence clustering is the distance metric and the initial K centroids chosen. The distance metric chosen for this dataset is the square of euclidean distance since the data is two dimensional. For choosing the initial K centroids, k-means++ algorithm is used, which initializes the centroids randomly far from each other. However for choosing the value of K itself, again, gap statistics is used (shown in Fig. 5.). According to the gap statistics K value chosen is five. The obtained clusters are shown in Fig. 6.

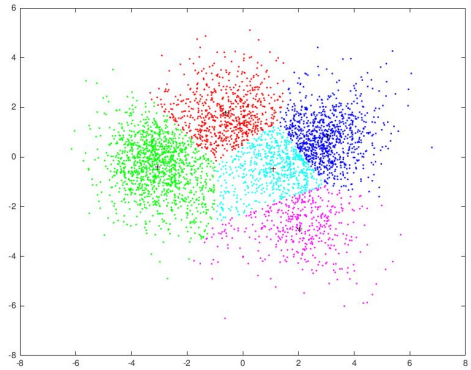

Fig. 6. Clusters using K-Means Clustering

III. Results

Data is first transformed into two dimensional data using PCA which retains 70 percent of the variance. In both methods, the number of clusters to be determined is found out to be five using gap statistics and the distance metric chosen for both methods is euclidean distance. Distance between clusters is found out using ward method in HAC while K initial centroids are chosen using k-means++ algorithm for K-Means clustering.

IV. Conclusion

The shape and size of clusters largely depended on the distance metrics. Choosing the number of clusters which was obtained using gap statistics but better techniques can be used.

References

[1] MATLAB Statistics and Machine Learning Toolbox Documentation