ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Variance-Inter-Quartile Range Hierarchical Clustering Method with Applications (VIQR)

Eriobu N. O[1], Abidoye A. O[2]

[1]Department of Statistics, Nnamdi Azikiwe University, Awka, Anambra State.

[2]Department of Statistics, University of Ilorin, Ilorin, Kwara State.

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150400060

Received: 13 April 2026; Accepted: 17 April 2026; Published: 08 May 2026

ABSTRACT

Clustering is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups (clusters). In this paper, a new procedure of linkage hierarchical clustering method (Variance-inter-quartile Range Hierarchical clustering method) was developed to reduce the effect of extreme values in the classification of groups into clusters. The new proposed method makes use of the variance to get the variance matrix and inter-quartile range for the construction of the dendogram which is used for the comparative analysis of the newly proposed method and the existing methods. A performance evaluation was carried out using the visual dendogram inspection and silhouette score on the newly proposed method VIQR (Variance-inter-quartile Range Hierarchical clustering method) and the existing methods (Single and Complete linkage). From the result, using visual dendogram inspection, It was observed that VIQR Hierarchical clustering method performs better than the existing methods for the classification of objects as it has similar dendogram with fewer steps in the classification procedure. Using Shilouette score as a method of validation of the clusters, from the result, the score for the proposed method is 0.53 which is higher than the scores of the existing methods; single linkage (0.23) and complete linkage linkage (0.30). In summary, VIQR is more efficient and robust classification method in terms of shorter algorithm, better control of extreme values, and less complexity. Therefore the VIQR should be adopted as a better method for classification of observations or variables especially when dealing with problems that have to do with classification.

Keywords: Clustering, Variation, Similarity, Dendogram, Hierarchical, Dispersion

INTRODUCTION

Nature of similarity measure plays an important role in the failure or success of clustering method (May and Moe, 2016). Inter object similarity is a measure of correspondence or resemblance between objects to be clustered. Just as correlation matrix between variables is used in factor analysis, the characteristics are combined into a similarity measure calculated for all pairs of objects. Clustering is one of the most common methods of unsupervised learning in which data is segmented based on the similarity of instances (Hamidi, Akbari, and Motameni 2019). Cluster analysis procedure then proceeds to group similar objects together into clusters. Inter object similarity can be measured in a variety of ways but three methods dominate the applications of cluster analysis (Krishnamurthy, 2006); correlation measures, distance measures and association measures. Each of the methods represents a particular perspective on similarity, dependent on both its objectives and the types of data. Both the correlation and distance measures require metric data, while the association measures are for non-metric data.

Correlation Measures can also be adopted. The inter object measure of similarity that probably comes to mind first, is the correlation coefficient between a pair of objects measured on several variables. The correlation between the two columns of numbers is the correlation (or similarity) between the profiles of the two objects. High correlation indicates similarity and low correlations denote a lack of it. A correlation measure of similarity does not look at the magnitude of the values but instead the patterns of values. But correlation measures are

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

rarely used because emphasis in most applications of cluster analysis is on the magnitudes of the objects, not the patterns of values.

Distance $(x, y) = \sum_{i=1}^{n} \frac{(x_i - \mu_x)(y_i - \mu_y)}{(n-1)\sigma_x \sigma_y}$, where μ_x is the mean of x and σ_x is its standard deviation.

Distance measures are normally used to measure the similarity or dissimilarity between two data objects. Joining or tree clustering methods use the dissimilarities or distances between objects when forming the clusters. These distances can be based on a single dimension or multiple dimensions. Clusters based on correlation measures may not have similar values but instead have similar patterns. Distance-based clusters have more similar values across the set of variables but the patterns can be quite different. Some of the distance measures are:

- Euclidean distance:
  Euclidean distance is a standard matrix for geometrical problems and widely used in clustering problems. It is the ordinary distance between two points and can be easily measured with a ruler in a two or three-dimensional space. It works well when a data set has compact or isolated clusters (Sruthi and Reddy, 2013).
  Distance $(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
  This is computed from the raw data and not from consistent data.
- Squared Euclidean distance:
  One may want to square the standard Euclidean distance in order to place gradually greater weight on objects that are further apart. This distance is computed as:
  Distance $(x_i, y_i) = \sum_{i=1}^{n} (x_i - y_i)^2$
- City-Block (Manhattan) distance:
  This distance is simply the average difference across dimensions. In most cases, this distance measure yields results similar to the simple Euclidean distance. However, it is noted that in this measure, the effect of single large differences (outliers) is reduced since they are not squared. The city-block distance is computed as:
  Distance $(x_i, y_i) = \sum_{i=1}^{n} |x_i - y_i|$

There are various types of clustering;

i. Hierarchical clustering or Hierarchical cluster analysis: is a method of cluster analysis which seeks to build a hierarchy of clusters *(*Rokach, Lior, and Oded, 2005*)*. It is also constructing the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. Hierarchical clustering produces a hierarchical tree of clusters called dendrogram. A dendrogram describing the relationships between all observations in a data set is useful for understanding the hierarchical relationships in the data, in many situations a discrete number of specific clusters is needed (Myatt, 2009). Examples of Hierarchical method include Agglomerative and Divisive. The weakness of hierarchical clustering is instability (unsteadiness) to scale well and can never undo what was done previously that is there is no back-tracking capability.
ii. Partitional clustering: is a method that divides the large object into (groups) clusters and each cluster containing at least one element. It follows an iterative process by use of this process; we can relocate the object from one group to another more relevance group. This method is effective for small to medium sized data sets. Examples of partitioning methods include k-means and k-medoids (Jiawei and Micheline, 2007).
iii. Density-based method: This method assumes that the points that belong to each cluster are drawn from a specific probability distribution (Banfield and Raftery, 1993). Density based method provide high quality performance but it depends on two specified parameters, r and Minpts (Bharat and Manan, 2012).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

LITERATURE REVIEW

Martin et al (1996) proposed a clustering technique called DBSCAN. The method was developed in discovering clusters of arbitrary shape because the shape of clusters in spatial datasets may be spherical, draw-out, linear, elongated etc. The method DBSCAN is from the word Density based spatial clustering of applications with Noise. The researchers use neighborhood and distance in the clustering of object for effective heuristic parameters. The method was compared with CLARANS and it was concluded that DBSCAN performed better in handling large dataset and also significantly more effective in discovering clusters of arbitrary shape.

Tian, Raghu and Miran (1996) proposed a clustering technique called BIRCH. The method was referred to as an efficient clustering method that can be used for a very large dataset. The method birch is from the word balanced iterative reducing and clustering using hierarchies. It was stated that the proposed method can handle noise which is a situation where data points are not part of the underlying pattern or structure of the given set of observations. The method was compared with an existing clustering method called CLARANS and it was concluded that Birch is better than Clarans in handling large dataset.

Ding and He (2002) provide a comprehensive analysis and experiments on divisive and agglomerative clustering. Similarities were used instead of the traditional distances. For divisive clustering, 4 new cluster selection criteria was introduced, the average similarity, the cluster cohesion, avg-cohesion, and temporary objective. Based on the 4 new cluster selection criteria introduced on internet newsgroups, it is shown that average similarity and similarity-cohesion selection perform well. For agglomerative clustering, MinMax linkage was introduced and compared with single-linkage, complete linkage and average linkage. MinMax linkage were found to be most effective than other in merging clusters.

Almeida et al 2007 presented a new method of hierarchical cluster analysis capable of detecting outlier and automatic clustering. This technique is based on agglomerative hierarchical clustering which is one of the most frequent approaches in unsupervised clustering. The algorithm comprises of three steps namely; treatment of outlier (outlier control), blocks identifier and group of blocks using similarity approach. The resulting classification method was validated by comparing it with some of the traditional methods. It was observed that the new method outperform some of the traditional methods in terms of clustering objects.

Basu and Murty (2013) developed am hierarchical clustering method called CUES. The method CUES is from the Clustering Using Extensive Similarity. The method were developed using a new cluster distance measure to identify two dissimilar clusters, it will never merger them but can stopped the algorithm if the distance between two clusters became very high. The method was compared with existing methods (single-link hierarchical, average–link hierarchical, bisecting k-means, buckshot, k nearest neighbor and it was found that CURE perform significantly better than other existing clustering.

Yogita and Harish (2013) stated some improved hierarchical clustering algorithms like CURE, BIRCH, CHEMELEON, Linkage, Leaders–Subleaders, Bisecting k-Means that overcome the limitations that exist in pure hierarchical clustering algorithms. They went further to give some criteria on the basis of which one can also determine the best among these mentioned algorithms.

May and Moe (2016) presented a modified agglomerative hierarchical clustering (MAHC) which is an improvement on commonly used hierarchical clustering method. The modification was done to address the shortcomings of AHC such as precision, measurement and dimensionality. Instead of frequency of occurrency, the modified method used item based collation for the construction of distance matrix. It was concluded that MAHC has shorter algorithm and grouped items in clusters better than the AHC method.

Pelin and Derya (2017) proposed a hierarchical clustering method called K- Linkage which evaluates the distance between two clusters by calculating the average distance between k pairs of observations. The researchers introduces two concepts k-min linkage considers k minimum (closet) pairs from points in the first cluster to points in the second cluster, k-max linkage takes into account k maximum (farthest) pairs of observations. The improved hierarchical clustering algorithm based on k-linkage was executed on five well-known benchmark

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

datasets with varying k values to demonstrate its efficiency. The results show that the proposed k-linkage method can often produce clusters with better accuracy, compared to the single, complete, average and centroid linkages.

In more recent years, the literature has shifted significantly toward enhancing hierarchical clustering through advanced structural representations and improved strongest. Cabezas, Izbicki, and Stern (2023) demonstrated that hierarchical clustering is no longer limited to simple dendrogram construction but can also support feature importance analysis and model selection. This reflects a growing recognition that clustering should not only produce groups but also provide interpretability and decision support.

Further advancements emphasize the integration of graph-based and structural learning techniques. Chen et al. (2025) introduced a crystallized neighborhood graph approach, showing that incorporating graph structures significantly improves clustering performance in complex datasets. Similarly, Jin et al. (2026) proposed granular-ball computing, which captures local data structures more effectively than traditional point-based representations. These approaches highlight a transition from purely distance-based clustering toward structure-aware methods that better reflect real-world data complexity

SUMMARY OF LITERATURE

The literature demonstrates clustering's broad applications as it has been used in different ways including grouping of locations or discovery of most common crime in locations. Also, comparison of the methods of clustering revealed that the strength of the methods varies significantly. This implies some of the existing methods can be referred to as better method than some other existing methods.

Since it is possible to detect a better method among existing hierarchical clustering methods, there is room for an improvement especially on the weak methods (see Olowoyori, 1994) for better performance.

Algorithm of hierarchical clustering can also be used in rating the methods as short steps with same achievement can be deemed to be better than a lengthy step of algorithm if both can perform the same assignment.

METHODOLOGY

The data collected is a secondary data from the work of Alvin (2002) where the researcher clustered cities with respect to crime rates. The formula for variance and inter-quartile range are shown in equations (1) and (2)

$$D_{\text{var}} = \sum_{i=1}^{n} \left(\frac{x_i - y_i}{n-1} \right)^2 \quad (1)$$

Inter-quartile range = $Q_3 - Q_1$ (2)

The new proposed method is Variance inter-quartile range classification technique. The algorithm of the proposed method is thus below

Algorithm of the Proposed Methods

Variance-Inter-quartile Range Classification Technique

1. Calculate variance to get variance matrix.
2. Locate the smallest value in the matrix.
3. Combine the point (say AB) with other points to get two values.
4. Calculate the inter-quartile range of the two values obtained.
5. Do same for all the points to get a reduced matrix.
6. Repeat step 2, 3 and 4, compile the new reduced matrix till the lowest possible matrix is formed.
7. Then, form the dendogram.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Existing Hierarchical Clustering Method

Single linkage Clustering: Single linkage clustering is the oldest and simplest algorithm (Stuetzle and Nugent, 2010). In Single linkage, the distance between two clusters is the minimum distance between an observation in one cluster and an observation in the other cluster.

The distance between two clusters is defined as $D_{ki} = Min \;\; d(x_i, x_j)$

$$i \varepsilon c_k, \;\; j \varepsilon c_l$$

Complete Linkage methods: Complete linkage method is the opposite of single linkage. The distance of two clusters is defined as the distance between their farthest objects which may be considered as the diameter (maximum within-cluster distance) of the new cluster. (Brian, Sabine and Morven, 2001).

The distance between two clusters is defined as $D_{ki} = Max \;\; d(x_i, x_j)$

$$i \varepsilon c_k, \;\; j \varepsilon c_l$$

Shilouette Score

The above method was proposed by Peter J. Rousseeuw in 1987 as a method for validating the consistency of clustering results. The Shilouette Score is one of the methods for comparison of classification techniques via the resulting dendogram and distance matrix**.** The Silhouette score formula is

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \qquad (3)$$

Where $a(i)$ = average distance from point i to all other points in the same cluster

$b(i)$ = smallest average distance from point i to points in the nearest other cluster.

The silhouette score takes values between −1 and +1. A value close to +1 indicates that an observation is well matched to its assigned cluster and poorly matched to neighboring clusters. A value around 0 suggests that the observation lies near the boundary between two clusters. Negative values (close to −1) imply that the observation may have been assigned to the wrong cluster. Hence the silhouette score will be used to compare the proposed method and the existing methods.

Analysis And Result

Comparison Of the Proposed and Existing Methods

The data in Table 1 was extracted from the work of Alvin (2002) where the researcher clustered cities with respect to crime rates.

Table 1: City Crime Rates per 100,000 Population

City	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto theft
Atlanta	16.5	24.8	106	147	1112	905	494
Boston	4.2	13.3	122	90	982	669	954
Chicago	11.6	24.7	340	242	808	609	645
Dallas	18.1	34.2	184	293	1668	901	602
Denver	6.9	41.5	173	191	1534	1368	780
Detroit	13	35.7	477	220	1566	1183	788

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Using the data in Table 1, the dendogram for both proposed method and the existing methods (Single linkage and Complete linkage) are as shown below;

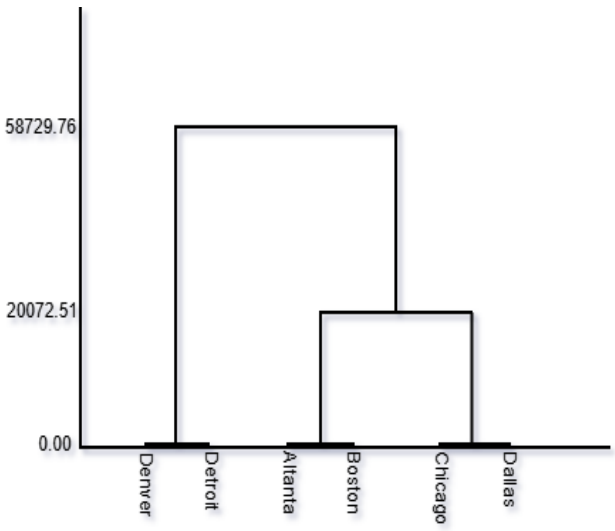

Figure 1: Dendogram of Clustering using Variance Inter-quartile Range (VIQR)

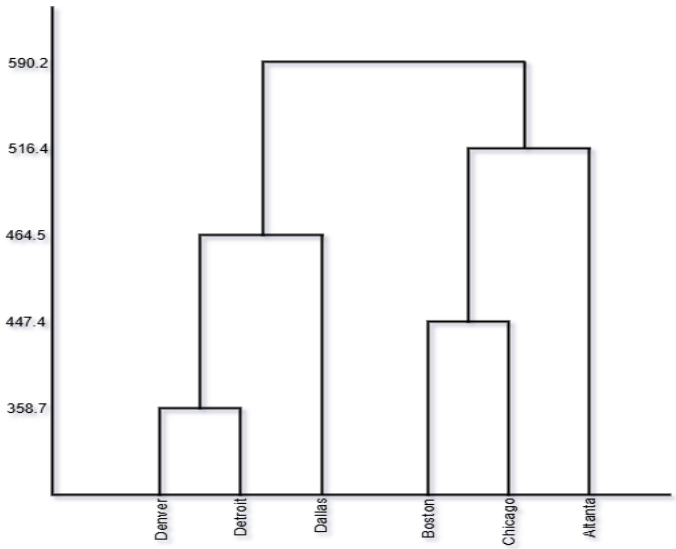

Figure 2: Dendogram of Clustering using Single linkage

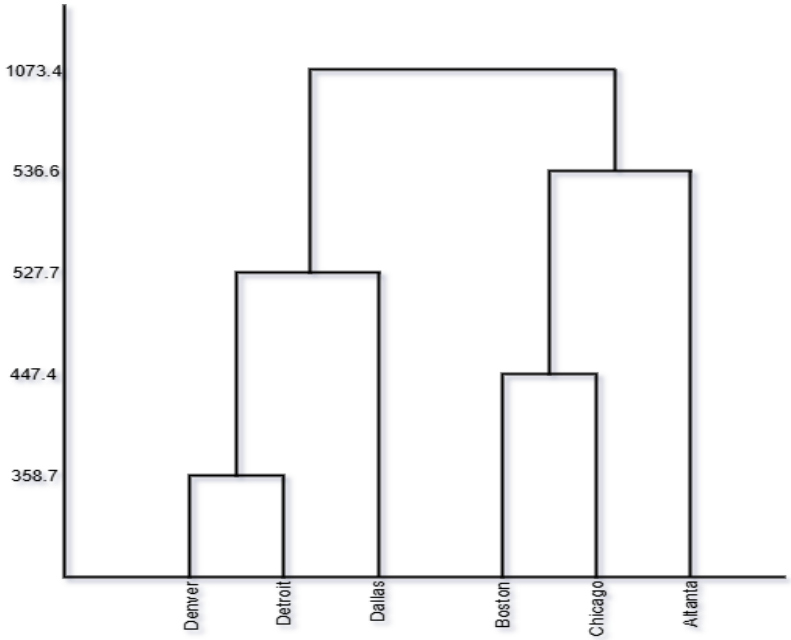

Figure 3: Dendogram of Clustering using Complete linkage

As shown in figure 1, the dendogram has two major clusters; Denver and Detroit as a cluster and Altanta, Bosto n, Chicago and Dallas as the second major cluster. It can be observed that the grouping has two major partition/ cluster using Variance Inter-quartile Range method. On the same data, as shown in figure 2, using Single linka ge, two major clusters were also formed; Denver, Detroit and Dallas as a cluster and Boston, Chicago and Alta nta as the second major cluster. Finally, as shown in figure 3, using Complete linkage, two major clusters were also formed; Denver, Detroit and Dallas as a cluster and Boston, Chicago and Altanta as the second major clust er.

Using Shilouette score as a method of validation of the clusters, see Table 2 below;

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Table 2: Comparison of Proposed and Existing Methods for Secondary Score

Hierarchical clustering	Shilouette Score
Variance inter-quartile range	0.53
Single Linkage	0.23
Complete linkage	0.30

As shown in the above Table 2, the score for the proposed methods are higher than the existing methods which imply the proposed methods adequately classified the variables of interest better than considered existing methods. From the scores, the existing methods also classified the variables appropriately since the values are higher than 0.5. Among the existing methods, Variance-Weighted Mean Method has the highest score of 0.53. This implies is the best and most appropriate among the methods used for the data. Single Linkage has the lowest score of 0.23 which implies the classification by the methods should be used with caution.

DISCUSSION OF RESULT

As observed from the results both existing and proposed method generated similarity dendogram which is an indication that the proposed method is capable of classifying the variables or observations the same way as the existing method does. The existing method used Euclidean distance as a measure of similarities but the proposed method used variance as substitute in the algorithm. Inter-quartile Range was used for the determination of values used for the dendogram. This implies in the proposed method, measures are variance and inter-quartile range.

SUMMARY OF FINDINGS AND CONCLUSION

In the classification techniques, as observed in the literature, there exist numerous methods with limitations in terms of accuracy, efficiency, length of algorithm, sensitivity to extreme values as well as complexity in the computation. A more efficient and robust classification method can be in terms of shorter algorithm, better control of extreme values, less complex etc. |In this study, a more efficient with shorter algorithm method of classification was proposed. The method utilizes variance and range which can easily be computed and less rely on statistical assumptions such as normality etc. This implies the method can be referred to as robust because it relies on less number of assumptions which makes it more useful among the existing methods. The suitability and superiority of the method was shown using data extract from the work of Alvin (2002). From the result of the analysis, it can be observed that the proposed method favourably competed with the existing methods which make it useful for classification purpose. Therefore, Variance-Range Classification technique can be adopted in the classification of observations or variables especially when dealing with problems that have to do with classification.

REFERENCES

1. Alvin C. R. (2002). Methods of Multivariate Analysis. John Wiley and Sons, Inc. publication Second Edition. Page 456
2. Banfield J. D. and Raftery A. E. (1993): “Model-based Gaussian and non-Gaussian clustering”. Biometrics, 49:803-821.
3. Bharat Chaudhari, Manan Parikh (2012): “A Comparative Study of clustering Algorithms”. International Journal of Application or Innovation in Engineering & Management.
4. bezas, L. M. C., Izbicki, R., & Stern, R. B. (2023). Hierarchical Clustering: Visualization,Feature Importance And Model Selection. *Applied Soft Computing, 141*, 110303.
5. Brian S. Everitt., Sabine Landau and Morven Leese (2001). Cluster Analysis (Fourth ed.) London: Arnold.
6. Chen, H., Zhang, R., Duan, Y., Wang, R., & Nie, F. (2025). Fast Multi-View Clustering Via Anchor Label Transmit With Tensor Structure Constraint. *Expert Systems With Applications, 274*, 126878.
7. Chris ding and Xiaofeng He (2002): “Cluster merging and Splitting In hierarchical clustering Algorithms”, International Conference on Data Mining (ICDM 2002) IEEE, pg139.
8. Hamidi S.S, Akbari. E, Motameni .H. (2019): Consensus clustering algorithm based on the automatic partitioning similarity graph, Data Knowl.Eng., 124, p.101754

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

9. J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais and S.J. Formosinho (2007): "A new method with outlier detection and automatic clustering. Chemometrics and Intelligent Laboratory Systems 87, pp. 208–217
10. Jiawei Han and Micheline Kamber, Jian Pei, B Data Mining: Concepts and Techniques, 3rd Edition, 2007.
11. Jin, T., Sun, X., Huang, B., & Wang, T. (2026). A Novel Hierarchical Clustering Approach Based On Granular-Ball Computing. *Neurocomputing*, *665*, 132200.
12. Krishnamurthy, P. (2006): Approaches to clustering Gene expression Time course Data, M.Sc Thesis: Department of Computer Science and Engineering, State University of New York at Buffalo.
13. Martin Ester, Hans-Peter Kriegel, J&g Sander and Xiaowei Xu (1996): "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Institute for Computer Science, University of Munich. Pp. 226-231.
14. May Thu Lwin and Moe Moe Aye (2016): "Modified Agglomerative Hierarchical Clustering (MAHC) Algorithm for Document Clustering", International Journal of Advances in Electronics and Computer Science, Vol. 3, Issue-8.
15. Myatt, Glenn J, (2009): "Making Sense of Data II", 2009, Wiley, Canada.
16. Onyeagu S. I (2003): A first course in Multivariate Statistical Analysis, Textbook.
17. Pelin Yildirim and Derya Birant (2017): K-Linkage: A New Agglomerative Approach for Hierarchical Clustering, Advances in Electrical and Computer Engineering, 17(4) pp 77-88.
18. Stuetzle, w and Nugent, R. (2010): A generalized Single linkage method for estimating the cluster tree of a destiny. Technical Report. University of Washington, Washington, USA.
19. Sruthi, K and B.V. Reddy (2013), "Document clustering on various similarity measures", International Journal of Advanced Research in Computer Science and Software Engineering, IJARCSSE, Vol.3, Issue-8.
20. Tian Zhang, Raghu Ramakrishnan and Miron Livny (1996): "Birch: An Efficient Data Clustering Method for Very Large Databases. Computer Sciences Dept. Univ. of Wisconsin-Madison. Pp. 103-114
21. Rokach, Lior, and Oded Maimon (2005),"Clustering methods." Data mining and knowledge discovery handbook. Springer US, pp 321-352.
22. Yogita Rani and Harish Rohil (2013): "A Study of Hierarchical Clustering Algorithm. International Journal of Information and Computation Technology, 3(10), pp. 1115-1122.