I J C T A, 9(34) 2016, pp. 317-323

A Comprehensive Study on Various Clustering Techniques

K. Maheswari[1] and M. Ramakrishnan[2]

ABSTRACT

Clustering is a challenging technique in data mining which tries to group huge amount of data. It is a difficult task since finding a suitable cluster for a data item is tough and it is unsupervised also. Care must be taken in clustering such that inter-cluster similarity should be high and intra-cluster similarity should be low. In this paper, a survey on three categories of clustering algorithms viz. partitional, hierarchical and density based have been carried out. The paper also presents the pros and cons of various clustering techniques in a table.

Keywords: Clustering, Hierarchical, Partitional, Density based, DBSCAN, CLARA, OPTICS, CURE, BIRCH, DENCLUE, ROCK, CLUBS, CHAMLEON

1. INTRODUCTION

Mining useful information from huge volume of data is said to be data mining and it plays a vital role in business intelligence. Data mining analyses huge repository of data in various aspects to generate useful information. Different methods are used in data mining to analyze the data for extracting knowledge include clustering, regression, classification, association and sequential pattern matching. Each technique is having its own importance in data mining process. Among them, clustering is an important technique which tries to grout the set of data items into clusters having similarity [1].

Clustering is an unsupervised learning technique which provides a structure for a collection of unlabelled data. It is the process of organizing the data items into groups where members of the group are similar in some way and are dissimilar with other groups. Clustering is said to be good quality if intra-cluster similarity measure is high and inter-cluster similarity is low[2]. Clustering is having wide applications in various fields such as psychology, statistics, medicine, biology, etc.

Clustering methods are broadly classified into six types and they are partitioning methods, hierarchical methods, density based methods, grid based methods, model based methods and constraint based methods. Partitioning methods divide the whole set of items into different set of groups based on distance between two objects[3]. Data are organized as groups based on the density value function in density based methods. This type of clustering uses density of the data in a particular region. Each instance of a cluster, the neighbourhood has to contain minimum number of points. DBSCAN is the popular density based clustering algorithm. This algorithm separates data points into three types viz. core points, border points and noise points[4].

Grid based clustering divides the clustering data points into finite number of cells in the form of rectangles and perform operations. The cells that contain more number of data points are dense and they form the cluster. The small cells that are close to the dense cells are integrated to form more densed cluster[5]. Model based clustering starts with a random initialization of parameters, and iteratively adjusts so that a

[1] Research Scholar, Department of Computer Science, Bharathiyar University, Coimbatore, Tamil Nadu, India, *E-mail: magi.research2014@gmail.com*

[2] Chairperson, School of Information Technology, Madurai Kamaraj University, Madurai, Tamil Nadu, India, *E-mail: ramkrishod@gmail.com*

group of likelihood is formed. In model based methods, data items are grouped that best fit the given model[6].

Hierarchical clustering methods group data items into a tree of clusters. Hierarchical methods are efficient and simple. The output of hierarchical clustering is a tree structure that is more informative. Hierarchical clustering is having advantage that specifying the number of clusters to be formed at the beginning of the clustering process is not needed[7]. It uses less computation interms of combinatorial number of data points. Hence this paper provides an overview on various hierarchical clustering algorithms. The paper is organized as follows: Section 1 provides basics about data mining and importance of clustering method. Section 2 elaborates about hierarchical clustering. Discussion on various clustering techniques has been narrated in section 3 and comparison of discussed clustering techniques interms of their advantages and disadvantages are discussed in section 4. Paper ends by providing a short conclusion in section 5.

2. BACKGROUND

As said above, hierarchical clustering groups the data into a tree of clusters and the structure is called dendrogram. Formation of clusters can be done either in top-down or bottom-up methods[8]. Root contains one cluster representing all the data points and the leaves containing one data point. Hierarchical clustering can be agglomerative or divisive[8].

Agglomerative method uses bottom-up strategy to form clusters. Initially it considers each data item as a singleton clusters. Then it merges the similar pair of clusters until all the data items form one single cluster or few clusters specified by the user. A dendrogram or tree graph is generally used to represent the output.

Divisive clustering is a top-down clustering method where all the items form one single cluster in the beginning. The cluster is divided recursively until individual data items form clusters or specific number of clusters as mentioned by the user. Both agglomerative and divisive methods of clustering terminates quickly[9].

Hierarchical clustering uses various techniques to decide which clusters should be joined or splitted during each iteration. Hierarchical clustering holds several merits.

3. HIERARCHICAL CLUSTERING

3.1. BIRCH

BIRCH stands for Balanced Iterative Reducing and Clustering using Hierarchies. It follows agglomerative hierarchical clustering and best suited for huge amount of metric data[10]. Data to be clustered is not uniformly distributed and hence all the data are not equally important for clustering. It represents the output clusters in a tree structure called Cluster Feature (CF) tree[10].

BIRCH works in two phases. It scans the data set for building initial in-memory CF tree in the first phase and applies any clustering algorithm to cluster the leaf nodes of the generated CF tree in the second phase[11]. This method is well designed to minimize memory requirement and I/O operations. Moreover, this method is capable of handling noisy data effectively.

3.2. ROCK

RObust Clustering using linKs, abbreviated as ROCK, is a hierarchical clustering algorithm which follows agglomerative style of clustering. Like BIRCH, this method is also well suited for clustering large volume of data, which contains categorical and Boolean attributes. ROCK method uses the combination of nearest neighbor, relocation and agglomerative methods[12]. Here cluster similarity is based on clusters having

common neighbours. It works in three phases viz. drawing random sample from population, forming clusters with links and labeling the data in the disk. Clusters are generated by sample points. Performance of ROCK algorithm is well appreciated in categorical, Boolean and time series data[13].

3.3. CLUBS

Clustering using Binary Splitting is the expansion for CLUBS and adopts both agglomerative and divisive approaches. The performance of CLUBS in terms of speed is better than k-means algorithm. The algorithm works in two phases: the first phase of the algorithm is divisive where original data set is split recursively under binary fashion to form the mini clusters[14]. These mini clusters are combined to form the main cluster and this step is agglomerative. Splitting the data into mini clusters under binary fashion is done using least quadratic distance criterion. This algorithm can also be used to refine other algorithm's performance.

3.4. CURE

It stands for clustering using representatives. This algorithm is very useful in discovering groups and identifying distributions in the data. This method is more robust towards outliers and forms clusters having non-spherical shapes. CURE works by employing random sampling and partitioning methods for handling large databases.

CURE chooses fixed number of representative points in data space. These representative points are generated by selecting well scattered objects and moving them towards cluster centre by shrinkage factor[15]. Overall working of CURE is as follows: draws random sample from data space, partition the data sample, eliminate outliers, form clusters and label the data on disk. Experimental results show that CURE outperforms existing techniques since it uses random sampling and partitioning. CURE is well suited for large databases without affecting cluster quality[15].

3.5. CHAMLEON

It is a type of hierarchical clustering using dynamic modeling. It uses the characteristics of the data to form the natural clusters. The clusters are of different shape and size. CHAMLEON determines the pair of similar sub-clusters by using relative inter-connectivity and relative closeness measures of cluster[16]. The relative inter-connectivity is the measure representing absolute inter-connectivity between two normalized clusters.

Relative closeness represents absolute closeness between two normalized clusters with respect to internal closeness. This algorithm works in two phases: In the first phase, it uses graph-partitioning algorithm to cluster the objects into relatively small sub clusters[17]. In the second phase, it adopts agglomerative hierarchical clustering algorithm to find the genuine clusters by repeatedly combining the small sub-clusters.

3.6. CLARA

Clustering large applications, abbreviated as CLARA, is a clustering method based on PAM and it tries to perform clustering on large dataset applications. CLARA overcomes the drawbacks of PAM and k-means algorithms since they are slow, and not fit for practical usage. In CLARA, clustering is performed in two steps[18]. A sample is chosen from the data space and divided into k-clusters. Dividing the data space into k-clusters is done by the algorithm which is used in PAM. In the first phase, successive mediods are selected with the objective of obtaining smallest possible average distance between the objects of sample.

This phase is said to be BUILD phase. Second phase is called SWAP, where an attempt is made to minimize average distance among objects by replacing representative objects. The efficiency of clustering

is calculated as average distance between each object and its mediod. This procedure is done five times and clustering with lowest average distance is retained.

3.7. CLARANS

Clustering large applications based on randomized search is the expansion for CLARANS, which combines sampling techniques with PAM. The process is generally depicted in a graph and each node in the graph is a solution. The clustering obtained after replacing a mediod is called the neighbor of the current clustering. It dynamically draws sample of neighbours. If the local optimum is found, it starts with new randomly selected node in search of new local optima. If a better neighbor is found, it moves to neighbour's node and process starts again.

CLARANS is more efficient and highly scalable when compared to PAM and CLARA[19]. The difference between CLARA and CLARANS is that former works only with part of the data set whereas the later works with all the data objects.

3.8. DBSCAN

DBSCAN stands for density based spatial clustering of application with noise. This algorithm forms clusters from large spatial datasets by referring the local density of the data items. High density regions designate a cluster where as low density regions indicate noise or outliers. As name suggests, DBSCAN is capable of handling large datasets with noise and generates clusters with different size and shapes[20].

DBSCAN needs three input parameters for its functioning viz. k, the neighbours list size, Eps distance that delimits the neighbourhood area of a point and Minpts minimum points that must exist in Eps neighbourhood. DBSCAN starts with the arbitrary point and all points that are reachable from arbitrary point with density value. If the arbitrary point is a core point, then a cluster is formed. If the arbitrary point is a border point, the algorithm visits next arbitrary point of the dataset since no point is reachable. A point to note about DBSCAN algorithm is that it does not do well with clusters of different densities[21].

3.9. DENCLUE

Density based clustering that works on overall density of the data points. The density used in this approach is kernel density estimation, which describes the distribution of data by a function. The framework of DENCLUE algorithm is build upon Schnell's algorithm. Clusters are defined by local maxima of the density estimation. The data points are assigned to local maxima by using hill climbing algorithm.

DENCLUE is capable of performing clustering on huge volume of data with noise[22]. This algorithm allows arbitrarily shaped clusters with high dimensional data to be represented in a compact mathematical way. DENCLUE uses two parameters viz. N, which determines the influence of a data point in their neighbourhood and one more parameter that describes whether a density-attractor is significant or not. This allows reduction of density attractors and helps improve the performance.

3.10. OPTICS

OPTICS stands for ordering points to identify the clustering structure and it is the extension of DBSCAN[23]. DBSCAN is having the difficulty that one set of global parameters is used in cluster analysis. To overcome this difficulty, OPTICS is proposed. OPTICS computes an ordering of points augmented by other attributes such as reachability distance, cluster structure and cluster ordering.

OPTICS does not produce clusters of data. It outputs linear list of all objects under analysis and represents density based clustering structure of data. OPTICS does not require specific density threshold as input[24]. To select an object that is density reachable, OPTICS needs two important parameters. They are core

distance and reachability distance. Ordering of a database along with a reachability value for each objects for the seedlist.

4. COMPARISON OF VARIOUS CLUSTERING TECHNIQUES

In this section, we give below the comparative statement of various clustering techniques that are discussed in this paper.

Method	*Clustering type*	*Advantages / Disadvantages*
BIRCH	Agglomerative	**Advantages:** ▪ Uses single scan over the database to find clusters. ▪ Quality of clustering is improved if more scans are performed. ▪ Computational complexity is O(n) **Disadvantages:** ▪ Only numeric data is used. ▪ Clusters should be spherical in shape and size need to be similar.
CURE	Agglomerative	**Advantages:** ▪ It can recognize arbitrarily shaped clusters. ▪ Uses space that is linear to input size. **Disadvantages:** ▪ It ignores information about aggregate inter-connectivity of objects in two clusters.
CLUBS	Agglomerative / Divisive	**Advantages:** ▪ Robust and impervious to noise. ▪ It provides better speed and accuracy when compared to BIRCH, k-means etc. **Disadvantages:** ▪ Time consuming since it is both agglomerative and divisive
ROCK	Agglomerative	**Advantages:** ▪ Well suited for real categorical and time-series data. **Disadvantages:** ▪ Time consuming as it works in three phases
CHAMLEON	Agglomerative	**Advantages:** ▪ Well suited for huge volume of data **Disadvantages:** ▪ Inability to adopt high dimensional data. ▪ Time complexity is $O(n^2)$
CLARA	Partition based	**Advantages:** ▪ Better than k-means algorithm and very fast. **Disadvantages:** ▪ If sampling is biased, clustering is very bad ▪ Trade off efficiency
CLARANS	Partition based	**Advantages:** ▪ Very effective than PAM and CLARA. ▪ Handles outliers **Disadvantages:** ▪ Computational complexity is $O(n^2)$ ▪ Quality of clustering depends on sampling method
DBSCAN	Density based	**Advantages:** ▪ Discovers clusters of arbitrary shapes ▪ Robust towards outliers and noisy data

contd. table

Method	*Clustering type*	*Advantages / Disadvantages*
		Disadvantages: ▪ Fails to form clusters if density varies and the data is too sparse. ▪ Sampling affects density measures
DENCLUE	Density based	**Advantages:** ▪ It allows a compact mathematical description of arbitrarily shaped clusters in high dimensional data sets. ▪ It uses grid cells and only keeps information about grid cells that actually contain data points. **Disadvantages:** ▪ Method requires careful selection of the density parameter and noise threshold
OPTICS	Density based	**Advantages:** ▪ Overcomes the difficulty of DBSCAN in terms of global parameters. **Disadvantages:** ▪ It does not produce clusters. It outputs linear list of all objects under analysis.

5. CONCLUSION

Since clustering is an important activity in data mining, a study on various clustering algorithms is carried out in this paper. We discussed general behaviour, functional advantages and disadvantages of various dominant clustering algorithms. We have selected three categories of clustering algorithms viz. Partitional, Hierarchical and Density based. We concluded the paper by presenting a comparative table, giving merits and demerits of various clustering techniques. It is observed from the table that several methods are capable of handling noisy data.

REFERENCES

[1] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." *Information Sciences* 275 (2014): 314-347.

[2] Anderberg, Michael R. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. Vol. 19. Academic press, 2014.

[3] Kaufman, Leonard, and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[4] Kisilevich, Slava, Florian Mansmann, and Daniel Keim. "P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos." *Proceedings of the 1st international conference and exhibition on computing for geospatial research & application*. ACM, 2010.

[5] Ilango, M. R., and V. Mohan. "A survey of grid based clustering algorithms." *International Journal of Engineering Science and Technology* 2.8 (2010): 3441-3446.

[6] Meila, Marina, and David Heckerman. "An experimental comparison of several clustering and initialization methods." *arXiv preprint arXiv:1301.7401* (2013).

[7] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *Unsupervised learning*. Springer New York, 2009.

[8] Langfelder, Peter, Bin Zhang, and Steve Horvath. "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R." *Bioinformatics* 24.5 (2008): 719-720.

[9] Franti, Pasi, Olli Virmajoki, and Ville Hautamaki. "Fast agglomerative clustering using a k-nearest neighbor graph." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28.11 (2006): 1875-1881.

[10] Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." *Neural Networks, IEEE Transactions on* 16.3 (2005): 645-678.

[11] Berkhin, Pavel. "A survey of clustering data mining techniques." *Grouping multidimensional data*. Springer Berlin Heidelberg, 2006. 25-71.

[12] Premalatha, K., and A. M. Natarajan. "A literature review on document clustering." *Information Technology Journal* 9.5 (2010): 993-1002.

[13] Cesario, Eugenio, Giuseppe Manco, and Riccardo Ortale. "Top-down parameter-free clustering of high-dimensional categorical data." *Knowledge and Data Engineering, IEEE Transactions on* 19.12 (2007): 1607-1624.

[14] Masciari, Elio, Giuseppe Massimiliano Mazzeo, and Carlo Zaniolo. "Analysing microarray expression data through effective clustering." *Information Sciences* 262 (2014): 32-45.

[15] Rauber, Andreas, Dieter Merkl, and Michael Dittenbach. "The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data." *Neural Networks, IEEE Transactions on* 13.6 (2002): 1331-1341.

[16] Karypis, George, Eui-Hong Han, and Vipin Kumar. "Chameleon: Hierarchical clustering using dynamic modeling." *Computer* 32.8 (1999): 68-75.

[17] Kotsiantis, Sotiris, and Panayiotis Pintelas. "Recent advances in clustering: A brief survey." *WSEAS Transactions on Information Science and Applications* 1.1 (2004): 73-81.

[18] Kaufman, Leonard, and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[19] Nagpal, Arpita, Aman Jatain, and Deepti Gaur. "Review based on data clustering algorithms." *Information & Communication Technologies (ICT), 2013 IEEE Conference on*. IEEE, 2013.

[20] Borah, B., and D. K. Bhattacharyya. "An improved sampling-based DBSCAN for large spatial databases." *Intelligent Sensing and Information Processing, 2004. Proceedings of International Conference on*. IEEE, 2004.

[21] Erman, Jeffrey, Martin Arlitt, and Anirban Mahanti. "Traffic classification using clustering algorithms." *Proceedings of the 2006 SIGCOMM workshop on Mining network data*. ACM, 2006.

[22] Pilevar, Abdol Hamid, and M. Sukumar. "GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases." *Pattern recognition letters* 26.7 (2005): 999-1010.

[23] Han, Jiawei, Jae-Gil Lee, and Micheline Kamber. "An overview of clustering methods in geographic data analysis." *Geographic Data Mining and Knowledge Discovery, 2nd edn, H. Miller, and J. Han, Eds. Taylor and Francis, FL, USA* (2009): 149-187.

[24] Ankerst, Mihael, et al. "OPTICS: ordering points to identify the clustering structure." *ACM Sigmod Record*. Vol. 28. No. 2. ACM, 1999.