A Generalized Study on Data Mining and Clustering Algorithms

Syed Thouheed Ahmed, S. Sreedhar Kumar, B. Anusha, P. Bhumika, M. Gunashree, and B. Ishwarya

1 Introduction

Structured Data, data which is easily organized and deposited in databases. Example, data stored in RDBMS. *Semi Structured* data is a data pattern which is unorganized but has some association within the data. For example: Log files, xml files. *Unstructured* data is a data which does not have lucid format in storage. For example: Image file, video files, audio files, etc. Big data analytic is a proficiency of examining the stored data to locate some schemes and interdependency among the data [1]. It can be applied in several fields where huge amount data is induced. Big data is stipulated by three attributes basically known as three V's. *Velocity* is the rate at which the data comes into an Organization. *Variety*, relates many types of data. *Volume* projects the magnitude of information which is uncluttered into the organization [1].

Data mining is operation used in big data scrutiny in identifying concealed correlations, framework and statistical information from data repository that contradicts to prevail by using traditional proficiencies [2]. It is intended to traverse enormous amount of information within probe of accordant framework where as to affirm the consequence by the detected framework for an advanced fragment in statistics. One of the Data Mining process is Clustering.

Clustering process for computer systems conceived in anthropology driver and Kroeber which was published in the year 1932 and initiated to psychology by Zubin in the year 1938 and Robert Tryon in the year 1939, and most prominently used by Chattel conceived in the year 1943 for eccentricity theory organization in personality psychology. Clustering is organizing data in such groups called clusters,

S. T. Ahmed (✉) · S. Sreedhar Kumar · B. Anusha · P. Bhumika · M. Gunashree · B. Ishwarya
Department of Computer Science and Engineering, Dr. T. Thimmaiah Institute of Technology, KGF, Karnataka, India

S. Smys et al. (eds.), *New Trends in Computational Vision and Bio-inspired Computing*, https://doi.org/10.1007/978-3-030-41862-5_114

in which there is high intra-cluster similarity. The cardinal abstraction of cluster scrutiny is segregating an entrenched deposit of information entity into sections. Every section is prominently isolated and comparatively entities in one cluster are interchangeable to another one, additional contradictory to entities to another cluster. It is significant process necessary in various applications such as, machine learning, data mining, pattern recognition, Image analysis and Bioinformatics, etc.

1.1 Need for Clustering?

Clustering helps in organizing huge voluminous data into clusters and displays interior structure of statistical information [2]. Clustering is the intent of segregating the data into clusters. Clustering improves the data readiness towards AI techniques [3]. Process for clustering, exhibits knowledge discovery in data, It is used either as a stand-alone tool to get penetration into data distribution or as a pre-processing step for other algorithm.

2 Clustering and Its Types

Clustering is stipulated as the sorting of indistinguishable text report into clusters that is reports within the clusters have high parallelism when compared to other but heterogeneous to reports in other clusters [4]. Since all the information have been added to the World Wide Web it becomes very consequential to graze or probe the pertinent information dramatically. The identification of appropriate algorithms for clustering induces the optimal clustering techniques, and inclines imperative to possess the contraption for differentiating the consequences of heterogeneous clustering techniques. Several heterogeneous clustering process to retain stipulation in directive to decode the issue from distinct approach, that is

- Partitioned clustering
- Density based clustering
- Hierarchical clustering

2.1 Partitioned Clustering

Partitioning methods breaks down the data into a set of different clusters. Give n objects, this method produces k clusters of data where $k < n$ clusters of data and using an iterative relocation method [1]. Algorithms used in partitioned clustering are *k-means algorithm, k-medoids algorithm.*

2.1.1 k-Means Algorithm

This algorithm can be used to cluster the data set thus forming clusters repeatedly. It is one of the unsupervised and iterative algorithms. The main aim of this algorithm is to find out the location of the clusters thus minimizing the distance between the cluster and the data set. This algorithm also called "Lloyd's algorithm" where m data set are clustered to form some number of clusters say k, where each of the data set belongs to the closer mean cluster.

Algorithm

1. Define the number of clusters (k) to be produced and identical data point centroids.
2. The distance from every data point to all the centroids are calculated and the point is assigned to the cluster with a minimum distance.
3. Follow the above step for all the data points.
4. The average of the data points present in a cluster is calculated and can set new centroid for that cluster.
5. Until desired clusters are formed repeat Step 2.

The initial centroid is selected randomly and thus the resulting clusters have larger influence on them. Complexity of k-means algorithm is $O(tkn)$ where n—total data set, k—clusters formed, t-iterations in order to form cluster [1].

Advantages

- Effortless implementation process.
- Dense clusters are produced when clusters are spherical when compared to hierarchical method.
- Appropriate for large databases.

Disadvantages

- Inappropriate for clusters with different density and size.
- Equivalent results are not produced on iterative run.
- Euclidean distance measures can weigh unequally due to underlying factors
- Unsuccessful for non-linear data set and categorical data
- Noisy data and outliers are difficult to handle.

2.1.2 k-Medoids Algorithm

In this algorithm, each of the cluster is described by one of the objects which is located near the centroid of the cluster. The repeated process of changing described objects by non-described objects continues until the resulting cluster is improved. The value is predicted using cost function which measures the variance between an object and described object of a cluster. The algorithm is implemented in two steps:

Build: Initial medoids are innermost objects.

Swap: A function can be swapped by another function until the function can no longer be reduced [5].

Algorithm

1. Initially choose m random points as initial medoids from given data set.
2. For every data point assign a closest medoid by distance metrics.
3. Swapping cost is calculated for every chosen and unchosen object given as *TCns* where s is selected and n is non-selected object [5].
4. If *TCns* < 0, s is replaced by n
5. Until there is no change in medoids repeat 2 and 3.

Four characteristics to be considered are

- Shift-out membership: Movement of an object from current cluster to another is allowed.
- Shift-in membership: Movement of an object from outside to current cluster is allowed.
- Update the current medoids: Current medoid can be replaced by a new medoid.
- No change: Objects are at their appropriate distances from cluster.

Advantages

- Effortless understanding and implementation process.
- Can run quickly and converge in few steps.
- Dissimilarities between the objects is allowed.
- Less sensitive to outliers when compared to k-means.

Disadvantages

- Initial sets of medoids can produce different clustering's. It is thus advisable to run the procedure several times with different initial sets.
- Resulting clusters may depend upon units of measurement. Variables of different magnitude can be standardized.

2.2 Hierarchical Based Clustering

Hierarchical methods, decomposes ***n*** number of data set into groups forming hierarchy of clusters. The tree like structural representation of hierarchical based clustering is called Dendrogram diagram [6]. The root of dendrogram represents the entire dataset and leaves represents each individual cluster present in the dataset. The clustering results are obtained by taking dendrogram at different levels. The two approaches of hierarchical based clustering, *Agglomerative (Bottom-up), Divisive (Top-down).*

2.2.1 Agglomerative Clustering Algorithm

This Algorithm is also referred as Bottom-up approach. This approach treats each and every data point as a single cluster and then merges each cluster by considering the similarity (distance) in each individual cluster until a single large cluster is obtained or when some condition is satisfied [7–14].

Algorithm

1. Initialize all n data points into N individual clusters.
2. Find the cluster pairs with the least distance (closest distance) and combine them as one single cluster.
3. Calculate pair-wise distance between the clusters at present that is the new formed cluster and the priority available clusters.
4. Repeat steps 2 and 3 until all data samples are merged into a single large cluster of size N

Advantages

1. Easy to identify nested clusters.
2. Gives better results and ease in implementation.
3. They are suitable for automation.
4. Reduces the effect of initial values of cluster on the clustering results.
5. Reduces the computing time and space complexity.

Disadvantages

1. It can never undo what was done previously.
2. Difficulty in handling different sized clusters and convex shapes lead to increase in time complexity
3. There is no direct minimization of objective function.
4. Sometimes there is difficulty in identifying the exact number of clusters by the Dendrogram.

2.2.2 Divisive Based Clustering

This approach is also referred as the top-down approach. In this, we consider the entire data sample set as one cluster and continuously splitting the cluster into smaller clusters iteratively. It is done until each object in one cluster or the termination condition holds [8]. This method is rigid, because once a merging or splitting is done, it can never be undone.

Algorithm

1. Initially, initiate the process with one cluster containing all the samples.
2. Select a largest cluster from the cluster that contains widest diameter.
3. Detect the data point in the cluster found in step 2 with the minimum average similarity to the other elements in that cluster.
4. The first element to be added to the fragment group is the data samples found in step3.
5. Detect the element in the original group which has the highest average similarity with the fragment group

6. If the average similarity of element obtained in step 5 with the fragment group is greater than its average similarity with the original group then assign the data sample to the fragment group and go to step 5; otherwise do nothing;
7. Repeat the step 2–6 until each data point is separated into individual clusters.

Advantages

1. It produces more accurate hierarchies than bottom-up algorithm in some circumstances.

Disadvantages

1. Top down approach is computationally more complex than bottom up approach because we need a second flat clustering algorithm.
2. Use of different distance metrics for measuring distance between clusters may generate different results.

2.3 Density Based Clustering

Density based clustering is clustering of database based on the densities (i.e. local cluster criterion). It has major features such as.

- It discovers the arbitrary shape cluster.
- It handles the noise data.
- It examines only the local region to justify the density.
- To termination the process it requires the density parameters.

It is categorized into two types namely:

(a) Density Based Connectivity: This includes clustering technique such as DBSCAN and DBCLAD.
(b) Density based Function: DENCLUE method density clusters are obtained based on some functions.

But here lets us see DBSCAN clustering in detail.

2.3.1 DBSCAN (Density Based Spatial Clustering of Applications with Noise)

In DBSCAN, a cluster is defined as group of data that is of highly dense. DBSCAN considers two parameters such as:

Eps: the maximum value of radius from its neighborhood.

MinPts: The Eps is surrounded by data points (i.e. Eps-Neighborhood) that should be minimum.

To define Eps-Neighborhood it should satisfy the following condition, $N_{Eps(q)} : \{ p \; belongs \; to \; D|(p,q) \leq Eps \}$.

In order to understand the Density Based Clustering let us follow few definitions:

(a) Core point: It is point which lies within Eps and MinPts which are specified by user. And that point is surrounded by dense neighborhood.
(b) Border point: It is point that lies within the neighborhood of core point and multiple core points can share same border point and this point does not contains dense neighborhood.
(c) Noise/Outlier: It is point that does not belongs to cluster.
(d) Direct Density Reachable: A point p is directly Density Reachable from point q with respect to Eps, MinPts if point p belongs to $N_{Eps(q)}$ and Core point condition i.e.$|N_{Eps(q)}| \geq MinPts$
(e) Density Reachable: A point p is said to Density Reachable from point q with respect to Eps, MinPts if there a chain points such as $p_1,p_2, \ldots\ldots p_n,\ p_1 = q$, $p_n = p\ such\ that\ p_{i+1}$ is directly reachable from p_n.

Algorithm

1. In order to form clusters, initially consider a random point say point p
2. The second step is to find the all points that are density reachable from point p with respect to Eps and MinPts. The following condition is checked in order to form the cluster
   (a) If point p is found to be core point, then cluster is obtained.
   (b) If point p is found to be border point, then no points are density reachable from point p and hence visit the next point of database.
3. Continue this process until all the points is processed.

Advantages

- It can identify Outlier.
- It does not require number of clusters to be specified in advance.

Disadvantages

- If the density of data keeps changing then efficiency of finding clusters is difficult.
- It does not suit for high quality of data and the user has to specify the parameter in advance.

3 Conclusion

The paper has incorporated detailed clustering techniques with algorithmic flow and features. The study provides a milestone for researchers to pause and consider the most suitable clustering form [15–19]. The paper is concluded with a comparative study of various clustering techniques in Table 1.

Table 1 Classification of clustering algorithm

Name	Algorithm	Author	Year	Key idea	Type of data
Partition based	k-Mean	MacQueen	1967	Mean centroid	Numerical
	k-Medoids	Kaufman and Rousseeuw	1987	Medoid centroid	Special
Hierarchical based	Agglomerative	S.C Johnson	1967	Complex	Special
	Divisive	Guha, Rastogi and Shim	1998	Partition samples	Numerical
Density based	DBSCAN	Ester et al.	1996	Fixed size	Numerical

References

1. Venkatkumar, Iyer Aurobind, and Sanatkumar Jayantibhai Kondhol Shardaben. "Comparative study of data mining clustering algorithms." *Data Science and Engineering (ICDSE), 2016 International Conference on*. IEEE, 2016.
2. Narang, Barkha, Poonam Verma, and Priya Kochar. "Application based, advantageous k-meansclustering Algorithm in Data Mining-A Review." *International Journal of Latest Trends in Engineering and Technology (IJLTET), ISSN* (2016).
3. Syed Thouheed Ahmed S., Sandhya M., Shankar S. "ICT's Role in Building and Understanding Indian Telemedicine Environment: A Study." In: Fong S., Akashe S., Mahalle P. (eds) Information and Communication Technology for Competitive Strategies. Lecture Notes in Networks and Systems, vol 40. Springer, Singapore 2019
4. Sisodia, Deepti, et al. "Clustering techniques: a brief survey of different clustering algorithms." *International Journal of Latest Trends in Engineering and Technology (IJLTET)* 1.3 (2012): 82–87.
5. Syed Thouheed Ahmed, Kiran Kumari patil "An Investigative study of Motifs extracted features on real time big-data signals" in Proceedings of ICETT – Kollam, pp 1–4, Kerala, IEEE, 2016
6. Saraswathi, S., and Mary Immaculate Sheela. "A Comparative Study of Various Clustering Algorithms in Data Mining." *International Journal of Computer Science and Mobile Computing* 11.11 (2014): 422–428.
7. Popat, Shraddha K., and M. Emmanuel. "Review and comparative study of clustering techniques." *International journal of computer science and information technologies* 5.1 (2014): 805–812.
8. De Silva, Pavani Y., et al. "Recursive Hierarchical Clustering Algorithm." Chitra, K., and D. Maheswari. "A Comparative Study of Various Clustering Algorithms in Data Mining." (2017).
9. Nagpal, Pooja Batra, and Priyanka Ahlawat Mann. "Comparative study of density based clustering algorithms." *International Journal of Computer Applications* 27.11 (2011): 421–435.
10. Nagpal, P. Batra, and P. Ahlawat Mann. "Survey of Density Based Clustering Algorithms." *International journal of Computer Science and its Applications* 1.1 (2011): 313–317.
11. Sandra Sagaya Mary.D.A, Tamil Selvi.R "A Study of k-means and cure clustering algorithms" International journal of Engineering Research & Technology (IJERT), vol.3 Issue 2,(2014):2278–0181.
12. Rashi Chauhan,Pooja Batna, Sarika Chaudhary "A survey of density based clustering algorithm" International journal of computer science & technology (IJCST) , vol.5,Issue 2,(2014): 0976–8491

13. Pooja Batra Nagal, Priyanka Ahlawat Mann "Comparative Study Of Density Based Clustering Algorithm" in International journal of computer science application , Volume 27-No.11,(2011): 0975–8887
14. Sikka, Sunil, and Juhi Singh. "A Comparative Analysis of Clustering Algorithms." M. A. Deshmukh, R. A. Gulhane "Importance of Clustering in Data Mining*" International Journal of Scientific & Engineering Research* (IJSER),Volume7,Issue 2 (2016):2229–5518
15. Gopinathan .S, Pandiyan .M, Thangavel .P "The New Clustering Approach Using Extreme Learning Techniques" *International Journal of Computational Intelligence Research* ISSN 0973-1873,Volume 13, Number 7(2017), pp.1669-167
16. S Syed Thouheed Ahmed, K. Thanuja, Nirmala S Guptha, Sai Narasimha, "Telemedicine approach for remote patient monitoring system using smart phones with an economical hardware kit", In proceedings of International conference on computing technologies and intelligent data engineering, pp 1-4, IEEE 2016
17. Syed Thouheed Ahmed, "A Study on multi objective clustering techniques for medical datasets" In proceedings of International Conference on Intelligent Computing and Control Systems, pp 174-177, IEEE, 2017
18. S Syed Thouheed Ahmed, Kiran Kumari Patil, "Novel Brest Cancer detection technique for TMS –India with dynamic analysis approach" In proceedings of International Conference on Communications and Signal Processing, IEEE, 2016
19. Thouheed Ahmed S., Sandhya M. "Real-Time Biomedical Recursive Images Detection Algorithm for Indian Telemedicine Environment". In: Mallick P., Balas V., Bhoi A., Zobaa A. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 768. Springer, Singapore 2019