Volume 9, Issue 06, June 2024 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

Analysis and Implementation of Density-Based DBACAN Clustering to Create More Accurate Clusters for Dense Dataset

Shivani Raghuvanshi
ME(Software Engineering) 4th semester
Computer Science and Engineering Department
Jawaharlal Institute of Technology Borawan (Khargone)

Mr. Kamlesh Patidar
Assistant Professor
Computer Science and Engineering Department
Jawaharlal Institute of Technology
Borawan (Khargone) M.P.

Abstract: ***Clustering techniques has may challenges like handling Outliers, dealing with high dimension data. Multiple Solutions provided by different methods, evaluating and identifying accurate solutions and finally computation limits. In the proposed work we used Density-based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN clustering provides several advantages over other clustering techniques like it Handles irregularly shaped and sized clusters, does not require the number of clusters to be specified, Less sensitive to initialization conditions, Robust to outliers and Relatively fast. We analysis the DBSCAN on real life data set. We have taken real life data set of Mall Customer and created clusters using different parameters like different age groups mostly visited the mall, how they spend money in mall based on income and age groups. We have implemented DBSCAN using python language. By the implementation we found that females are highly visited mall as compared to male Peoples of age between 25 to 40 are mostly visiting mall than other age groups. We also analysis the number of outliers present in the data set and number of objects in each cluster***

Keywords: ***Density-based, Clustering, Customers, Spatial Clustering, Noise, Core, Border point***

I. BACKGROUND

Clustering is a type of unsupervised learning method of machine learning. In the unsupervised learning method, the inferences are drawn from the data sets which do not contain labelled output variable. It is an exploratory data analysis technique that allows us to analyze the multivariate data sets. Clustering is a task of dividing the data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics. Clusters are nothing but the grouping of data points such that the distance between the data points within the clusters is minimal. Clustering is done to segregate the groups with similar traits. In other words, the clusters are regions where the density of similar data points is high. It is generally used for the analysis of the data set, to find insightful data among huge data sets and draw inferences from it. Generally, the clusters are seen in a spherical shape, but it is not necessary as the clusters can be of any shape. It depends on the type of algorithm we use which decides how the clusters will be created. The inferences that need to be drawn from the data sets also depend upon the user as there is no criterion for good clustering. Clustering analysis is an unsupervised learning method that separates the data points into several specific bunches or groups, such that the data points in the same groups have similar properties and data points in different groups have different properties in some sense.

All clustering methods use the same approach i.e. first we calculate similarities and then we use it to cluster the data points into groups or batches. Here we will focus on the Density-based spatial clustering of applications with noise **(**DBSCAN**)** clustering method. "Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu introduced an unsupervised machine learning algorithm called density-based clustering in their 1996 paper 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.'" The algorithm, known as DBSCAN, was designed to handle large spatial databases with noise and to discover clusters of arbitrary shape[9,10].

II. TYPES OF CLUSTERING METHODS

Partitioning-based Clustering

Partitioning objects into k number of clusters where each partition makes/represents one cluster, these clusters hold certain properties such as each cluster should consist of at least one data object and each data object should be classified to exactly one cluster. These methods are broadly classified to optimize a targeted benchmark similarity function such that distance becomes a significant parameter to consider first. The examples are

- K-means clustering, (understand K-means clustering from here in detail)

- CLARANS (Clustering Large Applications based upon Randomized Search)

Moreover, Partitioning clustering algorithms are the form of non-hierarchical that generally handle statics sets with the aim of exploring the groups exhibited in data via optimization techniques of the objective function, making the quality of partition better repeatedly. Partitioning-based clustering is highly efficient in terms of simplicity, proficiency, and easy to deploy, and computes all attainable clusters synchronously.

Hierarchical-based Clustering

Depending upon the hierarchy, these clustering methods create a cluster having a tree-type structure where each newly formed clusters are made using priorly formed clusters, and categorized into two categories: **Agglomerative (bottom-up approach) and Divisive (top-down approach).** The examples of Hierarchical clustering are

- CURE (Clustering Using Representatives)
- BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies)

The agglomerative clustering method is achieved by locating each point in a cluster, initially and then merging two points closest to it where points represent an individual object or cluster of objects. The divisive clustering first considers the complete population as one cluster and then segments into smaller groups.

Density-based Clustering

These methods of clustering recognize clusters of dense regions that possess some similarity and are distinct from low dense regions of the space. These methods have sufficient accuracy and the high ability to combine two clusters. Its examples include

- DBSCAN (Density-based Spatial Clustering of Applications with Noise)
- OPTICS (Ordering Points to Identify Clustering Structure)

These methods implement distance measures between the objects in order to cluster the objects. In most of the cases, clusters, produced using this method, are spherical in shape, so sometimes it becomes hard to identify arbitrary shaped clusters. Moreover, clusters are produced in all directions as long as the density, residing neighbourhood, surpass some threshold. Density-based methods save data sets from outliers, the entire density of a point is treated and deciphered for determining features or functions of a dataset that can impact a specific data point. Some algorithms like OPTICS, DenStream, etc deploy the approach that automatically filtrates noise (outliers) and generates arbitrary shaped clusters.

Grid-based Clustering

This method follows a grid-like structure, i.e, data space is organized into a finite number of cells to design a grid-structure. Various clustering operations are conducted on such grids (i.e quantized space) and are quickly responsive and do not rely upon the quantity of data objects. Its examples are;

- STING (Statistical Information Grid),
- Wave cluster,
- CLIQUE (Clustering In Quest)

Computing statistical measurements for the grids consequently increasing the speed of method extensively. Also, the performance of grid-based methods is proportional to the grid-size and demands very less space than the actual data stream[11,12,13].

III. DBSCAN CLUSTERING

DBSCAN stands for **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise.

It was proposed by Martin Ester et al. in 1996. DBSCAN is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density It groups 'densely grouped' data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points. The most exciting feature of DBSCAN clustering is that it is robust to outliers. It also does not require the number of clusters to be told beforehand, unlike K-Means, where we have to specify the number of centroids. DBSCAN requires only two parameters:

epsilon and *minPoints. Epsilon* is the radius of the circle to be created around each data point to check the density and *minPoints* is the minimum number of data points required inside that circle for that data point to be classified as a **Core** point. In higher dimensions the circle becomes hypersphere, *epsilon* becomes the radius of that hypersphere, and *minPoints* is the minimum number of data points required inside that hypersphere.

DBSCAN creates a circle of *epsilon* radius around every data point and classifies them into Core point, Border point, and **Noise**. A data point is a **Core** point if the circle around it contains at least '*minPoints'* number of points. If the number of points is less than *minPoints*, then it is classified as Border Point, and if there are no other data points around any data point within *epsilon* radius, then it treated as Noise[14,15].

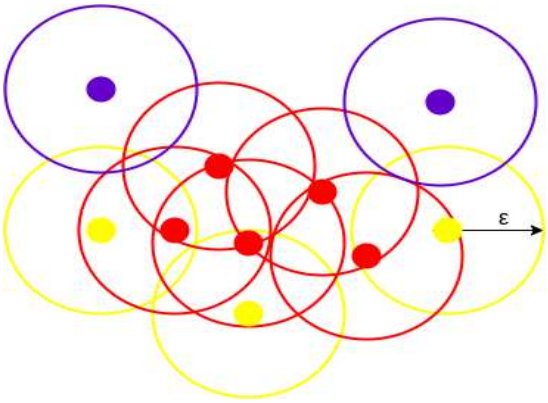

Figure 1 **Core** point, **Border** point, and **Noise**

DBSCAN creates a circle of *epsilon* radius around every data point and classifies them into **Core** point, **Border** point, and **Noise**. A data point is a **Core** point if the circle around it contains at least '*minPoints'* number of points. If the number of points is less than *minPoints*, then it is classified as **Border** Point, and if there are no other data points around any data point within *epsilon* radius, then it treated as **Noise**. The above figure shows us a cluster created by DBCAN with *minPoints = 3*. Here, we draw a circle of equal radius *epsilon* around every data point. These two parameters help in creating spatial clusters.

IV. LITERATURE SURVEY

In 2017 K. Chitra et al proposed "A Comparative Study of Various Clustering Algorithms in Data Mining". Clustering is a significant task in data analysis and data mining applications. They focused on a keen study of different clustering algorithms in data mining. They discussed a brief overview of various clustering algorithms . The objective of the data mining technique is to mine information from a large data set and make it into a reasonable form for the supplementary purpose. Clustering algorithms can be classified into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. Under partitioning method, a brief description of k-means and k-medoids algorithms have been studied. In hierarchical clustering, the BIRCH and CHAMELEON algorithms have been described. The DBSCAN and DENCLUE algorithms under the density based methods have been studied. Finally, under grid-based clustering method, the STING and CLIQUE algorithms have been described[1].

In 2018 Yewang Chen et al proposed "A Fast-Clustering Algorithm based on pruning unnecessary distance computations in DBSCAN for High-Dimensional Data". They proved theoretically and experimentally that ρ-Approximate DBSCAN degenerates to an O(n 2) algorithm in very high dimension such that 2D >> n. They proposed a novel local neighborhood searching technique, and apply it to improve DBSCAN, named as NQ-DBSCAN, such that a large number of unnecessary distance computations can be effectively reduced. Theoretical analysis and experimental results show that NQ-DBSCAN averagely runs in O(n ∗ log(n)) with the help of indexing technique, and the best case is O(n) if proper parameters are used, which makes it suitable for many real-time data . They proposed a clustering algorithm, named NQ-DBSCAN which may return the exact result as DBSCAN, to improve DBSCAN, by using neighbor searching technique and indexing technique to filter great number of unnecessary density computations. Although, the worse complexity of NQ-DBSCAN is still O(n 2), but its average complexity is about O(n ∗ log(n)) with the help of indexing technique, 540 and the best case is O(n) if proper parameters are used[2].

In 2018 Mingrui Zhang et al proposed "Use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm to Identify Galaxy Cluster Members" They ensured the correctness of classification? Based on the results of CoDECS numerical simulation and combining DBSCAN algorithm, they attempted to classify the data and compare and explain the results of the three methods. Then, based on the data of Abell 383 cluster, further comparison and analysis of the three methods were made. By comparing with the correct figure, they found that the DBSCAN algorithm can accurately identify all classes and eliminate noise interference to a certain extent, which is impossible to be achieved by the K-Means algorithm and the Decision Tree algorithm. DBSCAN method's advantage is its classification standard is classified by density, not limited to the shape of the classification, and so on these data can get very good result, but according to the real Abell data[3].

In 2019 Vikram Neerugatti et al proposed "Density Based Spatial Clustering Application with Noise by Varying Densities. They proposed algorithm and developed a model based on the existing DBSCAN algorithm. In the proposed algorithm they focus mainly on the epsilon parameter value. Whenever the DBSCAN algorithm fails to form a cluster, they increase the epsilon value by half of its original size. They repeated this step until a cluster is formed. Whenever a cluster is newly formed, we change existing epsilon parameter value by adding the 10 percent of the previous used epsilon parameter value. They used epsilon for varying the density of a cluster. So, they used the DBSCAN algorithm with the varying density values for developing a cluster. They applied this algorithm on the various datasets Cluster algorithms are used for grouping up of similar points to form a cluster. The most popular density-based algorithm is DBSCAN. By applying this algorithm on the three different size of datasets we have achieved a better silhouette score from these three datasets than that is achieved by the existing dbscan algorithm[4].

In 2020 G C Pamuji et al proposed "A Comparison study of DBScan and K-Means Clustering in Jakarta rainfall based on the Tropical Rainfall Measuring Mission (TRMM) 1998-2007". They proposed comparison between two different of cluster analysis algorithm in data mining on the Tropical Rainfall Measuring Mission (TRMM). They used rainfall data in Jakarta based on TRMM was analyzed and compared in the efficiency and the accuracy using each algorithm. The comparison results of the two algorithmic processes can be seen from several parameters, especially from the number of clusters formed and the time needed to process the model. Using the data mining techniques K-Means and DBSCAN the rainfall datasets analyzed. Each algorithm formed different number of clusters. K-Means used a predetermined variable (k) with a value of 3, and DBSCAN has the ability to determine how many clusters, which can be formed based by data points. The size of the datasets is affecting the process time. K-means is faster in processing larger data sets. Each of the algorithm have their own strengths and weakness. Due to the nature of the datasets, which are from data-sets type and size, K-Means produced a more efficient and accurate results than DBSCAN[5].

In 2020 Jashraj Gandhia et al proposed "Comparative Study on Hierarchical and Density based Methods of Clustering using Data Analysis". Data Mining is extraction of useful patterns using different techniques. Clustering is one such technique. They extracted all clusters from the data to find some useful information. Hierarchical clustering and density-based clustering are two such methodologies to find clusters. Hierarchical gives its results in a dendrogram using hierarchical clustering algorithm. DBSCA N is a density-based algorithm which helps find arbitrary shaped clusters. They implemented a more efficient. They concluded by observing both efficient hierarchical clustering and efficient DBSCAN using CLARANS that, DBSCAN is more efficient to use on a dataset than hierarchical clustering, as DBSCAN has well defined concept of noise and very well outlier points whereas hierarchical clustering may take lot of time to process said data[6].

In 2021 Jing Yang1 et al proposed "Research and Application of Business Ability Evaluation Based On DBSCAN Algorithm and Entropy Method". They constructed customer commissioner's business capability evaluation models of different business types. Firstly, through DBSCAN algorithm we mine the potential category characteristics of customer commissioners. Then, entropy method is used to score the customer commissioners under the target category comprehensively. Finally combining the clustering results and entropy score, the customer commissioners with relatively strong and weak business capabilities are identified accurately. By marking the business capacity labels for the corresponding customer commissioners, it has guiding significance for the post adjustment and business training, which can improve the quality of power supply service and customer satisfaction for customer service center[7].

In 2022 Jayasree Ravi et al proposed "A Critical Review on Density-Based Clustering Algorithms and Their Performance In Data Mining". They gave about various density-based clustering algorithms, their domain-specific applications, datasets used, methods of data extraction. They focused on the performance evaluation of the algorithms which are part of this survey. They illustrated the working of DBSCAN and compares with other three algorithms. While VDBSCAN addresses the varying density issue, AGED, AE-DBSCAN and ADAPTIVE DBSCAN algorithm calculate density parameters automatically. DBSTexC and VDCT focus on identifying quality clusters by considering textual information also. DBSCAN++ aims at reducing the run-time complexity. They observed that there have been enough studies done on calculating ε values. But there has not been any study that has handled the other important parameter which is the MinPts[8].

V. PRAPOSED APPROACH

Let X = {x_1, x_2, x_3, ..., x_n} be the set of data points. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

- Start with an arbitrary starting point that has not been visited.
- Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).
- If there are sufficient neighborhood around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).
- If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster is determined.
- A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.
- This process continues until all points are marked as visited.

.

VI. COMPARATIVE ANALYSIS

Number of objects in each cluster

we have 5 clusters and 1 outlier. The `0` cluster has the largest size with 112 rows. The visualization clearly shows how each customer is part of one of the 5 clusters, and we can use this information to give high-end offers to customers with purple clusters and cheaper offers to customers with dark green clusters.

Table 1 number of objects in each clusters

Cluster No	Number of Objects
-1	18
0	112
1	8
2	34
3	24
4	4

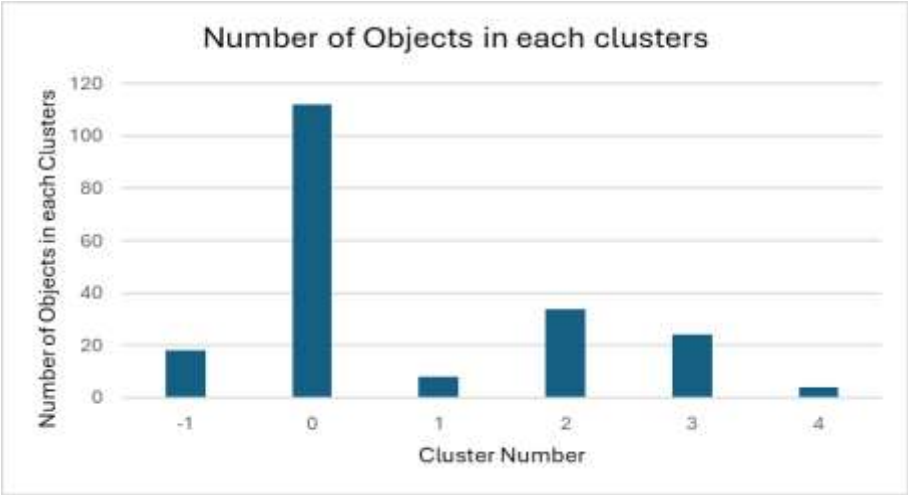

Figure Number of objects in each cluster

VII. CONCLUSION

In this paper we used Density-based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN clustering provides several advantages over other clustering techniques like it Handles irregularly shaped and sized clusters, does not require the number of clusters to be specified, Less sensitive to initialization conditions, Robust to outliers and Relatively fast. We analysis the DBSCAN on real life data set. We have taken real life data set of Mall Customer and created clusters using different parameters like different age groups mostly visited the mall, how they spend money in mall based on income and age groups. We have implemented DBSCAN using python language. By the implementation we found that females are highly visited mall as compared to male Peoples of age between 25 to 40 are mostly visiting mall than other age groups. We also analysis the number of outliers present in the data set and number of objects in each cluster**.**

REFERENCE

1. K. Chitra1 , Dr. D.Maheswari2 A Comparative Study of Various Clustering Algorithms in Data Mining International Journal of Computer Science and Mobile Computing, Vol.6 Issue.8, August- 2017, pg. 109-115
2. Yewang Chen, Shenyu Tang, Nizar Bouguila, Cheng Wang, Jixiang Du, HaiLin Li A Fast Clustering Algorithm based on pruning unnecessary distance computations in DBSCAN for High-Dimensional Data PII: S0031-3203(18)30210-3 DOI: 10.1016/j.patcog.2018.05.030 Reference: PR 6574 To appear in: Pattern Recognition Received date: 28 April 2017 Revised date: 1 February 2018 Accepted date: 31 May 2018
3. Mingrui Zhang Use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm to Identify Galaxy Cluster Members ESMA 2018 IOP Conf. Series: Earth and Environmental Science 252 (2019) 042033 IOP Publishing doi:10.1088/1755-1315/252/4/042033
4. Vikram Neerugatti, Mokkala Kiran Moni, Rama Mohan Reddy A Density Based Spatial Clustering Application with Noise by Varying Densities International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019
5. G C Pamuji 1.* , H Rongtao 2 A Comparison study of DBScan and K-Means Clustering in Jakarta rainfall based on the Tropical Rainfall Measuring Mission (TRMM) 1998-2007 INCITEST 2020 IOP Conf. Series: Materials Science and Engineering 879 (2020) 012057 IOP Publishing doi:10.1088/1757-899X/879/1/012057
6. Jashraj Gandhia , Radhika Goya Comparative Study on Hierarchical and Density based Methods of Clustering using Data Analysis International Conference on IoT based Control Networks and Intelligent Systems (ICICNIS 2020)
7. Jing Yang1,*, Kunpeng Liu1 , Mingjie Zhang2 , QingXu2 and Peng Jin1 Research and Application of Business Ability Evaluation Based On DBSCAN Algorithm and Entropy Method The 2nd International Conference on Computing and Data Science (CONF-CDS 2021) Journal of Physics: Conference Series 1881 (2021) 032064 IOP Publishing doi:10.1088/1742-6596/1881/3/032064
8. Jayasree Ravi1 , Sushil Kulkarni2 A CRITICAL REVIEW ON DENSITY-BASED CLUSTERING ALGORITHMS AND THEIR PERFORMANCE IN DATA MINING 2022 IJRAR March 2022, Volume 9, Issue 1 www.ijrar.org (E-ISSN 2348-1269, P- ISSN 2349-5138) International Journal of Research and Analytical Reviews (IJRAR).
9. Yan Zhang DBSCAN Clustering Algorithm Based on Big Data Is Applied in Network Information Security Detection Hindawi Security and Communication Networks Volume 2022, Article ID 9951609, 8 pages https://doi.org/10.1155/2022/9951609
10. Nidhi Suthar , Prof. Indr jeet Rajput , Prof. Vinit kumar Gupta "A Technical Survey on DBSCAN Clustering Algorithm" International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013 1775 ISSN 2229-5518.
11. Saif Ur Rehman, Kamran Aziz Simon Fong "DBSCAN: Past, Present and Future" Center of Excellence in Data Engineering Mohammad Ali Jinnah University Islamabad, Pakistan Saifi.ur.rehman@gmail.com, kamrandik@gmail.com Department of Computer and Information Science University of Macau Taipa, Macau SAR ccfong@umac.mo 2014
12. Dongkuan Xu1,2 · Yingjie Tian2, A Comprehensive Survey of Clustering Algorithms Ann. Data. Sci. (2015) 2(2):165–193 DOI 10.1007/s40745-015-0040-1 Received: 25 May 2015 / Revised: 18 July 2015 / Accepted: 31 July 2015 / Published online: 12 August 2015 © Springer-Verlag Berlin Heidelberg 2015
13. Pranjal Dubey, 2anand Rajavat Comparative Study Between Density Based Clustering - DBSCAN AND OPTICS International Journal of Advanced Computational Engineering and Networking, ISSN: 2320-2106, Volume-4, Issue-12, Dec.-2016.
14. K. Nafees Ahmed1 , T. Abdul Razak2 An Overview of Various Improvements of DBSCAN Algorithm in Clustering Spatial Databases International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016
15. Joasang Lim1 , Joongjin Kook2 and Jinman Kim3* DBSCAN-D: A Density-Based Clustering Method of Directionality International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 13 (2017) pp. 3927-3932 © Research India Publications. http://www.ripublication.com