A Survey of Various Clustering Techniques

[1]Brahm Prakash Dahiya, [2]Dr. Shaveta Rani, [3]Dr. Paramjeet Singh,
[1]Research Scholar, CSE, I.K.G. Punjab Technical University, Punjab, India.
[2]Professor, CSE, Giani Zail Singh Campus College of Engineering and Technology, Punjab, India.
[3]Professor, CSE, Giani Zail Singh Campus College of Engineering and Technology, Punjab, India.

Abstract- *Clustering is collection of objects. It is unsupervised classification that uses no label for patterns. It is also known as superset of unsupervised classification. It is divided into two categories such as hierarchical and partitioning techniques respectively. Hierarchical techniques are combination of layers nodes and each node represents a cluster. Clusters are represented in top-down or bottom-up style. Partitioning techniques represent natural grouping of data by single partition. Further each clustering techniques are subdivided in various categories and also discuss the various clustering algorithms that are applicable in clusters selection and creation.*

Keywords- Clustering, Hierarchical clustering, Partitioning clustering, ROCK and CURE.

I. INTRODUCTION

Clustering is grouping a set of data in a way that maximizes the matching within clusters and minimizes the matching between two different clusters. Grouping of object are applicable in different field such as engineering technology, health care, agriculture, weather forecasting, medical science and bioinformatics. Similar application in Wireless sensor network, the data together by many nodes will be similar. In such cases, duplicity of data transmission can be avoided by forming grouping of sensor nodes called clusters and perform election to elect cluster head among the nodes in the cluster. On the basis of various previous researches in many scenarios, no such grouping information is provided in advance. To resolve this problem, we come on classification based on grouping. There are supervised and un-supervised classifications based on grouping. The objectives of classification are discovered tool, novel techniques and algorithm. It is known as classifier. In classification process objects are represented by instance or pattern. The pattern combination of number of classification defined with attributes (elements). Classifier accuracy measurement depends upon successfully testing of patterns. Many supervised classification studied and discovered. In supervised classification given label pattern. In other side unsupervised assigned no label for any pattern. Clustering is superset of unsupervised classification. In which patterns are without labeling that’s why, working with clustering is more difficult as compare to supervised learning. In supervised classification becomes an idea to combining data objects as a whole. But in clustering, it is very difficult to identify pattern of group without label. There can be features or parameters which could be suitable for clustering [1-9].The paper will focus on various clustering techniques in different section. In Section-2, we will be discussed various clustering techniques. In last section conclusions find out of the summary.

II. CLUSTERING TECHNIQUES

Based on various literature surveys, many clustering techniques have discussed. But it is very difficult to find out an exact definition of cluster based on the various studies. Various clustering techniques have been proposed with different set of rules and principals [9, 10].To makes identification to each clustering techniques, many suggestion have come in knowledge. That is why; clustering approaches are divided into two different categories: hierarchical and partitioning techniques that will discuss in details.

2.1 Hierarchical clustering (HC) methods:

Hierarchical techniques are combination of layers nodes and each node represents a cluster. Clusters are represented in top-down or bottom-up style. Further hierarchical methods are divided into two forms i.e. agglomerative and divisive. The agglomerative is the bottom-up approach. Clusters are starting with individual object and combine these single objects into larger clusters until termination condition do not meet.

Divisive hierarchical approach is using top-down fashion. It divide large cluster in several smaller clusters and cluster division continuous until it satisfies certain closure situations. In hierarchical approach cluster is formatted in dendrogram style. The hierarchical clustering methods are divided in three classifications based on correlation or connections. Following classification of the hierarchical clustering methods is discussed below.

2.1.1 Single-linkage clustering: Single-linkage is very simple agglomerative hierarchical clustering method. It is also called as nearest neighbour method or connectedness or minimum method. Based on distance factor between two clusters, it defines minimum distance between an object in inner cluster and an object in outer cluster. In Figure 2.1 shows an example of the single linkage distance between two clusters. When small number of object exists between two relatively distinct clusters called chaining. In which similarity is counted between two or more clusters [9].

$$\hat{\text{ɗ}},T\text{single} \quad (d1,d2)=\arg \min_{\substack{x1\ \varepsilon\ Td1 \\ x2\ \varepsilon\ Td2}} \text{ɗ}\,(x1,x2) \qquad (1)$$

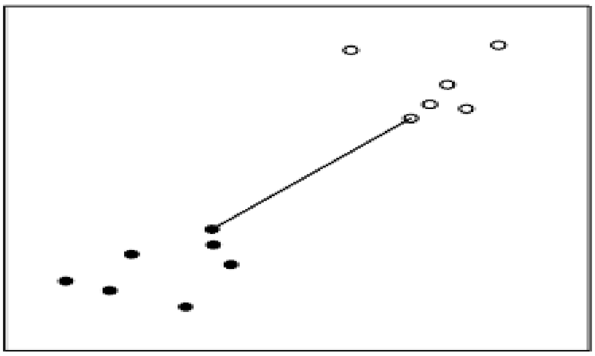


Figure 1. Single linkage distance between two clusters

Where Td defined as set of training objects allocated to cluster. When cluster with many distinct objects may be choose for combining with single linkage.

2.1.2 Complete-linkage clustering:

Complete-linkage clustering depends upon the dissimilarity between two objects instance from clusters. It counts distance between two group is maximum distance between an object instance in one cluster and an object instance in the other cluster. It is also identified with roughly diameters [17].

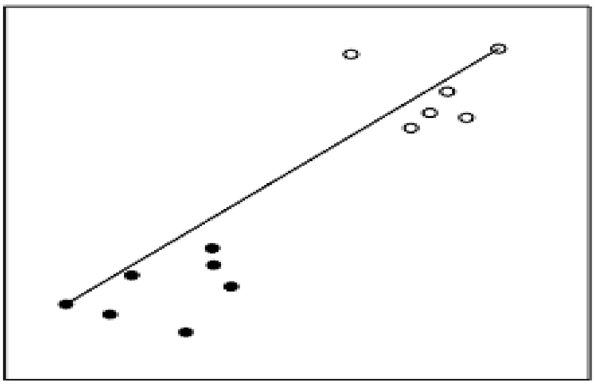


Figure 2 Complete linkage distance between two clusters

Figure 2 describes the distance factor between two clusters. It is more useful in real world application. It is very complex to even temperate outliers. It formed tightly coupled, compact and more suitable clusters. It combines two clusters to create a large cluster if size is below to predefined threshold values.

2.1.3 Average-linkage clustering:

Average-linkage clustering is working on average distance factor between two clusters. It is based on minimum variance approach. It is calculated average distance between two pairs of the objects in which each pair include one cluster. Average-linkage clustering is classified in two categories: UPGMA (Unweight pair group method with arithmetic mean) and WPGMA (weighted pair group method with arithmetic means). WPGMA is used in variant calculation [10].

2.1.4 Enhanced hierarchical clustering:

Hierarchical clustering [11] is suffered with next movement in cluster hierarchy when two points of clusters are connected to each other. So enhance version of hierarchical clustering has introduced. It is divided in following categories:

2.1.4.1 Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH):

It makes hierarchical data structure called cluster features (CF). It performs partition of incoming data points using ascending and automatic methods. It is also known as height-balanced hierarchy and used CF based on two factors: branching factor B and threshold T which define diameter of cluster and each cluster should be less than T. Threshold T manages the compress ratio of data. Computational complexity of BIRCH is O (N) [12, 13].

2.1.4.2 Clustering Using Representatives (CURE):

In CURE clusters are represented in various shapes and size using number of disperse points. It is identified scope of clusters and worked with large scale database. It is commonly used for two dimensional databases. It gives good result and quality using outliers as compare to BIRCH. CURE complexity is 0 (N2 log N). Based on computational performance BIRCH is better than CURE. It represents clusters with all point and solves the problem of BIRCH that uses single centroids to represent clusters. It is reduced noise point [14].

2.1.4.3 ROCK (Robust Clustering using links):

It is class of class of agglomerative hierarchical clustering algorithms. It use links instead of distance factors. It creates better quality clusters than other existing algorithms. It is applicable for large datasets [15].

2.1.4.4 CHAMELEON:

This is agglomerative hierarchical clustering. :

In which clusters are combined only if the connectivity and similarity between of two clusters are highly correlated with that interconnectivity of the clusters. Using graph separation algorithm divided the data into small clusters and applied agglomerative hierarchical algorithm to looping combining cluster and produced final output. It produced high quality clusters with perfect shapes [16].

2.2 Partition clustering methods:

Partition clustering methods are totally different from hierarchical clustering. It represented natural grouping of data by single partition. These methods are classified in following categories: k -means, k-medoids, and Fuzzy c - means etc [18].

2.2.1 k -means clustering:

It is more applicable partition clustering algorithm. It is also called as bench marked and simplest clustering that solve clustering problem. It works based on K centroids and data set represents using user defined number of K-clusters. Initial state is started after selecting K centroids points and Clusters centroids points are updated after formation of cluster in every cycles. It is also known as greedy algorithm and applied iterative methods to change membership's function. To find out closest centroid many proximity measurements are used in K-means algorithm. Main selection depends upon quality of cluster [18] [19].

Algorithm:

1. Initialization: Choose k-cluster as initial centroids.
2. From k clusters by assigning each point to its closest centroid.
3. Re-compute and recalculate the positions of the k centroids.
4. Perform repetition until convergence criterion is met.

Many choices are used in K-means algorithms such as Manhattan distance (L1 norm), Euclidean distance (L2 norm) and cosine. Euclidean distance (L2 norm) is most advanced and popular choice of K-means algorithms. Objective function is residual sum of squares (RSS). The mathematical representation is discussed below.

$$\begin{aligned} SSE(C) &= \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - c_k \|^2 \\ c_k &= \frac{\sum_{x_i \in C_k} x_i}{|C_k|} \end{aligned} \tag{2}$$

Dataset is defined D={x1,x2,…………xN} contains of N points. Next cluster is formatted after applying K-means clustering by C={C1,C2,….,Ck……,CK}. RSS is represented in equation …….. where Ck is the centroid of cluster CK. In K-means selection the initial centroids and estimating the number of clusters k is more challenging issues.

2.2.2 k-medoids clustering:

The K-medoids clustering algorithm is also known as partition around medoids. It provides clustering solution that minimizes a predefined objective function. It selects the actual data points as the rules. It is more robust to noise. It works with minimization of the absolute error criterion. According to K-medoids algorithm [20].

Algorithm:

1. Select K points as the initial representative objects.
2. Repeat
3. Assign each point to the cluster with the nearest representative object.
4. Randomly select a non-representative object xi.
5. Compute the total cost S of swapping the representative object m with xi.
6. If S<0, then swap m with xi to form the new set of K representative objects.
7. Until Convergence criterion is met.

2.2.3 The K- Medians Clustering:

It is different side of K-means to calculate the median for every cluster. It selects K cluster centers to minimize the sum of a distance measure between each point and closest cluster center. Here L1 norms are used to measure distance. The K- medians clustering equation define below [21]:

$$S = \sum_{k=1}^{K} \sum_{x_i \in C_k} |x_{ij} - med_{kj}| \tag{3}$$

The K-means clustering equation is represented by S. where x_{ij} represents the jth attribute of the instance xi and med_{kj} represents the median for the jth attribute in the kth cluster Ck. It is more robust to outliers as compared to K-means.

2.2.4 The K- Modes Clustering:

K-means is unable to work with non-numerical attributes. Based on some data transformation methods, k-means algorithm is applied to find new clusters. The k-means algorithm is less effective and did not produce best cluster. So, the k-modes clustering algorithm has proposed to avoid the limitation of k-means clustering [22].

Algorithm k-modes clustering:

1. Select K initial modes.
2. Reparation
3. From K clusters by assigning all the points to the cluster with nearest mode using the matching metric.
4. Re-compute the modes of the clusters.
5. Until convergence criterion is met.

2.2.5 Fuzzy k-means clustering:

It is also known as Fuzzy C-Means clustering. The existing algorithms are not capable to handle complex datasets where there are overlapping clusters. Using fuzzy K-means clustering algorithm resolve complex datasets and overlapping clustering. In which membership of points to different clusters can vary 0 to 1 [23].

$$E(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} w_{xik}^{\beta} \parallel x_i - c_k \parallel^2$$

$$w_{xik} = \frac{1}{\sum_{j=1}^{K} \left(\frac{x_i - c_k}{x_i - c_j}\right)^{\frac{2}{\beta - 1}}}$$

$$c_k = \frac{\sum_{x_i \in C_k} w_{xik}^{\beta} x_i}{\sum_{x_i \in C_k} w_{xik}} \quad (4)$$

Where w_{xik} is the membership weight of point xi belonging to Ck. Modified steps of Fuzzy K clustering is calculated using membership weight based on centroid Ck.

2.2.1.6 Intelligent k-means clustering:

This idea based on principal component analysis (PCA).It selects those points farthest from the centroid that correspond to the maximum data scatter. The cluster derived from anomalous pattern clusters point. It used in extracting clusters. It can also be used for initial centroid selection [24].

Algorithm:

1. Calculate the center of gravity for the given set of data point’s cg.
2. Repeat
3. Create a centroid c farthest from Cg.
4. Create a cluster S_{iter} that is closer to c compared to Cg by assigning all the remaining data points x_i to S_{iter} if $d(x_i,C) < d(x_i,Cg)$.
5. Update the centroid of Siter as Sg.
6. Set Cg=Sg.
7. Discard small cluster (if any) using a pre-specified threshold.
8. Until stopping criterion is met.

2.2.7 Bisecting K-means clustering:

It uses K-means repeating on the parent cluster C to determine the best possible split to obtain two child clusters C1 and C2. It provides uniform sized clusters [25].

Algorithm:

1. Repeat
2. Choose the parent cluster to be split C.
3. Repeat
4. Select two centroids at random from C
5. Assign the remaining points to the nearest sub-cluster using a pre-specified distance measure.
6. Recomposed centroids and continue cluster assignment until convergence.
7. Calculate inter-cluster dissimilarity for the 2 sub-cluster using the centroids.
8. Until / iterations are composed.
9. Choose those centroids of the sub-clusters with maximum inter-cluster dissimilarity.
10. Split C as C1 and C2 for these centroids.
11. Choose the largest cluster among C1 and C2 and set it as the parent cluster.
12. Until K clusters have been obtained.

III. CONCLUSION

Clustering is more useful in cluster selection and formation. Clustering is divided into two categories, hierarchical and partitioning techniques respectively. Hierarchical techniques are combination of layers nodes and each node represents a cluster. Clusters are represented in top-down or bottom-up style. Partitioning techniques represent natural grouping of data by single partition. Further each clustering techniques are subdivided in various categories and also discuss the various clustering algorithms that are applicable in clusters selection and creation. In future work we will implement different clustering algorithm for suitable cluster head selection in Wireless Sensor Networks.

ACKNOWLEDGEMENT

The authors acknowledge I. K. Gujral Punjab Technical University, Kapurthala for providing research facilities.

REFERENCES

[1] Rumelhart , J.L. McClelland , Parallel Distributed Processing, MIT Press, Cambridge, 1986.

[2] W. Zhou , Verification of the nonparametric characteristics of back-propagation neural networks for image classification, IEEE Trans. Geosci. Remote Sens. vol no 37 (2),pp.771–779, 1999.

[3] G. Jaeger , U.C. Benz , Supervised fuzzy classification of SAR data using multiple sources, IEEE Int. Geosci. Remote Sens. Symp ,1999.

[4] F.S. Marzano , D. Scaranari , G. Vulpiani , Supervised fuzzy-logic classification of hydrometeors using C-band weather radars, IEEE Trans. Geosci. Remote Sens. vol-45(11), pp. 3784–3799,2007.

[5] B. Xue , M. Zhang , W.N. Browne ,Particle swarm optimization for feature se- lection in classification: a multi-objective approach, IEEE Trans. Cybern. vol no-43(6), pp. 1656–1671,2007 .

[6] Saxena , M. Vora , Novel approach for the use of small world theory in par- ticle swarm optimization, in: Proceedings of the Sixteenth International Conference on Advanced Computing and Communications, 2008 .

[7] Z. Pawlak , Rough sets, Int. J. Comput. Inf. Sci. vol no.-11(5), 341–356,1982.

[8] S. Dalai , B. Chatterjee , D. Dey , S. Chakravorti , K. Bhattacharya , Rough-set-based feature selection and classification for power quality sensing device employing correlation techniques, IEEE Sens. J. vol no 13(2), pp. 563–573,2013.

[9] P. Sneath , R. Sokal , Numerical Taxonomy, W.H. Freeman Co, San Francisco, CA, 1973 .

[10] J.H. Ward , Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. vol no.-58(301),pp. 236–244,1963 .

[11] A . Nagpal , A . Jatain , D. Gaur , Review based on data clustering algorithms, in: Proceedings of the IEEE Conference on Information and Communication Technologies, 2013.

[12] Periklis , Data Clustering Techniques, University of Toronto, 2002 .

[13] T. Zhang , R. Ramakrishnan , M. Linvy , BIRCH: an efficient method for very large databases, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, 1996 .

[14] S. Guha , R. Rastogi , S. Kyuseok , CURE: An Efficient Clustering Algorithm For Large Databases, ACM, 1998.

[15] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: An e_cient data clustering method for very large databases. In Proceedings of the ACM SIGMOD Conference on Management of Data, pp.103-114, Montreal, Canada, June 1996.

[16] K. George , E.H. Han , V. Kumar , Chameleon: a hierarchical clustering algorithm using dynamic modeling, IEEE Comput. vol no. 32(8),pp. 68-75,1999.

[17] N. Abd Rahman, Z. Abu Bakar and N. S. S. Zulkefli, "Malay document clustering using complete linkage clustering technique with Cosine Coefficient," 2015 IEEE Conference on Open Systems (ICOS), Melaka, pp. 103-107,2015.

[18] D. Lam , D.C. Wunsch , Clustering, academic press library in signal processing, Signal Process. Theory Mach. Learn. Vol no-1, pp. 115–1149.

[19] J.B. MacQueen , Some methods for classification and analysis of multivari- ate observations, in: Proceedings of the Fifth Symposium on Mathematical Statistics and Probability, vol. 1, Berkeley, University of California Press, pp. 281–297,1967 .

[20] M. O. Shafiq and E. Torunski, "A Parallel K-Medoids Algorithm for Clustering based on MapReduce," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA,pp. 502-507,2016.

[21] M. S. Premkumar and S. H. Ganesh, "A Median Based External Initial Centroid Selection Method for K-Means Clustering," 2017 World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, Tamil Nadu, India,pp. 143-146,2017.

[22] L. Tao-ying, C. Yan, J. Zhi-hong and L. Ye, "Initialization of k-modes clustering for categorical data," 2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings, Harbin, pp. 107-112, 2013.

[23] C. T. Baviskar and S. S. Patil, "Improvement of data object's membership by using Fuzzy K-Means clustering approach," 2016 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC), Chennai, pp. 139-147,2016.

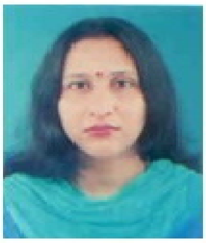
[24] M. A. Ma'sum, I. Wasito and A. Nurhadiyatna, "Intelligent K-Means clustering for expressed genes identification linked to malignancy of human colorectal carcinoma," 2013 International Conference on Advanced

Computer Science and Information Systems (ICACSIS), Bali, pp. 437-443,2013.

[25] S. Banerjee, A. Choudhary and S. Pal, "Empirical evaluation of K-Means, Bisecting K-Means, Fuzzy C-Means and Genetic K-Means clustering algorithms," 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Dhaka, 2015, pp. 168-172, 2015.

Author's Profile:

Mr. Brahm Prakash Dahiya got his Mr Brahm Prakash Dahiya pursed B.Tech degree (Information Technology) from P.T.U. University Punjab, India, in 2009 and received his M.Tech in (Computer Science and Engineering) from Maharshi Dayanand University, Rohtak, Haryana, India, in 2011. He is pursuing his Ph.d Degree from I. K. Gujral Punjab Technical University, Jalandhar, India. He has published more than 25 research papers in reputed national and international journals and conferences. His research interests include in Soft-Computing, Image Processing, and Wireless Sensor.



Dr. Shaveta Rani pursed B.Tech. (CSE) and M.S. (Software Systems) from BITS Pilani in 2009. She has completed Ph.D. from BITS Pilani in 2009 and currently working as Professor in Department of Computer Science & Eng., Giani Zail Singh Campus College of Engineering and Technology, Punjab. She has published more than 80 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE, Her main research work focus on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT, Wireless Sensor Network and Computational Intelligence based education. She has 18 years of teaching and Research Experience.

Dr. Paramjeet Singh pursed B.Tech. (CSE) and M.S. (Software Systems) from BITS Pilani in 2009. He has completed Ph.D. from BITS Pilani in 2009 and currently working as Professor in Department of Computer Science & Eng., Giani Zail Singh Campus College of Engineering and Technology, Punjab.He has published more than 80 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE, His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT, Wireless Sensor Network and Computational Intelligence based education. He has 18 years of teaching and Research Experience