Chapter 5

Proposed HK-Hybrid Clustering Technique

The problem stated in above chapter is solved by a HK-Hybrid approach that is based on k-means and Hierarchical. It takes advantages of both algorithms and analyses Big Data in an efficient way. The proposed technique is considered whose purpose is to generate accurate clusters in less processing time.

5.1 HK-Hybrid Technique

Clustering is a technique which is very important and popular technique used for Big Data mining. Clustering is basically a part of unsupervised leaning. Clustering is a process of organising similar objects into groups. It does not only organize, but also identifies structure in the given set of unlabeled data. It is a technique to group number of systems in such a way to work together like a single system. K-means is simple and an efficient method used in data clustering technique. Hierarchical technique is also important and useful in data clustering.

The proposed method is a HK-Hybrid technique based on K-Means and Hierarchical that combines the benefits of both K-Means and Hierarchical algorithms. The benefit of K-Means is that it is not complex and forms clusters in less execution time while the advantage of Hierarchical is that hierarchical clustering you can stop at whatever level (or clusters) you wish.

The main procedure of proposed method that works in four stages:

Stage 1: The first stage is the elbow stage that calculates the optimal minimum number of clusters.

Stage 2: In the second stage a cluster vector of the size of the total data points.

Stage 3: The third stage is the distance matrix calculation after which closest pair are assigned to one cluster.
Stage 4: Stage 2 and 3 are repeated until all data points have been assigned to a cluster and total clusters formed are greater than the minimum number of clusters calculated at Stage 1.

5.1.1 Proposed Algorithm

The design of proposed HK-Hybrid algorithms is as follows:

Input: D: Dataset having p data points.
K: Minimum number of clusters to be formed.

Output: K clusters having Cm centre mean as cluster id.

Let X = {x1, x2, x3, ..., xn} be the set of data points.

1) Calculate the optimal value of k with ELB method
2) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
3) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.
4) Assign the cluster pair to same cluster
5) Deleting the rows and columns which are corresponding to cluster (r) and (s) and then adding of a row and column which is corresponding to newly formed cluster, then update the distance matrix D accordingly.
   The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = mean (d[(k),(r)], d[(k),(s)]).
6) If all the data points are assigned to a cluster and total unique clusters is greater than optimal value of k then stop, else repeat from step 2.
5.1.2 Proposed Algorithm: Execution Stages

The detailed explanation of how the proposed algorithm works is given here.

Stage 1: Calculation of \( k \)

For some values of \( k \) (for example 2, 4, 6, 8, up to 15), compute the sum of squared error (SSE). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid. Mathematically, it is represented as:

\[
SSE = \sum_{i=1}^{K} \sum_{x \in c_i} \text{dist}(x, c_i)^2
\]

As illustrated by the following plot between \( k \) and the SSE, when \( k \) (number of clusters) gets larger this (SSE) error decreases, so distortion is also decreases. The elbow method is used to choose the \( k \) at which the SSE decreases abruptly. This produces an "elbow effect" in the graph, as shown in the following picture:
In this case, $k=6$ is the value that the Elbow method has selected.

Stage 2: Cluster Vector

For this stage a cluster vector is declared of the size of the total data points and each point is assigned a default cluster number.

Stage 3: Distance matrix

Distance matrix is a square matrix (two-dimensional array) containing the distances, taken pair wise, between the elements of a set.

The minimum of this distance matrix gives the closest pair of data points from the data set. The centroid of these dataset is then calculated and the in the original dataset the corresponding rows for the calculated closest pair is replaced by the mean of these data points.
Stage 4: After stage 3 we are left with a new different dataset than the original with mean value for the closest pair of data points calculated. If after stage 2 and 3 all data points have been assigned to a cluster and the total number of unique clusters is found greater than the value of $k$ from Stage 1, we stop the execution. Else we repeat Stage 2 and 3 until the desired result is achieved.

### 5.2 Conclusion

In this chapter the design of proposed algorithm was presented. The detailed explanation of proposed method was given along with its execution diagram and flow chart. In the next chapter the implementation of this proposed approach is conducted on different datasets and results of the experiment would be gathered.