Influence of an efficient Hierarchical Clustering Algorithm in analyzing Cancer affected DNA Dataset

E. Kiruba Nesamalar¹, J. Satheesh Kumar², T. Amudha³

¹Research Scholar, Ph.D. Research & Development Centre, Bharathiar University, Coimbatore, India
²Associate Professor, Department of Computer Applications, Bharathiar University, Coimbatore, India
³AmudhaSwamyNathan@Buc.Edu.In

E-mail: ¹kirubanesamalar@gmail.com, ²j.satheesh@buc.edu.in, ³AmudhaswamyNathan@buc.edu.in

Abstract. This research work presents an influence of hierarchical clustering approach in analyzing cancer affected DNA data set. The primary objective of this research work is to identify the best clustering algorithms to group cancer-affected DNA datasets. Data analysis shows an important role in bioinformatics. Data analysis technique used for grouping the data objects is based on unsupervised learning. Clustering is an unsupervised learning technique in data mining. It groups a set of clusters from the entire dataset. In this research work, 700 cancer-affected DNA datasets are considered for analysis. This research work compares three types of Clustering techniques, K-Means (KM), K-Medoids (KMS), and Hierarchical Clustering (HC), to group cancer-affected DNA. Each algorithm has some strengths and weaknesses. These clustering algorithms are compared in detail based on various parameters. Results prove that the hierarchical clustering algorithms show lesser execution time and increased accuracy than other KM and KMS algorithms.

Keywords – Cancer, DNA, Clustering, Data Mining, Bioinformatics

1. Introduction

Clustering is a Data Mining (DM) task to group a set of similar patterns without prior knowledge about the data. It groups the data objects into clusters based on the similarity of the data and other data objects are grouped into another set of clusters [1]. Clusters are defined as groups having identical data points. The significant difference from other machine learning analysis such as classification and regression is that the type of data set used to train the models [2]. This defines whether the machine learning technique is a supervised or unsupervised approach. In supervised learning techniques, the data works on labeled data. The unsupervised learning technique groups the data based on unlabelled data [3]. In Data Mining, the unsupervised learning technique clustering is used to group the data. This research work compares three different clustering algorithms such as K-Means, K-Medoids, and Hierarchical clustering, to find the most suitable clustering technique. These
clustering algorithms can be classified into two broad categories such as partitioning and Hierarchical methods [4]. Similarity or distance partitioning algorithms have been used in this research work to separate the data elements into the specified cluster numbers. Hierarchical algorithms combine the clusters in the hierarchical format. The Keras and Scikit-learn frameworks have been used for implementing the clustering algorithms [5].

Worldwide, Breast Cancer (BC) may occur commonly and popularly among women. According to global statistics, cancer deaths and new cancer cases are high nowadays. Through early detection of breast cancer, they can provide clinical treatment on time. This treatment helps to prevent cancer among patients by experiencing treatments in the classification of tumors. An advantage of detecting tumors can be determined by applying machine learning techniques for pattern matching among data grouping. Unsupervised learning algorithms have been considered for analysis in this research work since the data set consists of unlabelled DNA cancer records.

The primary objective of this research work is to identify the best clustering approach for clustering DNA cancer-affected data sets. This article also focuses on the importance of hierarchical clustering for the efficient grouping of cancer-affected DNA data.

2. Related Works on Clustering

M. A Qing et al. proposed a granular computing based on new K-Medoids clustering algorithm. Similarity function was used in datasets, and the granules are produced based on the equivalence relationship. According to the number of samples, the density of granules was defined. The closest to the granules were selected as initial centers for the K-Medoids algorithm [6]. X. GAO et al. proposed K-Medoids clustering based on incremental clustering. The updates of incremental clustering are collected and applied to the databases, and these patterns were derived from the databases. The updates were done incrementally since the size of the database was very large [7]. X. Zhang et al. proposed a parallel K-Medoids algorithm based on a map-reduce programming model. The map-reduce function was used to calculate the distance of each data object to the center point of the cluster, and the cluster reduce function was used to calculate the new center point of each cluster according to the intermediate results of the map section [8]. A.M. Fahim et al. proposed an efficient enhanced k-means clustering algorithm. K-Means minimize the mean squared distance from each data point to its nearest center. A simple data structure was used in clustering to keep the information in each iteration. The computational speed of the K-Means algorithm was improved by the total number of distance calculations [9].

K.G. Soni et al. proposed the analysis of K-means and k-medoids algorithm on iris data. The results obtained that k-medoids algorithm has the ability to provide better scalability than K-means algorithm. Hence K-medoids is more superior than K-means which has a execution time towards the data with outliers [10]. H.Drias et al. proposed a hybrid clustering algorithm based on K-Means and K-Medoids to make a comparison of K-means and K-medoids algorithms. Results obtained from hybrid clustering K-medoids is more efficient and effective compared to K-means, also that has the ability for the clusters to process faster and efficient in nature [11]. P. Arora et al. proposed the analysis of K-means and K-medoids algorithm to provide a comparison based on K-means and K-medoids of clusters in terms of selection and space complexity. By means of overlapping with clusters, K-medoids
is better in terms of execution time compared to K-means, and also it reduces outliers of data and minimizes the sum of dissimilarities of data objects[12].

3. Methods used for clustering

The benchmarking methods used for clustering include the K-Means (KM) and K-Medoids clustering (KMS) algorithms.

3.1. K-Means (KM)

The K-Means clustering is considered one of the widely used unsupervised algorithms. In a partitioning algorithm, the data points are used to form clusters. The cluster numbers can be represented by the value of “K” where “K” is the pre-determined number of clusters. It works iteratively on data elements based on the cluster numbers defined by the user. Initially, the “K” value is randomly selected through clusters known as centroids that find the closest data points. It continues iteratively until an optimum value is determined. K-Means clustering seems to work well for numerical data having lower dimensions [13].

The algorithm works with K points that are initialized randomly as centroids of clusters based on the pre-defined value of “K”. To form the “K” clusters, every data set is assigned to the nearest centroid by the distance. The Euclidean Distance measure calculates the distance between each data point and the initialized centroids. The closest Euclidean distance among the data objects can be found for better clustering since several previous research about clustering analysis gained excellent outcomes using the best Euclidean distance. The centroids are recalculated to reduce the whole intra-cluster change, and this process continues until specific criteria have been reached [14]. The total sum of all the distances between the data element is calculated when there are no changes in the centroid values. Cluster centroid has been updated continuously until the data points allocated to the clusters are similar to the previous assignment or the determined iteration number has been reached [15]. Advantages of the K-Means clustering algorithm include the minimal number of computations and grouping the clusters based on the minimum distance between data points. Hence, this algorithm can be computationally faster than hierarchical clustering. The time complexity of K-Means algorithm is O (n), and “n” is the cluster number. Additionally, this algorithm can easily adapt to new data samples. The drawback of this algorithm is the total number of clusters “K” which has to be indicated manually. The clustering results can vary depending on initial values of “K”. K-Means also randomly select the initial centroids for “K” clusters. Hence, the results can be different from one execution to another. Drawbacks of the K-Means algorithm include outlier detection, and it is not suitable for varying size data. The outliers or noises of the data set can affect the clustering process since the cluster and outliers themselves can form a cluster [16].

3.2. K-Medoids (KMS)

The K-Medoids algorithm is used to compute medoid values of items in a group of reference points located in the cluster. KMS process takes place by applying a partitioning method between the data point and the reference point. This approach is to find the k-number of clusters for “N” objects in
each set. This algorithm takes “K” as an input parameter [17]. The clusters are formed by partitioning among a set of “N” objects. Input to this algorithm are a number of clusters (K) and a data set (D) containing a set of “N” objects. The output provides a set of “K” clusters determined by decreasing the sum of objects where dissimilarity among the data objects is found by KMS clustering. This algorithm randomly selects the “K” number of objects, and “D” is considered an initial set of objects. Repeat Clusters are formed by the residual object with the nearest medoid, and a set of iterations occurs by randomly selecting a non-medoid object [18]. The exchange of objects is done at a random interval to form a new set of KMS. It is used to determine the “K” number of partitions among “N” objects. This algorithm iteratively performs a random selection of initial data, which improves the selection of the best set of medoids for effective clustering [19].

4. Efficient Hierarchical Clustering algorithm based on optimal cluster number for DNA grouping

4.1. Dataset

In this research work, the dataset used is openly accessible from UCI machine Learning repository [20]. It takes a set of samples among patients with solid breast masses. It contains a set of cancer data which is stored in CSV format with 31 attributes (p_id, p_diagnosis, p_radius_mean, p_texture_mean, p_perimeter_mean, p_area_mean, p_smoothness_mean, etc.) with a collection of 700 records. Figure 1 shows the top five data of the cancer-affected DNA dataset.

Figure 1. Top five data of cancer affected DNA dataset

In this research, considered 700 rows and 31 columns, in which diagnosis is an attribute that determines whether the cancer is Malignant and Benign. In a set of 700 datasets, 357 persons are categorized as Benign, and 212 persons are categorized as Malignant.

4.2. Hierarchical Clustering based on optimal cluster number for DNA grouping

This research work has been implemented in python. Hierarchical Clustering algorithms try to build a hierarchy of clusters. It starts with some progressively joins and primary groupings to the solution. This algorithm has two categories such as agglomerative and divisive. The agglomerative method first takes each data element as a distinct cluster. It iteratively merges the clusters until the final cluster comprises all data points in it. According to this approach, it combines the clusters based on a bottom-up approach. Agglomerative clustering, will work based on top-down flow from a single cluster with all data points and iteratively split the group into smaller ones until each cluster contains one data point. In this research work, the agglomerative approach has been used for clustering cancer data set.
As an initial step of the Agglomerative hierarchical clustering algorithm takes every element as an individual cluster. A proximity matrix is used to determine the distance between the groups. There are four distance functions available for the proximity matrix such as single linkage (min), average linkage, complete linkage, and ward (max) [21]. Single link means the distance among two or different clusters is well defined as the minimum distance between the first cluster group and the second cluster group. The maximum space of two data points value as the distance among two or different clusters is called complete linkage. The average link calculates the distance between the first cluster’s data points with all others from the second cluster. It takes the median value as the distance between the clusters. Ward is similar to the average linkage except that it uses the sum of squares to calculate the distance between the points. This research work uses “ward” as a distance function. The hierarchical clustering algorithm is shown in Algorithm 1.

Algorithm 1: Hierarchical Clustering

Step 1: A set of s objects \((s_1, s_2, s_3, \ldots s_n)\)

Step 2: Distance Measure \(\text{dist}(c1, c2)\)

For \(i=1\) to \(n\)

\(C_i = \{s_i\}\)

End for

Step 3: Find the closest pair of clusters

\(c = \{C_1, C_2, C_3, \ldots C_n\}\)

\(i = n + 1;\)

Step 4: Similar adjacent clusters are merged to form a group.

While \(c.\text{size} > 1\) do

\(\text{Mindist}(C_o, C_i)\)

add \(C_{\text{min1}}, C_{\text{min2}}\) from \(c\)

End while

Step 5: Update Distance matrix

updateclust \((C_{\text{min1}}, C_{\text{min2}})\) to \(c\)

\(i = i + 1;\)

In general, the hierarchical cluster algorithm will form a single tree of clusters. Each node represents the clusters, and each data point starts as a tree leaf. The tree root is the final cluster containing all of the data points. Hierarchical clustering cut out the “K” clusters from the last group. The algorithm successively forms a single tree of clusters. It then cuts at a certain level “K”, resulting in different clusters. The advantage of this hierarchical algorithm is that the total number of clusters is not essentially required. Like the K-Means approach, hierarchical clustering algorithms are easy to implement.

The choice of an appropriate metric is considered vital because it influences the clusters’ shape. Some elements close to one another according to one distance to another. Some commonly used metrics for hierarchical Clustering are shown in Equation 1, Equation 2 and Equation 3. In this research work Euclidean distance measure has been used.
Manhattan Distance  \[ \|a - b\|_1 = \sqrt{\sum |a_i - b_i|^2} \] \hspace{1cm} (1)

Euclidean Distance  \[ \|a - b\|_2 = \sqrt{\sum (a_i - b_i)^2} \] \hspace{1cm} (2)

Maximum Distance  \[ \|a - b\|_{\infty} = \max |a_i - b_i| \] \hspace{1cm} (3)

Equation 4 clustering criterion function is used to group the adjacent clusters

\[ d_{\text{min}}(D_i, D_j) = \min \| x - x' \|, \text{ where } x \in D_i \text{ and } x' \in D_j \] \hspace{1cm} (4)

The merging of the two clusters corresponds to adding an edge between the nearest pair of nodes in \( D_i \) and \( D_j \).

Advantage of a hierarchical algorithm is that it results in a tree-based representation called dendrogram for identifying groups in the dataset. In this Hierarchical approach, there are no backtrackings which means once one cluster is created, then the membership data points cannot be moved around. This algorithm is sensitive to noises and outliers depending on the choice of the distance matrix. Besides, this algorithm can face difficulty in handling the different size and convex-shape clusters [22].

4.3. Parameter Specification for clustering

Feature size and optimal cluster number are two important parameters that need to be set before implementing K-Means, K-Medoids, and Hierarchical clustering algorithms.

4.3.1. Feature size

Clustering may work by grouping a set of cancer data on the partitioning of data. It may confuse the model that does not tend to provide better accuracy. Hence, K-Means clustering is used to determine the size of cancer data for analysis [23]. Cancer data has a set of “K” number of clusters from 10 to 100. Comparison among data points of different sizes is determined that may affect the clustering performance of data. Based on the comparison, better data size is selected by the number of clusters with different data sizes.

4.3.2. Optimal Cluster Numbers K

K-Means algorithm is a kind of partitioning method where “K” is the pre-defined cluster numbers. An optimum cluster number has been calculated since this value plays a vital role in deciding the cluster efficiency. K-values can be approximately found for different kinds of shaped and sized data [24]. The elbow method’s optimal K-number of clusters can be determined, which seems to find the best measure for visualizing the data. The Elbow method is calculated by measuring the sum of clusters among a different set of clusters. The sum of squared errors is identified by finding the sum of the squared distance of data points from the centroid of a cluster [25].
5. Experiments on cancer Dataset

Jupyter Notebook has been used to import the necessary set of libraries for the entire dataset. Once importing the dataset, the Pandas library has been used to examine the real data manipulation. Data visualization works well in Python Programming language. Libraries such as seaborn and matplotlib, and Panda. These three visualization library is used for data analysis. For clustering, p_radius_mean and p_texture_mean feature are used in this research work. Figure 2 shows the visualization of p_radius_mean and p_texture_mean before clustering. The data plot without a diagnosis label is shown in Figure 3. The distance between two data points can be determined by Euclidean distance.

![Figure 2. Visualization of Radius and Texture mean before clustering](image)

Within Cluster Sum of Squares (WCSS) is a metric used for the k value selection process. After the selection process of the WCSS elbow rule is applied for the k value. Clusters among different data objects can be formed by measuring the sum of squared errors using the elbow method.  

![Figure 3. Data plot without diagnosis label](image)
Figure 4 shows the optimal “K” using the elbow method. The sum of square errors measuring the sum of the squared distance of each data points from its centroid of a cluster. K-Means popularly uses the Euclidean distance metric. After plotting the sum of squares at each number of groups matched with the respective number of clusters, the graph decreases the error sum. Final Optimal “K” is determined using the elbow method.

6. Results and Discussion

In this research, the dataset does not have a dependent variable or outcome variable. It uses unsupervised learning techniques such as KM, KMS, and HC to group cancer-affected DNA. The Pearson correlation function is used to find the correlation among the data points in the data frame. Row variable correlation with the column variable is the value of the cell. The correlation of the diagonal variable is 1.

Figure 5 shows the correlation matrix of the Pearson method. In this research work, data depends on independent variables. The correlation matrix of the heat map is shown in Figure 6. The heat map shows the cancer data that depends on two independent variables as a color-coded image plot. Figure 7 and Figure 8 show the data after K–Means and K-Medoids Clustering.
Each data point is transformed into a cluster using the closest two data points and repeat this process until the final cancer-affected DNA dataset. Table 1 shows the radius and texture of the cancer dataset. Figure 9 shows the dendrogram diagram of cluster data. Figure 10 shows the data after Hierarchical Clustering.

Table 1. Radius and texture mean of cancer DNA dataset

|   | Radius_mean | Texture_mean |
|---|-------------|--------------|
| 0 | 17.99       | 10.38        |
| 1 | 20.57       | 17.77        |
| 2 | 19.69       | 21.25        |
| 3 | 11.42       | 20.38        |
| 4 | 20.29       | 14.34        |

Figure 6. Heatmap of the Correlation matrix

Figure 7. Data after K-Means

Figure 8. Data after K-Medoids

Figure 9. Dendrogram of Hierarchical Clustering

Figure 10. Data after Hierarchical clustering
The Execution time of the Hierarchical clustering Algorithm is calculated using the Equation 5. Table 2 and Table 3 shows the comparison of performance based on Execution time and accuracy.

\[ O(T*K*N*D) \] \hspace{1cm} (5)

T is the number of iteration, and K is the number of centroids. N denotes the data point, and D is the dimension of each data point. In terms of accuracy, correctly clustered clusters are calculated using Equation 6.

\[ \text{Accuracy} \% = \frac{\text{Total Correctly Clustered Instances}}{\text{Total instances}} \times 100 \] \hspace{1cm} (6)

| Datasets | Instances | Time (MS) |
|----------|-----------|-----------|
|          | Hierarchical | KM | KMS |
| Cancer   | 100        | 250.4 | 281.4 | 309.2 |
|          | 150        | 375.3 | 384.2 | 399.2 |
|          | 200        | 500.2 | 516.7 | 539.1 |
|          | 250        | 624.5 | 642.8 | 679.4 |
|          | 300        | 751.1 | 820.4 | 849.7 |

Purpose of this research involves DNA affected cancer dataset which is used to analyse a set of data and also checks whether the data is cancerous or not which is done by means of clustering. Based on comparison with performance of execution time, duration varies based on the instances of data. In cancer data of 100 instances, Hierarchical clustering seems to take 250.4 Millisecond (MS) which has a lesser execution time compared to K-means and K-medoids clustering techniques. In 150 instances, Hierarchical clustering seems to take 375.3 Milli Second which has a lesser execution time compared to K-means and K-Medoids clustering. Hence, by the analysis of cancer data, the processing time of the hierarchical clustering algorithm is reduced compared with other clustering algorithms such as K-Means and K-Medoids.

| Datasets | Instances | Accuracy |
|----------|-----------|----------|
|          | Hierarchical | KM | KMS |
| Cancer   | 100        | 99.9% | 99.8% | 99.7% |
|          | 150        | 99.8% | 99.5% | 99.4% |
|          | 200        | 99.8% | 99.3% | 99.4% |
|          | 250        | 99.7% | 99.3% | 99.1% |
|          | 300        | 99.7% | 99.1% | 98.3% |

In cancer data of 100 instances hierarchical clustering seems to have 99.9% which is higher accuracy rate compared to K-means and K-medoids techniques. With 300 instances, hierarchical clustering seems to have 99.7% which is higher accuracy rate compared to other benchmarking.
methods. Hence, by the analysis of cancer data, Hierarchical clustering seems to gives a better accuracy rate compared to K-means and K-medoids clustering techniques.

7. Conclusion

In this research work, clustering algorithms are compared to identify the best clustering for grouping cancer-affected DNA datasets. The outcome of the research work shows that Hierarchical clustering algorithms are very much helpful to group cancer-affected DNA effectively. In KM and KMS algorithm, this research found a high error rate, so it is not convenient for medical data. Based on the analysis of data, a popular cluster method partitioning-based method is very much suitable for medical data. Clustering methods such as KM, KMS, and HC are used to find the similarity among the data objects to form a group of clusters. Compared to the clustering techniques, KM and KMS techniques take much time to perform grouping. The outcome of the HC algorithm shows higher consistent in grouping the DNA cancer datasets when compared with the KM and KMS algorithms. Accuracy can be extended by using optimized algorithm. Optimization plays a vital role the better optimization algorithm will enhance the accuracy of Hierarchical clustering to 100% for all the instances.

References

[1] R. Xu, D. Wunsch. "Survey of Clustering Algorithms." IEEE Transactions on Neural Networks, vol. 16, pp. 645-677, May. 2005.
[2] S. Bubeck and U. Von Luxburg. "Nearest Neighbor Clustering: A Baseline method for Consistent Clustering with Arbitrary Objective Functions." Journal of Machine Learning Research, vol. 10, pp. 657-698, Mar. 2009.
[3] D. Sisodia, L. Singh, and S. Sisodia, “Clustering Techniques: A Brief Survey of Different Clustering Algorithms”, (2012).
[4] Chavent, M., de Carvalho, F. D. A., Lechevallier, Y., & Verde, R. (2006). New clustering methods for interval data. Computational statistics, 21(2), 211-229.
[5] Handl, J., & Knowles, J. (2004, September). Evolutionary multiobjective clustering. Parallel Problem Solving from Nature (pp. 1081-1091). Springer, Berlin, Heidelberg.
[6] Qing, M. A., & Juanying, X. (2012). New K-medoids clustering algorithm based on granular computing. Journal of Computer Applications, 7.
[7] GAO, X., FENG, Y., & FENG, X. (2005). Incremental clustering algorithm based on k-medoids. Computer Engineering, S1.
[8] Zhang, X., Gong, K., & Zhao, G. (2013). Parallel K-Medoids algorithm based on MapReduce. Journal of Computer Applications, 33(4), 1023-1025.
[9] Fahim, A. M., Salem, A. M., Torkey, F. A., & Ramadan, M. (2006). An efficient enhanced k-means clustering algorithm. Journal of Zhejiang University-Science A, 7(10), 1626-1633.
[10] Soni, K. G., & Patel, A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. International Journal of Computational Intelligence Research, 13(5), 899-906.
[11] Drias, H., Cherif, N. F., & Kechid, A. (2016). K-MM: A hybrid clustering algorithm based on k-means and k-medoids. In Advances in Nature and Biologically Inspired Computing (pp. 37-48). Springer, Cham.
[12] Arora, P., & Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. Procedia Computer Science, 78, 507-512.
[13] Wu, J. (2012). Cluster analysis and K-means clustering: an introduction. In Advances in K-means Clustering (pp. 1-16). Springer, Berlin, Heidelberg.

[14] Jain, A. K. (2008, September). Data clustering: 50 years beyond k-means. Machine Learning and Knowledge Discovery in Databases (pp. 3-4). Springer, Berlin, Heidelberg.

[15] Shakeel, P. M., Baskar, S., Dhulipala, V. S., & Jaber, M. M. (2018). Cloud based framework for diagnosis of diabetes mellitus using K-means clustering. Health information science and systems, 6(1), 1-7.

[16] Tokushige, S., Yadohisa, H., & Inada, K. (2007). Crisp and fuzzy k-means clustering algorithms for multivariate functional data. Computational Statistics, 22(1), 1-16.

[17] Sheng, W., & Liu, X. (2006). A genetic k-medoids clustering algorithm. Journal of Heuristics, 12(6), 447-466.

[18] Li, C., Li, J., & He, M. (2016). Concept lattice compression in incomplete contexts based on K-medoids clustering. International Journal of Machine Learning and Cybernetics, 7(4), 539-552.

[19] Hu, J., Xu, K., Wang, G., Liu, Y., Khan, M. A., Mao, Y., & Zhang, M. (2021). A novel landslide susceptibility mapping portrayed by OA-HD and K-medoids clustering algorithms. Bulletin of Engineering Geology and the Environment, 80(2), 765-779.

[20] O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

[21] D. L. Yse, “A complete guide to K-means clustering algorithm”, https://www.kdnuggets.com/2019/05/guide-k-means-clustering-algorithm.html, May 2019.

[22] K-Means Advantages and Disadvantages, https://developers.google.com/machine-learning/clustering-algorithm/advantages-disadvantages.

[23] Ghosh, Soumi, and Sanjay Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms", IJACSA, Vol.4, No.4, 2013.

[24] C. Reddy. Understanding the concept of Hierarchical clustering Technique, (2018, Dec 10), https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec.

[25] A. Naik. Hierarchical clustering algorithm, https://sites.google.com/site/dataclusteringalgorithms/hierarchical-clustering-algorithm