Volume: 07, Issue: 01, June 2018, Page No.13-20
ISSN: 2278-2419

Review on Density Based Clustering Algorithms for Big Data

T.Miranda Lakshmi[1], R.Josephine Sahana[2], V.Prasanna Venkatesan[3]
[1]Department of Computer Science, Research and Development Centre, Bharathiyar University, Coimbatore, India
[2]Research Scholar, PG & Research Department of Computer Science, St.Joseph's College of Arts and Science, Cuddalore, Tamilnadu. India
[3]Department of Banking Technology, Pondicherry University, Puducherry, India
Email: cudmiranda@gmail.com, sahanarajendran1994@gmail.com, prasanna_v@yahoo.com

Abstract - Big data is termed as huge volume of both structured and unstructured dataset. When dealing with these huge dataset several challenges are encountered by the users such as analysis, capture, storage, search, transfer, sharing, and visualization. To handle these complex data sets numerous technologies have been developed such as data mining, web mining, machine learning and optimization methods. From the above technologies data mining technology is considered in this paper. Data mining technology contains various techniques such as classification, prediction and clustering etc., This paper reviews the basic concepts and clustering approaches used for big data and also a comparison study is carried out in DBSCAN, DENCLUE, and OPTICS algorithms. Each of these methods has its own pros and cons. This paper reveals that DBSCAN is efficient than the other algorithms because of its simplicity.

Key words: Big data, Clustering, DBSCAN, DENCLUE, OPTICS.

I. INTRODUCTION

Big data is termed as huge volume of both structured and unstructured dataset. When dealing with these huge dataset several challenges are encountered by the users such as analysis, capture, storage, search, transfer, sharing, and visualization. To handle these complex data sets numerous technologies have been developed. With the developing technologies and all its connected devices, it is predicted that vast amount of data is produced in the last couple of years [1]. When dealing with these huge dataset several challenges are encountered by the user such as analysis, capture, storage, search, transfer, sharing, and visualization. To handle this complex big data challenges numerous technologies and tools have been developed. In today's environment, massive amount of data is generated regularly from various sources as given in Table I.

Table I : Data generated in Internet in a minute in 2017

Sources of Data	Amount of data generated
Google	3.5 million search queries
Facebook	900,000 logins
YouTube	4.1 Million videos viewed
Google play, App Store	342,000 Apps Downloaded
Instagram	46,200 Posts downloaded
Twitter	452,000 Tweets sent
Mail	156 million Emails Sent
LinkedIn	120 New Accounts created
Messenger	15000 GIFs Sent via Messenger
Amazon echo	50 Voice – First Devices Shipped

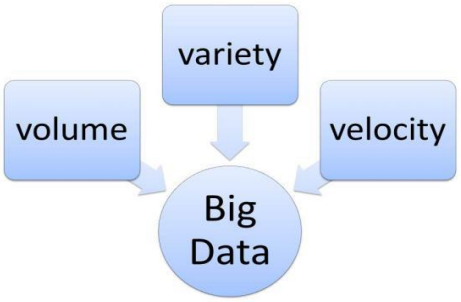
The large volume of data which is consider as big data is continuously changing factor and newer tools are being developed to handle the big data. In 2001, Gartner defines big data as follows: "Data with high volume,

velocity and with high variety that requires cost – effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation is known as big data". The above Table I. Shows that social media sites and the amount of data generated in the year 2017. From the table it is clear that the social media sites gives the major contribution in big data analysis. Many data scientists and experts characterize the Big Data by 3V's: Volume, Variety and Velocity as in Figure-1.

Volume (Data at Rest): It refers to quantity of data. Volume of data has grown from Megabytes and Gigabytes to Petabytes and Exabyte of data.

Velocity (Data in Motion): Velocity means how speed the data is generated and processed.

Variety (Data in Many Forms): It means data of various formats. It may be structured, semi-structured and unstructured. (Spreadsheet, Relational Database, ERP, Email, SocialMedia, Audio, Video).



ig.1: 3V's of Big Data

Now more dimensions are added to better understand the big data such as Veracity (data in doubt), Vision (purpose), Verification, Validation, Value, Visualization and Complexity. To overcome the difficulties in processing, big data has to be clustered in a compact format.[2]. Clustering is a data mining technique which is an essential for investigating big data where large volume of data should be grouped [3]. This paper is structured as follows: section 2 provides the various works related to density based clustering algorithms. Section 3 provides the fundamental information about clustering; Section 4 illustrates the basics of density based clustering algorithms. Section 5 concludes the paper with future work.

II. LITERATURE REVIEW

There are many algorithms available for clustering the big data. This section focus on the density based clustering algorithms and its related works. In recent years 2.5 quintillion bytes of data is generated from IoT devices such as sensors, RFIDs, Remote sensing Satellites etc., [7].Big data processing is aimed at data mining which refers to process of mining of useful information from bulky amount of data. It is classified into three types (i) Batch Processing (ii) Stream Processing (iii) Hybrid Processing [8]. Big Data Management also gaining importance because extracting the significant value out of huge amount of data is important to make different decision [9]. However big data also carries lots of inaccuracy in the quality of data such as wrong collection of data with lots of missing values [10]. Consolidation of cloud computing and Big Data can enhance the process of big data mining which enables businesses to improve decision making process [11]. The infinite capacity of cloud provide the basis of Big Data Applications such as managing an enterprise, IoT, Social Network, Medical application, & smart grids [12].

Data mining methods are used to extract the meaningful information from the large amount data [13]. Clustering is one of the most significant methods of extraction of knowledge [14]. It also plays a very important role in investigating big data [15].The clustering is a process where large volume data should be grouped. [3]**.** Clustering is a method to organize the things into a set in a way that they have high intra-cluster similarity and low inter cluster similarity [16]. Since clustering is the key of Big Data analytics, it is measured as an vital preprocessing step for the detection of information from enormous data [17]**.** ST-DBSCAN is an algorithm for

clustering spatial–temporal data based on DBSCAN. This is to determine the similar set of objects on spatial–temporal data and to discover noise objects when grouping of dissimilar densities exist. It introduces a novel notion known as density factor that is magnitude of the density of the cluster [18]**.** PACADBSCAN algorithm is a combination of partitioning-based DBSCAN and customized ant clustering algorithms that can splits the database into N groups based on the density of data, and then cluster each partition with DBSCAN. [19]**.** A novel and well-organized grid density based clustering approach consider the objects as a tiny unit and it apply the i-th order neighboring technique and do the density recompense to enhance the accuracy and finally the measure of minimal subspace distance is used discover clusters in subspaces [20]**.** DBSCAN is a foundation for clustering algorithms which are based on density. The benefit of this method is it does not need the user to specify the number of clusters prior and it can identify the clusters of arbitrary shapes.

OPTICS creates an improved cataloging of the database on behalf of its density based clustering structure. [21]. G-DBSCAN is a graph based DBSCAN which make use of a graph G (V, E) for data indexing, where V indicates the objects to be clustered and E indicates the edges then BFS method is applied on data for clustering. [22]**.** DBCURE MR is an improvement of existing DBCURE algorithm. The conventional density-based algorithms find each cluster separately and DBCURE-MR finds several clusters simultaneously [23]**.** The massive raise in the amount of data generated by all the sources leads to the proposal of highly scalable clustering algorithms. HiClus is such type of algorithm which is based on heterogeneous cloud computing [24]**.** Efficient incremental density-based algorithm enhances the incremental clustering process by limiting the search space to partitions rather than the whole dataset[25]**.**

FSFDP and DBSCAN and many others are restricted by its computation time of mutual distances between points .DGB is a clustering method which eliminates the calculation of mutual distances [26]. Density based algorithm is not appropriate for data with high variance in density and also it needs two parameters r and MinPts. An Efficient And Scalable Density-Based Clustering Algorithm takes one parameter and can handle high dimensional information set. It also reduces the number of iteration [27]**.** DENCLUE (DENsity-based CLUstEring) is a method that is based on the concept of density and the Hill Climbing algorithm. To increase the performance of DENCLUE the Hill Climbing method can be replaced by Simulated Annealing (SA) and by a Genetic Algorithm (GA). DECNLUE-SA shows its improvement in terms of fast execution time where the DENCLUE – GA reduce the time to indexing the clusters [28].DENCLUE-IM is an enhanced edition of DENCLUE. It reduces the time and enhances the speed by keep away from the difficult step in existing DENCLUE that is hill Climbing step [29]**.**

CFSFDP is an algorithm that can effectively find clusters with arbitrary shapes. [30]**.** RTDBC is a Real Time Density-Based Clustering Algorithm for Big Data which is developed to decrease the problems in DBSCAN. In this process data points are selected into clusters using labels representatives which produce fast result. However, RTDBC is faster when compared to DBSCANR and DBSCAN [31]. Enhanced Density-based Data Stream (EDDS) is developed to overcome the limitations of existing DBSCAN algorithm. The algorithm identifies the clusters and outliers in the new incoming data and merges them in to the existing clusters. This algorithm projected a new approach of representing output clusters using only surface-core points [32].

III. CLUSTERING ALGORITHMS

Clustering is a method for assembling the data object. So that things in the identical clusters are grouped together and things in the distinct clusters are grouped together. It can be categorized into five types as indicated in the Table. II.

A. Partitioning Based Clustering

Partitioning method segregates the given data objects into number of divisions known as clusters. In this approach each cluster requires to contain a minimum of one data object and each data object should belong to exactly one group. There are several algorithms for partitioning the data objects. K-means, K-medoids, k-mods, PAM, CLARA, CLARANS and FCM are examples of partitioning based clustering approach.

Table II: Categorization of Clustering Algorithm

	Clustering Approach	Algorithm
CLUSTERING	1.Partition Based	k-means, k-medoids, k-mods, PAM, CLARA,CLARANS and FCM
	2.Hierarchical Based	AGNES,DIANA,BIRCH & Chameleon
	3.Density Based	DBSCAN, DENCLUE and OPTICS [4](Ankerst, Breunig, Kriegel, & Sander, 1999)
	4.Grid Based	STING and CLIQUE
	5.Model Based	MCLUST, EM and COBWEB

B. Hierarchical Based Clustering

In hierarchical based clustering data objects are clustered in a hierarchical manner. There are two approaches in hierarchical partitioning. They are (i) Top – Down approach (Agglomerative) (ii) Bottom – Up approach (Divisive). Top down approach starts with single data object and then merges into the complete cluster. Bottom up approach starts with the entire cluster and splits them into single data object. AGNES (AGglomeratice NESting) is an example of top down approach where DIANA (DIvisive ANAlysis) is an example of bottom up approach. BIRCH and Chameleon are some other types of hierarchical approach.

C. Density Based Clustering

The prior clustering methods are not capable of finding clusters of arbitrary shapes. To satisfy this condition density based clustering algorithms are evolved which performs grouping that depends on density. Density based clustering method has the capability of finding the clusters of arbitrary shapes. It also prevents from outliers. DBSCAN[5], DENCLUE [6] and OPTICS [4] are example of density based clustering algorithm.

D. Grid Based Clustering

It divides the space of data objects into predetermined number of cubicles that forms an organization of grids. The major benefit of this method is its high-speed processing time because it only depends on numbers of cells and independent of number of things. It is an efficient method to many spatial data mining problems. STING and CLIQUE algorithms are example of this method.

E. Model Based Clustering:

Model based clustering approach is based on the improvement of relationship between the predefined mathematical model and the specified model. MCLUST, EM and COBWEB are example of such type of algorithms.

IV. DENSITY BASED CLUSTERING

This section focused on the three main and popular density based algorithm. They are DBSCAN, OPTICS, and DENCLUE.

A. DBSCAN (Density Based Spatial Clustering Application with Noise):

DBSCAN is a clustering algorithm that is based on density. It uses two parameters such as **ε** and **MinPts**. **ε** stands for Eps – neighborhood of a point and **MinPts** denotes least amount of points in an Eps-neighborhood.Both of these parameters are specified by the user. Density means number of points in the circle. There are three points as in Figure 2. They are Core Point, Border Point and Noise Point

Core Point: In Eps – neighborhood, if a point has greater value than a precise number of points that is MinPts then that particular point is said to be a core point. They are residing at the internal of a cluster.

Border Point: In Eps – neighborhood, if a point has less than a precise number of points that is MinPts then that particular point is said to be a border point. They are residing at the border of a cluster.

Noise Point: A point which does not comes under in core or border is said to be a noise point.

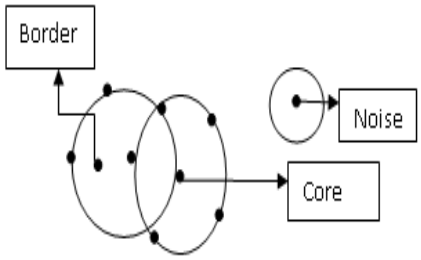


Fig.2. Density Points

Directly Density Reachable: A point **x** is directly density reachable from a point **z** as in Figure 3, with respect to Eps, MinPts if x belongs to $N_{Eps}(z)$ and it should satisfy the following condition that is : $| N_{Eps}(z) | \geq$ MinPts.

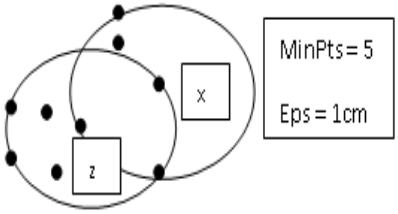


Fig.3. Directly Density Reachable

Density Reachable: If there are points like $x_1, x_2 \ldots . x_n$ and x_1 is equal to z, and x_n is equal to x and x_{i+1} is directly density reachable from x_i then that point x and z are said to be density reachable with respect to Eps, MinPts as in Figure 4.

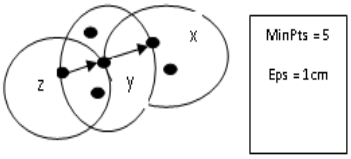


Fig.4. Density Reachable

Density Connected: If there is three points as x, y, z and if both x and z are density reachable from **y** then a point **x** is density connected to a point **z** with respect to Eps, MinPts as in Figure 5.

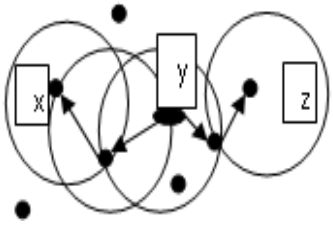


Fig-5 Density Connected

ALGORITHM

- Arbitrarily selects an unvisited point P
- Selects each and every point's density reachable from P wart Eps,Minpts.
- If P is a core point, form a cluster
- If P is a border point and no points are density reachable from P
- Visit the next unvisited point of the database
- Else add to Noise
- Until no object is unvisited.

Advantage

- Find out clusters of uninformed shapes
- Handle noise and outliers
- One Scan method

Disadvantage

- Sensitive to density parameters
- Not suitable when various density involved

B. Ordering Points to Identify Clustering Structure (OPTICS)

OPTICS is an extension to DBSCAN. In this method points with the higher density are processed first and discover clusters of higher density first. It organizes each and every object in the database and stores the values of core and reachability distance. It maintains a record known as Order Seeds to construct the output ordering.

Core Distance: Core distance is nothing but the least value of ε that should be present in the ε - neighborhood of a P. That means it must hold at least MinPts objects.
Reachability Distance: Reachability distance between p and q is defined as the least radius value that formulates p density reachable from q.

ALGORITHM

- Randomly selects an unvisited point P
- Selects all point's density reachable from P w.r.t Eps, Minpts.
- Assign core distance & reachability distance = NULL
- If P is not a core point
- Move next point in the order Seeds list
- If P is a core point
- For each object q, in the ε - neighborhood of P
- UPDATE reachability distance from P
- If q is unvisited INSERT q into Order Seeds
- Until no object is unvisited.

Advantage

- It does not require density parameters.
- Clustering order is useful in extract the basic clustering information.

Disadvantage

- It only produces a cluster ordering.
- It can't handle high dimensional data

C. Density – based Clustering (DENCLUE)

DENCLUE is useful for datasets with great quantity of noise. It uses grid cells to form a cluster and manage these grid units in a structure called tree. Clusters can be defined accurately by classifying density attractors and two functions called influence function, density functions.

Density Attractors: They are extreme values of the entire function.
Influence function: It expresses the effect of data points within its neighborhood.
Density Function: It is termed as the sum of all the influence functions.

Algorithm

- Take dataset in Grid whose each side is of 2σ
- Discover cells with higher density
- Calculate mean of highly dense cells
- If d (mean (c1), mean (c2)) < 4a then two cubes are said to be connected.

- Then cube that are connected to largely populated cells will be taken into account in defining clusters.
- Find density attractors using a Hill Climbing method
- Arbitrarily selects point r
- Calculate local 4σ density
- Choose a further point (r+1) close to calculated density.
- If den(r) < den (r+1) climb
- Place points within (σ/2) of path into cluster
- Connect the density attractor based cluster.

Advantage

- Faster than existing algorithms
- Good for database that contains huge amount of noise
- Produce precise result

Disadvantage

- Needs a large number of parameter
- Density parameter should be selected carefully
- It needs strong mathematical foundation

D. COMPARISON OF ALGORITHMS

The following Table III summarizes the comparison of the above density based algorithms.

Table III: Comparison of various density based algorithms

Algorithm	**DBSCAN**	**OPTICS**	**DENCLUE**
Size Of the Dataset	Large	Large	Large
Handling High Dimensional Data	No	No	Yes
Handling Noise	No	Yes	Yes
Complexity	O(n log n) if a spatial index is used otherwise $O(n^2)$	O(n log n)	O (log | D |)

The above table shows that all these algorithms have the capability of handling huge dataset. DBSCAN and OPTICS are not capable of handling high dimensional data; in such situation DENCLUE can be used. Both OPTICS and DENCLUE has the capability of handling outliers. Even though all these algorithms have its own pros and cons, DBSCAN is considered as an efficient method than the other algorithms because of its simplicity.

V. CONCLUSION

This paper illustrates the basic concepts of clustering algorithms to process the big data. It also discusses the advantages and disadvantages of DBSCAN, OPTICS and DENCLUE. It concludes that density based algorithms are suitable for finding clusters of arbitrary shapes and DBSCAN is considered as an efficient method than the other algorithms because of its simplicity. In future the problems encountered in the existing methods can be overcome by developing a hybrid based density algorithm.

References

[1] O. Y. Al-jarrah, P. D. Yoo, S. Muhaidat, and G. K. Karagiannidis, "Efficient Machine Learning for Big Data : A Review ☆," Big Data Res., vol. 1, pp. 1–7, 2015.
[2] A. Fahad et al., "A Survey of Clustering Algorithms for Big Data : Taxonomy & Empirical Analysis," 2014.
[3] O. Kurasova, V. Marcinkevi˘, V. Medvedev, and A. Rape˘, "Strategies for Big Data Clustering," 2014.
[4] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander, "OPTICS : Ordering Points To Identify the Clustering Structure," 1999.

[5] M. Ester, H. Kriegel, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise."
[6] A. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise," no. c, 1998.
[7] R. Ranjan, D. Thakker, A. Haller, and R. Buyya, "A note on exploration of IoT generated big data using semantics," Futur. Gener. Comput. Syst., vol. 76, pp. 495–498, 2017.
[8] N. Miloslavskaya and A. Tolstoy, "Big Data , Fast Data and Data Lake Concepts 2 Big Data Concept," Procedia - Procedia Comput. Sci., vol. 88, pp. 300–305, 2016.
[9] N. Koseleva and G. Ropaite, "Big data in building energy efficiency : understanding of big data and main challenges," Procedia Eng., vol. 172, pp. 544–549, 2017.
[10] J. Liu, J. Li, W. Li, and J. Wu, "ISPRS Journal of Photogrammetry and Remote Sensing Rethinking big data : A review on the data quality and usage issues," ISPRS J. Photogramm. Remote Sens., 2015.
[11] M. Balachandran, "ScienceDirect ScienceDirect ScienceDirect Challenges Deploying Challenges and and Benefits Benefits of of Deploying Big Data Data Analytics Analytics in in the the Cloud Cloud for for Business Business Intelligence Intelligence Big," Procedia Comput. Sci., vol. 112, pp. 1112–1122, 2017.
[12] A. Leva, A. Vittorio, A. V. Papadopoulos, A. Leva, A. V. Papadopoulos, and A. V. Papadopoulos, "ScienceDirect and Control of Modelling and Control of Modelling and Control Modelling and Control of of Frameworks Frameworks Frameworks Frameworks Big Big Big Big Data Data Data Data," IFAC-PapersOnLine, vol. 50, no. 1, pp. 6110–6115.
[13] I. Yaqoob et al., "International Journal of Information Management Big data : From beginning to future," Int. J. Inf. Manage., vol. 36, no. 6, pp. 1231–1247, 2016.
[14] A. C. Jinyin et al., "A Novel Cluster Center Fast Determination Clustering Algorithm," Appl. Soft Comput. J., 2017.
[15] P. Arora and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm For Big Data," Procedia - Procedia Comput. Sci., vol. 78, no. December 2015, pp. 507–512, 2016.
[16] A. E. Karrar and M. Mutasim, "Comparing EM Clustering Algorithm with Density Based Clustering Algorithm Using WEKA Tool," vol. 5, no. 7, pp. 2015–2017, 2016.
[17] M. S. Hidri, M. A. Zoghlami, and R. Ben Ayed, "Speeding up the large-scale consensus fuzzy clustering for handling Big Data," Fuzzy Sets Syst., vol. 1, pp. 1–25, 2017.
[18] D. Birant and A. Kut, "ST-DBSCAN : An algorithm for clustering spatial – temporal data," vol. 60, pp. 208–221, 2007.
[19] H. Jiang, J. Li, S. Yi, X. Wang, and X. Hu, "Expert Systems with Applications A new hybrid method based on partitioning-based DBSCAN and ant clustering," Expert Syst. Appl., vol. 38, no. 8, pp. 9373–9381, 2011.
[20] Y. Zhao, J. Cao, C. Zhang, and S. Zhang, "The Journal of Systems and Software Enhancing grid-density based clustering for high dimensional data," J. Syst. Softw., vol. 84, no. 9, pp. 1524–1539, 2011.
[21] R. Awasthi, A. K. Tiwari, and S. Pathak, "Analysis of Mass Based and Density Based Clustering Techniques on Numerical Datasets," vol. 3, no. 4, pp. 29–35, 2013.
[22] G. Andrade, G. Ramos, and D. Madeira, "G-DBSCAN : A GPU Accelerated Algorithm for Density-based Clustering," Procedia Comput. Sci., vol. 18, pp. 369–378, 2013.
[23] Y. Kim, K. Shim, M. Kim, and J. Sup, "DBCURE-MR : An efficient density-based clustering algorithm for large data using MapReduce," Inf. Syst., vol. 42, pp. 15–35, 2014.
[24] C. Chen and M. Chen, "HiClus : Highly Scalable Density-based Clustering with Heterogeneous Cloud," Procedia - Procedia Comput. Sci., vol. 53, pp. 149–157, 2015.
[25] A. M. Bakr, N. M. Ghanem, and M. A. Ismail, "Efficient incremental density-based algorithm for clustering large datasets," Alexandria Eng. J., vol. 54, no. 4, pp. 1147–1154, 2015.
[26] B. Wu, S. Member, B. M. Wilamowski, and L. Fellow, "Transactions on Industrial Informatics A Fast Density and Grid Based Clustering Method for Data with Arbitrary Shapes and Noise," vol. 3203, no. c, 2016.
[27] K. A. Patel, "An Efficient And Scalable Density-Based Clustering Algorithm For Normalize Data," Procedia - Procedia Comput. Sci., vol. 92, pp. 136–141, 2016.
[28] A. Idrissi, H. Rehioui, A. Laghrissi, and S. Retal, "An Improvement of DENCLUE Algorithm for the Data Clustering."
[29] H. Rehioui, A. Idrissi, M. Abourezq, and F. Zegrari, "DENCLUE-IM: A New Approach for Big Data Clustering," in Procedia Computer Science, 2016.
[30] L. Bai, X. Cheng, J. Liang, H. Shen, and Y. Guo, "Fast density clustering strategies based on the k -means algorithm," vol. 71, pp. 375–386, 2017.
[31] B. R. Prasad, "Real Time Density-Based Clustering (RTDBC) Algorithm for Big Data," 2017.
[32] A. Al and A. Alazeez, "EDDS : An Enhanced Density-based Method for Clustering Data Streams," 2017.