Approaches for finding Optimal Number of Clusters using K-Means and Agglomerative Hierarchical Clustering Techniques

Punyaban Patel
Department of Computer Science and Engineering,
CMR Technical Campus,
Kandlakoya, Hyderabad, India
E.Mail:
punyaban@gmail.com

Borra Sivaiah
Department of Computer Science and Engineering,
CMR College of Engineering & Technology, Kandlakoya,
Hyderabad, India,
E.Mail: sivabetld@gmail.com

Riyam Patel
Department of Computer Science and Engineering (AI & ML),
SRM Institute of Science & Technology, SRM University,
Kattankulathur, Chennai, India
E.Mail: riyampatel2001@gmail.com

2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP) | 978-1-6654-7258-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICCSP53532.2022.9862439

Abstract- **Machine learning and pattern recognition both benefit greatly from the study of clustering techniques.** ***Choosing the right number of clusters in cluster analysis is quite tough.*** **Quality of the cluster depends on Optimal number clusters. In this paper, we used three methods such as elbow method, gap statistics method and Silhouette method to find the optimal number of clusters. Furthermore, the agglomerative hierarchical clustering (AHC) algorithm and K-means methods has used for calculating the appropriate number of quality clusters for the data set with the optimal k-value. Both algorithms are evaluated on a given data set using the validation measures such as connectivity, Dunn, and Silhouette to find the optimal number of clusters.**

Keywords- K-means, Agglomerative Hierarchical Clustering (AHC), Connectivity, Dunn, Silhouette, Gap Statistic.

I. INTRODUCTION

Clustering is a hot issue in several fields of study, including image processing, pattern recognition, and machine learning. Many clustering techniques, such as partitioned and hierarchical, have been developed by researchers. Partitioned algorithms convert the entering data set into clusters by dividing it into partitions. On the other hand, hierarchical algorithms create a cluster hierarchy, which is a nested partition set.

Divisive clustering begins with a single cluster containing all of the points and continually splits the clusters, whereas agglomerative clustering begins with a single cluster containing all of the points and builds a hierarchy by merging clusters one by one [1]. Despite the fact that hierarchical techniques can provide more clustering results than partitioned algorithms, choosing the optimum partition from a large number of clustering results is a fascinating topic.

978-1-6654-7258-6/22/$31.00 ©2022 IEEE

The methods for determining the ideal number of clusters are divided into two categories: (i) Direct approaches and (ii) Statistical testing methods.

(i) Direct methods: It consist of maximising a criterion, such as the sums of squares inside clusters or the average silhouette. Elbow and silhouette techniques are the names of the comparable approaches.

(ii) Statistical testing procedures: These approaches include comparing evidence to the null hypothesis. The gap statistic is one example.

In this paper, we have described all the above three methods and compared among them.

This paper is organised as where the Section I: Introduction, Section II: Related works, Section III: Matrices used for performance Measures, Section IV: Proposed Methodology, Section V: Result and Analysis and Section VI: concluded the paper.

II. RELATED WORKS

Many research articles are available in the web related to finding the optimal cluster, few of them has been explained briefly as below.

Hierarchical clustering's optimal partition was first proposed in [1] by combining the expanded partition set technique with a novel measure of cluster validity known as context-independent optimality and partiality. By specifying the compactness and separability of clustering results, the authors in [2] developed a novel measure of clustering validity to discover the optimal number of clusters in hierarchical clustering. We can't verify the experiment's findings because the similarity measure method used to create the similarity matrix was not supplied in the paper.

By first scanning the data set and creating hierarchical partitions of the data set agglomeratively in [3], the authors have proposed a hierarchical method that incrementally constructs a curve of clustering quality for various partitions and finally estimates the optimal number of clusters using partitions corresponding to the extremum of this curve. [3]: Nonconvex data might be measured using the proposed method, they said. K-means and Fuzzy C-means Clustering methods, on the other hand, did not have great centroids.

To evaluate hierarchical clustering techniques, the authors in [4] used the ordered weighted averaging (OWA) operator in conjunction with hierarchical clustering to measure distance between groups, and then calculated the root-mean square deviation called the standard deviation, and R-squared validity indices. To their dismay, they didn't devote enough time to fine-tuning the OWA operator to get the greatest clustering results.

According to [6], the authors presented a new clustering method that separates the sequence list dataset into k clusters and then uses the sum of distances between each pattern and its centre as the goal function to simulate the results of an annealing approach. It's a decent algorithm when compared to others. Using a combination of the density estimation segmentation approach and the initial value limitation large-scale data group segmentation algorithm, [7] the authors proposed an initialization strategy for the cluster centre. This new algorithm is faster and more accurate than previous methods.

The authors [8] provided an automated method for estimating the optimal number of clusters when using a classification method based on k-means. Finding immutable clusters during several testing is the basis for selecting the best one. [9] The authors of this paper [9] analysed numerous methods for establishing the appropriate number of groups, and the best clustering algorithms. Researcher in [10] evaluated the ability of four validation strategies in terms three clustering methodologies, as well as the behaviour and applied the greedy method to the K-means clustering and specific preconditions of this research. As a last step, the optimal validation and clustering techniques are explored.
as the Centroid Auto-Fused Hierarchical Fuzzy c-

The authors claim in [19] that there is a method known if the data set has clusters that are relatively small and clearly defined from one another, it is anticipated that the diameter of the clusters will be smaller, while the spacing between the clusters would be on the larger. Therefore, the Dunn index ought to be increased to its maximum. Clusters that are less compact or well-separated are indicated by a Dunn index that is lower, whereas clusters that are compact and well-separated are indicated by an Index that is higher.

means method (CAF-HFCM) whose optimization technique can automatically agglomerate to form a cluster hierarchy, and more notably, yields an optimal number of clusters without having to resort to any validity index.

An effective cluster validity index (CVI) has been proposed by the authors [20]. This index is called the validity clustering index based on finding the mean of clustered data (VCIM), and it incorporates the characteristics of the score function index and the mean in order to locate new cluster centroid positions. The results indicate that the VCIM has worked more successfully than the other CVI.

The author states in [21] that a hierarchical clustering technique has been proposed for the purpose of clustering multiple nominal data streams. This technique involves the splitting and merging of clusters within a hierarchical structure, with the decision to split or merge being relying on the entropy measure of the data streams.

In [22], the authors proposed a clustering index that discovers the optimum number of clusters based on entropy measure from eigenvalue analysis of consumption time series correlation matrix. It uses genetic algorithms to choose features for clustering.

III. PERFORMANCE MEASURES

The following metrics has been used for performance measures.

(i) Connectivity

Connectivity [16] refers to the degree to which objects are grouped together in the same cluster as their neighbours who are physically located closest to them in the data space. The connectivity has a value that ranges from 0 to ∞ (infinity), and it should be reduced as much as possible.

The two commonly used indices for assessing the goodness of clustering: The Dunn index and the silhouette width

(ii) Dunn Index

Equation (1), which is utilised as a validation measure in internal clustering, can be used to determine the value of the Dunn index denoted by "D" [15].

$$\mathrm{D} = \frac{\text{min. separation}}{\text{max. diameter}} \tag{1}$$

(iii) Silhouette Index

The silhouette index [14] analysis examines how well an observation is grouped and it determines the average distance between clusters.
The silhouette width (S_i) of the observation i can be defined by the formula;

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{2}$$

Where , $b_i = \min_C d(i, C)$

IV. PROPOSED METHOD

The proposed method is shown in Fig. 1

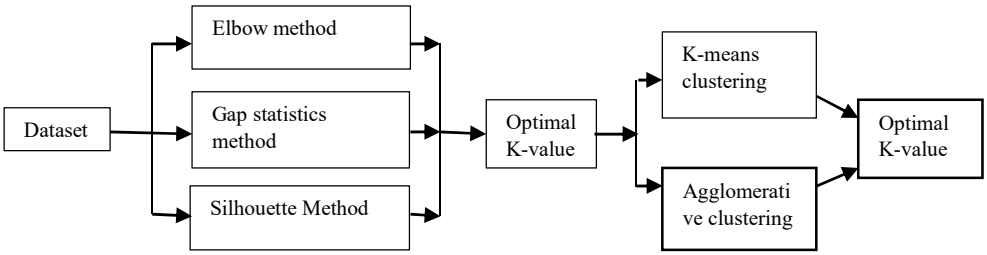

Fig. 1: Proposed method for finding optimal number of clusters

The following approaches has been used for finding the optimal value of K.

(i) Elbow Method

The elbow method [17] is explained in Fig. 2, in which the sum of squares at each number of clusters is calculated and graphed, and there is a variation of slope from steep to shallow (an elbow) to discover the best number of clusters.

(ii) Gap Statistic

The gap statistic [11] equates the total intra-cluster variance for several choices of k to their predicted values under a null reference distribution. The value that maximises the gap statistic can be used to estimate the best clusters, as shown in Fig. 3. The gap stats plot shows the statistics by number of clusters (k), with vertical segments representing standard errors and a vertical dashed blue line representing the ideal value of k. The ideal number of clusters in the data, according to this study, is $k = 2$

(iii) Silhouette Method

The average silhouette technique [13, 14] calculates the average silhouette of observations for a range of different values for the parameter k. The Fig. 4 shows the best number of clusters k, which maximises the average silhouette throughout a series of feasible values for k and also proposes an optimal of two clusters.

There are many clustering algorithms are available, but we used the K-means clattering algorithm and
agglomerative hierarchical clustering algorithm in this paper. The methods are presented below:

(i) Agglomerative hierarchical clustering [18]

Agglomerative processes begin with n singleton clusters and then merge clusters one by one to form a single cluster. Consider the data set is $\{ x_1, x_2 , x_3 \ \ldots \ x_n \}$and G_i is the i^{th} cluster in the k^{th} merging process. Single linkage, average linkage, and complete linkage are the three basic distance measurements used by the AHC algorithm [5]. The AHC technique is also known as the nearest neighbour clustering algorithm because the distance between clusters is calculated using a single linkage measure, A bottom-up approach is used in an agglomerative hierarchical clustering algorithm. It usually begins by allowing each item to establish its own cluster, then combines clusters into larger clusters repeatedly until all objects are in a single cluster or specified end necessities conditions are met.

The single cluster serves as the primary root node in the hierarchy. During the merging stage, based on some similarity metric, it finds the two clusters that are the most closely related to one another, and then it combines those clusters into a single larger cluster. Since two clusters are merged during each iteration of an agglomerative approach, this method requires the greatest number of iterations. Each cluster, however, must include at least one item.

(ii) K-means clustering Algorithm [12]

The partitioning procedure known as k-means, in which the centre of each cluster is determined by the average value of the items contained in that cluster.

Input: D: a data set containing n objects, k: the number of clusters,
Output: A set of k clusters.

Method:
Step 1: arbitrarily choose k objects from D as the initial cluster centers;
Step 2: repeat
Step 3: (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
Step 4: update the cluster means, that is, calculate the mean value of the objects for each cluster;
Step 5: until no change.

V. EXPERIMENTAL RESULTS AND ANALYSIS

We used mammal data set with 25 objects and we implemented k-means and agglomerative clustering as an empirical study. We used elbow method, gap static method, and the Silhouette Method to find the optimal number of clusters. One of the

challenge in clustering is finding the optimal number of clusters. Clustering quality depends on the optimal number of clusters. The pearson correlation among five attributes water, protein, fat, lactose, and ash is shown in Fig. 5.

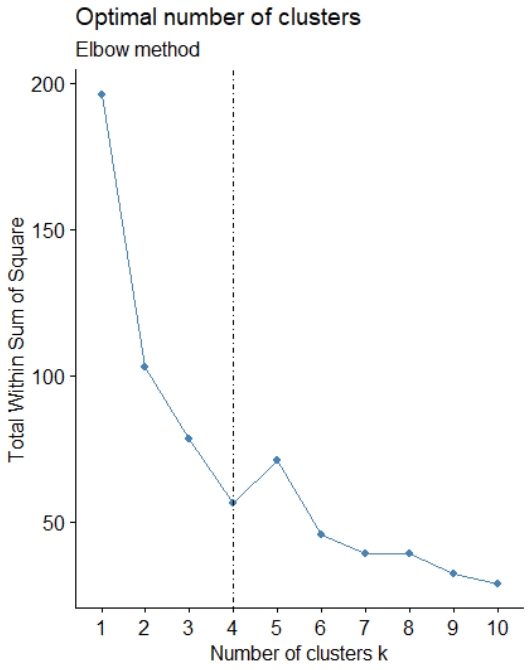

Fig. 2: Elbow method to find Optimal k value.

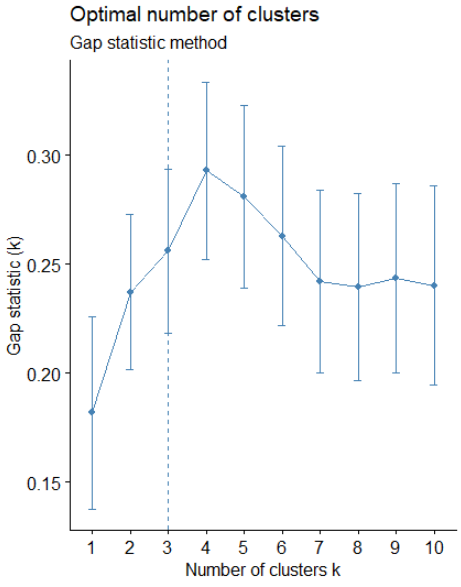

Fig. 3: Plot for Gap statistic method

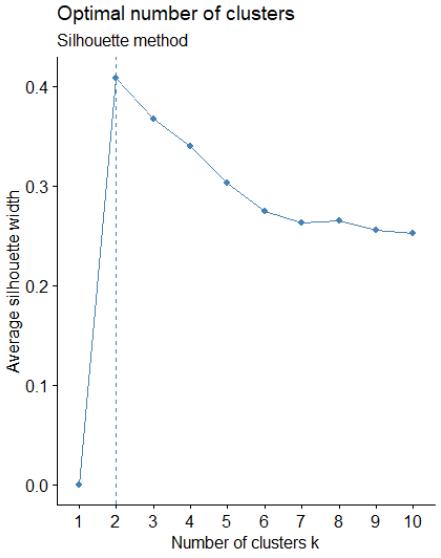

Fig. 4: Silhouette method

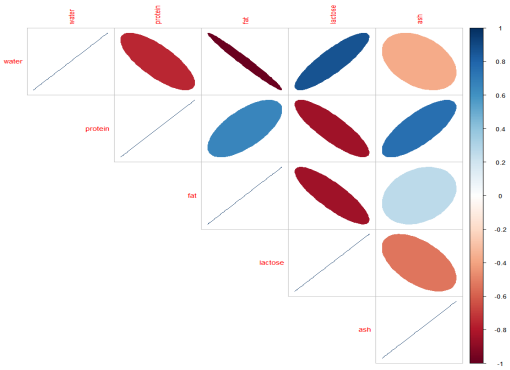

Fig. 5: Correlation relationship among five attributes water, protein, fat, lactose, and ash.

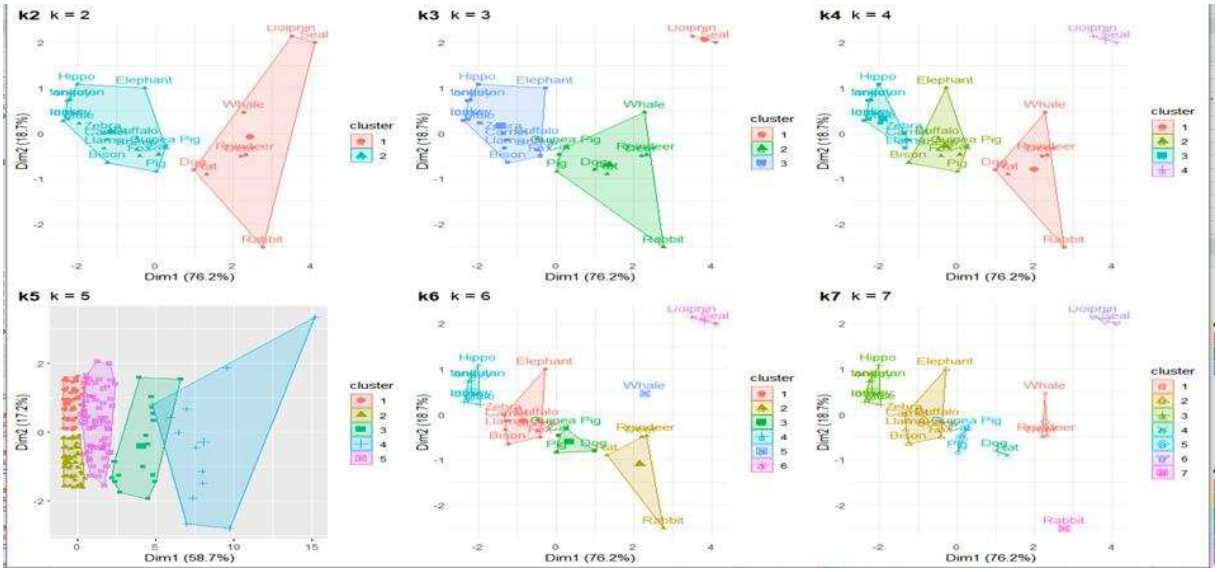

Fig. 6: K-Means algorithm with k=2 to 7 clusters.

TABLE I: K-MEANS ALGORITHM AND HIERARCHICAL CLUSTERING ALGORITHM WITH VALIDATION MEASURES

Number of Clusters (K)	Hierarchical clustering			k-Means algorithm		
	Connectivity	*Dunn*	*Silhouette*	*Connectivity*	*Dunn*	*Silhouette*
2	4.1829	0.3595	0.5098	7.2385	0.2070	0.5122
3	10.5746	0.3086	0.5091	10.5746	0.3086	0.5091
4	13.2579	0.3282	0.4592	15.8159	0.2884	0.4260
5	20.1579	0.2978	0.4077	20.1579	0.2978	0.4077
6	22.8508	0.3430	0.4077	22.8508	0.3430	0.4077
7	25.8258	0.3430	0.3664	25.8258	0.3430	0.3664
8	32.6270	0.4390	0.3484	33.5198	0.3861	0.3676
9	35.3032	0.4390	0.4060	35.3032	0.4390	0.4060
10	38.2905	0.5804	0.3801	38.2905	0.5804	0.3801
11	39.2405	0.5938	0.3749	39.2405	0.5938	0.3749
12	41.2405	0.5938	0.3322	41.2405	0.5938	0.3322
13	45.7742	0.8497	0.3646	45.7742	0.8497	0.3646
14	47.2742	0.8497	0.3418	47.2742	0.8497	0.3418
15	50.6075	0.5848	0.2650	51.8909	0.5866	0.2811
16	52.6075	0.5848	0.2317	53.8909	0.5866	0.2478
17	55.8575	0.4926	0.2166	57.1409	0.5725	0.2402
18	58.7242	0.9138	0.2469	58.7242	0.9138	0.2469
19	60.7242	0.9138	0.2213	60.7242	0.9138	0.2213
20	63.2242	0.8892	0.1659	63.2242	0.8892	0.1659
21	65.2242	0.9049	0.1207	65.2242	0.9049	0.1207
22	67.2242	0.9335	0.1050	67.2242	0.9335	0.1050
23	69.2242	1.0558	0.0832	69.2242	1.0558	0.0832
24	71.2242	2.1253	0.0691	71.2242	2.1253	0.0691

We applied k-means algorithm with $k = 2$ to 7 values and clusters are shown in Figure 6. Finding the optimal number is very difficult from all possible clusters and it requires more time to find optimal number of clusters.

The k-means clustering and hierarchical clustering algorithms with connectivity, Dunn index, and Silhouette index are computed and shown in in TABLE I and the optimal scores are computed and shown in TABLE II.

TABLE II: OPTIMAL SCORES FOR K-MEANS ALGORITHM AND HIERARCHICAL

Measure	Clustering Method	Score	Optimized number of Clusters
++Connectivity	hierarchical	4.1829	2
Dunn	hierarchical	2.1253	24
Silhouette	k-means	0.5122	2

VI. CONCLUSION AND FUTURE SCOPE

Business intelligence, picture pattern recognition, web search, biology, and security are just a few of the applications that employ cluster analysis. Clustering is a technique used in business intelligence to arrange a large number of customers into groups with substantial similarities in their features. This makes it easier t web search, biology, and security are just a few of o come up with corporate ideas to improve customer relationship management. We utilised three approaches to identify the ideal number of clusters in this paper, such as Elbow Method, Gap static method, and the Silhouette method, and we computed Connectivity measure, Dunn index, and Silhouette index to determine the optimal number of clusters. As a result, k-means and agglomerative clustering algorithms produce the same number of ideal clusters in both circumstances. In the future, we will employ a variety of indices to determine the appropriate number of clusters, as well as a variety of clustering methods.

REFERENCES

[1]. Gurrutxaga et al., "SEP/COP: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index,"

Pattern Recognition., vol. 43, no. 10, pp. 3364–3373, Oct. 2010.

[2] X.-Q. Hu, R.-N. Ma, and B.-J. Zhong, "Study on validity of hierarchical clustering," J. Shandong Univ., vol. 40, no. 5, pp. 146–149, Oct. 2010.

[3] L. Chen, Q. Jiang, and S. Wang, "A hierarchical method for determining the number of clusters," J. Softw., vol. 19, no. 1, pp. 62–72, Jan. 2008.

[4] E. Nasıbov and C. Kandemır-Cavas, "OWA-based linkage method in hierarchical clustering: Application on phylogenetic trees," Expert Syst. Appl., vol. 38, no. 10, pp. 12684–12690, Sep. 2011.

[5] C.-R. Lin and M.-S. Chen, "Combining partitional and hierarchical algorithms for robust and efficient data clustering with cohesion self-merging," IEEE Trans. Knowl. Data Eng., vol. 17, no. 2, pp. 145–159, Feb. 2005.

[6] Jinxin Dong, and Minyong, "Qi K-means Optimization Algorithm for Solving Clustering Problem", IEEE Computer Society, pp.53-55, 2009.

[7] Hui Ai and Wei Li et. al. ," K-means initial clustering center optimal algorithm based on estimating density and refining initial", 6th International Conference on New Trends in Information Science, Service Science and Data Mining (ISSDM2012), IEEE Explore digital library, pp. 603 – 606, Oct. 2012.

[8] Oussama Chabih et.al., "New approach to determine the optimal number of clusters K in unsupervised classification ", 6th IEEE Congress on Information Science and Technology (CiSt), 2020.

[9] Pugazhenthi A and Lakshmi Sutha Kumar, "Selection of Optimal Number of Clusters and Centroids for K-means and Fuzzy C-means Clustering: A Review", 5th IEEE International Conference on Computing, Communication and Security (ICCCS), pp. 1 – 4, 2020.

[10] Rajdeep Baruri et. al.,"An Empirical Evaluation of k-Means Clustering Technique and Comparison", IEEE International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), pp. 470 – 475, India, 2019.

[11] R. Tibshirani, G. Walther and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic", J. R. Statist. Soc. B-63, Part 2, pp. 411–423, 2001.

[12] Yuan C and Yang H, "Research on K-Value Selection Method of K-Means Clustering Algorithm", MDPI, 18 June 2019

[13] N. Kaoungku, et. al," The silhouette width criterion for clustering and association mining to select image features", International Journal of Machine Learning and Computing, vol. 8, pp. 69–73, 2018.

[14] Wang X and Xu Y, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index", IOP Conf. Ser., 2019.

[15] Oscar Miguel Rivera-Borroto et. al., "Dunn's index for cluster tendency assessment of pharmacological data sets", Canadian Journal of Physiology and Pharmacology, 90(4):425-33, 2012.

[16] Er. Arpit Gupta et. al.," Research paper on cluster techniques of data variations", International Journal of Advance Technology & Engineering Research (IJATER), Vol. 1, pp.39-47, Issue 1, November 2011.

[17] Hestry Humaira and Rasyidah,"Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm", IEEE Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA), 24-25th January, 2018.

[18] Smarika Sakshi et. al., "Agglomerative hierarchical clustering technique for partitioning patent dataset", IEEE 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2-4 Sept, 2015.

[19] Yunxia Lin and Songcan Chen," A Centroid Auto-Fused Hierarchical Fuzzy c-Means Clustering, IEEE Transactions on Fuzzy Systems, Volume: 29, Issue: 7, 2021.

[20] Ahmed KhaldoonAbdalameer et. al., "A new validity clustering index-based on finding new centroid positions using the mean of clustered data to determine the optimum number of clusters", Expert Systems with Applications, Elsevier, Volume 191, 1st April, 2022.

[21] Jerry W. Sangma et. al., "Hierarchical clustering for multiple nominal data streams with evolving behaviour Complex & Intelligent Systems, Springer Link, volume 8, Pp.1737–1761, 7 January, 2022.

[22] Nameer Al Khafaf, Mahdi Jalili and Peter Sokolowski, "A Novel Clustering Index to Find Optimal Clusters Size with Application to Segmentation of Energy Consumers", IEEE Transactions on Industrial Informatics, Volume:17, Issue:1, 2022.