**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**

**IMPACT FACTOR: 5.258**

***IJCSMC, Vol. 5, Issue. 5, May 2016, pg.633 – 638***

# A Study on Clustering Methods for Global Data Environments

**Preeti**

Student, M.Tech (CSE)
PDM College of Engineering
Bahadurgarh, Haryana
coolpreetidon@gmail.com

**Manju**

Asst. Prof, M.Tech (CSE)
PDM College of Engineering
Bahadurgarh, Haryana
Manju_engg@pdm.ac.in

***Abstract— Data clustering is the basic and the essential phenomenon applied to generate the interest region, data or the features. Clustering can be applied in different forms on different datasets, applications and environments. In this paper, an exploration to the clustering methods is provided based on the prior analysis. The paper has categorized the algorithms in three main categories. The paper has explored each of this category with process specification. The paper also defined the algorithmic specification of three of the effective algorithmic approaches. The strengths and characterization of these clustering methods are explored in this paper***

***Keywords : Clustering, FCM, KMeans, Hierarchical, Distance Driven***

## I. INTRODUCTION

Clustering is the task defined to divide the data in smaller segments or categories based on data analysis. The range driven clustering is the primary requirement of many of the applications to reduce the processing data size or to perform the effective data selection. While work on high dimensional data, large data repository and image processing, clustering is having the higher significant. Clustering also having the scope of document filtration, web mining and text driven analysis where the keyword strength based segmentation can be applied. In bioinformatics, where a large dataset is available in the form of microarray, the clustering can be applied to perform the data selection significantly. The clustering is having the higher scope in most of the applications as an essential preprocessing stage. It is considered as a unsupervised learning method that is able to take the predictive decisions. The outlier identification or the abnormal data filtration is also the application area of clustering. In more intelligent form, instead of applying the clustering on data values, the feature specific clustering can be applied. There are number of clustering algorithms available under different parameters and algorithmic approach specification. In the broader form, clustering algorithms are divided in three main categories called overlapped clustering, partitioned clustering and hierarchical clustering.

### A) Partition Clustering

As the name suggest in this clustering method, the data is divided in clear partitions so that no overlapping between the data items will occur. During the partitioning process, the data elements can move between the clusters. But as the process ends, each data item exist in exactly one clusters. The clustering process is center specific and the distance driven map is applied to identify the cluster member. The partitioned clustering is shown in figure 1.

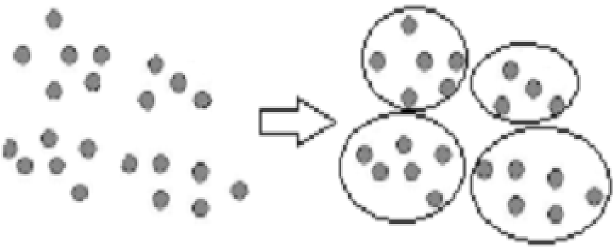

Figure 1 : Partition Clustering[3]

As shown in the figure, the data presented in the open space are qualified to the particular cluster. KMeans, Kmedoids are the most common form of these clustering. These methods are distance specific method and select the data eligibility to a cluster based on mean value observation. The clustering process is also sensitive to the outlier and able to reduce the effect of such abnormal data values. It is considered as the simplest clustering method and used as the integrated stage for many of the data processing and image processing applications.

### B) Overlapped Clustering

This clustering is applied on a large group of data values available in higher dimension space with dense data existence. As the distance between the data elements is lesser and larger data frequency, so that the election of data cluster is more critical. A data element can exist in more than one cluster because of which it is called overlapped clustering. In most cases, the numbers of clusters are not fixed at the earlier stage. Based on the dynamic analysis and feature specific observation, the clusters are identified. The clustering process of this method is shown in figure 2.

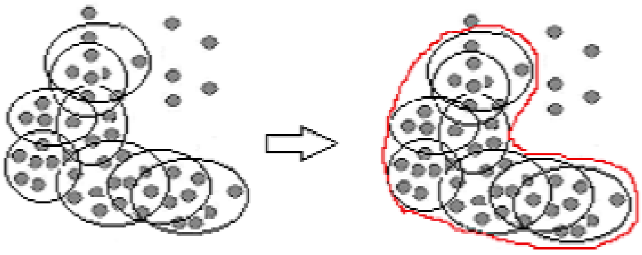

Figure 2 : Overlapped Clustering[3]

The figure shows that the method is showing the shape specific clustering with radial region specification. The data values that come under in the radial range of a cluster, are considered as the cluster member. This overlapped clustering method includes DBSCAN (Density Based Spatial Clustering of Applications with Noise), DENCLUE (Density Based clustering) etc. To apply the filter or the outlier detection, the threshold limits or the range specifications can be applied on radial parameter or on number of clusters.

### C) Hierarchical Clustering

As the name suggest, this clustering method is applied under cluster formation in a series by applying the cluster merge and split methods. This clustering technique is further classified in two main approaches called agglomerative clustering and divisive clustering. In first form of this clustering, the smaller clusters are first formed and later on the merge operation is applied on these clusters to generate the larger clusters. In the second clustering approach, the process starts from the larger cluster space and divison specific method is applied. The split method is applied to divide the clusters. The process of marge and split is also directed with specification of associated process and the stop condition specification to control the clustering. The criteria driven clustering is here applied in this stage. The basic process of this clustering method is shown in figure 3.

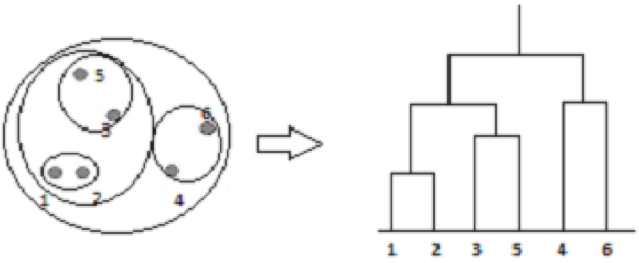

Figure 3 : Hierarchical Clustering[3]

The figure here is showing the stage driven clustering process which can either move from larger to smaller cluster or from smaller to larger. The clustering algorithms coms under this category includes BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) and CURE (Clustering using Representatives)

In this paper, a study on different clustering methods is provided. The work is here defined to identify scope of clustering along with method exploration. In this section, the clustering process is defined. The section has divided the available methods in three main categories. In section II, the work defined by earlier researchers is described. In section III, the method exploration of some of common clustering method is done. In section IV, the conclusion of this work is presented.

## II. RELATED WORK

Clustering is the base procedure defined in different applications and methods as individual process as well as the sub stage. In this section, some of the work defined by earlier researchers on clustering methods is explored. The section has discussed the contribution of earlier researchers on clustering methods.

The base concept of clustering method, cluster analysis and its primitive exploration was provided by Rui Xu et. al.[2]. Author has defined a knowledge driven method to provide the statistics driven clustering in the real environment. A knowledge driven analysis and the relative feature specific cluster formation is here defined in the paper. Author identified the intensive work efforts by applying work on some common benchmark databases. The cluster validation and proximity measures are also applied to improve the clustering results. Ahalya et. al.[7] has defined the exploration specific survey on different data clustering methods. Author applied the disparity analysis on different objects and group identification of these objects. A partition specific clustering method was here defined using self organized features. Author identified the emthod to improve the expectation maximization and to provide the effective cluster system generation. The method driven analysis and the quality and performance driven measures were discussed by the author. Zhang et. al[8] provided a study on reconsideration rules of clustering in comprehensive way. Author identified various clustering issues and associated application specific problems. The result specific observations are taken to provide the analysis under similarity measure, tendency analysis and the cluster validity formulation. A component driven graph measuring was provided by the author to improve the cluster formation. Amini et. al.[12] has provided a study on density driven and grid based clustering methods. Author provided the work on finite number of clusters and specific number of cluster formation in terms of algorithm specification. The area driven analysis is here provided for data structure observation. Based on this structural analysis, the clustering is applied. Ayed et. al.[13] has provided the study on different fuzzy clustering methods for big data. Author provided the Kmeans algorithm and mixture model analysis is here provided to generate the data variant. The hierarchical clustering method is here defined to generate the feature vector for insensitive measure is defined for effective cluster formation. The issues and characterization was provided relative to big data.

Different researchers provided the evaluation on different clustering method or some improvement in terms of parameter updation or the quality measures. Nisha et. al.[4] has provided a work on phase driven data mining method to improve the cluster formation. Author applied the work on hierarchical clustering to measure the cluster quality. The cohesion method, Elapsed time based analysis was provided by the author. Fenyi et. al.[5] has provided a work on grey integrated clustering method using comprehensive measure applied to explore the clustering results. The integrated environment is defined to provide the boundary driven estimation is provided for cluster formation. An instance specific analysis is here provided to improve the cluster formation and to generate the prominence difference measure specific cluster formation. An improvement to the spectral clustering was suggested by Yang et. al.[6] using multi task association specific work. Author generated the correlation analysis and provided the sample driven property exploration. A correlation analysis based improvement to intertask clustering is provided to form the relational learning. The cluster labels are assigned with regular control and coherence analysis. A task driven analysis is applied to predict the cluster label. The discriminative spectral clustering was here generated under extensive behavior analysis to achieve the task driven measure. Wang et. al.[9] has provided an unsupervised relational analysis and feature extraction method for defining a new clustering algorithm. An entity pair analysis and relationship observation was provided by the author. Author also applied the co-clustering theory and defined the characteristic driven relational observation to achieve the duality specific clustering. The result analysis under relational observation was provided by the author to generate the effective cluster formation. Data duality analysis was the key factor of this improved algorithmic approach. Huang et. al.[14] has provided the uncertain instance analysis for improving the clustering results. A heterogeneous clustering method is here defined for feature specific estimation. The probability driven frequency measures are applied to categorize the attributes and provided the cluster driven measures under uncertain behavior of data. The cluster quality analysis under different measures is observed. The micro data processing under this improved clustering form is provided for effective data clustering.

The basic clustering algorithms are also improved by including some optimization measure, method or constraints. A Study based on different evolutionary clustering methods was provided by Hruschka et. al.[1]. The clustering exploration is here not restricted to one technique instead, the multi objective and ensemble based methods are covered by the author. The context driven analysis on specification of different clustering parameters and their impact on clustering process was explored. The paper identified the scope of these algorithms in relation with different real time applications. The clustering issues and associated solutions are also explained by the author. Patel et. al.[3] has improved the partitioned clustering by combing the method with PSO approach. Author compares different constraints of PSO and integrated it with partition clustering to achieve the performance and quality driven optimization. The quality, performance and the issues of the partition methods are resolved by the author. Another optimization method using PSO approach for consensus clustering was provided by Esmin et. al.[10] Author has emerged the prominent method with stability and accuracy parameters. The work is applied on five different dataset to prove the robustness and the scope of the work. The method improved the accuracy and reduces the cluster loss from the work. Julrode et.al.[11] has provided the scope of PCA (Principal Component Analysis) and ACO (Ant Colony Optimization) with fuzzy clustering and K-harmonic means clustering methods. The optimization methods are defined to identify the effective number of clusters and elect the effective cluster members. The distance based feature mapping was applied to identify the effective cluster members. The distance driven optimization also reduce the error rate.

## III. CLUSTERING METHODS

The clustering is having the higher significance for knowledge filteration and improving the predictive results. Based on the data requirements, there are number of available clustering approaches to provided controlled segmentation. These approaches are based on the clustering category, algorithmic formulation, distance method etc. Some of the common clustering methods are discussed in this section.

### A) KMeans Clustering

KMeans clustering is one of the simplest algorithmic method to provide distance specific cluster formation. In many algorithmic approaches, it is used as the earlier stage to generate the interest data. The method is compact and based on hyper spherical features. The method uses the fixed cluster based distance analysis to identify the cluster members. It is effective for larger datasets with complexity O(NKd). The basic process of KMeans clustering is listed here under

1. Define the K-Partition based on data knowledge and specification of cluster centers
2. The distance comparison of data elements to the cluster center is done to identify the cluster member.
3. Estimation of the cluster center is done to identify the average of cluster member for prototype matrix
4. The process two and three till the stability in the clusters is not achieved.

The KMeans clustering is the simpler and robust method which provides the support to outlier and the noise identification. The distortion observation or the interest object identification can be identified easily using this method. The disadvantages of method include the inefficiency of method to generate the initial partition and optimal cluster member identification. The fix number of clusters is also a challenging which requires the earlier database knowledge and experience. KMeans clustering cannot guarantee the optimized cluster formation.

### B) FCM (Fuzzy CMeans)

FCM is another popular clustering method defined under the fuzzy characterization. The method is able to identify the clusters based on the fuzzy driven analysis. The method is able to minimize the cost of individual clusters as well as cluster formation. The method is variant to the clustering procss and applies the fuzzy rules to provide intensive investigation under distance measures. The weight exponent driven fuzzy control is provided to generate the partitions and to provide the effective partition formulation. The noise and outlier identification to the method is also formulated to estimate the cluster center effectively. The vertex driven analysis is defined based on the mountain function specification. The distance estimation between the object and the center is applied to observe the positive consent of the data existence in the center of the partition. The algorithmic specification of this method is given here under

1. Generate or select the effective clustering parameters including the cluster center, matrix prototype and the variation vector

2. The membership estimation based on the matrix derivation relative to center is evaluated.

3. The prototype matrix M is updated respective to the center and relatively data members are identified

4. Repeat the steps 2 and 3 till the clear cluster formulation is not done.

FCM is able to generate the clusters of different shapes such as ellipse, rings, rectangle etc. The data space utilization with surface formulation is provided by this method. The similarity measures with shape shell difference analysis are required to control the data existence behavior. The cluster structure observation is provided to effective cluster member election.

### C) Hierarchical Clustering

This clustering method basically organizes the data based on the hierarchical structure relative to the proximity matrix. The binary tree driven dendogram formulation is the key featured architecture of this clustering method. The method is able to process the data objects and provides the intermediate data formulation with structure specification. The data object analysis under distance vector is applied to generate the clear clusters with informative descriptions. The data visualization with potential structure analysis is provided to achieve data divisions. The method uses the merging or the split method for cluster formation. The subset level division is here the key factor. For N object there can be 2N-1-1 possible subset divisions. The expensive computation can be applied for cluster formation. A complete data linkage method is here provided to identify the data members for each clusters. The feature driven analysis is applied based on method summarization. The inter-cluster analysis with distance estimation is provided to control the clustering behavior. The noise estimation and the relative misclassification will be achieved to reduce the computational complexity. The method is based on the hierarchical structure so that the tendency of work algorithm will be reduced. The structured data analysis is here provided based on partition matrix to generate the cluster membership with shape robustness. The centroid specific shape analysis is here provided to achieve the cluster formation. The basic procedure of this method is listed here under

1. Process the algorithm with specification of each data member as a cluster and apply the proximity cluster formulation.

2. Apply the minimum distance analysis between the cluster centers and cluster members and identify the eligible clusters.

3. Apply the cluster distance analysis for cluster merge to combine the clusters

4. Update the proximity matrix for recomputation of distance between the clusters

5. Repeat steps 2 to 5, till the cluster switching exist.

In this section, three main clustering algorithms are explained with process level specification. The strengths and features of each method is also explored.

## IV. CONCLUSION

Clustering is the feature generation process applied on different data spaces to acquire the required information set. The clustering actually divides the data in smaller segments so that each segment represents a particular data form. In this paper, the scope and importance of clustering method is explained. The different categories of clustering algorithm are also defined. The paper also defined the algorithmic stages of three main clustering algorithms called KMeans, CMeans and hierarchical clustering.

## REFERENCES

[1] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas and A. C. Ponce Leon F. de Carvalho, "A Survey of Evolutionary Algorithms for Clustering," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 39, no. 2, pp. 133-155, March 2009.

[2] Rui Xu and D. Wunsch, "Survey of clustering algorithms," in IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, May 2005.

[3] G. K. Patel, V. K. Dabhi and H. B. Prajapati, "Study and analysis of particle swarm optimization for improving partition clustering," Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in, Ghaziabad, 2015, pp. 218-225

[4] Nisha and P. J. Kaur, "Cluster quality based performance evaluation of hierarchical clustering method," Next Generation Computing Technologies (NGCT), 2015 1st International Conference on, Dehradun, 2015, pp. 649-653.

[5] D. Fenyi, L. Junjuan and L. Bin, "Study on improved grey integrated clustering method and its application," 2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009), Nanjing, 2009, pp. 702-707.

[6] Y. Yang, Z. Ma, Y. Yang, F. Nie and H. T. Shen, "Multitask Spectral Clustering by Exploring Intertask Correlation," in IEEE Transactions on Cybernetics, vol. 45, no. 5, pp. 1083-1094, May 2015.

[7] G. Ahalya and H. M. Pandey, "Data clustering approaches survey and analysis," Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on, Noida, 2015, pp. 532-537.

[8] C. Zhang, Q. Xia and G. Yang, "Reconsideration about clustering analysis," Industrial Electronics and Applications (ICIEA), 2015 IEEE 10th Conference on, Auckland, 2015, pp. 1517-1524.

[9] J. Wang, Y. Jing, Y. Teng and Q. Li, "A novel clustering algorithm for Unsupervised Relation Extraction," Digital Information Management (ICDIM), 2012 Seventh International Conference on, Macau, 2012, pp. 16-21.

[10] A. A. A. Esmin and R. A. Coelho, "Consensus Clustering Based on Particle Swarm Optimization Algorithm," 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, 2013, pp. 2280-2285.

[11] P. Julrode, S. Supratid and U. Suksawatchon, "A performance comparison of using principal component analysis and ant clustering with fuzzy c-means and k-harmonic means," Computational Intelligence and Cybernetics (CyberneticsCom), 2012 IEEE International Conference on, Bali, 2012, pp. 123-128.

[12] A. Amini, T. Y. Wah, M. R. Saybani and S. R. A. S. Yazdi, "A study of density-grid based clustering algorithms on data streams," Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on, Shanghai, 2011, pp. 1652-1656.

[13] A. Ben Ayed, M. Ben Halima and A. M. Alimi, "Survey on clustering methods: Towards fuzzy clustering for big data," Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of, Tunis, 2014, pp. 331-336.

[14] G. Y. Huang, D. P. Liang, C. Z. Hu and J. D. Ren, "An algorithm for clustering heterogeneous data streams with uncertainty," 2010 International Conference on Machine Learning and Cybernetics, Qingdao, 2010, pp. 2059-2064.