CHAPTER 5
AN EFFICIENT HIERARCHICAL CLUSTERING ALGORITHM

This chapter presents a detailed description of various approaches in an Efficient Hierarchical Clustering (EHC) algorithm and it is the improvement part of classification in the area of mining high-speed data streams. This chapter also presents a detailed description of the results being produced by different methods. The proposed algorithm has also been implemented and evaluated for their performance in various metrics.

In data streams the huge amounts of data are arriving rapidly they are very difficult to process and storing aspects and these raise new challenges in the Research problems. Surely it is not possible to store the continuous arriving data in the storage systems. Preferably the stream data have been processed from online so that the results must be up-to-date and it must be processed using simple query method with time constraint. The high-speed data can change the behavior of data stream in the case of hierarchical clustering. When data arrives fast, the stream speed increases from low grade to high grade so that the clustering data makes fast. Suppose when data arriving speed is decreases from high grade to low grade then the stream data will become imperfect. Therefore the problem in clustering must be analyzed.

5.1 Existing Methodology

The different methods for finding clusters of random shape have been proposed. The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method could handle only the static environment but not able to handle the fast-changing streams. Data stream models are not possible for random access. The data streams models are varying from the relational model. Various reasons of variations of data
streams models: The incoming data are not under the system control. Mining data stream systems are latent of unbounded size. Stream data is very difficult to retrieve unless it is stored in memory. Stream files may be archived or discarded. A stream has limited memory, time constraint and produces approximate results by processing them. The stream data must be scan at once and clustering via local information. There are different methods for finding insight clusters of random shape have been proposed in the literature.

These methods assume that all the data are the occupant on a hard disk and one can get general information about the data at any time. Thus they are not applicable for processing data streams. The recent work constitutes the ranking based method that evolves the k-means clustering algorithm execution and exactness. In that the k-means clustering algorithm has been analyzed by two ways, one is they have an existence k-means approach that contained some starting point(threshold) values and the second one is ranking method applied on the k-means algorithm that equates the functioning of these two methods by using graphs.

5.2 Proposed Methodology

5.2.1 The behaviour of High-Speed Data

The high-speed data can change the behaviour of data in the case of clustering. When data arrives fast, the stream speed increases from low grade to high grade so that the behaviour of data clustering is done. This also helps to calculate the distance between the objects using some functions. These functions also support to compute the measurement.

5.2.2 Group the objects using EHC (Efficient Hierarchical Clustering)

This step helps to link pairs of the objects that are in proximity using the linkage functions. The linkage function uses the distance information which has generated in the above to find out the proximity of objects to each other. As objects are coupled by
binary clusters, the newly shape clusters are sorted into larger clusters until a hierarchical tree is formed.

5.2.3 Output Cluster

This is the last step that helps to use the cluster function to eliminate branches away from the bottom of the hierarchical tree and allocate the objects below for each cut to form a single cluster. This creates a partition of the data. The cluster function makes groupings in the hierarchical tree by separating off the hierarchical tree at an absolute point.

5.2.4 Various Clustering methods

Clustering is a method that groups the object of the datasets of the same group and it is one of the methods in mining data stream. The task of mining data streams is complex for certain types of clustering. The clustering algorithm is used to arrange the data into similar or dissimilar to the other group. To mining, the data streams the efficient algorithm is required. The various clustering methods have been explained below

Partition clustering - K-means

K-means algorithm constructs $k$ partitions of $n$ data. The larger cluster appears to be separated into a lower difference area and a higher difference area. This may point that the larger cluster is a two intersection cluster. In this, it is used k-means technique to compute the distances from each one Centroid to point on a grid.

Distance squared Euclidean

It is the method of every Centroid is the mean of the points in that cluster. The formula $d(x,c)=(x-c)(x-c)'$ is used to calculate the Euclidean distance where $d$ is the distance; $x$ and $c$ are the points of the cluster.
Sum of differences

Each one Centroid is the element-wise median of the points in that cluster. The formula is \( d(x,c) = \sum [x_j c_j] p_j=1 \) is used to calculate the sum of differences where \( d \) is the distance; \( j \) is the indicator function; \( x \) and \( c \) are the different points in the cluster.

Deciding on \( n \) Clusters

It is an idea of how well output clusters are classified known by silhouette plot method. The silhouette plot displays an amount of closeness of each one point in the nearby clusters. This ensures that the ranges from +1, identifying points that are very far from nearby clusters, through 0, suggests points that are probably allotted to the wrong cluster.

5.3 The Process and Technique of K-means

![Flow chart for the process of K-means](image)

Figure 5.1 Flow chart for the process of K-means
In numerical minimizations, often the k-means technique reaches a starting point. It is possible for k-means to attain local minima while allocating anyone point to a new cluster that would high the total sum of the point Centroid but where the best solution exists. However using 'duplicates' one can defeat the problem by taking the one with the lowest total amount of distances, over all duplicates as the final answer.

The figure 5.1 explains that the process of k-means for larger datasets. It is also called non-supervised learning algorithm. This algorithm is simple and used to solve the problem which is related to long known clusters. The K-means algorithm is the type of partition clustering used to classify the given data objects into $k$ number of clusters by iterative method. In this, each Centroid involves finding the average vector location for every cluster and bringing the distance between Centroid and calculates. This concept is implemented in section 5.3.1 as an algorithm.

In k-means algorithm, the database is assigned into $k$ clusters in which each register be owned by the nearest average value cluster. This starts with some collection of data then the mean value for each cluster will be calculated. These are described as an algorithm in section 5.3.1.

### 5.3.1 Algorithm for K-Means clustering

<table>
<thead>
<tr>
<th>Step1: Input K number of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datasets $D={d_1,d_2,d_n}$ // [d-distance]</td>
</tr>
<tr>
<td>Step2: Output Set of k clusters</td>
</tr>
<tr>
<td>Step3: Repeat</td>
</tr>
<tr>
<td>Form k clusters and assign the Centroid;</td>
</tr>
<tr>
<td>Calculate the distance between each object;</td>
</tr>
<tr>
<td>Re-compute the centroid of each cluster;</td>
</tr>
<tr>
<td>Until the centroid do not change;</td>
</tr>
</tbody>
</table>
5.4 EHC Method

It is an idea that aims to find the essential structure of data by grouping the data objects into similar groups. The high-speed data stream is data that arrives fast, continuously and transient. In order to combine the data structure safely, the efficient clustering algorithm is required. In this research work, it is Efficient Hierarchical Clustering (EHC) algorithm is proposed to analyze the streaming data. In parallel, it made the comparative analysis of the popular algorithms to produce the essential features of an efficient clustering algorithm. The figure 5.2 shows that the EHC algorithm is richer to identify the clusters having with non-domain data. It is an idea to ensure nearby points that end up in the same cluster and tested using different threshold values. This starts with a collection $C$ of the $n$ singleton clusters, where each cluster contains one data point called $C_i \{X_i\}$. These concepts are implemented in section 5.4.1 as an algorithm.

5.4.1 EHC Algorithm

| Input: $X=\{x_1,x_2,\ldots,x_n\}$ // $[x_1\ldots x_n]$ object sets |
| Output: dis_func($c_1,c_2$) // Distance function |
| Start |
5.5 Comparison between K-Means and EHC algorithm

Scan at once

This is an important constraint because of fast and huge arriving data. The need to read the same data again does not arise for computation in the k-means but EHC algorithm performs this fast.

Memory and CPU usage

In K-means, the system can be able to process very large data in CPU, but it has limited memory size. In EHC algorithm these files can be archived.

Streaming data vs. cluster

The user has no previous knowledge about the data clustered and its condition. In addition, with streaming data, the cluster alters the process over time.

Ability of an EHC algorithm

- Find an absolute attribute of clusters without an existing knowledge,
- Filters noise in streaming data so that clustering result can be good.
- This algorithm maintains the fixed number of clusters over a modern applicable part of the stream for memory constraint.

**Cluster compactness**

Generally, the system is being mined for the huge amount of data in which all the data cannot be clustered because of limited memory and limited space. Thus using EHC algorithm 20000 observations with the limited size of memory can be mined.

### 5.6 Results and Discussion

To study the execution of the proposed EHC algorithm, it is applied and tested on MATLAB 2015A with 2 GB main memory. The EHC algorithm is used on Iris database for clustering the efficiency of an algorithm and also used in various diagrams as an output. According to the information given above, the maximum number of top degree nodes keeps increasing until the total number of items reaches to one and yielding a better result.

![Figure 5.3 Screenshot of dendrogram](image)

The EHC algorithm that builds a cluster hierarchy which is shown as tree diagram is called dendrogram. In figure 5.3, the dendrogram starts with each object in
a separate cluster. In each step, the 2 clusters that are common are merged into a single new cluster. The horizontal axis represents the objects in the original dataset. The height of head ‘U’ represents the distance of various objects. For example in figure 5.3, object 2 groups with object 1, 3, 4 and 5 with the height of 3.5.

Figure 5.4 Screenshot of Binary Cluster Tree

Figure 5.4 shows the screenshot of binary cluster tree being generated by the proposed algorithm. This provides various numeric values which were generated by the EHC algorithm.

Figure 5.5 shows the screenshot of cluster result in the form of scatter diagram. It has different types of cluster-classifications each are represented by different colours. The biggest classification of a particular cluster represents the increase in the data.

Figure 5.5 Screenshot of clustering the scatter diagram
5.7 Summary

This chapter highlights that the Efficient Hierarchical Clustering (EHC) is an enhancement of agglomerative hierarchical clustering algorithm and it is widely used for clustering large sets of data. From evaluation of the experiments, it is concluded that the accuracy of EHC algorithm for Iris datasets bearing real attributes is better than the k-means and simple hierarchical clustering methods. The time taken for clustering is very less. The algorithm uses dendrogram and scatters diagrams. These are the good clustering methods that produce high-quality clusters. Therefore the EHC algorithm performs good mining high-speed data stream for larger datasets.