IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, June 2025
Impact Factor: 7.67

An Improved DBSCAN, a Density-Based Clustering Algorithm: A Comprehensive Review

Prabhu Patel and Dr. Abhishek Singh Rathore
Department of Computer Science
Shri Vaishnav Vidyapeeth Vishwavidyalaya (SVVV) Indore, India
prabhupatelmp55@gmail.com

Abstract: *A popular clustering method, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is well-known for its capacity to identify clusters of any shape and for successfully differentiating noise in datasets. Its fundamental mechanism depends on two parameters: MinPts, the smallest number of points needed to produce a dense zone, and epsilon (ε), which specifies the neighborhood radius. Because of these characteristics, DBSCAN is very helpful for a wide range of real-world applications, including anomaly detection, picture recognition, and spatial data analysis. Notwithstanding these advantages, DBSCAN's performance deteriorates in some situations, particularly when dealing with high-dimensional datasets and situations where cluster densities differ greatly.*

To overcome the aforementioned drawbacks of the conventional technique, we introduce Improved DBSCAN, an improved version of DBSCAN, in this study. Our approach's main innovation is the dynamic density threshold method we introduced. In contrast to typical DBSCAN, which employs a fixed and global ε value for every point, our approach calculates a local density threshold for every dataset region. Because the global ε parameter in the original DBSCAN is so stiff, clusters with different densities—which are frequently misrepresented or merged—can be handled more effectively by the algorithm thanks to this adaptive approach.

Keywords: Density based Algorithm, Clustering Algorithm, DBSCAN, Noise

I. INTRODUCTION

Clustering is a fundamental technique for uncovering meaningful patterns within large datasets. It is widely applied across various domains such as marketing, computer networking, image analysis, biological research, geographic monitoring, web data mining, and healthcare applications By measuring similarities between data points, clustering facilitates the automatic grouping of unlabeled data into coherent clusters. [1]The below figure 1. Shows the graphical representation of clustering.

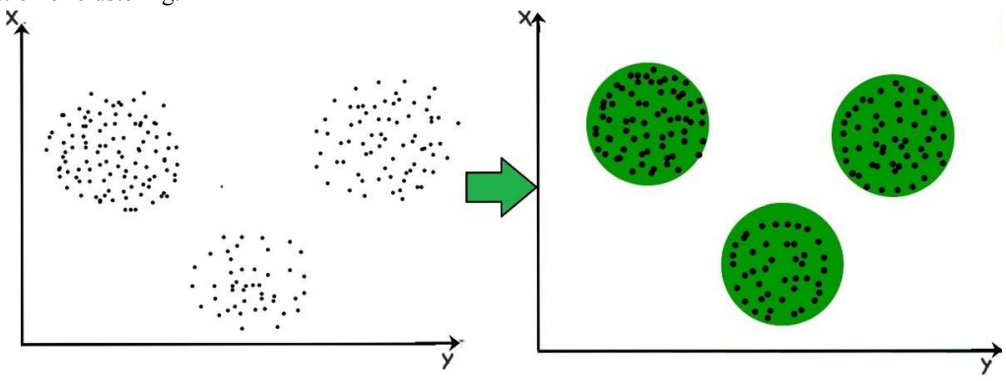


Figure 1. Graphical Representation of Clustering

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-28028

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, June 2025
Impact Factor: 7.67

There are several categories of clustering algorithms, each with its own approach to grouping data points into clusters. These categories include partitioning methods (e.g., Kmeans), hierarchical methods (e.g., agglomerative clustering), density-based methods (e.g., DBSCAN), grid-based methods (e.g., CLARANS), and model-based methods (e.g., Gaussian Mixture Models) [2]. The below figure 2. Shows the graphical representation of clustering.

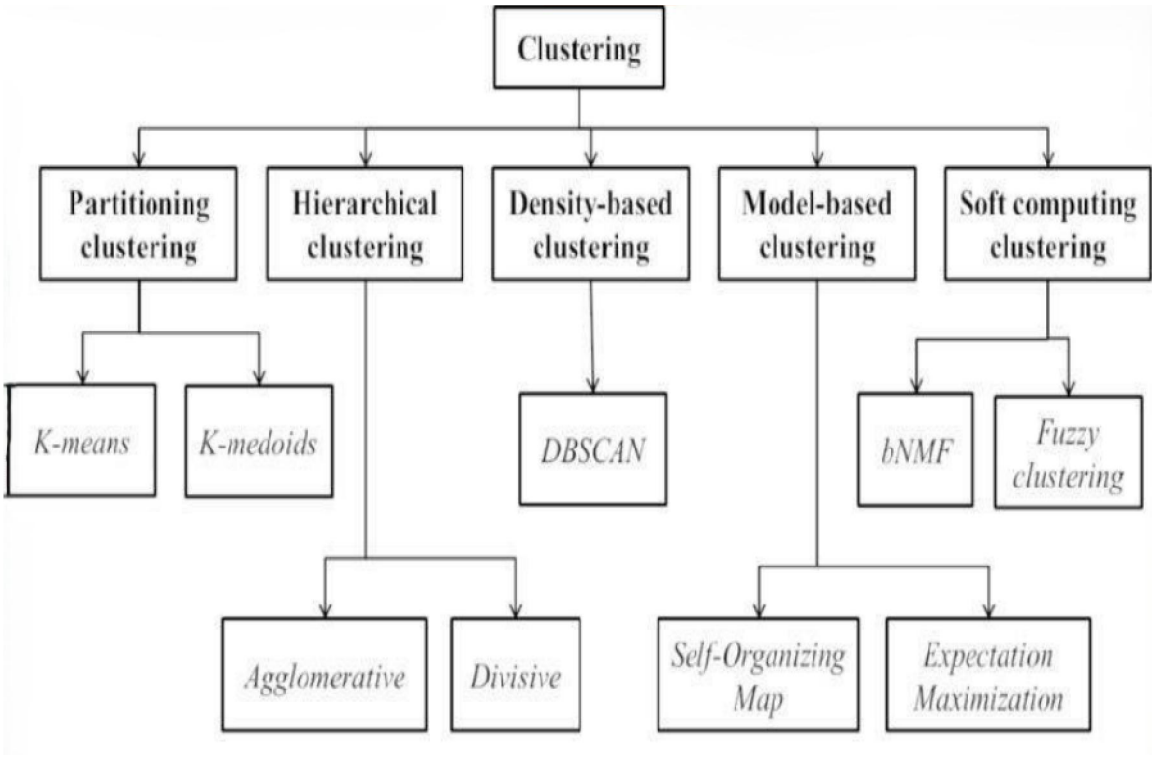
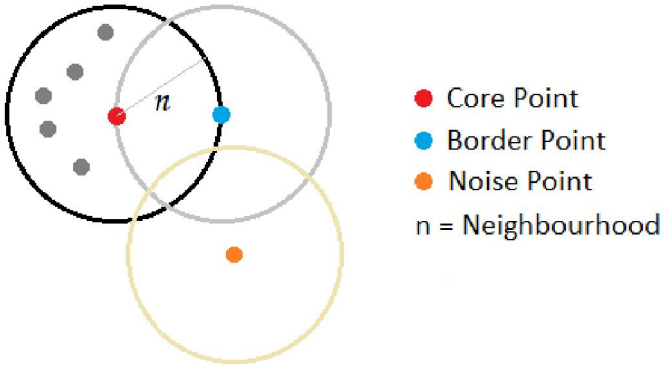


Figure 2. Shows the Graphical Representation of Cluster

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a prominent density-based clustering algorithm that effectively identifies dense regions in data to form clusters, while treating sparse regions as noise or outliers. It excels at grouping data points with similar characteristics into clusters and is particularly robust in detecting clusters of arbitrary shapes and varying sizes, even in the presence of noise.[3]

However, DBSCAN's clustering quality heavily depends on two crucial parameters: Eps (the neighborhood radius) and MinPts (the minimum number of points required to form a dense region), as illustrated in Figure 1. Setting Eps too high can lead to merging unrelated data into a single cluster, while a value that is too low may result in fragmented, overly small clusters. This parameter sensitivity makes it difficult to apply DBSCAN effectively to real-world datasets.The below figure 3. Shows the graphical representation of Dbscan.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, June 2025
Impact Factor: 7.67



DBSCAN CLUSTERING

Although previous studies have attempted to enhance DBSCAN by dynamically adjusting either Eps or MinPts a gap remains in methods that can simultaneously and adaptively tune both parameters. To address this limitation, this research introduces an enhanced DBSCAN variant that dynamically determines the initial values of both Eps and MinPts. The aim is to improve clustering quality and accuracy by optimizing parameter selection in a data-driven manner.

II. LITERATURE SURVEY

Fast DBSCAN is a density-based clustering method that groups data by identifying dense regions in a dataset. Unlike partitioning or hierarchical techniques, it excels at finding clusters of arbitrary shapes and handling noise or outliers. It is efficient for large datasets, requires minimal parameter tuning, and offers near-linear scalability. Its simplicity and ability to work without deep domain knowledge make it practical and user-friendly for real-world applications. [4]

This paper introduces an adaptive approach to select DBSCAN's parameters, Eps and MinPts, using the data's own distribution. By computing the average distance to each point's K-th nearest neighbor, candidate Eps values are generated. MinPtsis then derived as the average number of points within these Eps neighborhoods. As K increases, both parameters rise and the resulting cluster count stabilizes. The final parameters are chosen from this stable region to enhance clustering accuracy and reduce noise.To improve efficiency, the proposed K-DBSCAN algorithm processes only core points, eliminating non-core points early and merging clusters with overlapping neighborhoods. This strategy reduces the number of point evaluations and accelerates computation without affecting clustering quality.[5]

The M-DBSCAN algorithm employs Linear Congruential Method (LCM) techniques for random number generation to optimally select MinPts and Eps parameters. After modifying the original DBSCAN algorithm, M-DBSCAN was tested on five real-world datasets, successfully reducing outliers and improving runtime efficiency. Experimental results demonstrate that M-DBSCAN outperforms the standard DBSCAN across various datasets. Future work will focus on extending M-DBSCAN to effectively handle high-dimensional data for improved clustering performance.[6]

The research titled "A Faster DBSCAN Algorithm Based on Self-Adaptive Determination of Parameters" introduces enhancements to the standard DBSCAN method by resolving two primary issues: the reliance on user-defined parameters (Eps and MinPts) and the algorithm's intensive computational demands. To overcome these challenges, the authors develop a self-adaptive strategy that automatically estimates Eps and MinPts from the dataset by analyzing the average distance to each point's KKKth nearest neighbor. This dynamic approach improves clustering performance, particularly when dealing with datasets of uneven density.

To reduce execution time, the study proposes K-DBSCAN, a variant of DBSCAN that processes only the core points—those with high local density—while disregarding non-core points to minimize data traversal. Clustering is achieved by detecting overlaps between the neighborhoods of these core points and merging them accordingly. Although the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, June 2025
Impact Factor: 7.67

parameter estimation process requires repeated clustering runs, the overall design improves both computational efficiency and clustering quality, making the approach suitable for large-scale and complex data scenarios. [7].

Pramadihantoet al. (2024) propose CVR-DBSCAN, an improved DBSCAN algorithm aimed at clustering 3D point cloud data in challenging bin-picking scenarios involving piled pipes. The paper introduces two main enhancements: automated parameter estimation using k-nearest neighbor distances with elbow detection, and curvature-based core point identification through PCA. Two variants, CVR-DBSCAN-Avg and CVR-DBSCAN-Disc, refine clustering by incorporating surface normals and curvature data to better handle overlapping or irregular shapes. Experiments on Time-of-Flight 3D data show that the method achieves high accuracy (up to 99.67%) across different piling conditions. Although the method introduces additional computational steps, it remains efficient and well-suited for real-time industrial applications [8].

Al-Shaikhli et al. present SS-DBSCAN, a semi-supervised enhancement of the classic DBSCAN algorithm, aimed at improving clustering quality in datasets with varying densities their method introduces an additional "Is_important" constraint that allows users to guide the core-point selection based on domain-relevant criteria. This semi-supervised setup injects weak supervision—such as labeled points or feature-based conditions—to prevent inappropriate merging or splitting of meaningful clusters.[9]

Glory H. Shah (2012) introduced an improved DBSCAN algorithm to enhance clustering in high-dimensional datasets by automatically adjusting ε and MinPts based on data distribution. The method improves cluster detection in cases of varying density and nested clusters. It also compares distance metrics, finding that Euclidean distance provides more accurate results but with greater computational cost. The proposed approach shows better performance in terms of cluster quality, noise handling, and efficiency compared to the traditional DBSCAN [10]

III. METHODOLOGY USED

This study introduces a refined density-based clustering approach that enhances the conventional DBSCAN algorithm by addressing its key limitations, including fixed parameter dependency, inefficiency with large datasets, and poor adaptability to varying data densities. The proposed framework incorporates three major enhancements: automated parameter tuning, curvature-based refinement of core points, and optional semi-supervised input to guide clustering decisions.

A. Adaptive Parameter Tuning

To eliminate the manual selection of the ε (Eps) and MinPts parameters, the framework adopts a data-driven estimation technique. For every data instance, the distance to its K-th nearest neighbor is calculated using a KNN-based approach. These distances are analyzed to generate a distance distribution graph, and an elbow detection technique—leveraging curvature variation—is applied to estimate an appropriate ε value. Subsequently, MinPts is calculated as the average number of neighbors found within the estimated ε radius. This ensures that parameter values are tailored to local data structures, enhancing the model's adaptability to diverse cluster shapes and densities.

B. Curvature-Based Core Point Enhancement

For high-dimensional and 3D data, identifying accurate core points is critical. This method integrates curvature analysis using Principal Component Analysis (PCA) to compute surface normals and curvature values in the local neighborhood of each data point. Only those points that demonstrate both adequate local density and stable curvature patterns are selected as core points. This step enhances cluster boundary detection, especially in scenarios involving overlapping or densely packed objects. Two mechanisms are explored: averaging curvature across neighbors and threshold-based segmentation for core point selection.

C. Semi-Supervised Core Selection

To improve clustering alignment with domain-specific requirements, a semi-supervised mechanism is incorporated. A binary attribute, referred to as *Is_important*, allows limited human intervention by indicating key points based on labeled data or predefined rules. This ensures more meaningful core point designation and prevents misclassification that may arise in unsupervised settings, especially in data with ambiguous boundaries or mixed densities.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, June 2025
Impact Factor: 7.67

D. Cluster Construction

After core point determination, clusters are formed by connecting neighboring core points whose ε neighborhoods overlap. Non-core points that are density-reachable from any core point are assigned to the respective clusters, while noise points remain unclassified. To optimize results in various dimensions, the algorithm evaluates both Euclidean and Manhattan distance metrics, selecting the one that offers the best trade-off between computational cost and clustering performance.

E. Experimental Setup and Evaluation

The model is tested across various datasets, including synthetic 2D data, UCI high-dimensional benchmarks, and 3D point clouds obtained from Time-of-Flight sensors. Performance is assessed using standard clustering evaluation metrics such as Adjusted Rand Index (ARI), V-measure, silhouette coefficient, noise ratio, and runtime efficiency.

Here is a **comparison table** that provides context for each study, including the model/technique used, the nature and size of the dataset, performance metrics, and key observations:

Author(s)	Model / Technique	Dataset Details	Performance Metrics	Key Observations
Ayesha Agrawal et al. [4]	Fast DBSCAN ,Spatial indexing (kd-tree/grid)	2D synthetic datasets, UCI repository	Execution Time, Accuracy	Improved runtime over DBSCAN with comparable accuracy; reduced noise sensitivity
Bing Maa et al. [5]	K-DBSCAN with Self-Adaptive Parameterss	Real/synthetic varied-density data	Runtime, Clustering Accuracy	Dynamically selects ε and MinPts using k-distance; balances speed and accuracy.
Momtaz Begum et al. [6]	M-DBSCAN (Modified DBSCAN for Outlier Control)	Benchmark datasets with outliersIris, Seeds, Glass, etc	Noise Rate, ARI, Accuracy	Better outlier detection; retains clustering quality under noisy conditionswith YOLOv5.
Chenbet al. [7]	Self-Adaptive DBSCAN	Quantitative Management benchmark data	Accuracy, Execution Time	Focuses on automatic parameter tuning to reduce manual input.
Pramadihanto et al. [8]	CVR-DBSCAN, CVR-DBSCAN-Avg, Auto-DBSCAN	TOF 3D pipe clouds	Clustering Accuracy, Runtime	Uses curvature via PCA for accurate clustering of 3D data; handles irregular shapes well
AL-SHAIKHLI et al. [9]	SS-DBSCAN (Semi-Supervised DBSCAN)	Letter, wireless, Arabic datasets	Accuracy, V-measure, Noise Ratio	Lightweight; lacks deep learning comparison; effective for field-level area mapping.

IV. PROBLEM STATEMENT

Density-based clustering algorithms, such as DBSCAN, have proven effective in identifying clusters with arbitrary shapes and handling noise in spatial datasets. However, their performance significantly deteriorates when applied to complex data scenarios involving varying densities, high-dimensional spaces, and large-scale datasets. A key limitation lies in their reliance on fixed density thresholds and traditional distance metrics, which are not well-suited for datasets with non-uniform distributions or intricate structures. These challenges lead to inaccurate cluster formation, misclassification of noise, and decreased overall effectiveness. The problem becomes more pronounced in high-dimensional environments, where the "curse of dimensionality" hampers distance-based calculations. Consequently,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, June 2025
Impact Factor: 7.67

there is a growing need for improved density-based algorithms that can adapt dynamically to diverse data characteristics, maintain robustness against noise, and scale effectively with increasing dimensionality and data volume. Addressing these challenges is critical for advancing clustering techniques in real-world applications involving complex, high-dimensional data.

V. PROPOSED WORK

The proposed approach seeks to enhance the precision and efficiency of density-based clustering by overcoming challenges related to varying data densities, high-dimensional spaces, and noise. This is achieved by refining the foundational algorithms to deliver more robust performance in complex and heterogeneous data scenarios.

VI. CONCLUSION AND FUTURE WORK

This study introduces an improved variant of the DBSCAN algorithm aimed at overcoming key challenges faced by traditional density-based clustering approaches in complex datasets. By employing an adaptive density threshold and a refined distance metric, the proposed method is capable of handling clusters with diverse densities and enhancing performance in high-dimensional settings. These advancements contribute to more accurate detection of core and border points, reduced misclassification of noise, and greater overall clustering effectiveness. Experimental results suggest that the modified algorithm is better suited for handling real-world datasets characterized by heterogeneity and scale.

Future research will focus on assessing the algorithm's performance on large-scale data using parallel and distributed processing techniques. Additionally, we intend to explore advanced parameter optimization strategies, such as evolutionary computation and reinforcement learning, to further enhance adaptability. The integration of dimensionality reduction methods will also be considered to improve clustering results and interpretability in complex domains, including bioinformatics, image processing, and network security.

REFERENCES

[1]. Xiaogang Huang, Tiefeng Ma, Conan Liu, and Shuangzhe Liu"GriT-DBSCAN: A Spatial Clustering Algorithm for Very Large Databases‖, JOURNAL OFLATEX CLASS FILES,6 Nov 2022.

[2]. Pritika Talwar, 2Shubham, 3Komalpreet Kaur, —EXPLORING CLUSTERING TECHNIQUES IN MACHINE LEARNING‖,ijcrt.org, 3 March 2024 .

[3]. Md. Zakir Hossain, Md. Jakirul Islam, Md. Waliur Rahman Miah, Jahid Hasan Rony, Momotaz Begum—Develop a dynamic DBSCAN algorithm for solving initial parameter selection problem of the DBSCAN algorithm,‖Indonesian Journal of Electrical Engineering and Computer Science, 3 September 2021.

[4]. Ayesha Agrawal, Vinod Maan— A Novel Approach for Clustering Algorithm using Fast DBSCAN ‖ Jetir,September 2024.

[5]. Bing Maa, Can Yanga, Aihua Lia,*, Yuxue Chia, Lihua Chenb — A Faster DBSCAN Algorithm Based on Self-Adaptive Determination of Parameters||10th International Conference on Information Technology and Quantitative Managemen,.

[6]. Momotaz Begum, Mehadi Hussan Shuvo ,M-DBSCAN: Modified DBSCAN Clustering Algorithm for Detecting and Controlling Outliers, —Research gate, —April 2024.

[7]. Chenb et al "A Faster DBSCAN Algorithm Based on Self-Adaptive Determination ofParameters," Elsevier B.V, 10th International Conference on Information Technology andQuantitative Management (2023)

[8]. Pramadihanto et al, Improvement of DBSCAN Algorithm Involving Automatic Parameters Estimation and Curvature Analysis in 3D Point Cloud of Piled Pipe, Journal of Image and Graphics, 2024.

[9]. AL-SHAIKHLI et al , SS-DBSCAN: Semi-Supervised Density-Based Spatial Clustering ofApplications With Noisefor Meaningful Clustering in Diverse Density IEEE Acess, 2024

[10]. Glory H.Shah,et al An Improved DBSCAN, A Density Based Clustering Algorithm with ParameterSelection for High Dimensional Data Sets, NIRMA UNIVERSITY INTERNATIONALCONFERENCE ON ENGINEERING, NUiCONE-2012

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, June 2025
Impact Factor: 7.67

[11]. Momotaz Begumet al,M-DBSCAN: Modified DBSCAN ClusteringAlgorithm for Detecting and Controlling Outliers, Conference: 39th ACM/SIGAPP Symposium2024.

[12]. Issa et al , Improving Density-based Clustering using Metric Optimization , InternationalJournal of Computer Applications 2018.

[13]. Pooja Batra Nagpal, et al, Comparative Study of Density based ClusteringAlgorithms, International Journal of Computer Applications 2011.

[14]. Shyam Lal ,et al, Techniques to Enhance the Performance of DBSCANClustering Algorithm in Data Mining, International Journal for Research in Applied Science &Engineering Technology (IJRASET) 2022

[15]. Cooper, D. et al, "A comparative survey of VANET clustering techniques," IEEE Communications Surveys & Tutorials, vol. 19, no. 1, pp. 657–681, 2016.