International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 24 (2018) pp. 16787-16791

A Survey on Partitioning and Hierarchical based Data Mining Clustering Techniques

[1]M.Kiruthika, [2]Dr.S.Sukumaran

[1]Ph.D Research Scholar, [2]Associate Professor
Department of Computer Science,
Erode arts and Science College (Autonomous),
Erode, India.

Abstract

In Data Mining, Clustering is a general technique for statistical data analysis, which is used in dissimilar fields, including machine learning, pattern recognition, image analysis and bioinformatics. Clustering is an excellent data mining tool for a huge and multivariate database. It is the one of data mining techniques in which data is separated into the set of related objects. Clustering is an appropriate example of unsupervised classification. It means that clustering does not depend on pre-defined classes and training examples through classifying the data objects. A Partitioning and Hierarchical algorithm in data mining is the most active research algorithm among proposed algorithms. Several factors or themes determine the optimal actual clustering. The significant idea of this paper is classifying the methods on the bases of different themes so that it aids in choosing algorithms for some further improvement and optimization. In this survey paper, a review of clustering, partitioning and hierarchical based clustering techniques and evaluation metrics for clustering are discussed.

Keywords: Clustering, Clustering Technique (Partitioning, Hierarchical), Data Mining, Evaluation Metrics.

1. INTRODUCTION

Data mining is mainly meant for finding hidden details or information in the large databases. It involves different types of algorithms to accomplish different type of tasks and an attempt to make sense of the information explosion embedded in big volume of data [15]. Data mining consists of extract, transform, and load transaction information onto the data warehouse system, Store and manage the information in an exceedingly multidimensional database system, Provide data access to business analysts and information technology professionals, Analyze the information by application software and Present the information in an exceedingly useful format, like a graph or table. In data mining two learning approaches are used i.e. supervised learning and unsupervised learning [19].

Data mining tasks can be divided into two types - descriptive data mining and predictive data mining. Descriptive data mining is also called as pattern or relationships. It is used for finding interesting patterns or describing the relationships of data. Predictive data mining uses historical data or predefined data [13]. It classifies or predicting the model behavior based on the data availability.

The Data Mining tasks includes Clustering, Association rule, Summarization, Regression, Classification and Prediction. Clustering is also known as segmentation or unsupervised learning. It is usually practiced by finding the connection among the data on predefined attributes. Clustering defines the process of grouping physical objects into classes. The most similar data are group into clusters. There are many types in clustering such as Centroid-based, Density-based, Connectivity-based and Distribute-based [3]. An association rule is a model that identifies the special type of data associations. Associations are often used in many applications.

Summarization is the process of maps data into subsets with associated descriptions. Summarization derives or extracts needed information about the database. It summarily characterizes the contents of the database and it also known as generalization or characterization. Regression is mainly used to map between data item and prediction variable. It is used to predict a numeric values range in a particular dataset. Classification maps between data and classes or groups which are predefined. Classification is also known as supervised learning. Regression and classification are used to solve identical problems. Prediction is a type of classification. Prediction is mainly used to predict the future state than a present state.

In this paper, clustering analysis is done. Cluster Analysis, an automatic process to find related objects from a database. It is a fundamental operation in data mining.

Clustering

Clustering is the technique of combining a set of similar objects known as clusters. It is used in information retrieval, statistical data analysis, machine learning, pattern recognition, image analysis, medical analysis, spatial data analysis and bioinformatics. Clustering aims at discovering groups and identifying interesting allocation and model in data sets.

Generally, the outputs produced by a clustering algorithm are the assignment of data objects in dataset to various groups. In addition, it will be sufficient to identify each data object with a unique cluster label.

The Clustering objectives are:

- To reveal natural combinations.
- To set off premise about the data.
- To come across dependable and convincing organization of the data.

Clustering groups the objects depending upon the information established in the data describing the objects or their associations. The objective of clustering is that the objects in a group are related to one other and different from the objects in other groups [15]. A high-quality cluster contains high intra-class similarity and low inter-class similarity. The procedure of clustering is performed with four basic steps.

Feature selection or extraction

Feature selections decide characteristic features from a set of candidates'. The feature extraction also exploits some transformations on data to generate useful and novel features from the original ones. Both are very crucial to the efficiency of clustering applications.

Clustering Algorithm design or Selection

The step is generally combined with the selection of an equivalent proximity measure and the formation of a criterion function. It is important to carefully investigate the distinctiveness of the problem at hand, in order to select or design an appropriate clustering strategy.

Cluster validation

Different approaches usually lead to different clusters and even for the same algorithm, parameter identification or the presentation of input patterns may affect the final results. Therefore, effective evaluation standards and criteria are important to provide the users with a quantity of confidence for the clustering results.

Results interpretation

The vital aim of clustering is offering a user with novel data as a result, that they can successfully solve the problems encountered. Further analysis and experiments may be required to guarantee the reliability of extracted knowledge.

2. CLUSTERING TECHNIQUES

The clustering approaches can be sorted out as partition, hierarchical, density- based and grid- based. In this paper, we examined various partitioning clustering techniques and hierarchical clustering techniques.

2.1 Partitioning Clustering Algorithms

Partitioning clustering attempts to decompose the data set into a set of disjoint clusters. More specifically, it endeavors' to create an integer quantity of partitions that optimize a certain criterion function. The criterion function may emphasize the local or global arrangement of the data and its optimization is an iterative method.

There are various types of partitioning clustering algorithms are:

A. K–Means

K-Means algorithm is the well-liked clustering algorithm. It iteratively computes the clusters and their centroids. It is a top down approach to clustering. It is used for making and analyzing the clusters with 'n' amount of data points, point is separated into 'K' clusters supported the similarity measure criterion. The result generated by the algorithm generally depends on initial cluster centroids chosen. It is an unvarying clustering algorithm during which items are stirred among sets of clusters till the necessary set is reached. As such, it can be viewed as a sort of squared inaccuracy algorithm, though the convergence criteria need not be distinct supported the squared inaccuracy [18]. A high degree of association between components in clusters is obtained, whereas a high degree of variation between components in different clusters is achieved at the same time.

B. K-Medoids

The K-medoids algorithm also termed as PAM (Partitioning Around Medoids) algorithm mean a cluster by medoid. Mostly, a random set of k items is taken to be the collection of medoids. Then at every step, all items from the input dataset that are not presently medoids are examined separately to ascertain if they ought to be medoids [18]. That is, the algorithm determines whether or not there is an item that ought to replace one in all the prevailing medoids. Pam is a lot of robust than k-means within the presence of noise and outliers as a result of a medoid is less influenced by outliers or alternative extreme values than a mean. PAM works efficiently for small data sets, however does not scale well for huge data sets.

C. FCM

Fuzzy C-Means (FCM) is one in all the foremost standard fuzzy clustering algorithms [26]. Fuzzy C-Means (FCM) may be a technique of clustering that assign membership levels and exploit them to assign data components to one or additional clusters. It uses mutual distance to calculate fuzzy weights. Each part of the universe will belong to some fuzzy set with a level of membership that varies from 0 to 1. FCM introduces the fuzziness for the belongingness of every object and might retain a lot of data of the dataset. This algorithm works by conveying membership to every data point of datasets parallel to each cluster center on the source of distance among the cluster center and also the data point. The membership and cluster centers are updated after the every iteration.

D. CLARANS

CLARANS (Clustering Large Application Based upon Randomized Search) is one in all the partitioning techniques used for huge database. It is a clustering method which draws sample of neighbours dynamically by looking for the spatial clusters within the information. It takes every node in the graph as a feasible resolution [3]. It dynamically draws an arbitrary sample at each new search. It makes the foremost *"natural"* clusters with the assist of "*silhouette coefficient"* that tell the number of belongingness of a data point to an essential cluster. It outperforms PAM (also a Partitioning clustering method) in terms of run time and cluster feature. Combination of Sampling procedure and PAM is employed in CLARANS. CLARANS doesn't assure search to localized region. The minimum distance among Neighbours nodes enhance effectiveness of the algorithm [17].

2.2 Hierarchical Clustering Algorithms

Hierarchical based clustering is additionally brought up as connectivity based clustering. Hierarchical clustering algorithm basically creates collection of clusters. It differs in however the sets are created. A tree data structure, referred to as a dendrogram, may be accustomed to illustrate the hierarchical clustering method and the sets of different clusters [8]. The basis in a dendrogram tree contains one cluster wherever every component is together. The leaves within the dendrogram contain a single component cluster. Internal nodes within the dendrogram represent novel clusters shaped by merging the clusters that seem as its children within the tree. Every level within the tree is related to the distance measure that was accustomed merge the clusters. Every clusters created at a specific level were combined as a result of the children clusters had a distance among them less than the distance value related to the level within the tree. The hierarchical clustering algorithms, according to the technique that produce clusters can further are separated into Agglomerative algorithms and Divisive algorithms.

- Agglomerative - This algorithm is a bottom up approach. It produces a sequence of clustering schemes of decreasing amount of clusters at every measure. The clustering scheme produced at each measure results from the preceding one by integration the two neighbouring clusters into one.
- Divisive - This algorithm is a top down approach. It produces a sequence of clustering schemes of increasing amount of clusters at every measure [7]. Converse to the agglomerative algorithms, the clustering produced at each measure results from the preceding one by splitting a cluster into two.

A. BIRCH

BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) is an agglomerative hierarchical clustering algorithm that works best for huge amount of numerical information. The fundamental scheme is that a tree is created that captures required information. Clustering is performing on the tree itself; the nodes within the tree contain the information that is employed for the computation of distance values. It contains two novel features called Clustering Feature (CF) and Clustering Tree (CE). Both of the CF and CE summarize the cluster representations, and provide helps in achieving good speed and scalability for huge databases [17]. BIRCH procedure is fundamentally separated in four stages. It runs on the consideration that not every data points are equally significant. CF (Clustering Feature)-Tree is used in its first phase. It is then condensed in the second phase. Global Clustering phase is third phase and some simple traditional clustering of CF are performed and not data points. In last phase new clusters are formed and closest seed data points redistributed to get good to better clusters. BIRCH is reasonably fast: Inability to deal with non-spherical clusters of varying size and data order sensitivity [3].

B. CHAMELEON

Chameleon is an agglomerative hierarchical clustering algorithm which uses the dynamic modeling technique to find out the similarity among the pairs of clusters. CHAMELON is employed to determine the natural clusters of numerous sizes and shapes through automatically adapting the merging decision supported on different clustering model features. It deals with two stage method of clustering: Initial, a graph partitioning is used to divider the data points into sub-clusters [1]. In the second stage, frequently merging these sub-clusters till its find the actual clusters. The key feature in CHAMELEON is that it defines the pair of the foremost similar sub-clusters by two considering relative closeness and also the relative inter-connectivity of the clusters. The relative closeness between pair of clusters is that the absolute clones between two clusters normalized with respect to the internal closeness of them. The relative interconnectivity between pair of clusters is that the absolute inter-connectivity between two normalized clusters with respect to the internal inter-connectivity of them.

C. CURE

CURE (Clustering Using Representatives) is a Divisive hierarchical clustering algorithm to draw a random sample and additional partition the sample, so as to have partially clustered partitions. The data points taken for random sample can be well-scattered. The shape of the sample is decided by selected data points. Therefore, this algorithm forms a non-spherical shape of clusters [1]. Subsequently, these partitions are shrunk to remove outliers. During this Shrink factor between 0.2-0.7 is measured to find accurate clusters. It follows a central outlook between the centroid-based and every point extremes. This algorithm fundamentally addresses to the key issues of hierarchical clustering i.e. outliers and clusters of only spherical shape. It

ignores the interconnectivity among the clusters and provides preference to distance between the representatives points of two clusters [6]. Also, it fails to contemplate important features of a particular cluster therefore affecting the clusters merging decisions. CURE technique only works for metric data.

D. ROCK

ROCK (Robust Clustering using Links) is a hierarchical algorithm in that to form clusters it utilizes a link strategy. From bottom to top links are merging together to form a cluster. It introduced the concept of link and neighbour. Link incorporates comprehensive data of other similar sufficient neighbours in order that not only two points are measured whenever merging or splitting clusters [17]. Bigger is that the link, higher the likelihood of points being in same cluster. Traditional algorithms used functions for Categorical and Boolean attributes however here concept of links (common neighbours) is introduced. ROCK has demonstrated its power by being effectively utilized for real datasets.

3. EVALUATION METRICS FOR CLUSTERING

The process of validating the results of a clustering algorithm is called as cluster validity. The two cluster validation metrics are

1. **Internal measures:** The quality of clustering are measured using the basic information of internal measures. Connectivity, Silhouette Width and Dunn Index are the internal measures of clusters.
2. **Stability Measures*:*** It is a special version of internal measures, which assesses the reliability of a clustering outcome by matching it with the clusters obtained after every column is detached, one by one**.** Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance between Means (ADM) and Figure of Merit (FOM) are the stability measures of clusters.

As determined by the k-nearest neighbors the level of connectedness of the clusters is indicates the connectivity. It has a range between 0 and ∞ and should be minimized. The Silhouette width is the average of every observation's Silhouette value [21]. Silhouette validation is validating the outcome of clustering to find out the accuracy of the obtained outcomes from the cluster value. It has a range between -1 to 1. The Silhouette cluster interpretation result is,

- $\leq$0.25 horrible split
- 0.26 – 0.50 weak structure
- 0.51 – 0.71 reasonable structure
- 0.71 – 1.00 excellent split

The Dunn Index is the ratio between the minimum distances between observations not in the same cluster to the largest intra-cluster distance. It has a range between 0 and ∞ and must be maximized.

The cluster stability measures are based on the cross-classification table of the actual clustering of the complete data with the clustering based on the removal of one column. The values of APN, ADM and FOM ranges from 0 to 1, with smaller value corresponding to highly consistent clustering results. AD has a value between 0 and infinity, and smaller values are also preferred [21]. Some of other external measures including F- measures, Fair- Counting F-measures, Rand measures, Jaccard index, Fowlkes – Mallows Index, Confusion Matrix and Mutual Information's.

4. CONCLUSION

Clustering is significant in data analysis and data mining applications. This paper discussed the various partitioning and hierarchical based clustering techniques and evaluation metrics for clustering. Partition clustering algorithms are very helpful when the clusters are of bowed shape having similar size and the amount of clusters can be recognized earlier. Due to the disability in predicting the amount of clusters in advance Hierarchical clustering algorithms are used. They partition the dataset into numerous levels of partitioning termed as dendograms. These algorithms are very useful in mining but the cost of formation of dendograms is extremely high for huge datasets. All the clustering algorithms are validated using cluster validation metrics such as internal and stability measures.

REFERENCES

[1]. Abdullah. Z, A. R. Hamdan, "Hierarchical Clustering Algorithms in Data Mining", World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol:9, No:10, 2015.

[2]. Amandeep Kaur Mann and Navneet Kaur, "Survey Paper on Clustering Techniques", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[3]. Amudha. S, "An Overview of Clustering Algorithm in Data Mining", International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 12, Dec -2016.

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

[5]. Ding. K, C. Huo, Y. Xu, Z. Zhong, and C. Pan, "Sparse hierarchal clustering for VHR image change detection," Geoscience and Remote Sensing Letters, *IEEE,* 12 (3), pp. 577 – 581, 2015.

[6]. Fahad A, Alshatri N, Tari Z and Alamri A, "A survey of clustering algorithms for Big Data: Taxonomy and empirical analysis", IEEE Transactions on Emerging Topics in Computing, pp. 267–79, 2014.

[7]. Garima, Hina Gulati and P.K.Singh, "Clustering Techniques in Data Mining: A Comparison", International Conference on Computing for Sustainable Global Development (INDIACom), pp. 410 – 415, IEEE 2015.

[8]. Harshada S. Deshmukh and Prof. P. L. Ramteke, "Comparing the Techniques of Cluster Analysis for Big Data", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 12, 2015.

[9]. Namrata S Gupta, Bijendra S.Agrawal and Rajkumar M. Chauhan , " Survey on Clustering Techniques of Data Mining", American International Journal of Research in Science, Technology, Engineering & Mathematics, pp-206-111, 2015.

[10]. Lori Dalton, Virginia Ballarin and Marcel Brun, "Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics", Current Genomics, Vol. 10, No. 6, 2009.

[11]. Pandove D and Goel S, "A comprehensive study on clustering approaches for Big Data mining", IEEE Transactions on Electronics and Communication System; Coimbatore, pg.26-27, Feb-2015.

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

[13]. Pragati Shrivastava and Hitesh Gupta, "A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research, pp.2249-7277, Sep-2012.

[14]. Rashedi. E, A. Mirzaei, and M. Rahmati, "An information theoretic approach to hierarchical clustering combination," *Neurocomputing*, 148, pp. 487-497, 2015.

[15]. Rama Kalaivani.E, Suganya. G and Kiruba. J, "Review on K-Means and Fuzzy C Means Clustering Algorithm", Imperial Journal of Interdisciplinary Research (IJIR), Vol-3, Issue-2, 2017.

[16]. Saroj and Tripti Chaudhary, "Study on Various Clustering Techniques", International Journal of Computer Science and Information Technologies, Vol. 6 (3), pp.3031-3033, 2015.

[17]. Sajana. T, 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), DOI: 10.17485/ijst/2016/v9i3/75971, January 2016.

[18]. Sonamdeep Kaur, Sarika Chaudhary and Neha Bishnoi, "A Survey: Clustering Algorithms in Data Mining", International Journal of Computer Applications, ISSN: 0975 – 8887, 2015.

[19]. Soni Madhulatha. T "An Overview on Clustering Methods", IOSR Journal of Engineering, Vol. 2(4), pp: 719-725, Apr-2012.

[20]. Sukhvir Kaur, "Survey of Different Data Clustering Algorithms", International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, pg. 584-588, May- 2016.

[21]. Sukhdev Singh Ghuman, "Clustering Techniques- A Review", International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, pg. 524-530, May- 2016.

[22]. Suman and Pinki Rani, "A Survey on STING and CLIQUE Grid Based Clustering Methods", International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017.

[23]. Sunil Chowdary, D. Sri Lakshmi Prasanna and P. Sudhakar, "Evaluating and Analyzing Clusters in Data Mining using Different Algorithms", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.2, pg. 86-99, 2014.

[24]. Vaishali R. Patel and Rupa G. Mehta, "Clustering Algorithms: A Comprehensive Survey", International Conference on Electronics, Information and Communication Systems Engineering, 2011.

[25]. Vijayalakshmi. M and M.Renuka Devi, "A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets" , International Journal of Advanced Research in Computer Science and Software Engineering, pp.305-307, 2012.

[26]. Zeynel Cebeci and Figen Yildiz, "Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures", Journal of Agricultural Informatics, Vol. 6, No. 3, 2015.

AUTHOR PROFILE

M. Kiruthika received Bachelor of Computer Science (B.Sc.) degree from the Anna University, in 2014 and the Master of Computer Application (MCA) degree from the Anna University, in 2016. She also recieved the M.Phil degree from the Bharthiar Unviersity, Coimbatore, in 2018. Currently she is doing her Ph.D computer science in Erode Arts and Science College. Her research area includes Data Mining.

Dr. S. Sukumaran graduated in 1985 with a degree in Science. He obtained his Master Degree in Science and M.Phil in Computer Science from the Bharathiar University. He received the Ph.D degree in Computer Science from the Bharathiar University. He has 30 years of teaching experience starting from Lecturer to Associate Professor. At present he is working as Associate Professor of Computer Science in Erode Arts and Science College, Erode, Tamil nadu. He has guided for more than 55 M.Phil research Scholars in various fields and guided 13 Ph.D Scholars. Currently he is Guiding 3 M.Phil Scholars and 6 Ph.D Scholars. He is member of Board studies of various Autonomous Colleges and Universities. He published around 63 research papers in national and international journals and conferences. His current research interests include Image processing, Network Security and Data Mining.