RESEARCH ARTICLE OPEN ACCESS

Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering

M. Venkat Reddy [1], M. Vivekananda [2], R U V N Satish [3]

Junior Technical Superintendent [1]
IIT Tirupati, Andhra Pradesh
Scientific Assistant [2] & [3], INCOIS
Hyderabad - Telangana

ABSTRACT
To implement divisive hierarchical clustering algorithm with K-means and to apply Agglomerative Hierarchical Clustering on the resultant data in data mining where efficient and accurate result. In Hierarchical Clustering by finding the initial k centroids in a fixed manner instead of randomly choosing them. In which k centroids are chosen by dividing the one dimensional data of a particular cluster into k parts and then sorting those individual parts separately, then the middle elements id in each part is mapped to id of m-dimensional data. The m-dimensional elements whose ids are matched, taken as initial k centroids of any cluster. The applying the Agglomerative Hierarchical Clustering on the resultant each element has its own individual cluster, where the clusters are merger based on the centroid distance. Then finally obtaining k-clusters. A Divisive hierarchical clustering is one of the most important tasks in data mining and this method works by grouping objects into a tree of clusters. The top-down strategy is starting with all objects in one cluster. It subdivides the clusters into smaller and smaller pieces by k-means algorithm by choosing initial k centroids in a fixed manner to get an efficient result, until each object form a cluster on its own and by applying Agglomerative Hierarchical Clustering on the result to get the efficient k cluster with high accuracy.

Keywords:- Clustering, Centroid, Hierarchical Clustering, K-Means, One Dimensional, M-Dimensional, Centroid Distance, Agglomerative Hierarchical Clustering, Divisive, Efficient, Result, Cluster, Accuracy.

I. INTRODUCTION

1.1 WHAT IS CLUSTERING?

Clustering or Cluster Analysis is a procedure of organizing the objects in groups whose member exhibits some kind of similarity. Cluster is a collection of objects which are alike and are different from the objects belonging to other clusters. A cluster of data objects which are considered as one group and while performing cluster analysis, initially partitioning the set of data into several different groups based on data similarity and then assign the labels to each and every group. It is a vital task in the process of exploratory data mining and a simple technique for statistical data analysis, used in various fields, including pattern recognition, machine learning, bioinformatics, information retrieval and image analysis.

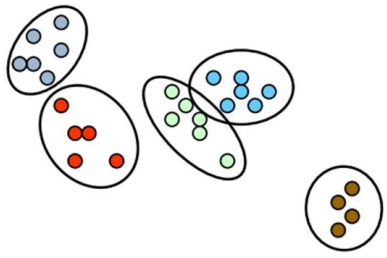

Figure 1.1: Simple Clustering based on Color

Cluster analysis as such is not an automatic task, but iterative process of knowledge discovery or interactive multi-objective optimization which involves trial and failure. Frequently it will be necessary to modify the data pre-processing and model parameters until the result satisfies the desired properties. The main advantage of clustering over classification is that, it is adaptable to the changes and helps single out useful features that distinguish different groups. These Clustering methods can be divided into eight different categories in which Hierarchical Clustering method creates a hierarchical decomposition of the given set of data objects. Hierarchical methods can be classified on the basis of how the hierarchical decomposition is formed. There are two approaches are Agglomerative approach and Divisive approach.

1.2 DIVISIVE HIERARCHICAL CLUSTERING

Divisive Hierarchical approach is commonly known as the top-down approach because in this, it generally starts with all of the objects in the same cluster. Then the continuous iteration, a cluster is split up into smaller clusters by the application of K-means Clustering. It is down until each object in one cluster or the termination condition takes holds. This method is rigid i.e., once a merging or splitting is done, it can never be undone

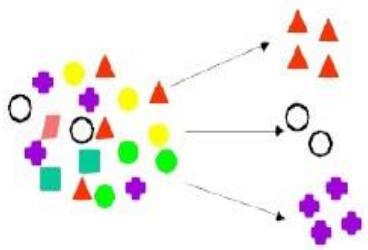

Figure 1.2: Clustering based on Color and Shape.

1.3 CLUSTERING METHODS

Clustering methods can be classified into the following categories.
• Hierarchical Clustering
• Centroid based Clustering
• Distribution-based Clustering
• Density-based Clustering
• Partitioning Method
• Grid-based Method
• Model-based Method
• Constraint-based Method

II. PROBLEM STATEMENT

The main problem focused here is combining of two algorithms i.e. divisive hierarchical clustering with K-means and agglomerative hierarchical clustering to increase the speed of clustering process and make data clusters more meaningful and highly similar data. The great challenge is to bring more closely connected data in single cluster and related clusters close to one another.

2.1 REQUIREMENTS OF CLUSTERING IN DATA MINING

• Scalability:
It is must to have highly scalable clustering algorithms that deal with large databases at any time.
• Ability to deal with different kinds of attributes:
Algorithms should be capable to be applied on any kind of data such as
• Discovery of clusters with arbitrary shape:
The clustering algorithm should be capable of detecting clusters of arbitrary shape.
These algorithms should not be restricted to only distance measures which tend to find spherical cluster of tiny sizes.
• High dimensionality:
The clustering algorithm should handle both low-dimensional data and at the same time it should handle high dimensional space too.
• Ability to deal with noisy data:
Some databases contain noisy, missing or erroneous data. And few algorithms are too sensitive to such data and which may cause poor quality clusters.
• Interpretability and usability:
The clustering results should be interpretable, comprehensible, and usable.
• It is essential to have minimal requirements of domain knowledge to determine input parameters
• Insensitivity to order of input records

2.2 *PROBLEMS WITH CLUSTERING*

- Dealing with large number of dimensions and large number of data items can be problematic because of time complexity.
- The effectiveness of the method used in the algorithm depends on the definition of distance for distance-based clustering.
- If an obvious distance measure doesn't exist, then it must defined, which is not always easy, especially in multi-dimensional spaces;
- The results of the clustering algorithm in many cases can be arbitrary in itself and can be interpreted in various ways.

2.3 HIERARCHICAL CLUSTERING

Hierarchical Clustering creates a hierarchical decomposition of the given set of data objects in the cluster. These methods are useful in classifying hierarchical methods on the basis of how the hierarchical decomposition is formed.

There are two approaches here.

2.3.1 *AGGLOMERATIVE APPROACH*

Agglomerative approach is popularly known as the bottom-up approach because in this, one starts with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds. The classic example of this is species taxonomy. The gene expression data might also shows the same hierarchical quality. Agglomerative hierarchical clustering starts with every single object or sample in a single cluster, then in every successive iteration, agglomerates the closest pair of clusters by satisfying some similarity criteria, unless all the data is in one cluster.

A. Process:
• Initially assign each and every object to different cluster.
• Evaluate all pair-wise distances between clusters distance metrics are described in Distance Matrices Overview.
• Construct a distance matrix using the distance values.
• Look for the pair of clusters with the shortest distance and remove this pair of clusters from the matrix then merge them.
• Evaluate all distances from this new cluster to all other clusters, and update the matrix.
• Repeat until the distance matrix is reduced to a single element.

B. Advantages:

• It can produce an ordering of the objects, which may be informative for data display.

• By using this approach smaller clusters are created which may be helpful for discovering similarity in data.

C. Disadvantages:

• No provision can be provided in this approach for relocation of objects that may have been incorrectly grouped at an earlier stages and the same result should be closely examined to ensure it have sense.

• Usage of various distance metrics for measuring distances between clusters may produce different results. Hence performing multiple experiments and then comparing the results is recommended to help the veracity of the original results.

2.3.2 DIVISIVE APPROACH

In Divisive approach one starts with all of the objects in the same cluster followed by continuous iteration, a cluster is split up into smaller clusters depending on their characteristics. This process continues until each object comes under one cluster or the termination condition holds. Here the method used in divisive approach is rigid, i.e., once a merging or splitting is performed, it can never be reverted.

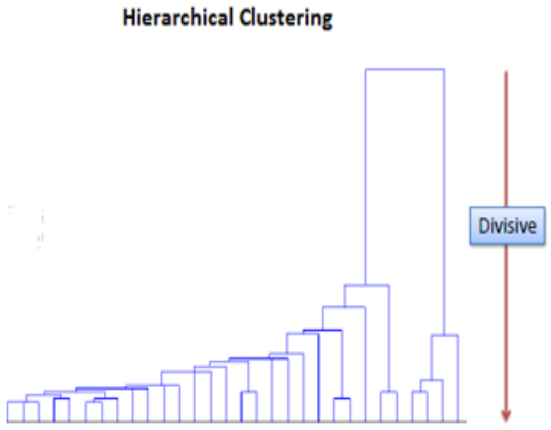

Figure 2.1: Divisive Hierarchical Clustering

Divisive Hierarchical Clustering with K-means:

Clustering is an important analysis tool in many fields, such as pattern recognition, image classification, biological sciences, marketing, city-planning, document retrievals, etc. Divisive hierarchical clustering is one of the most widely used clustering methods. Divisive hierarchical clustering with k-means is one of the efficient clustering methods among all the clustering methods.

In this method, a cluster is split into k-smaller clusters under continuous iteration using k-means clustering until every element has its own cluster. Here while using k-means clustering the initial centers are taken differently as by converting the m-dimensional data into one dimensional data, then dividing one dimensional data into k parts. Sorting those one dimensional in different parts and taking the middle element id and that particular ids one dimensional element is taken as centroid, these four centroids are taken as initial four centroids for the m-dimensional data . The architecture of the divisive hierarchical clustering with K-means clearly explains that it is working.

III. RELATED WORK

A Hierarchical Clustering Algorithm based on K-means with Constraints [1] by Guo Yan Hang, Dong Mei Zhang, Jia Dong Ren from China presents a research agenda to improve efficiency in Hierarchical Clustering. In their work they explained about importance of Hierarchical Clustering. In this paper, a Hierarchical Clustering Algorithm based on K-means with Constraints (HCAKC) is proposed. In HCAKC, in order to improve the clustering efficiency, Improved Silhouette i.e. CUCMC (Constraints-based Update of Cohesion Matrix between Clusters) is defined to determine the optimal number of clusters. Here silhouette is a function which measures the similarity of an object with the objects of its own cluster compared with the objects of other clusters. In addition, to improve the hierarchical clustering quality, the existing pairwise must-link and cannot-link constraints are adopted to update the cohesion matrix between clusters. In the algorithm proposed by them, the curve graph about average IS of the dataset and different partition number has been plotted. The optimal number of clusters is determined by locating the maximum of the curve graph. Because of this the complexity of the process when determining the number of clusters has been improved significantly. Thereafter, the existing constraints have been incorporated to complete the CUCMC during the hierarchical clustering process. In this experiment they also included a penalty factor to modify the similarity metric to address constraint violation. Hence, the clustering quality has been improved. This paper concluded that HCAKC has lower computational complexity and better clustering quality compared with the existing algorithm i.e. Cohesion based Self Merging.

[2]In this journal paper Khaled Alsabti et al, the authors introduce the novel algorithm for performing k-means clustering. The main aim of the author here was to study the computational aspects of the k-means method. Here the datasets are generated synthetically to understand the scaling properties of the algorithm used by them. They stated that their algorithm designed will significantly superior performance than the direct k-means algorithm in most cases of their experimental results. The scheme which they proposed is said to be improving the computational speed of the direct k-means algorithm by an order to two orders of magnitude in the total number of distance calculation and the overall time of computation.

[4] The journal paper titled, "Efficient High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm", the author says about high dimension data clusters. The author concerns about fulfilling the demanding task posed by large dimension clusters that are computation efficiency and result quality. Primarily, data clustering of high dimensions dataset is performed using Constraint-Partitioning K-Means clustering algorithm, which did not fit properly to cluster high dimensional data sets in terms of effectiveness and efficiency. So that the author followed the two step process were first dimensionality reduction on the high dimension dataset using principal component analysis and secondly is using of Constraint-Partitioning K-Means clustering algorithm which gave good and accurate clusters.

The authors of the paper, [7] Madjid Khalilian, Norwati Mustapha, MD Nasir Suliman, MD Ali Mamat "A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets", describes a method that uses divide and conquer technique with equivalency and compatible relation concepts to improve the performance of K-Means clustering method for using in high dimensional data sets in International Multi Conference of Engineers and Computer Scientists. This paper describes K-Means as staple of clustering methods, due to its ease of implementation. It works well for many practical problems, particularly when the resulting clusters are compact and hyper spherical in shape. The time complexity of K-Means is O(N .K .d .T) where T is number of iterations. Since K, d and T are usually much less than N, the time complexity of K-Means is approximately linear. Therefore K-Means is a good selection for clustering large-scale data sets. The main objective of their paper is to propose a framework of combining relational definition of clustering space and divide and conquer method to overcome the difficulties and improving efficiency and accuracy in K-Means algorithm to apply in high dimensional datasets. The experiment results of this method also showed appropriate accuracy along with speed up. In this study K-Means has been used for second phase whereas we can use other clustering algorithms e.g. genetic algorithm, Hierarchal Clustering algorithm, Ant Clustering, Self Organizing Maps etc. determining number of subspaces can be studied as important direction for the proposed method.

IV. SYSTEM DEVELOPMENT

Here in this section we will discuss the architecture of divisive hierarchical clustering with K-means.

From the below architecture the implementation of divisive hierarchical clustering with k-means is understood where first data taken which has objects and their measured features. First data will be read from the cluster list, where initially the whole data is taken as one big cluster which consists of all the objects. Then the next steps are given below.

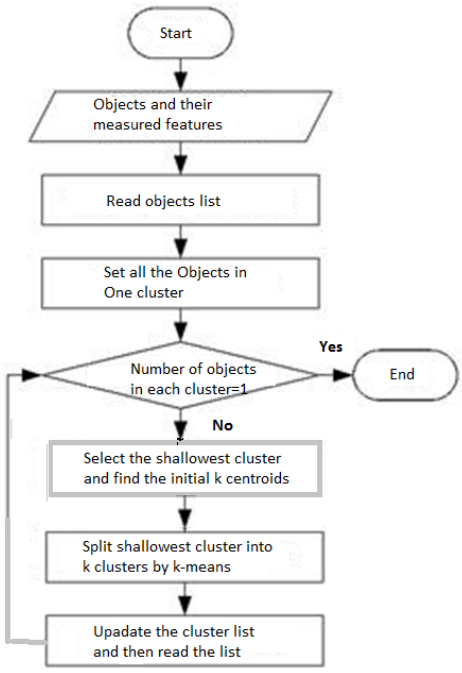

Figure 2.2: Architecture for Divisive hierarchical clustering with k-means

4.1 SELECTION OF INITIAL K CENTROIDS

Instead of selecting the initial centroids randomly here used some different process where first the m-dimensions corresponding one dimensional data is divided into k parts then the data is sorted in each individual part then take the middle elements id in each part, compare this ids with the m-dimensional data ids then these particular ids k elements are taken as initial k centroids. Then apply K-means clustering by considering these as initial k centroids.

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.

Figure 3.1: The above figure shows the selection process of initial centroids for any cluster, here c1,c2,c3,c4 are the initial centroids where K=4.

4.2 K-MEANS CLUSTERING

K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. This procedure in K-Means follows an easy and simple way to differentiate the given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.

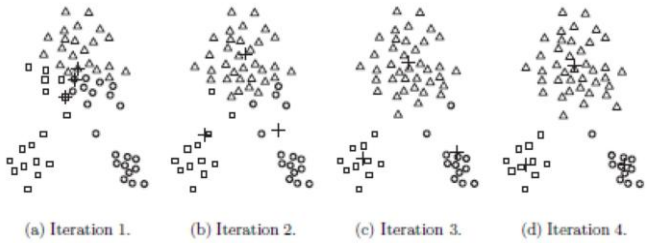

Figure 3.2: Architecture of K-means Clustering Algorithm

The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point one needs to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After acquiring these 'k' new centroids, a perfectly new binding has to be built in between the same data set points and the nearest new center. A loop has been generated. As a result of this looping may noticed that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

4.2.1 Algorithmic steps for k-means clustering

Suppose X= x1, x2, x3,....., xn be a set of data points given and let V=v1, v2,....., vc be a set of centers.

1. Find the initial k centroids by the above process.
2. Calculate the Euclidian distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster centre is least compared to all other cluster centers.
4. Recalculate the new cluster center using
5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3).

4.3 REPETITIVE APPLICATION OF K-MEANS

After applying k means clustering for the initial data, If considered initial data as one cluster then the initial cluster then split into k clusters, then if it is imagined to be in tree structure the next left most cluster will divide into k clusters as the above process will repeat where after finding initial centroids k-means clustering will be applied. This splitting process using k-means will repeat until each element is in one individual cluster. After the application of divisive hierarchical clustering one obtains the clusters given below figure. where the CLR1,CLR2,CLR18 are the all the clusters which obtained by divisive hierarchical clustering of the cluster CLR1 which is initial whole data. Every time every cluster is divided into 4 clusters. Finally each element has its own individual cluster. Example Tree Structure of Divisive Hierarchical clustering If original dataset contains 8 elements where k=2.The tree may looks like this.

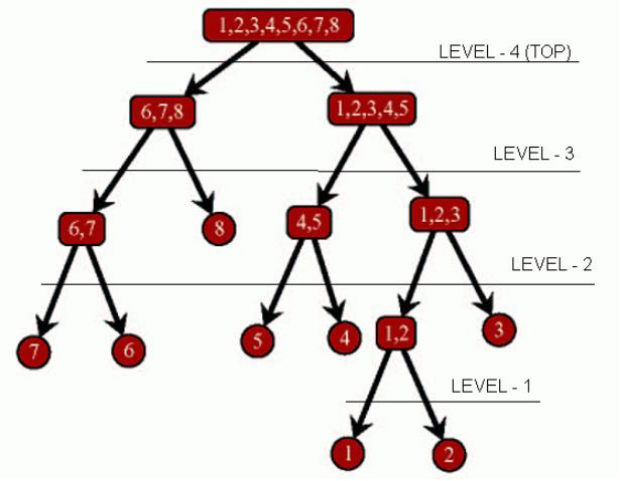

Figure 3.4: Architecture of K-means Clustering Algorithm

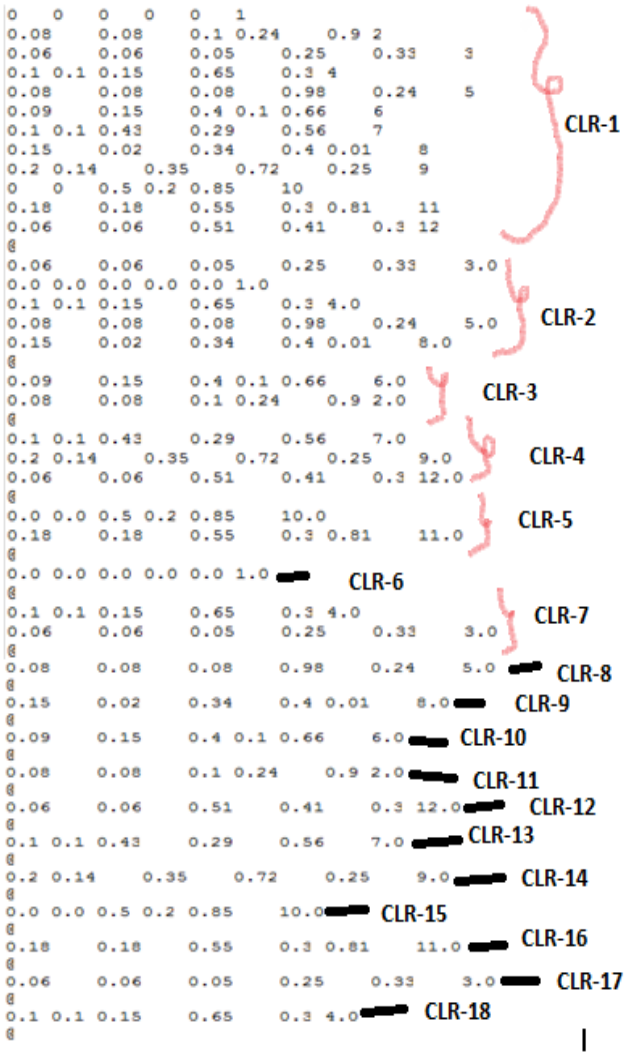

Figure 3.5: Architecture of K-means Clustering Algorithm

4.4 AGGLOMERATIVE CLUSTERING

Agglomerative Hierarchical Clustering:

Agglomerative Hierarchical Clustering is a bottom-up approach. This clustering starts with every single object (gene or sample) in a single cluster which is a result of Divisive Hierarchical Clustering with k means. Keeps on merging the objects or groups from the bottom of the tree of Divisive Clustering that are close to one another, until the number of groups or clusters are k. Gene expression data might also exhibit this hierarchical quality (e.g. neurotransmitter gene families).

4.4.1 Implementation of Agglomerative Hierarchical Clustering:

Get each object to a separate cluster by Divisive Hierarchical Clustering with k-means. Evaluate all pair-wise distances between the element and centroids of the clusters evaluate all distances from this new cluster to all other clusters by considering Euclidian distance between centroids. Look for the pair of clusters with the shortest distance merge them, and then update the cetroids. Repeat until the number of clusters is k. Finally calculate the accuracy of the clusters.

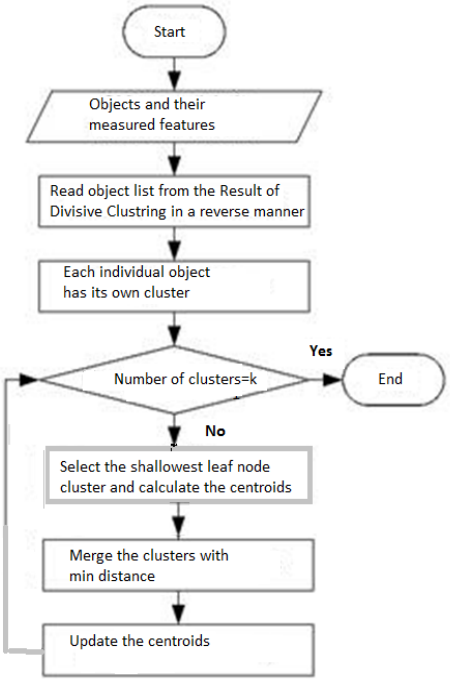

Figure 4.1: Architecture for Agglomerative Hierarchical Clustering on the Result of Divisive Hierarchical Clustering with K-means.

V. CONCLUSIONS

This algorithm implements Divisive Hierarchical Clustering with k-means efficiently, where the initial centroids for each cluster can be taken in a fixed manner instead of randomly choosing them. By choosing fixed centroids it gives an efficient result. Here implemented Agglomerative Hierarchical Clustering on the result to get efficient clusters high accuracy.

ACKNOWLEDGMENT

We thank each and every Colleague and Friend for their support and help in writing this paper. We especially thank our Parents and other Family Members for their encouragement without which it was highly impossible to take up this task. We thank our organization administration for providing sufficient tools to make use of them and get this activity successfully done. We would like to acknowledge every person who is directly or indirectly connected with this project and for the output of this paper.

REFERENCES

[1] Guo Yan Hang, DongMei Zhang,JiaDong Ren , A Hierarchical Clustering Algorithm based on K-means with Constraints (2009) ,pages(1479-1482).

[2] Khaled Alsabti, Sanjay Ranka and Vineet Singh "An efficient k-means clustering algorithm", Syracuse University SURFACE, L.C. Smith College of Engineering and Computer Science, 1997.

[3] Lor Rokach, Oded Mainmon (Tel-Aviv University), Clustering Methods pages.321-325.

[4] Aloysius George "Efficient High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm", The International Arab Journal of Information Technology, Vol. 10, Issue No. 5, September 2013.

[5] Ali Ghodsi, Dimensionality Reduction(2006) Department of Statistics and Actuarial Science , pages 1-17.

[6] https://en.wikipedia.org/wiki/Hierarchical_clustering

[7] Madjid Khalilian, Norwati Mustapha, MD Nasir Suliman, MD Ali Mamat, IMECS2010, A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets,,1-5 (2010).

[8] Fionn Murtagh and Pedro Contreras Methods of Hierarchical Clustering Science Foundation Ireland(2011) pages 1-21.