Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 6s (2024)

Fuzzy Membership Partition Based Effective Hierarchical Agglomerative Flat Clustering Method for High Dimensional Data

[1]S.Sivasankari, [2]Dr.S.Sukumaran

[1]Ph.D Research Scholar, Erode Arts and Science College, Erode, Tamilnadu, India

[,2]Associate Professor, Erode Arts and Science College, Erode, Tamilnadu, India

[1]sivamirtha@gmail.com

Article History:

Received: *20-06-2024*

Revised: *20-07-2024*

Accepted: *08-08-2024*

Abstract:

Introduction: Hierarchical clustering is an unsupervised powerful method for empirical knowledge interpretation from data. It has a fundamental role in understanding the complex pattern in huge datasets. It creates a hierarchical representation of data by forming clusters in two ways namely Agglomerative (Bottom-up) and Divisive (Top-Down). The main advantage is that it does not need to fix number of clusters.

Objectives: To handle the issues such as, the pertinence for enormous data is minimal as the computational complexity is high in using Hierarchical clustering, complication of fixing Threshold value in Dendrogram height while combining Flat clustering, and non existence of mathematical objective function to assess the Hierarchical clustering.

Methods: On focusing on these challenges, this work proposes (a) a liner split of data in order to reduce the computational complexity in Hierarchical Agglomerative clustering. (b) Fuzzy Partition matrix is created to enhance the cluster generation in Hierarchical clustering. (c) this work applies an objective function in Flat Clustering to ease the process of fixing threshold.

Implementation, Results: This work is implemented in Rapidminer tool. The Sum of Squares, Cluster Density and Processing Time is minimized in th e proposed work.

Conclusions: The proposed method handles enormous data effectively using linear split with sequential execution, the proposed usage of a sum of squares fixes a optimum threshold value in dendrogram height while transforming the dendrogram to flat clusters, the proposed method improves the existing Hierarchical clustering effectively.

Keywords: Cluster Density, Fuzzy Partition, Flat Clustering, Hierarchical Clustering, Sum of Squares.

1. Introduction

Clustering is an unsupervised process that split data into well defined groups that helps to know how a dataset is structured that oppose to classification that assigns labels to each data objects sometimes enhanced with [1] feature selection. The clustering method varies such as centroid based, Density based, Hierarchical based, [2][3]and Probability based algorithms. This paper focuses on Agglomerative Hierarchical Clustering method with Complete linkage criteria mainly on Transaction data. The classical hierarchical clustering algorithm (HAC) frames a hierarchical structure without the user parameter ‘number of clusters’. It sorts the data in two ways such as Agglomerative (top down) or Divisive (bottom-up).

Hierarchical Agglomerative Clustering

Agglomerative clustering initially the whole samples are clustered individually and pairs are merged as one moves up the hierarchical structure as opposed to this in Divisive all samples are in

single cluster initially divided into two smaller clusters in each successive calculation. The Linkage Criteria determines the distance between sets of observations among the clusters. [4]Below are the commonly used linkage criteria for the two samples (a and b) from two different clusters (A, B).

Single Linkage - The nearest interspaces between a pair of instance from two dissimilar groupings.

$$\mathrm{dist}_{\min}(C_i, C_j) = \min_{a \in A,\ b \in B} d(a, b) \quad (1)$$

Complete Linkage - The farthest interspaces between a pair of instance in two dissimilar groupings.

$$\mathrm{dist}_{\max}(C_i, C_j) = \max_{a \in A,\ b \in B} d(a, b) \quad (2)$$

Average Link - The average interspaces between pairs of instances from two dissimilar groupings.

$$\mathrm{dist}_{\mathrm{avg}}(C_i, C_j) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \quad (3)$$

Centroid Linkage- The mean (μ) distance between the centroids (μ_A , μ_B)of two clusters

$$\mathrm{dist}_{\mathrm{centroid}} = ||\mu_A - \mu_B|| \quad (4)$$

Ward Linkage – It measures the minimum increase of Sum of Squares.

$$\frac{|A|.|B|}{|A \cup B|} ||\mu_A - \mu_B||^2 = \sum_{x \epsilon A \cup B} ||x - \mu_{A \cup B}||^2 - \sum_{x \epsilon A} ||x - \mu_A||^2 - \sum_{x \epsilon B} ||x - \mu_B||^2 \quad (5)$$

The distance function calculates the [5][6] similarity between the data points. The functions are categorized into numerical function, nominal, mixed and Bregman divergences. The numerical functions include Euclidean, Manhattan, Cosine, Inner Product similarity, Chebychev, Camberra etc. [7] The important nominal functions include Dice, Jaccard, Simple matching simialrity, RusselRao etc., and Bregman divergences include Generalized divergence, Logarithmic loss, Logistic loss, Mahalonobis distance etc. The complete linkage produce more spherical c lusters than the single criteria.In order to visualize the clusters Dendrogram is used that shows the hierarchical relationship between objects. [8] The main usage of dendrogram is to allocate objects to clusters. In the dendrogram, points with similar characteristics will be allotted in the same branch the rest in the separate branch. It contains clades that further break down into smaller ones. At the bottommost, resides the individual elements after they are merged based on linkage criteria using distance functions. [9] The bottommost node is known as leaf and two leaves in the same clade are more similar than other. The vertical axis in the dendrogram depicts the height of the branch and the objects in the groups are dissimilar with the height. Generally, hierarchical structure is decomposed to compact clusters with Flat clustering model. [10] It can be in the form of fixing number of clusters or defining a threshold on the dendrogram height.

The traditional Hierarchical clustering has the time complexity of $O(n^3)$ and it needs memory of $\Omega(n^2)$ thus it is too slow for even medium sized datasets and makes it practically unusable. The main advantage is that the model doesn't require prefixing of number of clusters. It builds a hierarchical structure of the pattern. The disadvantage is that it produces numerous clusters that lead to complicate analysis. However, the implementation of flat cluster model ease the visualization of compact clusters there is no mathematical objective function while prefixing number of clusters or

threshold value to cut the dendrogram in the model. The foremost drawback is, even for a medium sized dataset it takes high processing time so the computational complexity is high.

The rest of this paper is organised as follows, section 1 explains about Hierarchical clustering model, Existing works are discussed in section 2, section 3 gives details about the proposed methodology. In section 4, results are discussed and compared with appropriate tables and charts finally section 5 concludes the findings.

2. Literature Review

Dajun Si *et al* [11] addressed skewed concept in the hierarchical agglomerative for organizing substations . t-SNE algorithm expressed the data from multidimensional to two-dimensional space. Single Linkage Criteria is followed to cluster the data and the Gini coefficient is calculated with the hierarchical clustering. A threshold value is set and it is checked with the Gini coefficient of the clustered set. If the Gini value is less than the threshold then the method is adopted or else fairness index is calculated with the proposed equation. From the results with contour coefficient, the proposed method gives better results.

Mohamed S. Halawa *et al* [12] did a research on the type of metric that suits better for classifying different types of jobs in the large performance computing. This work applied partition and hierarchical techniques from the Galician computation center. In Prior to the application of such algorithms, PCA was applied in order to reduce the dimensionality of the data. The results concluded that the best results were obtained from hierarchical Bottom upmethod with the use of Silhouette and Davies– Bouldin.

Lifeng Yin *et al* [13] proposed an improved hierarchical with the theme of fusion and population reproduction. The initial stage of proposed PRI-MFC order the data in batches to prevent running out of memory usage. In the fusion stage, Jaccard fusion was used to use only the number of labels to complete the statistics which then reduced the computational complexity of the hierarchical clustering process.

Joelson Antonio dos Santos *et al* [14] put forth an efficient Hierarchical-Density namely MR-HDBSCAN* with the implementation of MapReduce. Different parallelization scheme based on recursive sampling approach was proposed to make the algorithm much faster to run. This was evaluated with execution time and ARI thus proved its effectiveness and scalability.

Praveen *et al* [15] compared linkage criteria namely Single, Complete, Average, and Average Weighted in Hierarchical clustering on five different datasets in UCI. As there is no mathematical objective function for Hierarchical clustering, running time is taken into account in this work. From the results, it was noted that complete linkage criteria takes high processing time. As per the running time, single linkage is found to be best than other linkage criteria.

William W. Tsoa *et al* [16] developed a method using agglomerative hierarchical for energy storage applications to handle model complexity and optimize the system with representative time periods. This method is applied on a power system database and showed that 15 clusters are chosen to be good with less cluster distance.

Karli Eka Setiawana *et al* [17] compared Fuzzy-C-Means, K-Means for grouping hospital data. Euclidean, Manhattan and Hamming distance metrics was used. For dimensionality lessening, PCA

was applied. From the results, it was noted that SSE was reduced with increasing number of clusters or both methods.

Joaquín Pérez-Ortega *et al* [18] combined Fuzzy C-Means and K-Means namely HOFCM for large datasets. The work started with applying K-Means, select the best solution and then Fuzzy C-Means was implemented. From the observed outcome, it was revealed that HOFCM lessen the time as well as the quality of the solution was 93.94%.

Seyyed Mohammad Razavi *et al* [19] addressed time and space complexity analysis on clustering data while using big data. This work used a bee colony algorithm and Map-Reduce architecture. The Adjusted Rand Score (ARI) parameter was used to assess the quality of clustering and found the proposed method gives high ARI value.

Ahmad Afif Supianto *et al* [20] employed (FCMPSO) method with the combination of Fuzzy-C-Means, PSO to merge students on the basis of their learning activity. This work utilized the silhoutte coefficient as an evaluation method. Two clusters were formed with high and low performance. FCMPSO improved the original FCM with the assessment of silhoutte coefficient.

3. Objectives

To lessen the computational complexity in using Hierarchical clustering, to fix a optimum Threshold value in Dendrogram height while combining Flat clustering, to analyze the performance by using mathematical objective function in Hierarchical clustering.

4. Methods

Proposed Methodology - Fuzzy Partition based Hierarchical Agglomerative Clustering FPHAC

This portion introduces FPHAC algorithm that includes preprocessing, Fuzzy partitioning, effective Hierarchical Flat Clustering.

The proposed method FPHAC starts with Replacing missing values, Nominal to Numerical data transformation, Data Summarization with pivot table. After Summarization, the total samples in the dataset is minimized and aggregated based on the transaction and analysis. Then, split data operator (value with 0.2) is applied to segment the data for minimal memory and time consumption in Hierarchical clustering. Fuzzy clustering is implemented for obtaining the Fuzzy partition matrix. This output partition matrix is transformed to regular attributes, added with the transformed dataset in order to get the initial probability detail of being in the cluster. Hierarchical Agglomerative clustering (HAC) with Complete Linkage criteria is implemented in the new dataset and it produces hierarchy of clusters and output is visualized in dendrogram. To ease the analysis with the hierarchy, this work applies Flat clustering with the distance threshold value. Based on the threshold value, the clusters are formed. To fix optimum threshold value this work proposes an objective function SOS (Sum of Squares). On the basis of the user fixed SOS value, the distance threshold(height of the dendrogram) is determined. Finally, the method is evaluated with intra cluster distance measure.

4.1 Data Summarization

The data in a larger table is reorganized into groups by calculating sums, averages, or other statistics functions for each group based on the analysis. The pivot table is defined by three new attributes such as group by attributes for row, column grouping attributes, and aggregation attributes. The function used for this purpose creates a unique row, unique column and assign every transactions to the appropriate row and column by aggregating the transaction. The most common functions used are count, average, product, maximum, minimum, mode, median, standard deviation, sum, and variance.

Example: In Table 1, the original table is converted to pivot table by aggregating the year with region and sales. Here the number of column and rows are minimized and made simple which in turn ease the analysis.

Table 1. Original, Pivot Table

Original table			
Year	Region	Sales	Country
2022	North	1800	Pune
2022	South	1500	Kerala
2023	South	2000	Kerala
2023	North	900	Pune

Pivot table		
Year	North	South
2022	1800	1500
2023	900	2000

4.2 Linear Sampling

The classic Hierarchical Agglomerative clustering (HAC) has a time complexity of $O(n^3)$ and needs $\Omega(n^2)$ memory, hence it takes more processing time even for small datasets. To handle this issue, the dataset is applied with split data operator that creates subsets (partition) of the data as per the user prescribed ratio (0.2) and then the clustering is implemented in serial execution. This work applied split with Linear sampling. Each partition is evaluated for performance individually there by reduces the time and memory complexity while executing.

The samples are derived by using the formula,

$$\frac{N}{(n_1 + n_2)} * p \tag{6}$$

Where, N is the total number of samples, n_1, n_2 are the partitions, p is the ratio of the particular partition. The decimal can be rounded off.

Example: Lets have the split operator with two partitions (0.7, 0.3), total number of samples be 20. The samples for each partitions are: Partition 1: (20/1)* 0.7 = 14, Partition 2: (20/1)* 0.3 = 6

4.3 Fuzzy Membership Partitioning

In Fuzzy C-means clustering, membership grades are given to each data points that indicate the degree of partition of each data belonging to the particular cluster. The points on the edge of a cluster have less membership than points in the center of cluster. The Fuzzy clustering partition a database, Let ($E=\{e_1, e_2, ..., e_n\}$) into a Fuzzy Clusters ($G=\{g_1, g_2,..., g_n\}$) and a partition matrix ($X=x_{ij}\in[0,1]$) in where *i* ranges the number of instances and *j* ranges the number of clusters. This matrix gives the information about the data e_i belongs to which cluster g_j. Manhattan distance is used for clustering.

x_{ij} is defined with the hyper parameter *m* that controls the fuzziness by,

$$X_{ij} = \frac{1}{\sum_{k=1}^{g}\left(\frac{\|e_i - g_j\|}{\|e_i - g_k\|}\right)^{\frac{2}{m-1}}} \tag{7}$$

$$\text{Manhattan Distance} = \sum_{i=1}^{n}|p_i - q_i| \tag{8}$$

Where, p_i, q_i are the two points in the dataset .

4.4 Complete Linkage Criteria in HAC, Flat Clustering

At beginning, each element is in its own cluster and then it is sequentially combined into larger one til all element sends in same cluster, it is also known as farthest neighbor clustering. In sequence steps, the two clusters disjoint by the shortest distance are combined. In complete-linkage as per eq. (2) the interspace between clusters is equivalent to the distance between the two objects s i.e one from each cluster which are farthest. The minimal length of these at each step combines the two clusters. The method starts by combining rows and with the old clusters are merged with new ones. This matrix contains all distances between the data elements. The clustering are numbered in sequence (m) starts from 0,1,....., to ($n-1$) and $V(k)$ is the level of the k^{th} clustering with n elements. The proximity between the clusters {(c), (s)} is denoted d [(c),(s)].

Working Principle

- Starts with the disorganized cluster with level V(0), sequence no. m=0.
- Retrieve the most unique pair in the recent clustering provided that pair (c), (s) is the minimum over all pairs of clusters.
- Increase the sequence no. by one. Merge (c) , (s) into a single one and fix the level of this clustering to V(m)=d{(c), (s).
- Revise the Proximity Matrix with the deletion of the rows and columns corresponding to clusters (c) and (s) and adds new row and column. The proximity between (c, s), and an old cluster (k) is defined as d[(c, s), (k)] = max {d[(k), (c)], d[(k), (s)]}.
- Check if the entire objects are in single cluster, then halt otherwise continue with the second step.

Let explain with the following example that has five points *(p1, p2, p3, p4, p5).* The Table 2(a), represents the matrix of T_1 distances where the $T_1(p1, p2)$ is the smallest of T_1 hence first join *p1*and *p2*.

Table 2 (a) First Pairwise Distance

	p1	p2	p3	p4	p5
p1	0	16	20	30	22
p2	16	0	29	33	20
p3	20	29	0	27	38
p4	30	33	27	0	42
P5	22	20	38	42	0

Table 2 (b) Second Pairwise Distance

	(p1, p2)	p3	p4	p5
(p1, p2)	0	29	33	22
p3	29	0	27	38
p4	33	27	0	42
p5	22	38	42	0

Table 2 (c) Third Pairwise Distance

	((p1 ,p2), p5)	p3	p4
((p1 p2), p5)	0	38	42
p3	38	0	27
p4	42	27	0

Table 2 (d) Fourth Pairwise Distance

	((p1 ,p2), p5)	(p3, p4)
((p1 ,p2), p5)	0	42
(c p3, p4	42	0

First branch length – In Table 2(a) first clustering, to ensure the equidistant of the points p1 and p2 from u(node), set $\delta(p1, u) = \delta(p2, u) = T_1(p1, p2)/2$ and the value is 16/2 = 8. Now reduce the size of the matrix by one row, column each as the clustering of p1 and p2 is done. The remaining points are calculated in first clustering for new matrix T_2.

First distance matrix - T_2 ((p1, p2),c)= max(T_1(p1, c), T_1(p2, c) = max(20, 29) =29; T_2 ((p1, p2),p4)= max(T_1(p1, p4), T_1(p2, p4) = max(30, 33) =33; T_2 ((p1, p2),p5)= max(T_1(p1, p5), T_1(p2, p5) = max(22, 20) =22

Second Branch length - In Table 2(b) second clustering, let v be the node to which the (p1, p2) with p4 are connected. The cumulative length $\delta(p1, v) = \delta(p2, v) = \delta(p5, v)$ = 22 divided by 2 and get 11, this value updated for deducing the missing branch length as 11.0- 8.0(the prior value) =3. Following this, the matrix revised for T_3 with the clustering of (p1, p2) and p5 as,

Second Distance matrix - T_3(((p1, p2),p5),p3) = max(T_2((p1, p2),p3), T_2(p5, p3) =max (29,38) =38; T_3(((p1, p2),p5),p4) = max(T_2((p1, p2),p4), T_2(p5, p4) =max (33,42) =42.

Third branch length - In Table 2(c) third clustering, from the revised matrix T_3, the T_3(p3, p4) =27 is the least hence join p3 and p4. The length by joining p3 and p4 to w is $\delta(p3, w) = \delta(p4, w)$ = 27/2 =13.5. The entry is updated as,

Third distance matrix - T_4((p3, p4), ((p1, p2), p5)) = max(T_3(p3, ((p1, p2), p5)), T_3(p4, ((p1, p2), p5)))= max (38, 42) =42

Fourth branch length – In Table 2(d) final one, the T_4 is updated by joining ((p1, p2), p5), (p3, p4) to the root node r. Then the length is δ((((p1, p2), p5), r) = δ((p3, p4), r) = 42/2 =21.0, After diminishing the two remaining branch are δ(v, r) = δ(((p1, p2), p5, r) - δ(p5, v) =21.0-11.0 =10; and δ(w, r) = δ((p3, p4), r) - δ(p3, w) =21.0-13.5 = 7.5

The diagram relates to Complete Linkage with the e above values are represented as in Fig.1

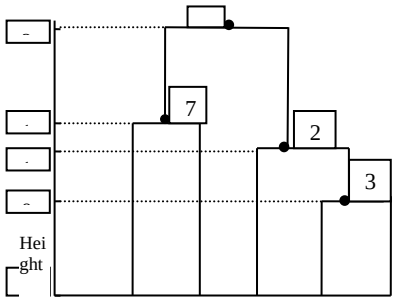

Figure 1. Complete Linkage Diagram

4.5 Flat Clustering, Sum of Squares (SOS)

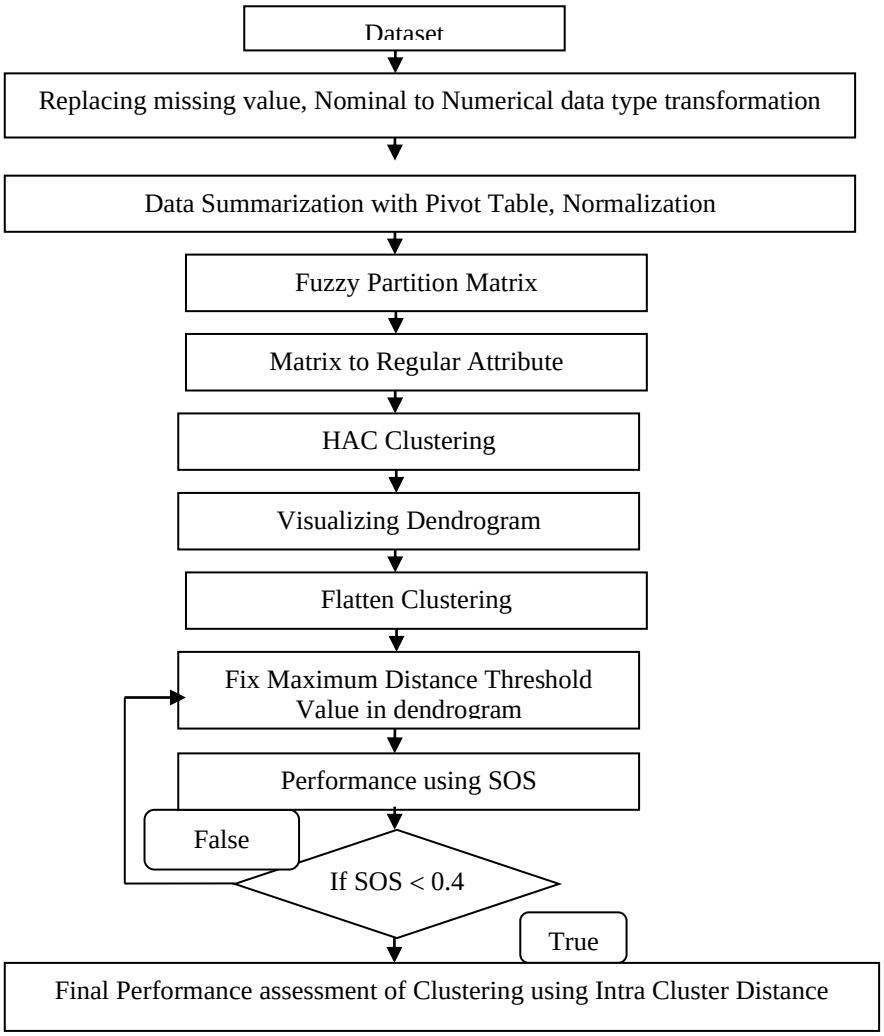

Figure 2. Flow of FPHAC

Fig. 2, shows the flow of work that starts with preprocessing, Data Summarzation, Fuzzy partitioning, Hierarchical clustering, Flat clustering with threshold value which is fixed using the objective function SOS. In final, the clusters are assessed with intra cluster distance (Cluster Density). Generally, traditional Hierarchical clustering does not include any objective function. It produces numerous clusters based on the linkage criteria. The inclusion of Flat clustering simplifies the hierarchical (numerous clusters) cluster by applying the parameter such as number of clusters or threshold value to cut the clusters from dendrogram. The Flat Clustering expands the nodes in the order of their distance in the dendrogram.

This hierarchy information is visualized in dendrogram. The flat clustering is used to define some cut-point level on the dendrogram height that leads to further comprehensible analysis. It is associated with a threshold on the dendrogram height. But, there is no defined procedure to fix this threshold value to a compact clusters. Hence, this work proposed an objective function named Sum of Squares on Flat Clustering to fix the optimum threshold value. This work applies a value for the objective function Sum of Squares and executes the clustering till it satisfies the objective function. It measures the deviation of data points from the mean value of the cluster. Generally, higher the sum of squares denotes higher variability while a lower implies less variability. The three types of sum of squares are total, residual, and regressive. In order to calculate the sum of squares, first subtract the mean from the data points then square the differences, and finally add them together. The SOS value will decrease as the number of clusters increases.

FPHAC Algorithm

Input	Unlabeled Transaction Datasets
Output	Cluster Hierarchy, Dendrogram, Flat Clustered set
Step 1	Preprocessing - Replacing missing value, Nominal to Numerical, Normalization
Step 2	Creation of Pivot Table using the attribute order id, product id, sales.
Step 3	Applying Fuzzy clustering to obtain Fuzzy Partition Matrix with Manhattan Distance using, $$Partition\ Matrix: w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$ $$Manhattan\ Distance = \sum_{i=1}^{n} \|p_i - q_i\|$$
Step 4	Transform Fuzzy Partition Matrix to regular attributes
Step 5	Apply Hierarchical Agglomerative clustering with complete linkage criteria using the steps,
	5.1 Initialize Level, Sequence number as V*(0)*, *m=0* respectively.
	5.2 Retrieve the most uniqueness pair, let d[(r), (s)] is min d[(i), (j)] over all the pairs. $$\mathrm{dist}_{\max}(C_i, C_j) = \max_{a \in A,\ b \in B} d(a, b)$$
	5.3 Increase the Sequence number m+1, merge clusters (r, s) as a new one.

5.4 Fix this level to V*(m)=d{cr), (s)]*

5.5 Revise the Proximity Matrix T with the new cluster (r, s) and the old one (k) as,

$$d[(c,s),(k)] = max\{d[(k),(c)],d[(k),(s)]\}$$

5.6 Continue from 5.2 till the entire points in the same cluster

Step 6 Visualize dendrogram and apply Flat clustering on dendrogram.

6.1 Fix maximum threshold distance *td* randomly

6.2 Apply Sum of Squares (SOS) using Manhattan distance

6.3 If $SOS = \sum_{i=1}^{n}(x_i - \overline{x})^2 < 0.4$

6.4 Stop, otherwise continue with 6.1

Step 7 Assess Cluster quality using Average within cluster distance for the final clusters with,

$$\Delta(S) = \frac{1}{|S|.(|S|-1)} \sum_{\substack{x,y \\ x \neq y}} \{d(x,y)\}$$

5. Implementation, Results and Discussion

The transaction datasets to experiment this proposed method is taken from UCI, and Kaggle repositories. The experimental research is implemented in Rapid miner tool.

5.1 Dataset - Superstore Sales Analysis (SSO)

It consists of 21 attributes, and 9988 instances. There are three numeric attributes such as sales, quantity, year, two alphanumeric attributes order id, product id, two date attributes such as order date, ship date, eleven categorical attributes such as ship mode, customer name, segment, state, country, market, region, category, sub category, product name and order priority. The attributes taken to create pivot table are product id, category and these two are aggregated with the attribute quantity using the mathematical function count.

5.2 Clustered Results

In order to handle memory and time management while using Hierarchical clustering the dataset is split into ratio and the proposed hierarchical Flat Clustering methodology is implemented in serial execution. For the total samples 9988, the ratio is fixed as 0.2 and five sets of data is generated. In Table 3, the number of samples in each split, number of cluster formed for the Hierarchical method and the Flat clustering method is listed. The cluster numbers varies as it depends on the samples (data points).

Table 3. Number of Clusters

Sample subset	Samples with 0.2 ratio	Number of Clusters	
		Existing HAC	Proposed FPHAC
Subset 1	1996	3991	7
Subset 2	1997	3995	7
Subset 3	1996	3988	7
Subset 4	1997	3998	5
Subset 5	1996	3985	8

In order to handle memory and time management while using Hierarchical clustering the dataset is split into subset with ratio and the proposed hierarchical Flat Clustering methodology FPHAC is implemented in serial execution. For the total samples 9988, the ratio is fixed as 0.2 and five sets of data is generated. In Table 4, Flat threshold distance with Sum of Square values, Cluster density (Average within cluster distance) is listed. These values solely depends the samples in each set.

Table 4. Flat Distance Threshold Value

Sample subset	Flat Distance Threshold value	SOS	Cluster Density
Subset 1	0.4	0.25	73.03
Subset 2	0.5	0.39	86.56
Subset 3	0.5	0.28	81.39
Subset 4	0.6	0.39	109.11
Subset 5	0.6	0.21	129.41
Average		0.31	479.50

5.3 Performance Measures

Sum of Squares – Measures the deviation of data points in a group. A less Sum of Squares implies a good model.

Cluster Density (Average within cluster distance) – Averages all the distances between each pair of samples in an individual cluster. Then calculate the average of all the clusters. The smaller the distance the higher the similarity of data points in a cluster.

Processing Time – The running time of the proposed method measured in seconds.

Table 5, shows the SOS and Cluster Density values for the existing methods namely HAC, HDBSCAN, PRI-MFC and the proposed method FPHAC. All the methods are the improvements of Hierarchical clustering. The experimental results shows that the proposed method has less SOS value

0.31 and Cluster density value 479.5 that implies the objects in the proposed method clusters are closer than the existing.

Table 5. Comparison of Existing versus Proposed Methods

Methods	SOS	Cluster Density	Processing Time (in milliseconds)
Existing HAC	0.61	728.2	1440000
Existing HDBSCAN	0.49	592.6	1455000
Existing PRI-MFC	0.57	621.3	955000
Proposed FPHAC	0.31	479.5	300000

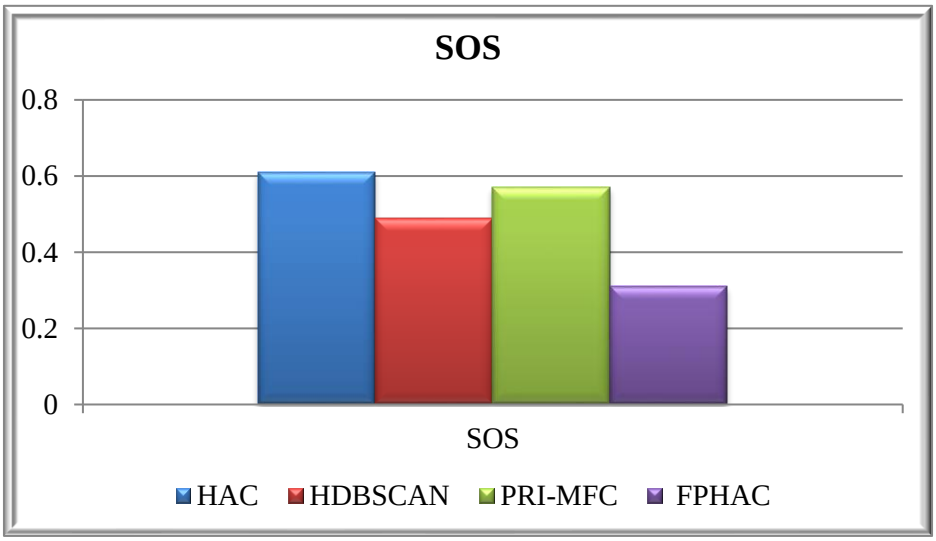

Figure 3. SOS Comparison Graph

Fig. 3, shows that the proposed method FPHAC is 0.3 less SOS value than HAC, 0.18 less than HDBSCAN, 0.26 less than PRI-MFC and thereby proves the efficiency.

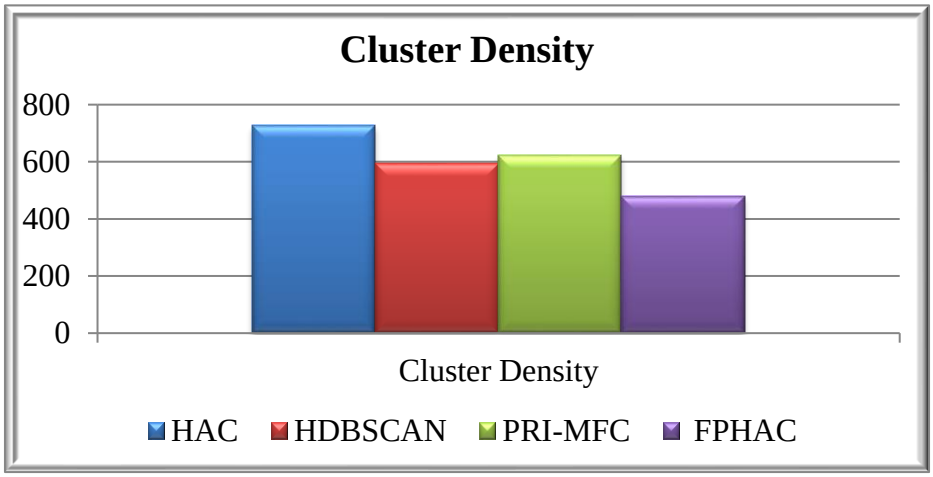

Figure 4. Cluster Density Comparison Graph

Fig. 4. shows the cluster density in terms of Average within cluster distance values. The proposed method FPHAC is 248.7 less value than HAC, 113.1 less than HDBSCAN, 141.8 less than PRI-MFC. Hence, the objects in the proposed method clusters are closer than other existing methods. Less cluster density (intra cluster distance) implies a good clustering by having the data points within the cluster more closer.

Fig. 5, shows the processing time analysis of the existing and proposed methods and it is confirmed that FPPHAC takes least processing time as the method includes data summarization using pivot table based on the analysis. This work implements the method on product analysis. The minimized processing time is due to the pivot table that groups only the attribute that has details about products.

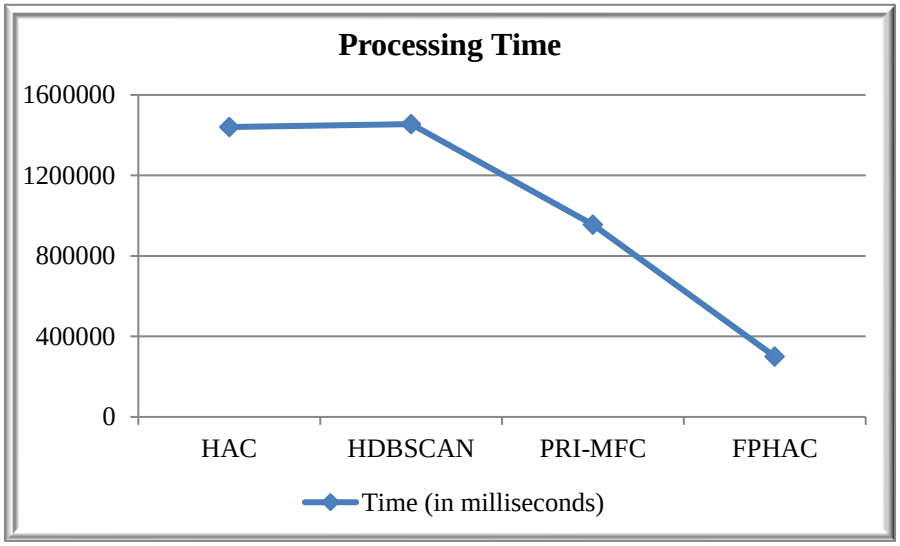

Figure 5. Processing Time Analysis Graph

Table 6, lists the values for number of clusters formed, Flat Threshold value, SOS and Cluster density using the proposed method on various UCI, Kaggle repository datasets. The experimental value confirms that all the values depend on the dataset and its samples.

Table 6. Performance of FPHAC for variant repository Transaction Datasets

Datasets	Clusters formed	Flat Threshold distance	SOS	Cluster Density
Products	3	3.1	0.34	137.6
Online Retail	5	0.9	0.35	196.8
Transactions	4	2.6	0.32	126.5
Online Store	4	1.5	0.26	112.8

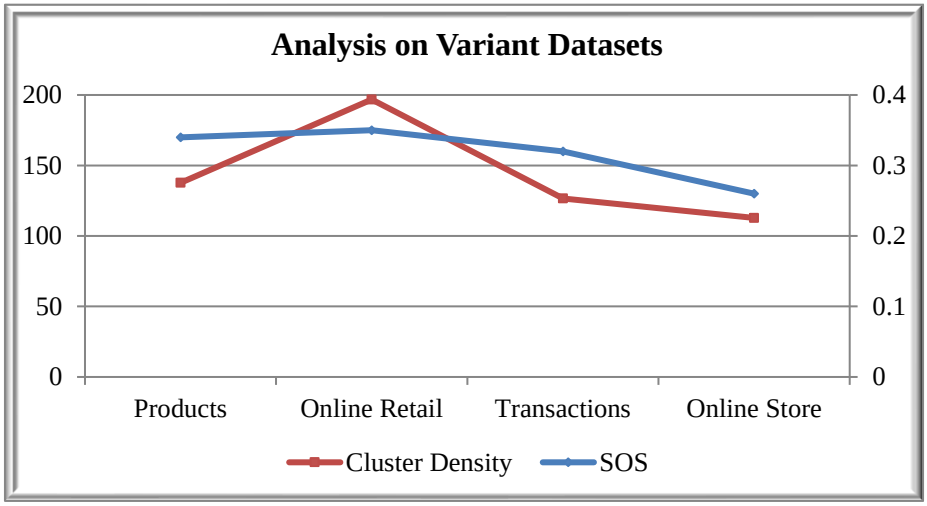

Figure 6. Analysis on Variant repository Datasets using FPHAC

Fig 6, shows the Analysis on Variant repository Datasets using FPHAC and the results varies on the basis of the dataset form. Among them, Online Store data has least SOS and Cluster Density values.

6. Conclusion

Hierarchical clustering generates numerous clusters based on the hierarchy level of the data points. Initially, the proposed method put forth a simple method to handle large dataset using linear sampling split and the original dataset is transformed to pivot table with information rich data that summarize the data with limited dimension in transaction data thereby lessen the computational and memory complexity. Fuzzy Partition matrix with normalization is applied to the dataset to get the partition score of being in the cluster that gives additional insight while the clusters in the hierarchy are merged. Generally, the hierarchical clusters in dendrogram are breakdown using Flat clustering with the user parameter number of clusters or with the distance threshold value. However, there is no standard objective function to fix the distance threshold value in existing Hierarchical Flat clustering. This research work proposed an objective function Sum of Squares (SOS) to fix the optimum distance threshold value that assures the quality of the cluster formed. The proposed method outperforms the existing methods and proves its efficiency by having less Cluster Density, SOS and Processing Time.

References

[1] Sivasankari .S, Dr.S.Sukumaran, Dr.S. Muthumarilakshmi, “Deep Learning based Weight Guided Wrapper Feature Subset Method for Multiclass Data Classification”, NeuroQuantology, Volume 20, Issue 16, pp.5675-5686, 2022. https:// www. neuroquantology. com/search?q= Deep% 20 Learning 20based% 20Weight% 20Guided%20Wrapper%20Feature%20Subset%20Method%20for%20Multiclass%20Data%20Classification

[2] Teng Li a, Amin Rezaeipanah, ElSayed M. Tag El Din, “An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement”, Journal of King Saud University – Computer and Information Sciences, Volume 34, Issue 6, Part B, pp.3828-3842, June2022. https:// www. sciencedirect. Com /science /article /pii/ S1319157822001380

[3] Thomas Märzinger , Jan Kotík, Christoph Pfeifer, “Application of Hierarchical Agglomerative Clustering (HAC) for Systemic Classification of Pop-Up Housing (PUH) Environments”, Applied Sciences, 11(23),pp.1-19, 2021. https:// doi.org/ 10.3390/app112311122

[4] Praveen P, Ranjith Kumar M, Mohammed Ali Shaik, R Ravikumar, R Kiran, "The comparative study on agglomerative hierarchical clustering using numerical data", IOP Conf. Series: Materials Science and Engineering, pp.1-10, 2020. doi:10.1088/1757-899X/981/2/022071

[5] Nicholas Monath et al, "Scalable Hierarchical Agglomerative Clustering", KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1245–1255, August 2021.. https://doi.org/10.1145/3447548.3467404

[6] Crisnandra Rahmita Mardiantien; Imelda Atastina; Ibnu Asror, "Product Segmentation Based on Sales Transaction Data Using Agglomerative Hierarchical Clustering and FMC Model (Case Study: XYZ Company)", 2020 3rd International Conference on Information and Communications Technology (ICOIACT), IEEE, pp. 223-228, 2020.DOI: 10.1109/ ICOIACT50329. 2020. 9332023

[7] Yue Wang et al , "PACk: an efficient partition-based distributed agglomerative hierarchical clustering algorithm for deduplication, ",Proceedings of the VLDB Endowment, Volume 15, Issue 6, pp 1132–1145, 2022. https:// doi.org/ 10.14778/ 3514061. 3514062

[8] Pranav Shetty and Suraj Singh, "Hierarchical Clustering: A Survey" , International Journal of Applied Research, 7(4), pp. 178-181, 2021. DOI: https://doi.org/10.22271/allresearch.2021. v7. i4c. 8484

[9] Farhad Zolfaghari, Hassan Khosravi, Alireza Shahriyari, Mitra Jabbari, Azam Abolhasani, "Hierarchical cluster analysis to identify the homogeneous desertification management units", PLOS|ONE, pp.1-21, 2019. https://doi. org/ 10. 1371/journal. pone. 0226355

[10] Junyi Zhou, Ying Zhang, Wanzhu Tu, "clusterMLD: An Efficient Hierarchical Clustering Method for Multivariate Longitudinal Data", Volume 32, Issue 3, pp. 1131-1144, 2023. https://doi. org/10.1080/ 10618600.2022.2149540

[11] Dajun Si , Wenyue Hu , Zilin Deng, Yanhui Xu, "Fair hierarchical clustering of substations based on Gini coefficient", Global Energy Interconnection, Elseiver Publication, Volume 4, Number 6, pp. 576-586, December 2021. https:// www. sciencedirect. com/ science/ article/pii/S2096511 722000093

[12] Mohamed S. Halawa, Rebeca P. Díaz Redondo, Ana Fernández Vilas , "Unsupervised KPIs-Based Clustering of Jobs in HPC Data Centers", Sensors, MDPI, pp.1-20, 2020. doi:10.3390/s20154111

[13] Lifeng Yin, Menglin Li , Huayue Chen, Wu Deng, "An Improved Hierarchical Clustering Algorithm Based on the Idea of Population Reproduction and Fusion, electronics, MDPI, pp.1-19, 2022. https:// doi.org/10.3390/electronics11172735

[14] Joelson Antonio dos Santos, Talat Iqbal Syed, Murilo C. Naldi, Ricardo J. G. B. Campello, Joerg Sander, "Hierarchical Density-Based Clustering Using MapReduce", IEEE Transactions on Big Data, Vol. 7, No. 1, pp. 102-114, January-March 2021,. doi: 10.1109 /TBDATA 2019.2907624

[15] Praveen .P, "An Efficient Linkage Criterion for Creating Clusters in Hierarchical Method", International Journal of Future Generation Communication and Networking, Vol. 12, No. 5, pp. 294- 300, 2019. https://www.researchgate. net/ publication/339828649_ An_Efficient_ Linkage_Criterion_for_Creating_Clusters_in_Hierarchical_Method

[16] William W. Tsoa, C. Doga Demirhana, Clara F. Heubergerc , Joseph B. Powelld , Efstratios N. Pistikopoulos, "A hierarchical clustering decomposition algorithm for optimizing renewable power systems with storage", Applied Energy, Vol. 270, Elseivier, pp.1-10, 2020,. https:// www. sciencedirect.com/ science /article /abs/ pii/ S0306261920307029?via%3Dihub

[17] Karli Eka Setiawana, Afdhal Kurniawana, Andry Chowandaa, Derwin Suhartono, "Clustering models for hospitals in Jakarta using fuzzy c-means and k-means, 7th International Conference on Computer Science and Computational Intelligence 2022, Elseiver, Procedia Computer Science 216, pp. 356–363, 2023, https:// doi.org /10.1016/ j.procs. 2022. 12.146

[18] Joaquín Pérez-Ortega , Sandra Silvia Roblero-Aguilar, Nelva Nely Almanza-Ortega, Juan Frausto Solís, Crispín Zavala-Díaz, Yasmín Hernández, Vanesa Landero-Nájera, "Hybrid Fuzzy C-Means Clustering Algorithm Oriented to Big Data Realms", Axioms, MDPI, pp.1-16, 2022,. https:// doi. org/ 10.3390 / axioms11080377

[19] Seyyed Mohammad Razavi, Mohsen Kahani, Samad Paydar, "Big data fuzzy C-means algorithm based on bee colony optimization using an Apache Hbase", Journal of BigData, Springer Open, pp.1-22, 2022,. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00450-w

[20] Ahmad Afif Supianto, Nur Sa'diyah , Candra Dewi , "Improvements of Fuzzy C-Means Clustering Performance using Particle Swarm Optimization on Student Grouping based on Learning Activity in a Digital Learning Media", Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology, pp.239–243, November2020. https:/ /doi.org/ 10.1145/ 3427423. 3427449

Author Profile

Dr. S. Sukumaran, working as Associate Professor, Department of Computer science (Aided) in Erode Arts and Science College, Erode, Tamilnadu, India. He is a member of Board of studies in various Autonomous colleges and universities. In his 35 years of teaching experience, he has supervised more than 55 M.Phil Research Works, guided 23 Ph.D research works and still continuing. He has presented, published around 80 research papers in National, International Conferences and Peer Reviewed Journals. His area of research interest includes Digital Image Processing and Data mining.

S.Sivasankari, has completed B.Sc (Computer science) in affiliated college of Madurai Kamaraj University, M.C.A at Bharathiar University, Coimbatore in distance education. She has awarded M.Phil in Data Mining from Bharathiar University, Coimbatore. At present, she is continuing her doctorate research work (Part time) in Data Mining at Department of Computer science (Aided), Erode Arts and Science College, Erode, Tamilnadu, India.