IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 1, August 2023

Exploring Clustering Algorithms for Parkinson's Disease Data: A Comparative Analysis

Dip Das[1] and Adepu Rajesh[2]

Assistant Professor, Computer Science and Engineering[1]
Associate Professor, Computer Science and Engineering[2]
Guru Nanak Institute of Technology, Hyderabad, India
hellodip2@gmail.com[1] and adeprajesh@gmail.com[2]

Abstract*:* *Clustering, an essential analytical approach utilized in data mining, encompasses the act of grouping alike data items into clusters. It is crucial to note that the clustering outcome is significantly impacted by the employed clustering algorithm. This research paper presents a thorough analysis of various clustering algorithms, such as k-means, hierarchical, and DB-scan clustering algorithms, among others, while simultaneously scrutinizing their strengths and limitations. Within each algorithm type, the computation of the distance between data objects and cluster cantres is executed in every iteration, which inevitably poses a challenge to the efficiency of clustering. This paper provides an extensive summary of the fundamental techniques and highlights the associated challenges with clustering algorithms, such as recall, precision, and f-measure, to produce superior outcomes under diverse circumstances. The paper concludes with a discussion of the results obtained from a high-dimensional dataset of Parkinson's disease.*

Keywords: Data mining, Clustering algorithm, k-means, Parkinson’s disease

I. INTRODUCTION

According to a survey conducted at UC Berkeley, it has been observed that the overall amount of generated data has shown an exponential growth throughout the past decade.[1]

The present surge in both the amount and diversity of data necessitates the advancement of methodological approaches for comprehending, processing, and categorizing such data. From a technical standpoint, understanding the structure of massive datasets that have arisen as a result of data proliferation is of paramount importance in the field of data mining. In this research, I have concentrated on data mining techniques, specifically the clustering of data analysis, in both the machine learning and bioinformatics domains. The distinctiveness of bioinformatics data, which sets it apart from machine learning data, is characterized by a significant amount of random noise, missing values, an expansion of the range of thousands, and a small sample size.

1.1 Knowledge Dicovery Process

It is of great importance to establish a robust method for analyzing copious amounts of data in order to uncover valuable insights, particularly given the exponential growth in the size of data files, storage, and repositories. This, in turn, can facilitate more informed decision-making. Referred to as "Knowledge Discovery in Databases" or simply KDD, data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns within data. Although often presented as distinct phenomena, data mining is in fact an integral part of the knowledge discovery process.

1.2 Clustering

This particular model employs a methodology that partitions an extensive collection of data into smaller subsets or clusters. Each cluster comprises a group of data objects that share a high degree of similarity with one another within the same cluster, yet remain distinct from the data objects in other clusters.[2][3][4]

Classification: Classification trees, also known as Decision trees, are a statistical tool that partitions a set of records into disjoint classes. These records are presented as tuples with various numeric and categorical attributes, along with an

DOI: 10.48175/IJARSCT-12473

additional attribute that specifies the class to be predicted. The Decision trees algorithm is characterized by its variability in selecting variables for partitioning, as well as its approach to identifying the dividing points.[2][3]
Association Mining: The analysis reveals intriguing correlation patterns among a vast collection of data elements by demonstrating the frequent co-occurrence of attribute value conditions.[2][3]

1.3 Clustering Algorithm

K stands for the number of clusters in the phrase "k-means clustering." Typically, k's importance is not known a priori and must be determined by the user. Each cluster has a centroid, which is often calculated as the cluster's mean of its feature vectors. According to the k-means clustering technique, each data point's cluster membership is determined based on the cluster-centroid that is closest to the point. The user specifies k initial values for the centroids at the start of the clustering process since centroids cannot be determined directly until clusters are created. Once clusters have formed, the actual centroid values are determined

The following stages are used by the k-means method to divide a dataset into k groups. [13]

1. Set k starting values for the cluster centroids.
2. Locate the nearest centroid for each data point in the dataset, then assign the point to the cluster corresponding to this centroid.
3. Based on the updated cluster memberships, determine the centroid for each of the k clusters.
4. Repeat procedures (2) and (3) as necessary until a termination condition is satisfied.

In step (4), a variety of termination criteria may be applied. Every condition compares the value(s) of the same measure calculated in the current iteration to the value(s) of the same measure calculated in the previous iteration. There are three often used conditions:

I. The centroids do not change;
II. The sum of squared distances from each of the data points to their respective centroids does not change;
III. The cluster membership of the data points does not change.

We use requirement (iii) as the k-means clustering completion requirement. The centroids calculated based on these memberships do not change when the participations of the data points do not. As a result, neither does the sum of squared distances between the data points and their centroids.

The closest centroid for each point is found by the k-means algorithm in Step 2 of the clustering process. Only when there is a predetermined distance measure is the word "nearest" useful. The Euclidean distance is the measure that we employ. Even though there are a variety of different metrics that may be utilised, we choose to use the Euclidean distance for both k-means and hierarchical clustering since it was applied in numerous early studies that used cluster analysis to identify PDD subtypes, which led to more accurate results.

1.4 Hierarchical Clustering

To perform hierarchical clustering on a set of n data points, a symmetric n x n distance matrix consisting of pair-wise distances between the data points is generated. The main steps of hierarchical clustering are [15].

- Assign each data point to a separate cluster to obtain n clusters, each of whichcontains one data point.
- Find the closest pair of clusters; merge the two clusters to form a new cluster.
- Compute distances between the new cluster formed in step (2) and each of theremaining clusters;
- Iterate through steps (2) and (3) until all of the n points in the dataset are merged intoa single cluster.

II. EXPERIMENTAL RESULTS AND COMPARATIVE STUDIES

Three datasets are utilised in the experiment to examine the effectiveness of the k-means and hierarchical clustering algorithms. The UCI Machine Learning Repository[17] served as the source for all three of these datasets. The characteristics of these three datasets are shown in Table 1.

Datasets	No. of samples	No. of features	No. of classes
PD_Speech_Features	756	754	2
PD_Voice_Features	196	23	2

DOI: 10.48175/IJARSCT-12473

Parkinson_Multiple_Sound	1040	29	2

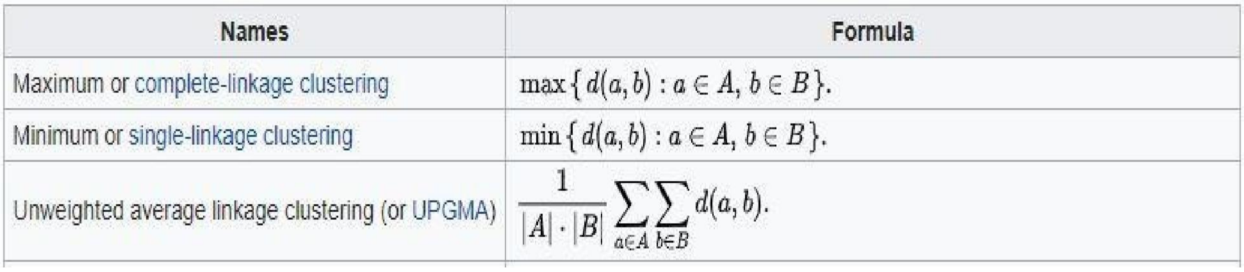
Table1 : Data Sets

2.1 K-means Algorithm:

The goal of the method is to split a set of n input data points into k clusters such that the clusters to which the points belong appear to be very similar to one another and the clusters to which they belong appear to be less similar.The output of the k-means clustering is shown first. All of these results are validated by using stability and fitness. The results of the stability validation are presented after the findings from the fitness validation for each technique. Once the cluster assignments of the data points stop changing, the k-means clustering technique utilised in this study is complete. We used launching points.

2.2 Hierarchial Algorithm

As a function of the pairwise distances between observations, the linkage criteria calculates the distance between sets of observations. The following are a few often used connection criteria between two sets of observations A and B:

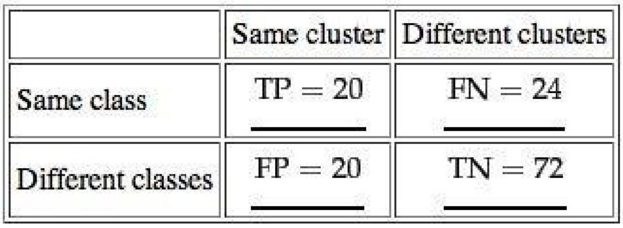


Names	Formula
Maximum or complete-linkage clustering	$\max\{d(a,b) : a \in A, b \in B\}.$
Minimum or single-linkage clustering	$\min\{d(a,b) : a \in A, b \in B\}.$
Unweighted average linkage clustering (or UPGMA)	$\frac{1}{\lvert A\rvert \cdot \lvert B\rvert} \sum_{a\in A}\sum_{b\in B} d(a,b).$

Strategies for hierarchical clustering generally fall into two types:

- **Agglomerative:** This is a "bottom-up" approach: each observation starts in its own cluster,and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive:** This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

F-Measure is a measure that combines precision and recall and is the harmonic mean of precision and recall.
With the samples of data set the confusion matrix is :



	Same cluster	Different clusters
Same class	TP = 20	FN = 24
Different classes	FP = 20	TN = 72

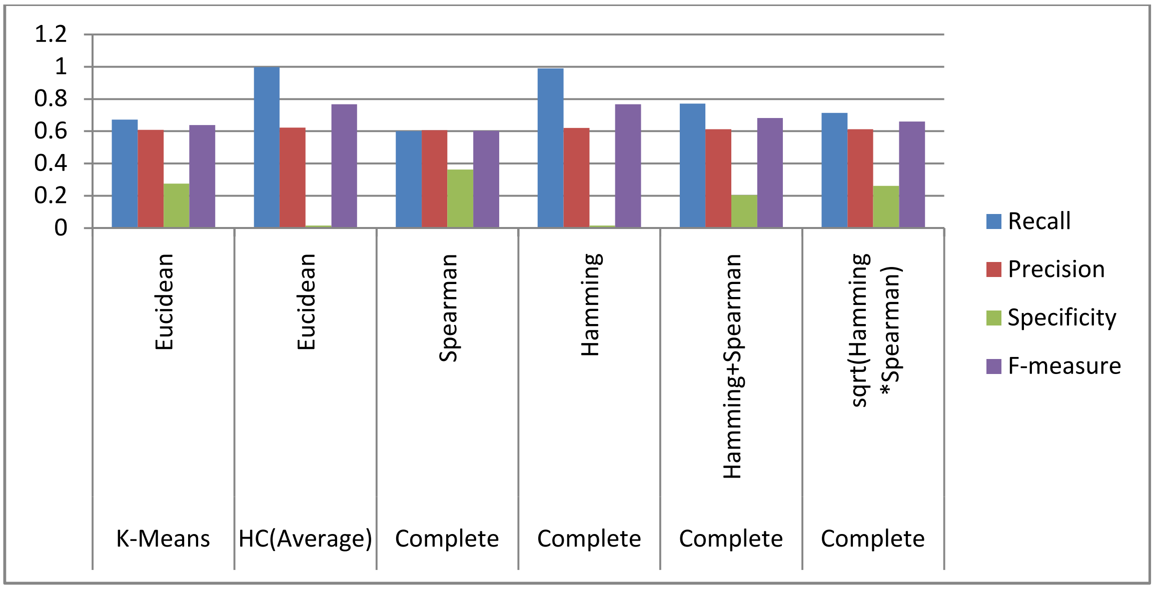
3.3 PD_Speech_Features dataset.:

For this dataset, we have used
Hamming + Spearman, and (ii) sqrt (Hamming * Spearman) Composite metrics
The hierarchical clustering technique with the complete linkage approach and (Spearman and hamming) metric performs well in terms of specificity. For this dataset, the hierarchical clustering algorithm performs well in terms of recall, accuracy, and f-measure. However, the gained Specificity of the approach is not very good

Method	Metric	Recall	Precision	Specificity	F-measure
K-Means	Eucidean	0.6716	0.6094	0.2765	0.639
HC(Average)	Eucidean	0.9968	0.6235	0.0156	0.7672
Complete	Spearman	0.6000	0.6062	0.3627	0.6031
Complete	Hamming	0.9894	0.6205	0.0156	0.7672

Complete	Hamming+Spearman	0.7706	0.6131	0.2047	0.6829
Complete	sqrt(Hamming *Spearman)	0.7144	0.6127	0.2613	0.6597

Table 2: Composition of different clustering algorithms in terms of four supervised cluster validity indexes for PD_Speech_Features dataset



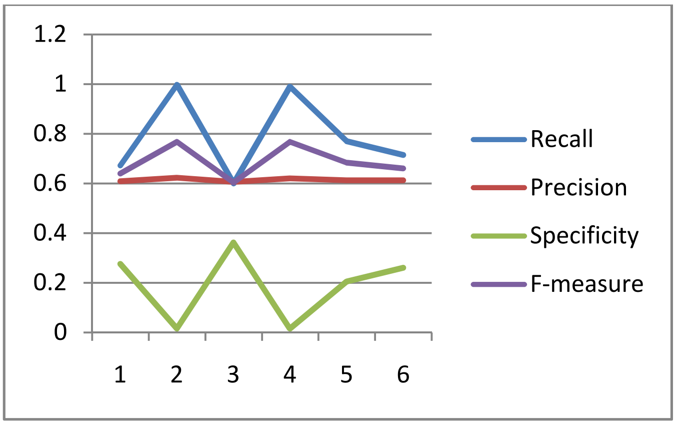


Figure 1: PD_Speech_Features dataset Visualization Analysis

3.4 PD_Voice_Features Data Set:

For this dataset, we have used

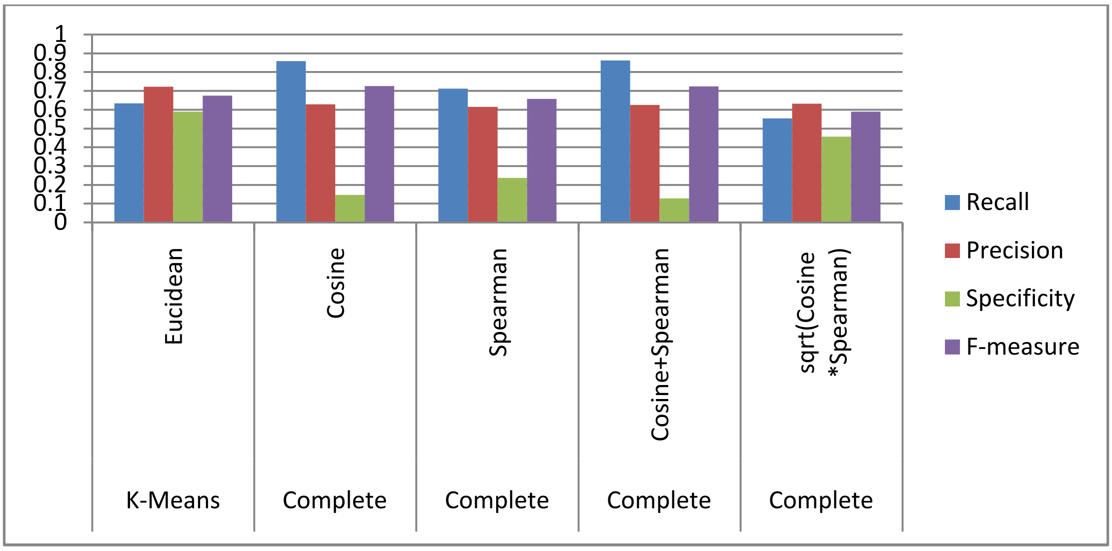
(i) Cosine + Spearman and (ii) sqrt (Cosine *Spearman) composite metrics

The hierarchical clustering technique that use the whole linkage approach and (cosine and spearman) metric performs well in terms of specificity. For this dataset, the hierarchical clustering algorithm performs well in terms of recall, accuracy, and fmeasure. However, the gained Specificity of the approach is not very good.

Method	**Metric**	**Recall**	**Precision**	**Specificity**	**F-measure**
K-Means	Eucidean	0.6329	0.7216	0.5896	0.6743
Complete	Cosine	0.859	0.6282	0.1457	0.7257
Complete	Spearman	0.7119	0.6148	0.2364	0.6573
Complete	Cosine+Spearman	0.8625	0.6244	0.128	0.7244
Complete	sqrt(Cosine *Spearman)	0.5538	0.6314	0.4566	0.59

DOI: 10.48175/IJARSCT-12473

Table 3: Composition of different clustering algorithms in terms of four supervised clustervalidity indexes for PD_Voice_Features dataset.



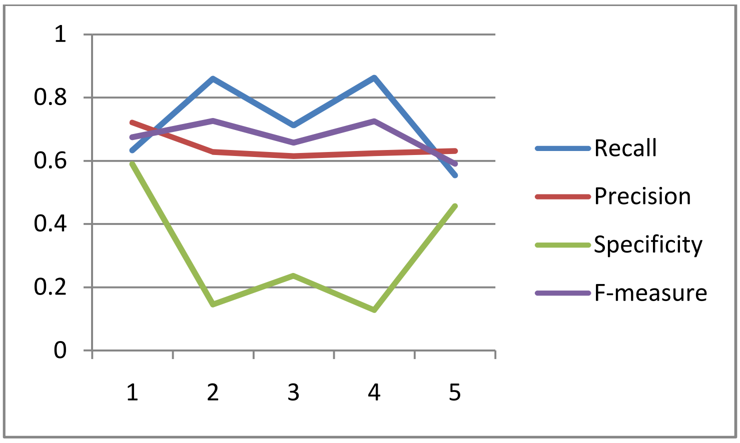


Figure 2: PD_Voice_Features Data Set Visualization Analysis

3.5 Parkinson_Multiple_Sound Data set:

For the this dataset, we have used

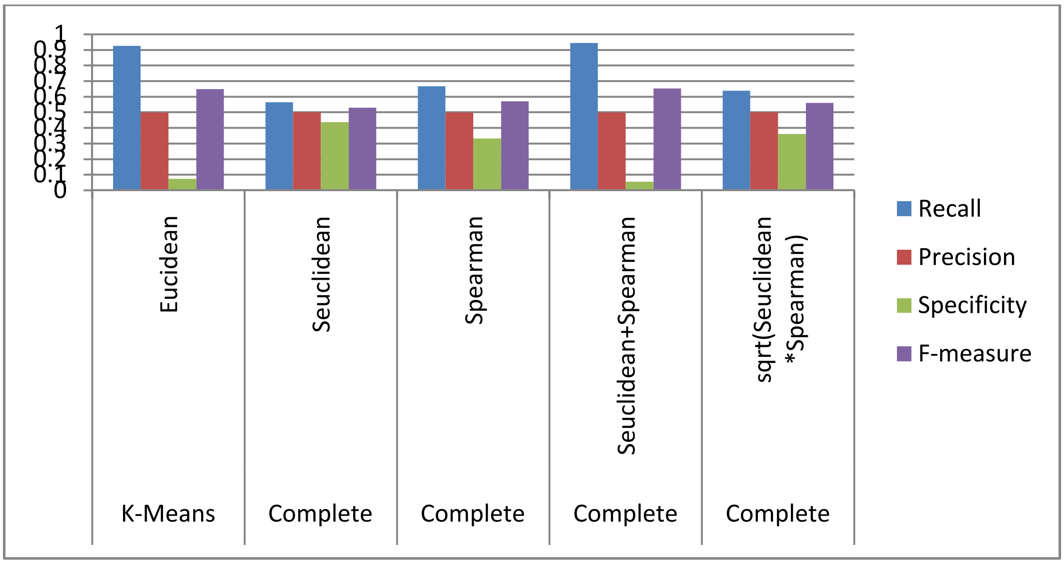
(i) Seuclidean + Spearmanand (ii) sqrt (Seuclidean * Spearman) composite metrics

The (seuclidean and spearman) metric and the hierarchical clustering technique with complete linkage both perform well in terms of specificity. For this dataset, the hierarchical clustering algorithm performs well in terms of recall, accuracy, and f-measure. However, the gained Specificity of the approach is not very good.

Method	**Metric**	**Recall**	**Precision**	**Specificity**	**F-measure**
K-Means	Eucidean	0.926	0.4995	0.074	0.649
Complete	Seuclidean	0.5652	0.5003	0.4367	0.5308
Complete	Spearman	0.667	0.4993	0.3326	0.5711
Complete	Seuclidean+Spearman	0.9439	0.4995	0.056	0.6533
Complete	sqrt(Seuclidean *Spearman)	0.6392	0.5001	0.3622	0.5611

Table 4: Composition of different clustering algorithms in terms of four supervised cluster validity indexes for Parkinson_Multiple_Sound dataset.

DOI: 10.48175/IJARSCT-12473



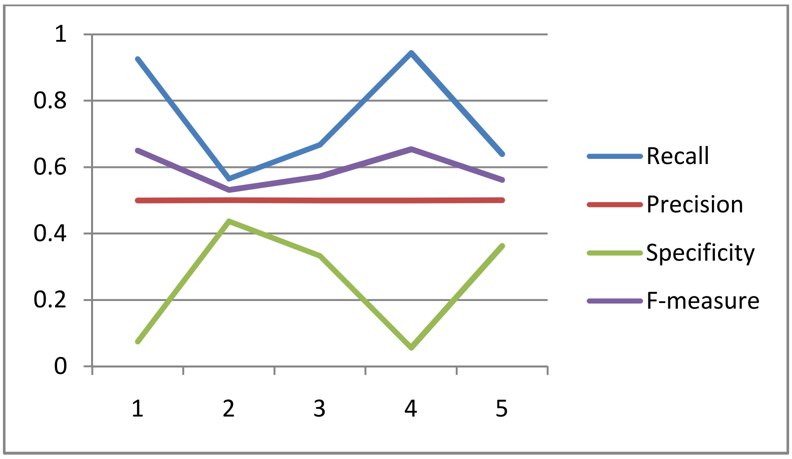


Figure 3: Parkinson_Multiple_Sound Data set Visualization Analysis

III. CONCLUSION

In this study, we compare several clustering methods using three datasets pertaining to Parkinson's illness. In order to compare two different clustering methods, k-means and hierarchical, a variety of distance measurements are used. For each of these three datasets, we assessed each algorithm's performance in terms of recall, accuracy, specificity, and F-measure. The experimental results demonstrate that the hierarchical clustering algorithm employing the full linkage technique offers effective results when compared to other algorithms

REFERENCES

[1]. “Universityofberkley,”http://www.sims.berkeley.edu/research/projects/how- much-info2003/.

[2]. N. Ye, The Hand Book of Data Mining. Mahwah, New Jersey: Lawrence ErlbaumAssociates, 2003.

[3]. J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. Morgan Kaufmann, Elsevier, 2006.

[4]. R. Dubes and A. Jain, Algorithms for Clustering Data. Prentice Hall, 1988.

[5]. R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd ed. New York: JohnWiley & Sons, 2001.

[6]. S. S. Stevens, “On the theory of the scales of measurement science,” Science, vol.103, no. , pp. 677–680, June 1946.

[7]. E. Sungur, “Overview of multivariate statistical data analysis,” http://www.mrs.umn.edu/ sungurea/multivariatestatistics/overview.html.

DOI: 10.48175/IJARSCT-12473

[8]. T. Joachims, “Transductive inference for text classification using support vector machines,” in Proc. 16th International Conference on Machine Learning. MorganKaufmann, 1999, pp. 200–209.

[9]. A. Zhang, Advanced analysis of gene expression microarray data: Science, Engineering, and Biology Informatics, Singapore, 596224:World Scientific Publishing Co. Pte .Ltd., 2006, vol. 3.

[10]. S. Bandyopadhyay, U. Maulik, and J.T.L.Wang, Analysis of biological data: as oft comp uting approach: Science, Engineering, and Biology Informatics. Singapore, 596224:World Scientific Publishing Co. Pte. Ltd.,2007, vol. 3.

[11]. I. S. Kohane, A. T. Kho, and A. J. Butte, Microarrays for an Integrative Genomics. Massachusetts, London, England: MIT Press Cambridge, 2003.

[12]. J. T. L. Wang, M. J. Zaki, H. T. T. Toivonen, and D. Shasha, Data Mining in Bioinformatics. London: Springer-Verlag, 2005.

[13]. J. MacQueen. Some methods for classification and analysis of multivariate observations. Proceeding of the 5th Symposium on Mathematical Statistics and Probability. 1: 281–297. 1967.

[14]. J. F. Curry, R. J. Thompson, Jr. Patterns of behavioural disturbance in developmentally disabled and psychiatrically referred children:A cluster analytic approach. Journal of Pediatric Psychology. 10:151-167. 1985.

[15]. S. C. Johnson. Hierarchical clustering schemes. Psychometrika. 32: 241– 254. 1967.

[16]. P. Sneath, R. Sokal. Numerical taxonomy. W. H. Freeman, San Francisco. 1973.

[17]. https://archive.ics.uci.edu/ml/datasets.php

DOI: 10.48175/IJARSCT-12473