International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

Comparative Analysis of Outlier Detection Methods with Emphasis on DBSCAN

Aditi Badhan[1], Anita Ganapati[2]
[1]Research Scholar, Department of Computer Science, Himachal Pradesh University, Shimla
[2]Professor, Department of Computer Science, Himachal Pradesh University, Shimla
Emails: *aditibadhan@gmail.com[1], anitaganpati@gmail.com[2]*

Abstract
Outlier detection, also called anomaly detection or novelty detection, is a cornerstone in data analysis and machine learning, aiming to identify data instances that do not conform to expected behavior. Numerous approaches have been widely used in numerical domains; their performance on categorical datasets remains limited due to the complexity of distance computation and density estimation. This paper presents a comparative review of outlier detection methods on multilevel data. First, the major families of outlier detection methods –statistical-based, clustering-based, density-based, and ensemble-based are compared based on their working principle, strength, and limitations. Second, Clustering-based methods of outlier detection are examined. Third, several clustering-based algorithms such as K-Means, K-Modes, ROCK, DBSCAN, LOF, Hierarchical, CLIQUE, STING, and GMM are compared. A detailed comparison of DBSCAN with the well-established categorical clustering approaches ROCK and K-Modes is also presented to determine their underlying clustering, applicability to categorical data and outlier detection mechanism. To contextualize the comparison, several benchmark datasets for detection of outliers in categorical datasets are presented along with their characteristics. The current study presents a comparative analysis of outlier detection methods at multiple levels, and among these, the DBSCAN seems particularly suitable for identifying outlier instances. Its density-based measure makes it better suited to handling noise and clusters of arbitrary shape, and is independent of predefined clusters. Although DBSCAN is primarily designed for numerical data, its density-based framework can be extended to categorical datasets by incorporating an appropriate categorical similarity or dissimilarity measure. Developing such an adaptation provides a potential direction for applying DBSCAN-based outlier detection to categorical data.

Keywords: *Outlier Detection, Categorical data, clustering algorithms, density-based methods, DBSCAN.*

1. Introduction
In Data mining, several analysis tasks are performed to extract valuable and hidden patterns from raw data, and Outlier detection is one of the most important tasks, which aims to identify anomalous observations, which is crucial in real-world applications. Identifying unusual instances in a dataset can lead to discovering patterns that do not match the most common patterns of the data; this has many critical applications and has been widely used in the literature. Identifying irregular patterns is a crucial task in domains such as healthcare, fraud detection, cybersecurity, and network analysis. With the advent of technology, as data is growing in complexity and dimensionality, identifying outliers or anomalous patterns across multiple structural and analytical levels has become an increasingly daunting and challenging task. [1] Proposed a definition of outlier: “An observation which deviates so much from other observations as to arouse suspicions that a different mechanism generated it”. (Knorr et al., n.d.) “An object ‘O’ in the dataset ‘T’ is a distance-based outlier if at least a fraction p of the objects in the T are farther than distance ‘D’ from ‘O’ [3] “An anomaly is a pattern in the data that does not conform to a well-defined notion of normal behavior”.[4] “Outlier detection is the identification of data points that do not fit the expected pattern or behavior of the dataset”. Mostly, the traditional methods are designed for numerical data, but real-world data also contain predominantly categorical features. Numerous methods have been proposed in the literature, categorized into Statistical-based [5],

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

clustering-based[6], density-based[7], distance-based[8], and ensemble-based[3], [9]. Recently, clustering has become a widely used unsupervised approach for outlier detection. The underlying assumption is that the normal data points form dense regions and meaningful clusters, whereas the outliers lie in the sparse area [6], [10]. The clustering algorithms differ in how they interpret similarity, density, proximity, and cluster boundaries. Therefore, their suitability for outlier detection also varies according to data type and structure. The ROCK and K modes are well-established and widely used approaches specifically developed for clustering of categorical data. The k modes are developed by extending the k-means paradigm by employing a categorical similarity measure, cluster mode and a frequency -based procedure for updating modes[11]. Robust Clustering Using Link (ROCK) uses the concept of links between neighborhood objects to determine the similarity and connectivity of categorical observations[12]. Therefore, these approaches provide important established approaches for analyzing inherent structure of categorical datasets. DBSCAN, a density-based clustering algorithm, provides a different mechanism which identifies clusters based on density and explicitly handles observations which do not belong to a sufficiently dense region as noise. This study compares the outlier detection methods and major clustering algorithms for outlier detection. The methods are compared on basic principles, strengths and limitations, and outlier detection strategy. A detailed comparison of DBSCAN with K-Modes and ROCK provides a broader understanding of partitioning-based, link-based and density-based approaches and their potential applicability to categorical clustering and outlier detection. Furthermore, the study incorporates the dataset used for outlier detection and a theoretical justification for using DBSCAN for the identification of outlier instances in categorical data, thereby laying a foundation for future research. This study presents a conceptual study which i) compares the major outlier detection methods, ii) compares the clustering-based outlier detection methods, iii) algorithm-level comparison of the clustering-based methods, iv) Comparative Analysis of DBSCAN, ROCK and K-Modes) and review of the benchmark dataset, and vi) theoretically justifies DBSCAN as the most suitable clustering-based method for outlier detection. As shown in Figure 1 Conceptual Illustration of Outliers in Clustering.

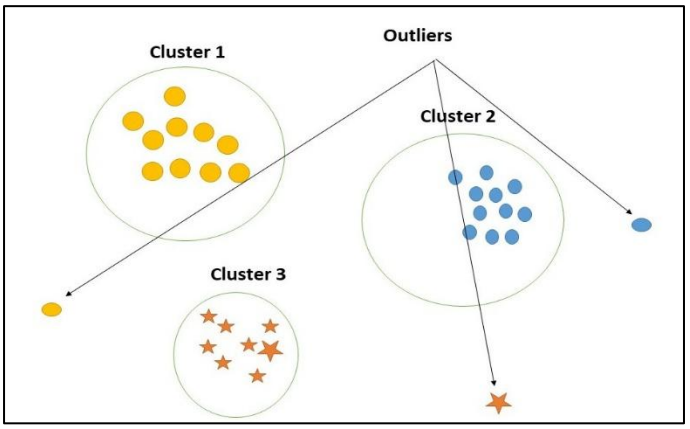


Figure 1 Conceptual Illustration of Outliers in Clustering

2. Literature Review

The significant impact of outliers across numerous data-intensive applications has driven extensive research and the development of outlier detection methods within the statistical, data mining, and machine learning communities. The outlier detection methods are divided into numerous categories, such as statistical-based[1], clustering-based[6], density-based[13], [14], distance-based[2], [15], and ensemble-based[16], [17], [18]. Generally, clustering-based methods are divided into five categories. Each class of clustering uses a different mechanism to form clusters and identify outliers.

Partition-Based Clustering: These clustering algorithms, including k-means, Medoids, and Modes, divide data into a fixed number of clusters by minimizing the intra-cluster similarity. Outlier detection typically involves identifying data points that are too far from cluster centroids, or that belong to small clusters [1-10].

Density-Based: Density-based methods define clusters as regions of high density separated by sparse regions. The most influential algorithm in this category is DBSCAN (Ester et al., 1996), which introduced the concepts of core, border, and noise points based on density reachability. Extensions such

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

as OPTICS (Sander et al., 1998) and LOF (Breunig et al., 2000) build upon density concepts by allowing variable-density detection and local density estimation.

Grid-Based: Grid-based methods, such as STING (Wang et al., 1997), CLIQUE (Agrawal et al., 1998), and WaveCluster, segment the feature space into grids and compute density at the grid-cell level. These methods scale exceptionally well for large datasets and high-dimensional spaces. Outliers are typically detected as points located in grid cells with low density [11 – 20].

Model-Based: Model-based algorithms, such as Gaussian Mixture Models (GMM) (McLachlan & Peel, 2000), Expectation–Maximization (EM) (Dempster et al., 1977), and One-Class SVM (Schölkopf et al., 1999), assume that data follows a probabilistic model or boundary. Outliers are points with low likelihood, or that fall outside the learned decision boundary [21-30].

Hierarchical-Based: Hierarchical clustering algorithms—such as AGNES and DIANA (Johnson, 1967)—create nested clusters through iterative merging or splitting. ROCK (Guha et al., 1999) is one of the few hierarchical algorithms explicitly designed for categorical data, using link similarity rather than distance.

3. Methodology

This study incorporates a conceptual research design to compare different outlier detection methods. In this section, a methodology for conducting a comparative analysis of outlier detection methods, explicitly concentrated on the clustering-based method for a categorical dataset, and a theoretical justification for using DBSCAN to identify outlier instances for further research. As shown in Figure 2 Research Procedure of the Study. Outlier detection is extensively studied across statistics, data mining, and machine learning. Outlier Detection methods are divided into different categories. Each category is based on a different principle of identifying outlier instances. This section compares numerous outlier detection methods across different levels. As shown in Table 1 Comparison Based on Categories.

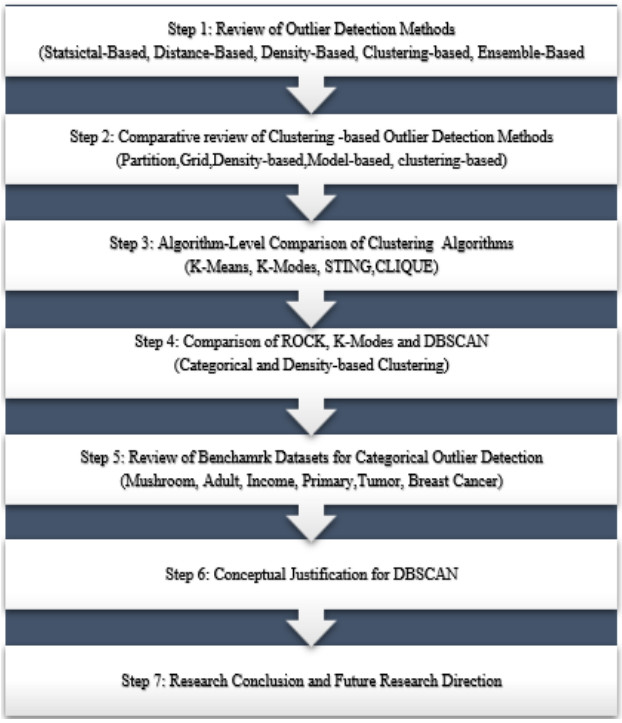


Figure 2 Research Procedure of the Study

Table 1 Comparison Based on Categories

Method Type	Core principle	Strengths	Limitations
Statistical-Based[19], [20], [21]	Assume data follows a known statistical distribution; points with low probability are considered outliers.	Simple and interpretable Based on a probability distribution Works well for low-dimensional numeric data	Assumes underlying data distribution and poorly performs on complex or non-Gaussian data, even sensitive to skewed and imbalanced data.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

Distance-based[2], [22]	Objects far from their nearest neighbors are considered outliers (based on a distance threshold)	No distribution assumption	Numeric datasets, low-dimensional anomaly detection
Density-Based[7]	Outliers lie in regions of significantly lower density than their neighbors.	Detects an arbitrary–shaped cluster Identifies both local and global outliers No need to specify the number of clusters Robust to noise	Sensitive to epsilon and Minpts and struggles with data of varying densities. Requires meaningful distance metrics.
Clustering-Based (Jainet al., 2000; Jiang & An, 2008; Pu et al., n.d.)	Normal data form clusters; objects that do not belong to any cluster or are in very small clusters are outliers.	Data points appear as small clusters or far from cluster centers. Works well for mixed data types with appropriate similarity measures	Requires the cluster assumption and is sensitive to initial parameters like k and the similarity threshold. Not ideal for datasets with overlapping clusters.
Ensemble-based [16], [17], [18]	Combine multiple outlier-detection models; anomalies are points consistently flagged by multiple detectors.	Combine multiple models for robustness, and works well with heterogeneous data types.	Computationally expensive for large datasets and difficult to interpret due to the combination of multiple models and complex parameter tuning.

4. Clustering-based Outlier Detection

Clustering is an effective method of identifying anomalous observations because outliers are the observation which do not confirm to the dominant structure of the dataset, whereas clustering identifies group of observations which have similar characteristics. [3] Clustering groups similar data instances and relies on the assumption that normal observations belong to dense clusters, whereas outliers may not belong to any cluster or may vary from small clusters. [25] Authors demonstrated the effectiveness of clustering empirically and identified outlier instances in categorical data. [26] Also state the usefulness of clustering for identifying outlier instances because it can identify inherent groups in data, thereby helping to identify observations which deviate from the normal data instances. The K modes and ROCK are well-established clustering techniques designed for categorical data [11], [12], [27], [28], [29], [30], [31]. As shown in Table 2 Comparison Based on Clustering. Table 3 Comparison of Clustering-Based Algorithms, Table 4 Comparative Analysis of DBSCAN, ROCK and K-Modes. Table 4 Comparative Analysis of DBSCAN, ROCK and K-Mode. Table 5 Mostly Used Datasets for Outlier Detection [31 – 40].

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

Table 2 Comparison Based on Clustering

Method Type	Core idea	Data Suitability	Cluster Shape	Strength	Limitations	Representative algorithm
Partitioning-based clustering[24], [32]	It divides the data into a fixed number of clusters. Points far from the cluster centers or in small clusters are marked as outliers.	Numeric and categorical (with modification)	Spherical	Fast, scalable, and easy to interpret, they work well for both numeric and categorical datasets.	It requires predefining k; it is sensitive to initialization; it is ineffective for non-convex clusters; and it is not robust to noise.	K-Means, K-Medoids, K-Modes
Density-Based clustering[13]	Data instances lie in the dense region of the cluster; points in the sparse region are treated as noise points.	Numeric and Categorical (with similarity)	Arbitrary	Detects a cluster of arbitrary shape and naturally identifies noise points. It has the ability to identify both local and global outliers, even without requiring k.	Performance is highly influenced by epsilon and Minpts, and issues in high-dimensional datasets; DBSCAN also struggles with varying densities.	DBSCAN, OPTICS, LOF
Hierarchical - based clustering [33], [34], [35]	It builds a tree structure. Outliers appear as singleton nodes, late merging points, or a cluster with a long linkage distance.	Numeric and Categorical	Variable	It reveals a multilevel structure and no need to specify k, works with multiple distance metrics, and provides an interpretable dendrogram.	High computational cost and sensitivity to linkage choices also pose difficulties while selecting the dendrogram cut-level,	Agglomerative, Divisive

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

					and are not suitable for massive datasets.	
Grid-based Clustering[24], [36], [37], [38]	The space is divided into finite grid cells; the density of each cell is evaluated to identify a sparse region, which indicates outliers.	Numerical and categorical (after discretizatio n)	Grid-Shaped	Efficient for large datasets and supports multi-resolution analysis.	It is susceptible to grid resolution and boundary distortion effects, and also struggles with arbitrary-shaped clusters.	STING, CLIQUE
Model-based clustering[24], [37], [38]	It assumes that data follows a statistical or mathematical model; the data instances having low probability or lying outside the learned decision boundary are treated as outlier instances.	(Numeric and Mixed)	Elliptical or probabili stic	It provides probabilistic scoring and is effective when normal data is available, and it also handles continuous distributions .	It is sensitive to model assumptions and initialization . It may not be suitable for large categorical datasets.	GMM, EM, One-class SVM

Table 3 Comparison of Clustering-Based Algorithms

Method Type	Working Principle	Category	Outlier Detection procedure
K-Means[39]	Minimizes within-cluster variance by updating centroids iteratively.	Centroid-based	Points far from the nearest centroid or belonging to tiny clusters are treated as anomalies.
K-Medoids[9]	Uses Medoids instead of centroids to reduce the effect of outliers.	Medoids - based	Points with high distance to the medoid or forming small clusters.
K-Modes[9]	Uses mode instead of mean; frequency-based distance.	Mode-based	Objects poorly matching the mode or in tiny clusters are flagged.
CLARANS	Randomized search-based clustering for large datasets using	Randomized K-Medoids	Same as k-Medoids—high distance from Medoids.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

	Medoids.		
STING	Divides space into hierarchical grid cells; computes cell statistics.	Grid-based	Low-density cells are treated as anomalous regions.
CLIQUE	Discovers dense subspace grids and clusters.	Grid-Based	Points not belonging to any dense grid are considered outliers.
DBSCAN[40]	Discovers dense regions using parameters (*eps*, *Minpts*).	Density-based	Noise points not belonging to any cluster are outliers.
OPTICS	Extended DBSCAN produces an ordering of points based on reachability.	Density-Based	High reachability-distance points become outliers.
LOF[9], [41]	Computes the relative local density of a point vs neighbor.	Density-based	A high LOF score indicates a strong local anomaly.
GMM	Data modeled as a mixture of Gaussians using the EM algorithm.	Model-Based	Points with low mixture likelihood outliers.
EM[9]	Iteratively estimates the parameters of probabilistic models.	Model-Based	Low probability under the learned model indicates anomalies.
One-Class SVM[42]	Learns the decision boundary around normal data using kernel functions.	Model-Based	Points outside the learned boundary are outliers.
Isolation Forest[43]	Random splits isolate points; fewer splits imply an anomaly.	Model-based	Points isolated quickly anomalies.
Agglomerative [33], [44]	Bottom-up merging using linkage (single/complete/average/Ward).	Hierarchical	Points merging late or forming singleton branches are anomalies.
Divisive[33], [44]	Top-down splitting of clusters.	Hierarchical	Points separated early in the divisive process are considered anomalies.
HDBSCAN[45]	Hybrid of density and hierarchy; minimum cluster size controls splitting.	Hierarchical and density-based	Noise points and low-density areas automatically become outliers.

Table 4 Comparative Analysis of DBSCAN, ROCK and K-Mode

Criterion	K-Modes	ROCK	DBSCAN
Primary purpose	Categorical data clustering [11]	Categorical data clustering [12]	Density-based clustering and noise identification [14]
Method type	Partitioning-based	Hierarchical/link-based	Density-based
Data Type	Specifically developed for categorical data [1]	Specifically developed for categorical and Boolean attributes [2]	Originally developed for spatial data [3]
Method Type	Partitioning-based clustering	Hierarchical/link-based clustering	Density-based clustering

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

Basic principle	Uses categorical modes to represent clusters[29], [31]	Uses links between neighboring objects to determine similarity [12], [46]	Uses density connectivity to form clusters [47], [48], [49]
Cluster representation	Categorical mode	Neighborhood links/connectivity	Dense regions
Requires number of clusters (K)	Yes	Number of clusters is used in the clustering procedure	No predefined number of clusters
Main parameters	K	θ / clustering parameters	ε and MinPts
Similarity mechanism	Matching dissimilarity	Link-based on share neighbors	ε (neighborhood radius) and MinPts or Distance/neighborhood density
Density information	No	No	Yes
Explicit noise identification	No explicit noise category in the basic algorithm	No explicit DBSCAN-style noise category	Yes; observations not belonging to sufficiently dense regions can be classified as noise [14], [47]
Outlier detection perspective	Can provide information through dissimilarity to cluster modes, but primarily a clustering method	Can provide information through weak connectivity, but primarily a clustering method	Naturally supports noise-based outlier identification
Neighborhood information	Based primarily on object-to-mode dissimilarity	Central to the link-generation process	Central to density estimation
Arbitrary-shaped clusters	Limited; clusters are formed around categorical modes	Hierarchical/link-based structure	Yes; can identify clusters of arbitrary shape [47], [49]
Sensitivity	K and initialization	Similarity threshold and clustering parameters	ε and MinPts
Categorical data applicability	Directly applicable [11], [29]	Directly applicable[12]	Requires an appropriate distance/similarity measure when adapted to categorical data [14], [50]
Computational consideration	Generally efficient	Link computation can be costly	Depends on neighborhood/distance computation

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

Purpose in this study	Established categorical clustering approach	Established categorical clustering approach	Density-based comparison approach for outlier detection

Table 5 Mostly Used Datasets for Outlier Detection

Dataset	Type	Outlier Type
Mushroom[51]	Categorical	Edible vs Poisonous
Adult Census[52]	Categorical	Income Anomalies
Breast Cancer[53]	Categorical	Malignant and Benign
Primary Tumour[54]	Categorical	Rare Tumour classes
Car Evaluation[55]	Categorical	Rare preference patterns

5. DBSCAN and its role in Outlier Detection

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular unsupervised machine learning algorithm that belongs to the clustering class of techniques. Clustering aims to separate data points into several groups, in such a manner that the data points in the same group have similar properties.[56], [57]. DBSCAN has the ability to detect clusters of arbitrary shape while simultaneously identifying outliers as data points located in a sparse region [47], [48]. Although DBSCAN was generally developed for numerical data, its underlying density-based framework can also be applied to categorical data by employing suitable categorical similarity or dissimilarity measures. Previous studies have demonstrated adaptations of DBSCAN for categorical datasets using measures appropriate to categorical attributes[50], [58], [59]. Our future work is focused on using DBSCAN, as it has emerged as one of the most powerful and versatile methods for identifying outlier instances in categorical data by adapting similarity and dissimilarity measures.

DBSCAN works on two parameters:

1. Epsilon (ε): it is the radius that defines a point's neighborhood.
2. MinPts: the minimum number of points required to form a dense region.

As shown in Figure 3 DBSCAN, Table 6 DBSCAN Parameter Impact Table.

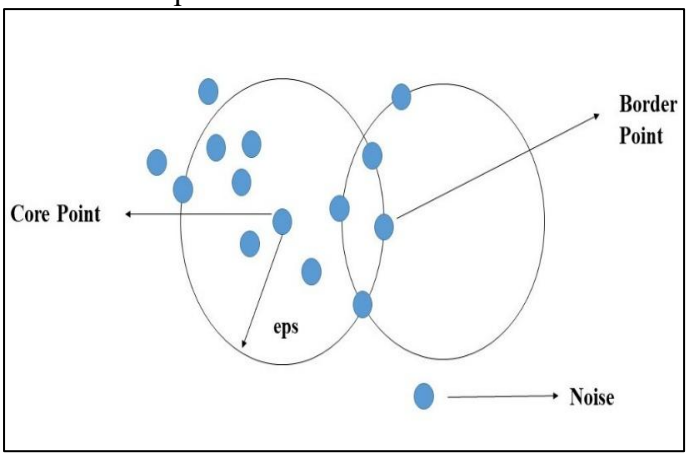


Figure 3 DBSCAN

Table 6 DBSCAN Parameter Impact Table

Parameter	Meaning	Impact	Use Case
ε (Epsilon and neighborhood Radius)	Radius of a neighborhood	Controls Density threshold	Intrusion detection and fraud detection. Outlier detection Noise filtering before classification.
Minpts (Minimum Points)	Min Neighbors	Influences core and border points	
Similarity Metric	Distance metric	Defines cluster boundaries	

DBSCAN Algorithm

1. Start
2. Select an arbitrary point p
3. Retrieve all points that are density reachable from p w.r.t Eps and Minpts
4. If p is a core point, form a cluster
5. If p is a border point, no points are density reachable from p, and DBSCAN visits the next point

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

6. Continues the process until each point of the dataset has been processed
7. End

Density-based clustering, DBSCAN (Density-based clustering Application with Noise), is considered a highly effective method for identifying outliers, due to its core mechanism aligning with the fundamental definition of anomalies as data instances occurring in sparse or low-density regions. [3], [14]. Unlike centroid-based clustering that forces every point to be in a cluster, DBSCAN explicitly allows the data instances to remain unassigned, labeling them as noise when they do not satisfy a density requirement. DBSCAN naturally identifies noise points. In DBSCAN, noise detection is not a by-product- it is a core component. If a point does not fall in a denser region, it is marked as an outlier without requiring additional thresholds or statistical assumptions. These characteristics make it conceptually robust, flexible, and naturally aligned with the objective of detecting categorical outliers.

Conclusion

Outlier detection is a crucial task in data mining and machine learning due to its capability of identifying the rare and unusual patterns in the data, which may carry significant practical importance. Numerous methods have been devised by researchers for detecting outliers in numerical data; real-world applications constitute categorical data also. This paper presents a conceptual comparative analysis of outlier detection methods at the multilevel, and a primary clustering-based outlier detection method is conducted. The benchmarked datasets were also reviewed. The comparative study demonstrates that clustering-based methods offer strong performance for outlier detection. Their unsupervised nature, adaptability to noise and high-dimensional data, and the capacity to capture both local and global outliers make them a recommended choice for a wide range of real-world applications. Furthermore, a comparative analysis of K-Modes, ROCK and DBSCAN highlights their distinct characteristics, clustering mechanism and applicability to categorical data. While ROCK and K-Modes are well established for categorical data clustering, DBSCAN offers an explicit noise identification mechanism based on density. This comparison also provides a base for adaption of DBSCAN for to handle categorical data. Furthermore, the study presented that DBSCAN establishes the strongest theoretical foundation, attributable to its density-based structure, its ability to find clusters of arbitrary shape, its natural identification of noise, and its independence from predefined clusters. Future research focuses on using DBSCAN, a density-based clustering method, to identify outliers in categorical datasets.

References

[1]. D. M. Hawkins, Identification of Outliers. Springer Netherlands, 1980. doi: 10.1007/978-94-015-3994-4.
[2]. E. M. Knorr, R. T. Ng, and V. Tucakov, “Distance-based outliers: algorithms and applications.”
[3]. V. Chandola and V. Kumar, “Outlier Detection : A Survey,” 2007. [Online]. Available: https://www.researchgate.net/publication/242403027
[4]. V. J. Hodge and J. Austin, “A survey of outlier detection methodologies,” Oct. 2004. doi: 10.1023/B:AIRE.0000045502.10941.a9.
[5]. P. J. Rousseeuw and M. Hubert, “Robust statistics for outlier detection,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 1, no. 1, pp. 73–79, Jan. 2011, doi: 10.1002/widm.2.
[6]. S. Y. Jiang and Q. B. An, “Clustering-based outlier detection method,” in Proceedings - 5th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008, 2008, pp. 429–433. doi: 10.1109/FSKD.2008.244.
[7]. L. Duan, “5 Density-Based Clustering and Anomaly Detection.” [Online]. Available: www.intechopen.com
[8]. E. M. Knorr, R. T. Ng, and V. Tucakov, “Distance-based outliers : algorithms and applications,” no. i, pp. 1–17, 1847.
[9]. C. C. Aggarwal, “An introduction to outlier analysis. In Outlier Analysis,” 2016. doi: 10.1007/978-3-319-47578-3.
[10]. B. V. Sánchez Vinces, E. Schubert, A. Zimek, and R. L. F. Cordeiro, “A comparative evaluation of clustering-based outlier

detection," Data Min. Knowl. Discov., vol. 39, no. 2, Mar. 2025, doi: 10.1007/s10618-024-01086-z.

[11]. Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," 1998.

[12]. S. : SUDIPTO Guha, R. Rastogi', K. Shims ', and S. Cjniversit, "ROCK: A R.OBUST CLUSTERING ALGORITHM FOR, CATEGORICAL ATTRIBUTES+," 2000.

[13]. H. P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 1, no. 3, pp. 231–240, May 2011, doi: 10.1002/widm.30.

[14]. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 1996. [Online]. Available: www.aaai.org

[15]. L. Tran, C. Shahabi, and L. Fan, "Distance-based Outlier Detection in Data Streams," 2016. [Online]. Available: https://www.researchgate.net/publication/307513287

[16]. X. Gao et al., "An ensemble-based outlier detection method for clustered and local outliers with differential potential spread loss," Knowl. Based. Syst., vol. 258, p. 110003, 2022, doi: 10.1016/j.knosys.2022.110003.

[17]. C. C. Aggarwal and S. Sathe, Outlier ensembles: An introduction. 2017. doi: 10.1007/978-3-319-54765-7.

[18]. L. Vanneschi and S. Silva, "Ensemble Methods," Natural Computing Series, pp. 283–288, 2023, doi: 10.1007/978-3-031-17922-8_11.

[19]. A. Zimek and P. Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms," Nov. 01, 2018, Wiley-Blackwell. doi: 10.1002/widm.1280.

[20]. H. Yan, L. Wang, and Y. Lu, "Identifying cluster centroids from decision graph automatically using a statistical outlier detection method," Neurocomputing, vol. 329, no. xxxx, pp. 348–358, 2019, doi: 10.1016/j.neucom.2018.10.067.

[21]. A. Zimek, E. Schubert, and H. P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," 2012, John Wiley and Sons Inc. doi: 10.1002/sam.11161.

[22]. E. M. Knox and R. T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Datasets."

[23]. G. Pu, L. Wang, J. Shen, and F. Dong, "A Hybrid Unsupervised Clustering-Based Anomaly Detection Method." [Online]. Available: http://creativecommons.org/licenses/by/4.0/

[24]. A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," 2000.

[25]. R. Pamula, J. K. Deka, and S. Nandi, "An outlier detection method based on clustering," in Proceedings - 2nd International Conference on Emerging Applications of Information Technology, EAIT 2011, 2011, pp. 253–256. doi: 10.1109/EAIT.2011.25.

[26]. N. N. R. Ranga Suri, N. Murty M., and G. Athithan, "Outlier detection in categorical data," in Intelligent Systems Reference Library, vol. 155, Springer Science and Business Media Deutschland GmbH, 2019, pp. 69–93. doi: 10.1007/978-3-030-05127-3_5.

[27]. M. Cendana and R.-J. Kuo, "Categorical data clustering: A bibliometric analysis and taxonomy," Mach. Learn. Knowl. Extr., vol. 6, no. 2, pp. 1009–1054, 2024.

[28]. F. Cao, J. Liang, D. Li, L. Bai, and C. Dang, "A dissimilarity measure for the k-Modes clustering algorithm," Knowl. Based. Syst., vol. 26, pp. 120–127, Feb. 2012, doi: 10.1016/j.knosys.2011.07.011.

[29]. J. Z. Huang, "Section: Clustering Clustering Categorical Data with k-Modes," 2009. [Online]. Available: http://www.

[30]. K. S. Dorman and R. Maitra, "An efficient k-modes algorithm for clustering categorical datasets," Stat. Anal. Data Min., vol. 15, no. 1, pp. 83–97, Feb. 2022, doi: 10.1002/sam.11546.

[31]. S. T. Gavva, K. C. S., and S. Punna, "Clustering Categorical Data: Soft Rounding k-modes," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2210.09640

[32]. M. C. W. J. Gan Guojun, "Data Clustering: theory, algorithms, and applications," 2020.

[33]. F. Nielsen, "Hierarchical Clustering," 2016, pp. 195–211. doi: 10.1007/978-3-319-21903-5_8.

[34]. N. Ishaq, T. J. Howard, and N. M. Daniels, "Clustered Hierarchical Anomaly and Outlier Detection Algorithms," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2103.11774

[35]. F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," Computer Journal, vol. 26, no. 4, pp. 354–359, 1983, doi: 10.1093/comjnl/26.4.354.

[36]. D. Gibson, J. Kleinberg, and P. Raghavan, "Clustering Categorical Data: An Approach Based on Dynamical Systems."

[37]. T. Dinh et al., "Data clustering: an essential technique in data science," Dec. 2024, [Online]. Available: http://arxiv.org/abs/2412.18760

[38]. W. Andreopoulos, "Clustering Algorithms for Categorical Data," 2006.

[39]. A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[40]. "DBSCAN — scikit-learn 1.7.0 documentation." Accessed: Jun. 17, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

[41]. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," 2000.

[42]. S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," Pattern Recognit., vol. 58, pp. 121–134, Oct. 2016, doi: 10.1016/j.patcog.2016.03.028.

[43]. F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation Forest," Proceedings - IEEE International Conference on Data Mining, ICDM, pp. 413–422, 2008, doi: 10.1109/ICDM.2008.17.

[44]. F. Murtagh and P. Contreras, "Methods of Hierarchical Clustering," Apr. 2011, [Online]. Available: http://arxiv.org/abs/1105.0121

[45]. R. Setiawan, B. Tjahjono, G. Firmansyah, and H. Akbar, "Fraud Detection in Credit Card Transactions Using HDBSCAN, UMAP and SMOTE Methods," International Journal of Science, Technology & Management, vol. 4, no. 5, pp. 1333–1339, 2023, doi: 10.46729/ijstm.v4i5.929.

[46]. B.-H. Kim and K. Kim, "Improvements of K-modes Algorithm and ROCK Algorithm," KOREAN JOURNAL OF APPLIED STATISTICS, vol. 15, no. 2, pp. 381–394, 2002.

[47]. E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," ACM Transactions on Database Systems, vol. 42, no. 3, Jul. 2017, doi: 10.1145/3068335.

[48]. D. Deng, "DBSCAN Clustering Algorithm Based on Density," in Proceedings - 2020 7th International Forum on Electrical Engineering and Automation, IFEEA 2020, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.

[49]. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 1996. [Online]. Available: www.aaai.org

[50]. S. Amiri, B. S. Clarke, and J. L. Clarke, "Clustering Categorical Data via Ensembling Dissimilarity Matrices," Journal of Computational and Graphical Statistics, vol. 27, no. 1, pp. 195–208, Jan. 2018, doi: 10.1080/10618600.2017.1305278.

[51]. "Mushroom - UCI Machine Learning Repository." Accessed: Jun. 20, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/73/mushroom

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5678-5690
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0739

[52]. "Adult - UCI Machine Learning Repository." Accessed: Feb. 03, 2026. [Online]. Available: https://archive.ics.uci.edu/dataset/2/adult

[53]. "Breast Cancer - UCI Machine Learning Repository." Accessed: Jun. 20, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/14/breast+cancer

[54]. "Primary Tumor - UCI Machine Learning Repository." Accessed: Feb. 04, 2026. [Online]. Available: https://archive.ics.uci.edu/dataset/83/primary+tumor

[55]. "Car Evaluation - UCI Machine Learning Repository." Accessed: Feb. 04, 2026. [Online]. Available: https://archive.ics.uci.edu/dataset/19/car+evaluation

[56]. Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," Pattern Recognit. Lett., vol. 24, no. 9–10, pp. 1641–1650, 2003, doi: 10.1016/S0167-8655(03)00003-5.

[57]. N. N. R. R. Suri, M. N. Murty, and G. Athithan, "Detecting outliers in categorical data through rough clustering," Nat. Comput., vol. 15, no. 3, pp. 385–394, Sep. 2016, doi: 10.1007/s11047-015-9489-2.

[58]. Z. G. Chen, H. S. Kang, S. N. Yin, and S. R. Kim, "An efficient privacy protection in mobility social network services with novel clustering-based anonymization," EURASIP J. Wirel. Commun. Netw., vol. 2016, no. 1, Dec. 2016, doi: 10.1186/s13638-016-0767-1.

[59]. B. Andreopoulos, A. An, X. Wang, and D. Labudde, "Efficient layered density-based clustering of categorical data," J. Biomed. Inform., vol. 42, no. 2, pp. 365–376, Apr. 2009, doi: 10.1016/j.jbi.2008.11.004.