A Detailed Review on Data Mining Clustering Algorithms

KolluruVenkata Nagendra[1], V.Bharati[2], N.Sivanagamani[3], R.Sivaiah[4]
[1, 4] Assistant Professor, Dept of CSE
[2, 3] Assoc.Professor, Dept of CSE
[1, 2] Geethanjali Institute of Science & Technology, Nellore, Andhra Pradesh, India

Abstract- *Clustering is a significant task in data analysis and data mining applications. It is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Data mining can do by passing through various phases. Mining can be done by using supervised and unsupervised learning. The clustering is unsupervised learning. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. In this survey paper, an analysis of clustering and its different techniques in data mining is done.*

Keywords- Data mining, Clustering, Types of Clustering, Supervised, Unsupervised.

I. INTRODUCTION

Clustering is a major task in data analysis and data mining applications. It is the assignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups. Cluster is an ordered list of data which have the familiar characteristics. Cluster analysis can be done by finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. Clustering is an unsupervised learning process. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. The superiority of a clustering result depends on equally the similarity measure used by the method and its implementation. [1].

The Cluster analysis groups data objects based only on the information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in the other groups. The greater similarity (or homogeneity) of clustering is within a group, and the greater the difference between groups, the better or more distinct the clustering [2]. Data may be thought of as points in a space where the axes correspond to the variables. The cluster analysis divides the space into regions, characteristic of the groups found in the data. The main advantage of a clustered solution is automatic recovery from failure, that is, recovery without user intervention. The disadvantages of clustering are complexity and inability to recover from database corruption. An ordered list of objects, which have some common characteristics of cluster. The objects belong to an interval [a, b], in our case [0, 1] [3]. The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed [4]. A similarity measure SIMILAR (Di, Dj) can be used to represent the similarity between the documents. Typical similarity generates values of 0 for documents exhibiting no agreement among the assigned indexed terms, and 1 when perfect agreement is detected. Intermediate values are obtained for cases of partial agreement [5]. If the similarity measure is computed for all pairs of documents (Di, Dj) except when i=j, an average value AVERAGE SIMILARITY is obtainable. Specifically, AVERAGE SIMILARITY = CONSTANT SIMILAR (Di, Dj), where i=1,2,….n and j=1,2,….n and i< > j. The lowest possible input value of similarity is required to join two objects in one cluster. The similarity between objects calculated by the function SIMILAR (Di,,Dj), represented in the form of a matrix is called a similarity matrix. The dissimilarity coefficient of two clusters is defined as the distance between them. The smaller the value of the dissimilarity coefficient, the more similar the two clusters are.
The first document or object of a cluster is defined as the initiator of that cluster, i.e., similarity of every incoming object's is compared with the initiator [6]. The initiator is called the cluster seed. The procedure of the cluster analysis with four basic steps is as follows:

- Feature selection or extraction.
- Clustering algorithm design or selection.
- Cluster validation.
- Results interpretation
-

In Data Mining the two types of learning sets are used, they are supervised learning and unsupervised learning.

a) *Supervised Learning:* In supervised training, data includes together the input and the preferred results. It is the rapid and perfect technique. The accurate results are recognized and are given in inputs to the model through the learning procedure. Supervised models are neural network, Multilayer Perceptron and Decision trees.

b) *Unsupervised Learning*: The unsupervised model is not provided with the accurate results during the training. This can be used to cluster the input information in classes on the basis of their statistical properties only. Unsupervised models are for dissimilar types of clustering, distances and normalization, k-means, self organizing maps.

II. GOALS OF CLUSTERING

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data but how to decide what constitutes a good clustering? It can be shown that there is no absolute "best criterion which would be independent of the final aim of clustering. Consequently, it is a user which must supply this criterion, in such a way that the results of clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction) , in finding " natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection). [7].

III. GENERAL TYPES OF CLUSTERS

Well-Separated Clusters: If the clusters are sufficiently well separated, then any clustering method performs well. A cluster is a set of node such that any node in a cluster is closer to every other node in the cluster then to any node not in the cluster.[8].

Center-Based Clusters: A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the "center" of a cluster, than to the center of any cluster other then it. The center of a cluster is often called as centroid, the average of all the points in the cluster, or a mediod, the most "representative" point of a cluster.

Contiguous Clusters (Nearest neighbor or Transitive): A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster.

Density-Based Clusters: A cluster is a dense region of points, which is separated by according to the low-density regions, from other regions that is of high density. Used when the clusters are irregular, and when noise and outliers are present.

Conceptual Clusters: Shared property or Conceptual Clusters that share some common property or represent a particular concept.

IV.TYPES OF CLUSTERING ALGORITHMS

Clustering is a division of data into groups of similar objects [9]. The clustering algorithm can be divided into five categories, viz, Hierarchical, Partition, Spectral, Grid based and Density based clustering algorithms. The overview of clustering algorithms is depicted in Figure 1.

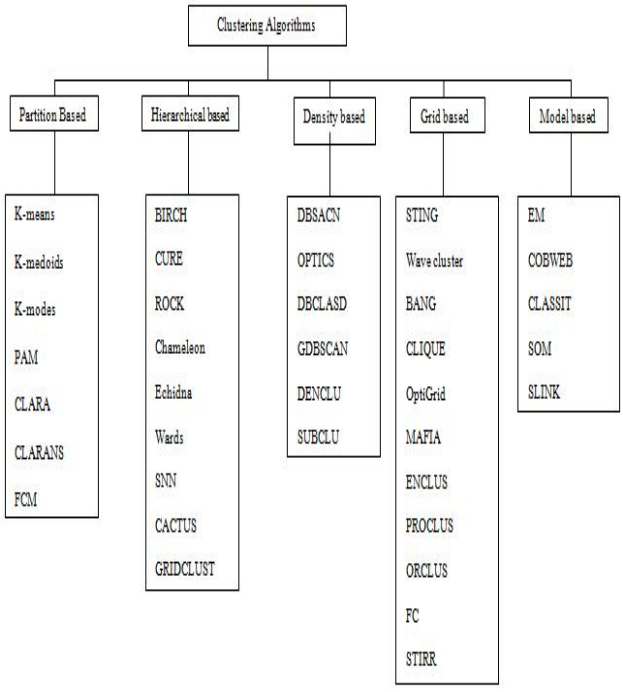

Figure (1). An overview of clustering algorithms.

The overall aim of the data mining process is to separate the information from a large data set and transform it into an understandable form for further use. The clustering algorithm can be divided into different categories and the description of each algorithm, advantages and disadvantages are described in the below table 1. The computational complexity of some typical and classical clustering algorithms in Table 1 with several newly proposed approaches specifically designed to deal with large-scale data sets. The cluster algorithm examines unlabeled data, by either constructing a hierarchical structure, or forming a set of groups, according to a pre specified number [10].

Table-1: Clustering Methods-Advantages, Disadvantages and its Complexities

Clustering Method	Description	Advantages	Disadvantages	complexity
K-Means	It is efficient in processing large data sets. It often terminates at a local optimum. The clusters have spherical shapes. It is sensitive to noise.	Ease of implementation and high-speed performance. Measurable and efficient in large data collection.	Selection of optimal number of clusters is difficult. Selection of the initial centroids is random	O(NKd) time O(N+ K) (space)
Fuzzy C Mean (FCM)	This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one	Gives best result for overlapped data set and comparatively better then k-means algorithm. Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.	Apriori specification of the number of clusters. With lower value of β we get the better result but at the expense of more number of iteration. Euclidean distance measures can unequally weight underlying factors.	Near O(N)
Hirarchical	It is of two types: Agglomerative Hierarchical clustering algorithm/AGNES (agglomerative Nesting) and Divisive Hierarchical clustering algorithm or DIANA (divisive analysis). Both this algorithm are exactly reverse of each other. So we will be covering Agglomerative Hierarchical clustering algorithm in detail.	No apriori information about the number of clusters required. Easy to implement and gives best result in some cases.	Time complexity of at least O(n2 log n) is required., where 'n' is the no. of data points. Based on the type of distance matrix chosen for merging different algorithms can suffer with one or more of the Sensitivity to noise & outliers. No objective function is directly minimized	O(N2) (time) O(N2) (space)
CLARA	The ClusteringLARgeApplication algorithm randomly chooses a small portion of the actual data as a representative of the data. If the sample is selected in a fairly random manner, it should closely represent the original dataset. CLARA draws multiple samples of the dataset, applies PAM to each sample, finds the medoids, and then returns its best clustering as the output.	It is simple to understand and implement. It takes less time to execute as compared to other techniques.	The drawback of this algorithm is the user has to provide pre-determined value of k and it produces spherical shaped clusters. It cannot handle with noisy data objects	O(K(40+K) 2+K(N-K))+ (time)
CLARANS	CLARANS is an efficient medoid-based clustering algorithm. In CLARANS, the process of finding k medoids from n objects is viewed abstractly as searching through a certain graph. In the graph, a node is represented by a set of k objects as	The advantage of the partition-based algorithms that they use an iterative way to create the clusters.	The number of clusters has to be determined in advance and only spherical shapes can be determined as clusters.	Quadratic in total performance

	selected medoids. Two nodes are neighbors if their sets differ by only one object. In each iteration, CLARANS considers a set of randomly chosen neighbor nodes as candidate of new medoids.			
BIRCH	IRCH (Balanced Iterative Reducing and Clustering using Hierarchies).It is a scalable clustering method. Designed for very large data sets. Only one scan of data is necessary. It is based on the notation of CF (Clustering Feature) a CF Tree. CF tree is a height balanced tree that stores the clustering features for a hierarchical clustering. Cluster of data points is represented by a triple of numbers (N, LS, SS) Where N= Number of items in the sub cluster.LS=Linear sum of the points.SS=sum of the squared of the points.	Finds a good clustering with a single scan and improves the quality with a few additional scans	It Handles only numeric data	O(N) (time)
DBSCAN	Density Based Clustering of Applications with noise. DBSCAN is a density-based algorithm. DBSCAN requires two parameters: epsilon (Eps) and minimum points (MinPts). It starts with an arbitrary starting point that has not been visited .It then finds all the neighbor points within distance Eps of the starting point. If the number of neighbors is greater than or equal to MinPts, a cluster is formed. The starting point and its neighbors are added to this cluster and the starting point is marked as visited. The algorithm then repeats the evaluation process for all the neighbors' recursively. If the number of neighbors is less than MinPts, the point is marked as noise.	DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-means. DBSCAN can find arbitrarily shaped clusters. It can even find a cluster completely surrounded by a different cluster. DBSCAN has a notion of noise. DBSCAN is designed for use with databases that can accelerate region queries, e.g. using an R* tree.	The quality of DBSCAN depends on the distance measure used in the function region Query. The most common distance metric used is Euclidean distance.	O(N log N) (time)
CURE	A new algorithm for detecting arbitrarily-shaped clusters at large-scale is presented and named CURE, for "Clustering Using Representatives". The algorithm works by pre-clustering a sample of the entire dataset, then using representative points within the sample to assign the remainder of	CURE ignores the information about the aggregate inter-connectivity of objects in two clusters. So it is introduced Chameleon algorithm. It appropriate for handling large data sets.	The algorithm is robust to the presence of outliers. The clustering algorithm can recognize arbitrarily shaped clusters.	O(N2sample logNsample) (time) O(Nsample) (space)

	the dataset. The sample is clustered using an agglomerative clustering algorithm which keeps track of the representative points in each cluster, as well as the nearest neighbor of each cluster at each step.			

V. PROPERTIES TO THE EFFICIENCY AND EFFECTIVENESS OF A NOVEL ALGORITHM

New technology has generated more complex and challenging tasks, requiring more powerful clustering algorithms. The following properties are important to the efficiency and effectiveness of a novel algorithm.

- Generate arbitrary shapes of clusters rather than be confined to some particular shape;
- Handle large volume of data as well as high-dimensional features with acceptable time and storage complexities;
- Detect and remove possible outliers and noise;
- Decrease the reliance of algorithms on users-dependent parameters;
- Have the capability of dealing with newly occurring data without relearning from the scratch;
- Be immune to the effects of order of input patterns;
- Provide some insight for the number of potential clusters without prior knowledge;
- Show good data visualization and provide users with results that can simplify further analysis;
- Be capable of handling both numerical and nominal data or be easily adaptable to some other data type.

VI. CONCLUSION

Clustering is an important task in Data analysis and data mining applications. Clustering is a task grouping a set of objects so that objects in the same group are more similar to each other than to those in other cluster. In this paper, an attempt has been made to give the basic concept of clustering, by providing different clustering algorithms with their Pros and Corns.

REFERENCES

[1] Amandeep Kaur Mann & Navneet Kaur, "Review Paper on Clustering Techniques", Global Journal of Computer Science and Technology Software & Data Engineering Volume 13 Issue 5 Version 1.0 Year 2013.

[2] Bruce Moxon "Defining Data Mining, The Hows and Whys of Data Mining, and How It Differs From Other Analytical Techniques" Online Addition of DBMS Data Warehouse Supplement, August 1996.

[3] P. Berkhin, 2002. Survey of Clustering Data Mining Techniques. Ttechnical report, AccrueSoftware, San Jose, Cailf.

[4] Jiawei Han, MichelineKamber, "Data Mining Concepts and Techniques" Elsevier Publication.

[5] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Trans. Pattern Anal. Mach.Intell., vol. 22, no. 1, pp. 4–37, 2000.

[6] P. IndiraPriya, Dr. D.K.Ghosh, " A Survey on Different Clustering Algorithms in Data Mining Technique", International Journal of Modern Engineering Research (IJMER), Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274.

[7] Ramandeep Kaur, Dr. Gurjit Singh Bhathal, "A Survey of Clustering Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013.

[8] K.Kameshwaran, K.Malarvizhi, "Survey on Clustering Techniques in Data Mining",IJCSIT- International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014.

[9] T. Sajana, C. M. Sheela Rani and K. V. Narayana, " A Survey on Clustering Techniques for Big Data Mining", Indian Journal of Science and Technology, Vol 9(3), January 2016.

[10] K. Chitra1, Dr. D.Maheswari," A Comparative Study of Various Clustering Algorithms in Data Mining", IJCSMC, Vol. 6, Issue. 8, pg.109 – 115, August 2017.

[11] K. Kameshwaran and K. Malarvizhi, "Survey on Clustering Techniques in Data Mining", International Journal of Computer Science and Information Technologies (0975-9646), Vol. 5(2), 2014

[12] Pradeep Rai and Shubha Singh (2010) A Survey of Clustering Techniques, International Journal of Computer Applications (0975 – 8887) Vol 7– No.12, pp. 1-5

[13] V.Kavitha, M.Punithavalli (2010) Clustering Time Series Data Stream – A Literature Survey, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 1, pp. 289-294.

[14] .S. Anitha Elavarasi and Dr. J. Akilandeswari (2011) A Survey On Partition Clustering Algorithms, International Journal of Enterprise Computing and Business Systems.

[15] S.Vijayalaksmi and M Punithavalli (2012) A Fast Approach to Clustering Datasets using DBSCAN and Applications (0975 – 8887) Vol 60– No.14, pp. 1-7.

[16] Preeti Baser and Dr. Jatinderkumar R. Saini, A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets, International Journal of Computer Science & Communication Networks,Vol 3(4),271-275.

[17] Oded Maimon, Lior Rokach, "Data Mining AND Knowlwdge Discovery Handbook", Springer Science+Business Media.Inc, pp.321-352, 2005.

[18] Arun K Pujari " Data Mining Techniques" pg. 42-67 and pg. 114-149,2006.

[19] Pradeep Rai, Shubha Singh" A Survey of Clustering Techniques"International Journal of Computer Applications, October 2010.

[20] Zheng Hua, Wang Zhenxing, Zhang Liancheng, Wang Qian,"Clustering Algorithm Based on Characteristics of Density Distribution" Advanced Computer Control (ICACC), 2010 2nd
International Conference on National Digital Switching System Engineering & Technological R&D Center, vol2", pp.431-435,2010.

[21] Anoop Kumar Jain, Prof. Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining", International Journal of Computer Science and Management Research, pp.72-78, 2012.

[22] P. Thangaraju, B.Deepa, T.Karthikeyan, Comparison of Data mining Techniques for Forecasting Diabetes
Mellitus, Vol. 3, Issue 8,
August 2014

[23] K.Kameshwaran, K.Malarvizhi, Survey on Clustering Techniques in Data Mining, Vol. 5 (2) 2014

[24] Megha Mandloi, A Survey on Clustering Algorithms and K-Means, July-2014

[25] Amandeep Kaur Mann, Survey Paper on Clustering Techniques, Vol- 2, Issue 4, Apr-2013.

[26] M. Anbarasi et. al. "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 5370-5376 ,2010.

[27] Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", IEEE, 2011.

[28] V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66.

[29] Quinlan J. Induction of decision trees. Mach Learn 1986; 1:81—106.

[30] Quinlan J. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.

[31] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/ Cole Advanced Books & Software; 1984.

[32] Anand Bahety, " Extension and Evaluation of ID3 – Decision Tree Algorithm". University of Maryland, College Park.

[33] S. K. Yadav and Pal S., "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology (WCSIT), 2(2), 51-56, 2012.

[34] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Autom. Control, vol. AC-19, no. 6, pp. 716–722, Dec. 1974.

[35] A. Alizadeh et al., "Distinct types of diffuse large B-cell Lymphoma identified by gene expression profiling," Nature, vol. 403, pp. 503–511,2000.