International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0027
e ISSN: 2584-2854
Volume: 04 Issue: 02
February 2026
Page No: 192 - 197

CatDBSCAN for Outlier Detection in Categorical Datasets

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

Abstract
Outlier detection is a critical task and presents unique challenges due to the lack of natural ordering or distance measure among categorical attributes. Numerous methods have been devised for outlier detection. DBSCAN has proven to be useful in numerical domains, identifying the noise points as outliers. In this paper, a modified approach for identifying outlier instances within categorical data sets has been proposed, named CatDBSCAN. The CatDBSACN is adapted for categorical data by incorporating a distance measure such as Hamming Distance. The CatDBSCAN also detects outliers in the small cluster, as data instances lying in low-density regions are prone to outliers. It employs a static parameter while recognizing noise and minor clusters as outlier points. Additionally, an outlier scoring mechanism is used to label noise points and cluster-based outliers. Experiments are conducted on the purely categorical mushroom dataset and the breast cancer dataset of the UCI ML repository.

Keywords: *Outlier**;**DBSCAN**; CatDBSCAN**; Categorical Outlier**; outlier Detection**; mushroom* ***and*** *breast**;***

1. Introduction
Outlier detection is the identification of anomalous patterns within a dataset. Outliers are the data instances that deviate so much from the general data distribution [1], [2], [3]. The outliers in the dataset indicate critical incidents such as fraudulent activities, system faults, serious medical conditions, etc. The presence of Outliers can distort the data analysis, leading to biased results. Outliers are caused by measurement error, data entry error, inherent data variability, and human mistakes. Outlier detection is used to identify the outlying or unusual observations. Several methods have been developed for outlier detection, including statistical-based, distance-based, density-based, clustering-based, and ensemble-based approaches.[1],[4].Statistical-based methods define outliers as data instances that do not conform to a general data distribution. Distance-based methods focus on the distance computation between data instances.[5]. A point is said to be an outlier if it is too far from its neighboring points. The core concept of density-based methods is that outliers reside in low-density regions, while inliers are found in dense neighborhoods.[6].Clustering-based methods group similar data instances and identify outliers that do not belong to any nearby or dense clusters. Ensemble-based methods combine multiple base detectors to improve detection results. Unlike numerical data, where outlier detection typically exploits geometric relationships, categorical data present unique challenges. Density-based spatial clustering application with noise (DBSCAN), is a popular density-based clustering algorithm that can identify cluster of arbitrary shape and does not require the number of clusters to be specified beforehand. As the performance of DBSCAN is affected by epsilon (ε) (the maximum distance between the data points to be considered as a part of neighborhood of other) and Minpts (it is the minimum number of points required to from a dense region). In this paper, a CatDBSCAN is proposed for identifying outlier instances within categorical datasets. CatDBSCAN expands the traditional definition of outliers. Rather than classifying the noise point as outliers, it also considers small clusters as potential outliers. This approach is particularly significant because small clusters often indicate the presence of anomalous patterns or behaviors within the data. In contrast, legitimate data points tend to form larger, denser clusters, which can be effectively detected using fixed parameters for epsilon (ε) and Minpts. The fixed parameters of epsilon (ε) and Minpts used to maintain the consistency across experiments and ensure fair comparison result. By recognizing small clusters alongside noise as outliers, the CatDBSCAN

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0027
e ISSN: 2584-2854
Volume: 04 Issue: 02
February 2026
Page No: 192 - 197

variant provides a more nuanced understanding of the data structure and enhances the capabilities for outlier detection. furthermore, The proposed method adapt a hamming distance with an outlier scoring mechanism addressing the limitation of existing DBSCAN in handling the categorical data. An experiment is conducted on purely categorical real-world datasets taken from the UCI Machine learning Repository. The Method is compared with DBSCAN; our method outperforms in terms of ROC-AUC and PR-AUC for both datasets. This paper is organized as follows: In Section 2, Approaches of outlier detection are discussed. In Section 3, the proposed approach, CatDBSCAN, is described, and in Section 4, Present and analyze the results on the mushroom and breast cancer datasets of the UCI ML repository. Section 5 concludes the work.

2. Related Work

Outlier detection is a widely studied in the area of data mining, machine learning, and data analysis due to its in identifying the usual instances. Numerous clustering-based method has been proposed to address the problem. Among them the density-based clustering methods have been attracted attention of researchers due to their ability of discovering cluster of arbitrary shape and naturally detect outliers in lower dense regions. The DBSCAN (Density-Based Spatial clustering if Application with Noise) is introduced by is one of the most influential and widely used density-based method. It forms cluster as area of high density separated by regions of low density and governed by two parameters: epsilon and Minpts. The DBSCAN performs well on numerical data, it depends on the distance metrics such as Euclidian distance limits applicability to purely categorical data. Author proposes LOF (Local Outlier Factor), it assess the local density of a data points with its neighbors. The method is effective in identifying outliers in numerical data but not perform well with categorical data.

3. Methods of Outlier Detection

3.1 Statistical-Based Techniques

Identifying the outlier instances using statistical methods can be applied in supervised, unsupervised, and semi-supervised ways [1]. These approaches depend on the underlying data distribution characteristics of the data to detect the anomalous patterns that deviate from the normal data distribution. Statistical methods are categorized into two main categories: parametric and non-parametric methods. The key difference among these methods is that the parametric method is based on the assumption about the underlying data distribution, and the non-parametric methods do not have assumptions or prior knowledge about data distribution. Instead, these methods rely on data-driven techniques such as density, distance, and rank measures to identify outliers within a dataset.

3.2 Distance-Based Techniques

Distance-based methods of outlier detection identify outliers based on their spatial relationship to other points in a metric space. The key idea behind these methods is that normal data instances lie in a dense neighborhood, while outlier instances lie far from the nearest neighbor. These methods identify anomalies by computing the distance between points [5].

3.3 Density-Based Techniques

The Density-based methods identify outlier instances by measuring the local density of data points. It assumes that normal data points lie in denser regions, while outlier instances lie in sparse regions. These methods estimate the density around each data point and compare it with the density of its neighborhood to detect deviations.

4. Clustering-Based Techniques

The clustering-based method of outlier detection identifies outliers by analyzing the structure of clusters formed within the data [7]. The basic assumption is that the normal data instances form large, dense clusters while outlier instances form small or sparse clusters. Clustering-based methods are generally divided into five categories: partition-based, density-based, Grid-based, Hierarchical-Based, and Model-based. Among others, Density-Based Spatial Clustering Application with Noise (DBSCAN) is the most extensively used, which detects outliers in low-density areas and can identify clusters of arbitrary shape [8].

5. Methodology

The study focuses on identifying the unusual or outliers in a categorical dataset by employing the customized Density-based Spatial Clustering

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0027
e ISSN: 2584-2854
Volume: 04 Issue: 02
February 2026
Page No: 192 - 197

Application with Noise (DBSCAN), named the CatDBSCAN approach. Traditional DBSCAN is designed for numerical data, relying on distance metrics such as Euclidean distance. However, categorical data require specialized handling due to the absence of inherent ordering and numerical distance. Therefore, this research proposes a modified DBSCAN algorithm suitable for categorical attributes. In this study, any points that do not belong to a cluster (Noise points), as well as points residing in the small cluster, are considered potential outliers. An outlier score is assigned to each point, quantifies the degree to which the point deviates from the dense regions of the dataset. To assess the performance of the standard DBSCAN and customized DBSCAN (CatDBSCAN), two widely used evaluation metrics for imbalanced datasets are employed: the ROC-AUC and PR-AUC, provides a measure of an algorithm's ability to distinguish between normal points and outlier instances.

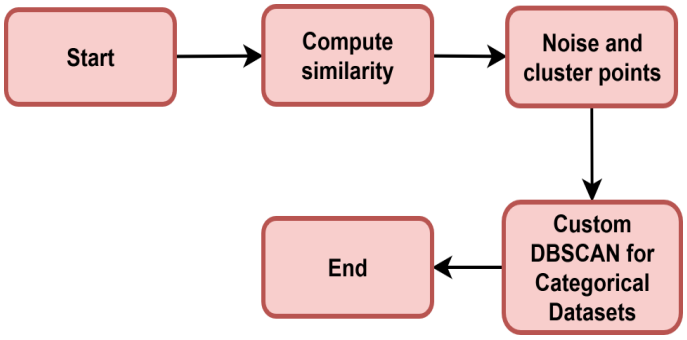

Figure 1 Roadmap of Methodology

6. Categorical Density-Based Spatial Clustering Application with Noise

An efficient and straightforward clustering-based approach for unsupervised outlier detection specifically designed for categorical datasets has been introduced. This technique utilizes DBSCAN for clustering and determines outlier scores based on the characteristics of the clusters. Noise points are labeled as 1, while points within smaller clusters are also evaluated for their outlier status, resulting in the calculation of an outlier score. CatDBSCAN stands for Categorical Density-Based Spatial Clustering Application with Noise. CatDBSCAN is an improved version of DBSCAN specifically created to detect anomalous instances within categorical datasets. CatDBSCAN is a specialized adaptation of DBSCAN, representing a modification of the foundational DBSCAN algorithm. As DBSCAN is not inherently suited for categorical data, a distance matrix is derived using the Hamming distance. Typically, DBSCAN considers noise points as outliers; however, in this tailored version, both noise points and small clusters are classified as outliers, since these smaller clusters often harbor anomalous patterns, while true data points tend to cluster into larger or denser formations. An outlier score is designated; noise points, indicating significant anomalies, receive a score of 1.0. Points situated in small clusters are deemed to be more anomalous. An outlier score is calculated, ensuring that points within small, potentially rare groups are marked with a higher outlier score. The static parameters eps and min_samples are utilized. While this approach is efficient, it tends to be less precise and adaptable. The CatDBSCAN outlier detection technique is lightweight, interpretable, and effective for initial outlier detection tasks, irrespective of the availability of labeled data. The CatDBSCAN merges the advantages of Density-Based clustering through DBSCAN with a straightforward scoring system that prioritizes noise and small cluster affiliations.

Algorithm

Input:
X ← Dataset with n samples and d categorical/binary features
Y ← True binary labels (optional, for evaluation)
D ← Distance metric (e.g., 'hamming')
eps ← DBSCAN epsilon parameter
min_samples ← DBSCAN minimum samples per cluster
Output:
Outlier score vector $s \in [0, 1]$ for each sample
Evaluation metrics (ROC AUC, PR AUC)

Step 1: Distance Matrix Computation
Compute a pairwise distance matrix using the specified metric (e.g., Hamming), which captures the dissimilarity between all sample pairs:

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0027
e ISSN: 2584-2854
Volume: 04 Issue: 02
February 2026
Page No: 192 - 197

dist_matrix = squareform(pdist(X, metric=D))

Step 2: DBSCAN Clustering

Apply DBSCAN with the precomputed distance matrix to identify dense clusters and noise:

labels = DBSCAN (eps=eps, min_samples=min_samples, metric='precomputed').fit_predict(dist_matrix)

Step 3: Outlier Score Assignment

Initialize an outlier score array s of length *n*, all values set to 0.

Count the number of points in each cluster:

label_counts = count of samples per label

For each point *i*:

If it is labeled as noise (label = -1):

s[i] = 1.0 (maximum outlier score)

Else:

s[i] = 1 - (cluster_size / total_samples)

(Smaller clusters yield higher outlier scores)

Step 4: Evaluation (Optional)

If ground truth labels *Y* are available, evaluate the outlier scores using:

- ROC AUC (Receiver Operating Characteristic Area Under Curve)
- PR AUC (Precision-Recall Area Under Curve)

Return:

- Outlier scores for all samples
- ROC AUC and PR AUC (if *Y* is given)

7. Experimental Result

Our experiment is conducted on the purely categorical "Mushroom" and "Breast Cancer" datasets taken from the UCI Machine Learning Repository [9], [10]. The mushroom dataset is a dataset of food safety that consists of 22 attributes, and the breast cancer dataset consists of 9 features. The minority class is considered an outlier instance on both datasets. In the mushroom dataset, the toxic classes are considered as outliers as the poisonous mushroom exhibits rare or unusual attribute combinations, and in the breast cancer dataset, rare patterns of events are considered outlier instances.

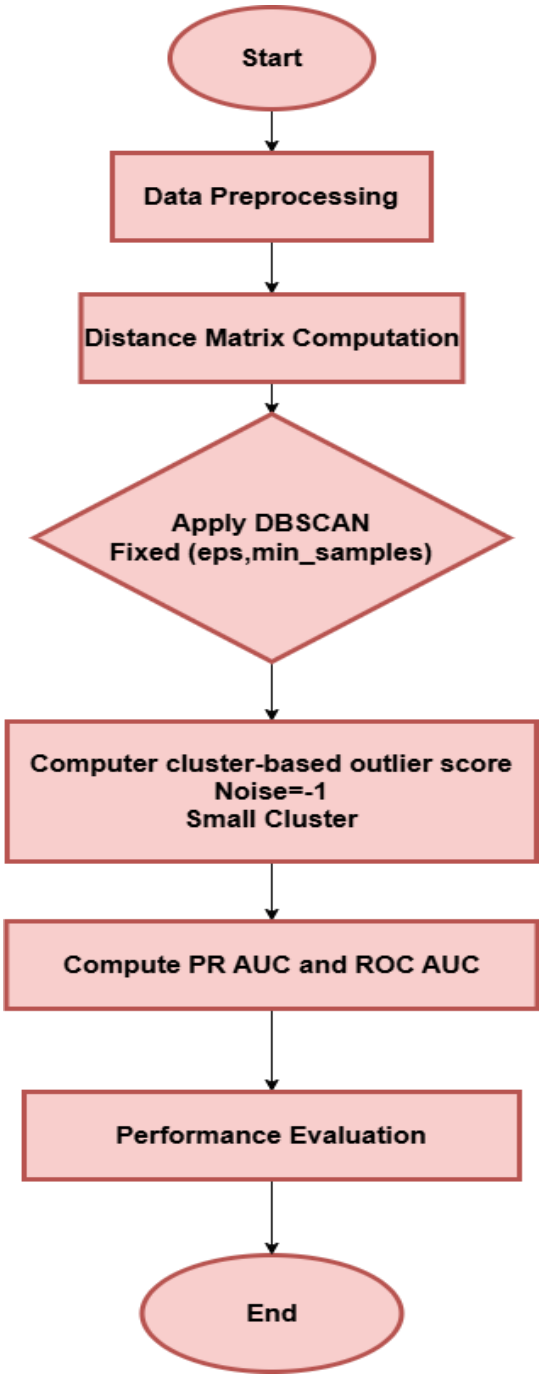

Figure 2 Workflow of CatDBSCAN

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0027
e ISSN: 2584-2854
Volume: 04 Issue: 02
February 2026
Page No: 192 - 197

Table 1 Result On Mushroom Dataset

	CatDBSCAN	DBSCAN
ROCAUC	0.98	0.50
PRAUC	0.98	0.0883

Table 2 Result On Cancer Dataset

	CatDBSCAN	DBSACN
ROCAUC	0.9994	0.50
PRAUC	0.9994	0.061

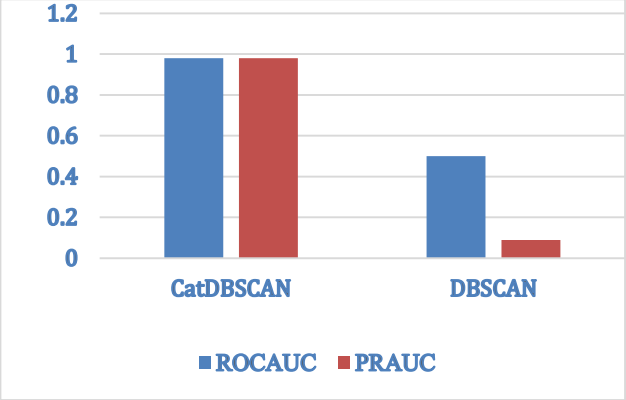

Figure 3 Analysis of CatDBSCAN VS DBSCAN on Cancer Dataset

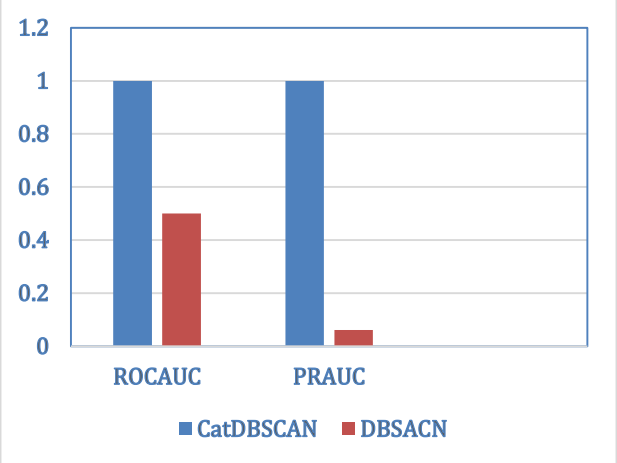

Figure 4 Analysis of CatDBSCAN VS DBSCAN on Mushroom Dataset

The results are evaluated using the Precision-Recall Area Under the Curve and ROC-AUC (Receiver Operating Characteristic Area Under the Curve)[11], [12]. The ROC-AUC score quantifies the model's ability to distinguish outliers from inliers by measuring how well it ranks true outliers above normal data points. PR-AUC focuses only on the positive class (outliers) and provides a clearer picture of how well the model identifies rare anomalies. The result demonstrated that our method provided the best result in terms of both metrics.

Conclusion

Outlier detection represents unique challenges due to the lack of natural ordering and inherent distance measures. In this paper, an innovative approach of DBCSAN has been introduced, called as CatDBSCAN, for identifying outliers within categorical datasets. As DBSCAN is naturally not designed to deal with numerical data, to apply it effectively, a distance matrix is computed using Hamming distance for each data point to quantify the dissimilarities between categorical data points. An outlier score is assigned to each data point. The noise points indicate a strong anomaly, and the outlier score is computed for the points in the small clusters. It ensures that points in small, potentially rare or unique groups are flagged with a higher outlier score. The effectiveness of the proposed CatDBSCAN is tested on two real-world datasets, and ROC-AUC AND PRAUC scores are calculated for both datasets. It is observed that CatDBSCAN performs best on both datasets. Our outlier detection policy can still be improved by a hyper parameter tuning process and applied on dataset of different domains of mixed data type.

References

[1]. H. Wang, M. J. Bah, and M. Hammad, "Progress in Outlier Detection Techniques: A Survey," IEEE Access, vol. 7, pp. 107964–108000, 2019, doi: 10.1109/ACCESS.2019.2932769.

[2]. 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.

[3]. N. N. R. R. Suri, N. Murty, and M. G. Athithan, Outlier Detection: Techniques and

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0027
e ISSN: 2584-2854
Volume: 04 Issue: 02
February 2026
Page No: 192 - 197

Applications - A Data Mining Perspective. 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-05127-3%0A

[4]. V. Chandola, "Outlier Detection : A Survey."

[5]. E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications."

[6]. L. Duan, "5 Density-Based Clustering and Anomaly Detection." [Online]. Available: www.intechopen.com

[7]. 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.

[8]. 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

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

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

[11]. J. Davis and M. Goadrich, "The Relationship between PR and ROC curves," ACM International Conference Proceeding Series, vol. 148, pp. 233–240, 2006, [Online]. Available: http://portal.acm.org/citation.cfm?doid=1143844.1143874

[12]. T. Fawcett, "An introduction to ROC analysis," Pattern Recognit Lett, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.