



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Application of Clustering-Based Outlier Scoring with GDBSCAN for Health Care and Food Safety Domain

Aditi Badhan¹, Anita Ganpati²

¹Department of Computer Science, Himachal Pradesh University, Shimla, India, Email: aditibadhan@gmail.com

²Department of Computer Science, Himachal Pradesh University, Shimla, India, Email: anitaganpati@gmail.com

Abstract

Outlier detection is an essential task of discovering unusual or rare patterns in a dataset that have exceptional behavior compared with the other records in the dataset. Outliers do not follow patterns with other objects in the dataset. Many approaches identify outliers in numerical data; our focus is on identifying the noise points and cluster-based outliers in the categorical data. Due to the scarcity of labeled data, high cost, and time required in data annotation of the labeled data, it is impractical to obtain labeled datasets. The challenges led to a growing interest in the unsupervised methods, which do not require labeled data and find patterns within the data. The Density-based spatial clustering Application with Noise (DBSCAN) is integrated with the grid search for hyperparameter tuning, and a special scoring mechanism for identifying the noise points and cluster-based outliers in a categorical dataset. The method CBOS-GDBSCAN is proposed in our earlier study. To demonstrate the practical applicability of the method, in this study, Clustering-based Outlier Scoring with Grid DBSCAN is implemented. The method CBOS-GDBSCAN effectively identifies outliers in the dataset characterized by high-dimensional and large-scale categorical features. The method is applied to the real-world dataset taken from the UCIML repository of diverse domains such as healthcare and food safety. The experimental analysis demonstrates that the method effectively identifies outliers effectively show that the method outperforms and detects outliers in a large-scale categorical dataset.

Keywords: Outlier Detection, DBSCAN, Grid DBSCAN, CBOS, and Clustering-based Outlier Detection.

Introduction

Outlier detection is an essential area of study in many applications, such as credit card fraud, intrusion detection, medicine, and public health. Outlier identification aims to identify unusual or rare instances in the dataset [1]. The infrequent items hold valuable items, may have the potential to reveal data entry errors, rare events, fraud, and emerging trends and interesting patterns that are worthy of investigation. Discovering and analysing the outlier instances helps the analysts and researchers to gain deeper insight, yielding more precise and reliable decisions informed by data[2]. Moreover, the importance of identifying outlier instances is rapidly increasing across different areas of application. For example, an unusual transaction on a credit card may indicate credit card theft and misuse. In the case of intrusion detection, anomalous traffic patterns within a network may indicate that the hacked computer is sending sensitive information to an unauthorized destination.

The healthcare and diagnostic domain uses outlier detection to identify rare patterns in the patient data. The rare patterns in the record may represent serious hidden health concerns, and the earlier identification facilitates timely treatment and improves patient outcomes. Identifying unusual instances and irregularities in the domain of food safety helps to ensure to maintain quality control standards and protect public health. Generally, the ability to detect outliers not only enriches our understanding of the data but also translates to actionable insights that can drive significant improvements across different sectors. The utilization of advanced outlier detection techniques helps the organization to enhance its operational efficiency, safeguard against risks, and ultimately make more informed decisions. Several techniques have been developed for identifying outliers or rare patterns within the datasets, such as statistical-based [3], Distance-based [4], Density-based [5], ensemble-based [5], and learning-based methods. Generally, on the basis of the availability of labeled data, it is classified into three categories, namely supervised methods, semi-supervised methods, and unsupervised methods [7]. The supervised methods of outlier detection require a labeled dataset [8]. In contrast, Semi-supervised methods where labels are available for normal instances only, and outliers are detected as deviations from the normal patterns. Since in real-world scenarios it's impractical to obtain the labeled dataset, led to an increase the interest towards unsupervised methods that does not require predefined labels and can discover hidden patterns in the data. These methods rely on the intuition that outliers are rare and different from the normal data points [9].

Cluster Analysis is a widely used unsupervised learning technique that groups similar instances of data into clusters [8][10]. The use of clustering is rapidly increasing with the development of data mining, machine learning, and artificial intelligence [11]. As an unsupervised classification method, the aim of clustering is to organize a set

of objects into distinct groups without any prior information, resulting in objects within the same cluster being more alike than those in separate clusters [11]. Clustering and Outlier Detection are essentially quite distinct from one another. While the former focuses on identifying groups with similar characteristics, the latter seeks to identify instances that differ significantly from all others in the dataset [8], [12]. However, methods for outlier detection based on clustering have been developed, leveraging the understanding that outliers do not fit into any cluster due to their scarcity and dissimilarity to typical instances [8]. DBSCAN, a Density-Based Spatial Clustering Application with Noise, is a density-based clustering technique introduced by Ester in 1996 [13]. It has the ability to effectively detect clusters of arbitrary shape and identify noise points in sparse or low-density regions. This capability is particularly advantageous in fields where data labeling presents significant challenges, such as in food safety datasets, which may contain numerous variables and complex relationships, or in healthcare datasets, where privacy concerns and the variability of patient data complicate labeling efforts. The method is proposed in our earlier study. In the present work, CBOS-GDBSCAN, a clustering-based outlier score with Grid DBSCAN, is implemented for the identification of outliers. Our method is tested using two real-world datasets of a food safety and a medical health dataset from the UCI ML repository, and compared with other techniques, which are DBSCAN, CatDBSCAN [14] INFS-DBSCAN [15], Adaptive KNN [16], Reverse KNN [17], ROCK and K- Modes [18]. The remainder of the paper is organized as follows. The background is presented in Section 2. The Methodology of the paper is discussed in Section 3. The evaluation metrics to assess the effectiveness of the method are discussed in Section 4. The detailed experiment on two types of UCI datasets is implemented in Section 5. The results and discussion are discussed in Section 6. The conclusion is provided in Section 7.

Background and Related Work

Outlier detection techniques are widely studied across various domains. Generally, Outlier detection methods are classified into Statistical-based, Density-based, Distance-based, Clustering-based, Ensemble-based, and Learning-based approaches [2] [7], [19]. The statistical-based techniques [20] assume a distribution and identify instances that deviate significantly. In these methods, a data point is considered an outlier if it is extremely deviated from a standard distribution [19]. The core principle of the density-based approaches [21] of outlier detection is that outliers are the data points that can be found in a low-density region, and the normal lies in a denser region. Distance-based techniques [4] identify outliers based on the distance of a data point from its nearest neighbors, assuming that outliers are far from the other points [4][22] methods that detect outliers are those data instances that do not belong to any cluster or lie in a sparse cluster [23]. Ensemble-based methods combine multiple models to improve the robustness of outlier detection [24].

Machine learning techniques, encompassing supervised, semi-supervised, and unsupervised categories, have become increasingly favored for detecting outliers. Supervised techniques like isolation forests and support vector machines necessitate labeled data, which is frequently difficult or costly to acquire. Unsupervised techniques such as K-Means, hierarchical clustering, and DBSCAN have gained popularity for their capability to identify patterns without any prior information. DBSCAN, introduced by Ester et al. (1996) [13], has emerged as a powerful clustering and outlier detection tool. Its effectiveness in detecting non-spherical clusters and handling noise makes it ideal for real-world applications. DBSCAN has been used to identify anomalies in ECG signals and unusual patient vital signs. Despite its proven efficacy, the application of DBSCAN on fully categorical datasets remains relatively underexplored. Many studies rely on numerical or mixed-type data, overlooking the challenges posed by high-dimensional categorical attributes. The performance of DBSCAN is influenced by the hyperparameters. The Clustering-based Outlier Scoring with Grid Density-based Spatial Clustering Application with Noise) CBOS-GDBSCAN uses a grid search method, so it helps to improve the global search ability and convergence accuracy, which helps in identifying the optimal parameter values of the (ϵ)epsilon and MinPts. Unlike traditional DBSCAN, this approach provides a more systematic and automated parameter selection. We validated our proposed method based on PRAUC, ROCAUC, Precision@k, and Recall@k with other outlier detection methods, by using the Real-world datasets of diverse domains taken from the UCI ML Repository with categorical attributes.

Applications of Outlier Detection

Outlier detection has numerous applications in a wide variety of domains, such as the financial industry, intrusion detection, food safety, and medical and public health. The application of outlier detection is diverse [9]. Here, we mention a few:

1. **Network Intrusion Detection:** Intrusion Detection System (IDS) has immensely benefited from outlier detection techniques, which help to identify malicious activities within computer systems and networks.
2. **Health Care:** The health care and medicine sectors use outlier detection to work with patient records. The methods enable to discover the rare and unusual patterns in the patient records. The earlier detection of disease helps to provide better treatments.
3. **Fraud Detection:** Financial organizations utilize outlier detection methods to recognize anomalous transactions. These unusual transactions indicate error, fraud, and suspicious activity. Similar to credit card fraud, where the wrong use of a credit card may be recognized upon observation of unusual patterns [25][8].
4. **Food Safety:** Detecting outliers is essential in food safety as it ensures to uphold the quality standards are upheld, as it represents critical health concerns. Contaminated food products may lead to serious diseases. A prominent example involves the utilization of the mushroom dataset of the UCI ML Repository, where a classification of poisonous and edible mushrooms is based on various features such as cap shape, color, aroma,

and gill size, etc. Identifying outliers in such a dataset helps to prevent foodborne illness by highlighting the toxic mushroom patterns that deviate from the established classifications.

5. **Text Data Detection:** A collection of information holds a substantial amount of data, and a portion of it consists of noisy data. This noisy data is identified through outlier detection methods applied to the dataset. The primary goal of outlier detection is to manage the noisy data within the extensive dataset [26].

6. **Internet of Things (IOT):** The advancement of technology and the rise of IoT devices have led to an escalation of connected devices. IoT devices are generating a stream of data in real time in large volume. Outlier detection in IoT environments is crucial for monitoring sensor networks, detecting equipment faults, and ensuring system reliability. For example, anomalies in smart home sensors can indicate security breaches or malfunctions in appliances.

7. **Environmental Monitoring:** The monitoring of air and water quality, the identification of weather anomalies, and the forecasting of natural disasters are all environmental applications and activities. Sudden surges in pollution levels, inconsistency in temperature, and seismic events can be identified through machine learning techniques. Data derived from satellites and remote sensing can be analyzed for identifying the anomalous patterns like deforestation, crop disease, and land cover changes. Outlier detection plays an important role in identifying early warning systems for natural hazards, including floods, wildfires, and hurricanes.

Methodology

In this study, the CBOS-GDBSCAN, a clustering-based method of identifying unusual instances called outliers within the categorical dataset, is implemented. The method is implemented on the real-world datasets of diverse domains of Health (Breast cancer dataset) and the Food safety domain (Mushroom datasets). The features are encoded using one-hot encoding to enable similarity computation suitable for clustering. A pairwise distance is computed, and GDBSCAN is applied to group the instances into clusters. The performance of the method is assessed using ROCAUC and PRAUC, Precision@k and Recall@k metrics, which effectively measure the ability to distinguish outliers from the normal instances in an imbalanced dataset. The methodology is described in Figure 1.

CBOS-DBSCAN Algorithm

CBOS-DBSCAN is a clustering-based method for detecting noise points and cluster-based outliers in a categorical dataset. The method adapts a Hamming distance measure for computing the similarity and dissimilarity among the categorical attributes and incorporates a grid search method to identify the optimal parameter values of epsilon (ϵ) and MinPts. Furthermore, the method employs a specific scoring process which quantifies the likelihood of data points being an outlier; data points identified as noise are assigned the highest score as they belong to a sparse region. The method is particularly effective at identifying outliers in categorical datasets, which is challenging for traditional clustering algorithms.

In the initial stage, the datasets have been pre-processed. Subsequently, the distance matrix is computed using the Hamming distance, a method well-suited for categorical data to measure the dissimilarity between the data points based on their categorical attributes. The distance matrix serves as a basis upon which the Grid DBSCAN method operates. It enables the GDBSCAN to identify the dense regions—areas with a high concentration of data points—and sparse regions—areas with fewer data points. Once the clustering procedure is executed, CBOS-GDBSCAN provides a score to each data point to assess the outlierness of data points. In contrast to traditional DBSCAN, where data points are classified as core, border, and noise points without providing a quantitative measure of outlierness, CBOS-GDBSCAN assigns a continuous score between 0 and 1 to each data point. The data points identified as noise points by DBSCAN receive the highest score, indicating a high potential of being an outlier. Furthermore, each data point within a cluster is also analyzed for its potential to be an outlier based on the distance from the cluster centroid. Here, the size of the cluster is also considered to measure the outlierness of clusters, and weights are assigned to clusters. The small clusters are assigned greater weight in the outlier scoring process, reflecting the intuition that they are more likely to contain unusual data points as compared to large or dense clusters. The performance of DBSCAN is greatly affected by the parameters ϵ (the maximum distance between two data points for them to be considered as in the same neighborhood) and MinPts (minimum number of points to form a dense region). To optimize the clustering process, a grid search is conducted over a spectrum of parameters. The hyperparameter tuning is executed through a performance-driven approach. Each combination of parameters is evaluated by utilizing the DBSCAN algorithm. The outlier score for each data instance is computed; subsequently, it is evaluated using the supervised metrics, assuming the ground truth is available for benchmarking purposes. The configuration yielding the maximum ROCAUC is retained as the optimal configuration. Using the optimal parameter configurations, the outlier scores are further evaluated to give a comprehensive assessment of the outlier detection performance. Overall, the CBOS-GDBSCAN not only improves on the standard DBSCAN by introducing a scoring procedure, but also provides an effective method for detecting outliers in categorical datasets.

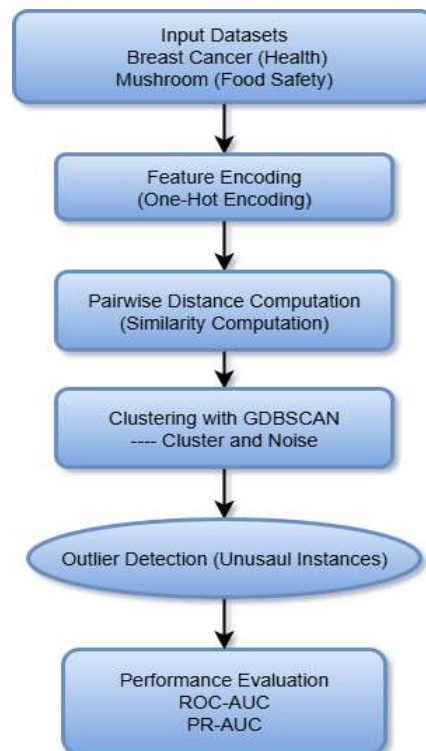


Fig. 1 Methodology Roadmap

Algorithm

Input:

$X \leftarrow$ Dataset (n samples $\times$ d features)
 $D \leftarrow$ Distance metric ('hamming' or 'jaccard')
 $E \leftarrow$ Set of DBSCAN eps values (Grid)
 $M \leftarrow$ Set of DBSCAN min_samples values (Grid)
 $Y \leftarrow$ True binary labels (for evaluation)

Output:

Best DBSCAN parameters (eps*, min_samples*)
 Outlier scores $s \in [0, 1]$ for each point
 Evaluation metrics (ROC AUC, PR AUC, etc.)

1. Compute the distance matrix dist_matrix from X using metric D
2. Initialize $\text{best_score} \leftarrow 0$, $\text{best_params} \leftarrow \text{None}$, $\text{best_scores} \leftarrow \text{None}$
3. For each eps in E :
4. For each min_samples in M :
5. Run DBSCAN on dist_matrix with eps, min_samples $\rightarrow$ get labels
6. If all points are in one cluster or each in its own cluster:
7. Continue
8. Initialize $\text{outlier_scores} \leftarrow \text{zeros}(n)$
9. For each unique cluster label c :
10. If $c == -1$: // noise points
11. Set score 1.0 for all points with label -1
12. Else:
13. Let $X_c \leftarrow$ points in cluster c
14. Compute centroid μ_c of X_c
15. For each point x in X_c :
16. Compute distance $d \leftarrow ||x - \mu_c||$
17. Normalize distances to $[0, 1]$
18. Compute $\text{size_weight} \leftarrow 1 - (|X_c| / n)$
19. For each point x in X_c :
20. $\text{outlier_score} \leftarrow \text{normalized_distance} \times \text{size_weight}$
21. Assign score to x
22. Evaluate scores against true labels (e.g., ROC AUC)
23. If performance improves:
24. Update best_score , best_params , best_scores
25. Return best_params , best_scores , $\text{evaluation_metrics}$

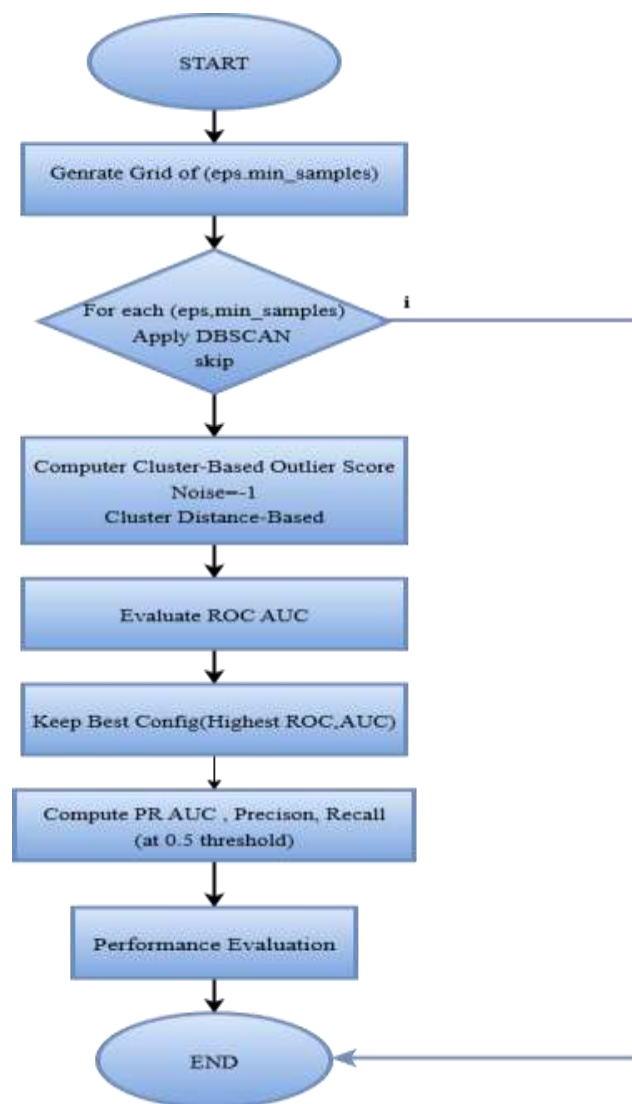


Figure 2. Workflow of CBOS-GDBSCAN

Datasets

In this study, two distinct datasets have been used to investigate the practical application of the proposed method.

Mushroom Dataset

Mushroom dataset is a popular dataset of food safety obtained from the publicly available UCI ML repository. The dataset. The dataset has been established as a benchmark for classification tasks and has also demonstrated its utility in detecting outliers. The dataset consists of 8124 instances, which describe 23 species of mushroom using 22 categorical attributes, such as cap shape, Odor, gill size, and habitat, along with a binary class label indicating whether the mushroom is edible or poisonous[27]. In the scenario of outlier detection, this dataset presents unique challenges and opportunities. Since poisonous mushrooms may exhibit rare or unusual attribute combinations, they can be interpreted as outliers.

Table1. Mushroom Dataset

Attribute	Values
Number of instances	8124
Number of attributes	23
Attribute Types	Categorical
Missing Values	yes
Target Variable	Class (edible or Poisonous)

Breast Cancer Dataset

Breast cancer is a dataset in the healthcare domain that comprises of categorical attributes obtained from the UCI ML repository. The dataset consists of records of breast cancer patients predicting the recurrence and non-recurrence of the disease. It has 9 categorical attributes and 286 instances. The dataset is suitable for outlier

detection as it consists of rare clinical profiles and rare combinations of categorical features that deviate from the other data instances [28].

Table 2. Breast Cancer Dataset

Attribute	Values
Number of instances	286
Number of attributes	9
Attribute Types	Categorical
Missing Values	yes
Target Variable	Class (no-recurrence-events or recurrence-events)

Evaluation Metrics for the Performance of Methods

Evaluating the effectiveness of outlier detection methodologies presents significant challenges, primarily due to the unsupervised nature of the problem and the frequent class imbalance observed in real-world datasets. To assess the effectiveness of the ranking generated by the anomaly scoring mechanism, in this study, the Receiver Operating Characteristic Area Under the Curve, Precision-Recall Area Under the Curve, Precision@k and Recall@k are used to measure the performance of Outlier detection methods.

Precision @k: This measure can be calculated on k^{th} value. This measures the proportion of true outliers among the top- k instances identified by the detection algorithm as most anomalous. It is defined as:

$$\text{Precision@k} = \frac{\text{Number of true outliers in top k}}{k} \quad (1)$$

This metric captures how well the model ranks the actual outliers at the top of the anomaly score list. It is especially valuable in real-world scenarios

Recall @k: This measures the proportion of all true outliers that are identified within the top- k ranked instances:

$$\text{Recall@k} = \frac{\text{Numer of true outliers in top k}}{\text{Total number of true outliers}} \quad (2)$$

This metric evaluates the coverage of true anomalies in the top- k results. High Recall@k [2] indicates that the algorithm successfully includes a large portion of the true outliers in its most anomalous predictions.

ROC-AUC: The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across varying decision thresholds[2], [29] AUC quantifies the overall performance of a classifier by measuring the area under the curve; it represents the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative instance. It provides insight into the trade-off between detecting true anomalies and incorrectly labeling normal instances as outliers. Formally, these rates are defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

It measures the proportion of actual positives correctly identified.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (4)$$

It measures the proportion of actual negatives incorrectly classified as positive.

The ROC-AUC represents the area under this curve and summarizes the model's ability to distinguish outliers above inliers [29]. It can handle an imbalanced dataset and has ranking ability rather than strict classification.

PR-AUC: The PR AUC stands for Precision-Recall Area Under the Curve, which is a metric used to evaluate classification models, particularly in cases where class imbalance is present. It provides a more focused evaluation by plotting Precision against Recall [29]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

It measures how many of the predicted positive instances are actually correct.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

It measures how many actual positive instances were correctly identified.

The PR-AUC, which is the area under this curve, is particularly suitable for imbalanced settings like outlier detection, where the number of true anomalies is small [2].

Experimental Analysis

In this study, a Clustering-based Outlier Scoring method with Grid DBSCAN (CBOS-GDBSCAN) is implemented to detect outliers within categorical datasets. The performance of the method is evaluated on two diverse real-world datasets from the health (Breast Cancer) and food safety (Mushroom Dataset) domains. To assess the effectiveness of the proposed method, it is compared against several existing methods designed for numerical, categorical, mixed-type and high-dimensional data. The results demonstrate that the proposed method performs better across multiple performance metrics, including Precision, Recall@k, ROCAUC, and PRAUC. Furthermore, to examine how well the methods perform if the number of attributes increases, an additional experiment is conducted under high-dimensional settings. High-dimensional data present challenges for clustering and outlier detection because increasing the number of attributes can reduce the distinction between normal and anomalous observations. It is commonly associated with the curse of dimensionality, which can adversely affect distance-based and density-based methods [30]. The findings demonstrate that the proposed methods attain stable detection performance even in high-dimensional settings. Scalability is investigated by progressively increasing the datasets synthetically and measuring the execution time and detection performance of the method.

Table 3: Results on the Mushroom Dataset

	DBSCAN	CatDBSCAN	CBOS-GDBSCAN (Proposed)	Reverse KNN	Adaptive KNN	K Modes	ROCK
ROC AUC	0.5	0.99	1	0.62	0.51	0.546	0.5115
PR AUC	0.0061	0.99	1	0.17	0.55	0.552	0.61
Precision@k	0.061	1	1	0.1016	0.58	0.549	0.6667
Recall@k	1	0.998	1	0.93	0.1358	0.81	0.52

Table 4. Results on the Breast Cancer Dataset

	DBSCAN	CatDBSCAN	CBOS-GDBSCAN (Proposed)	Reverse KNN	ROCK	K Modes
ROC AUC	0.5	0.98	1	0.79	0.5	0.47
PR AUC	0.883	0.98	1	0.18	0.81	0.636
Precision@k	0.7915	1	1	0.79	0.627	0.708
Recall@k		0.96	1	1	1.0	0.224

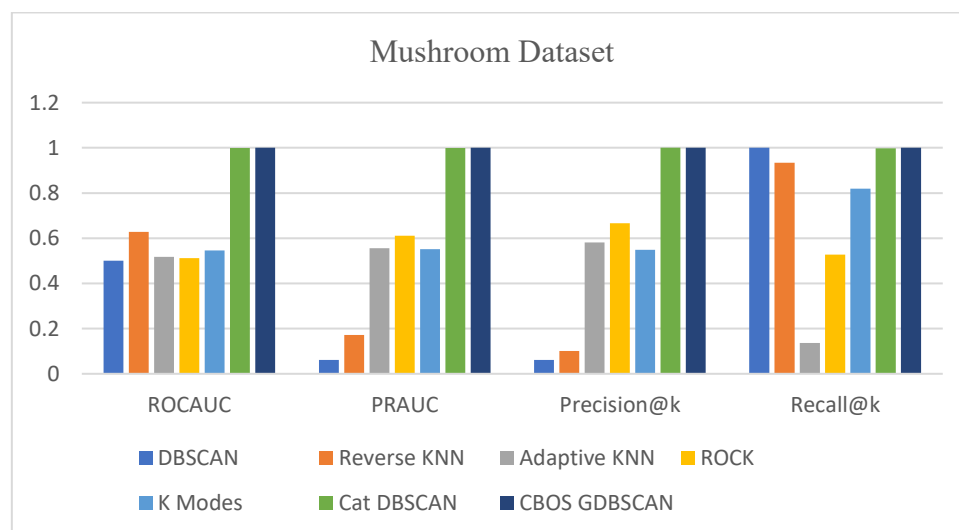


Figure 3. Analysis Result of Methods on Mushroom Dataset

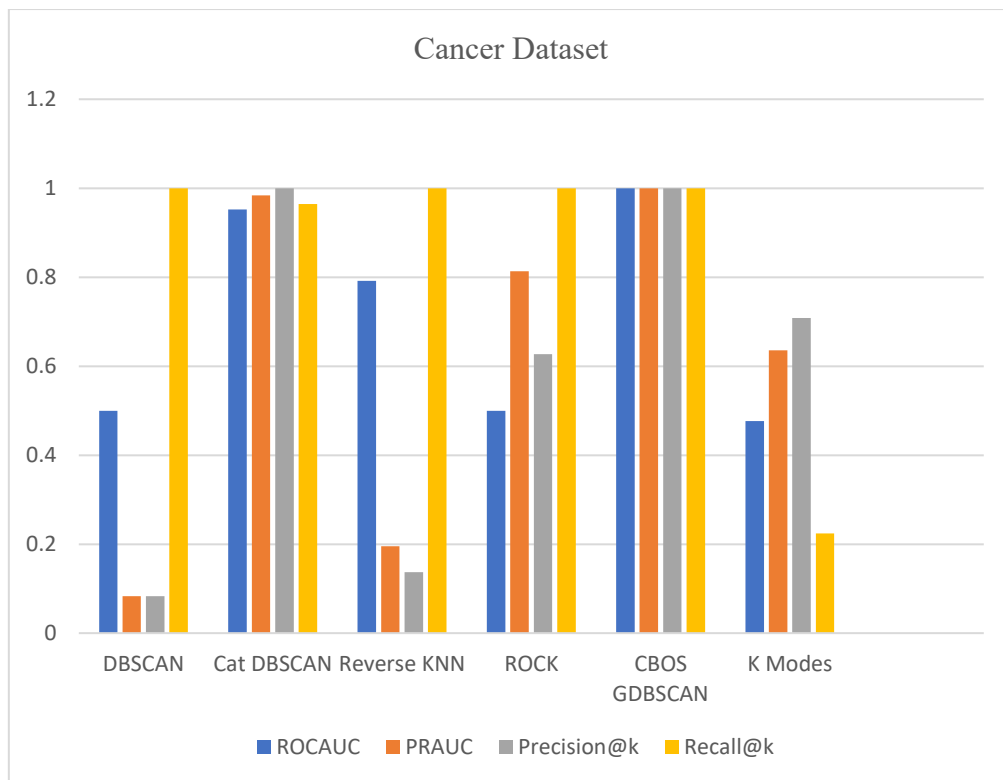


Figure 4. Analysis Result of Methods on Cancer Dataset

Experiment Result for High-Dimensional Data

DBSCAN tends to become unreliable when applied to high-dimensional data [31]. Outlier detection may yield poor results in high-dimensional contexts [15]. We assess our proposed CBOS-GDBSCAN methodology against INFS-DBSCAN [15]. In INFS-DBSCAN, to mitigate the curse of dimensionality, feature selection is performed, leading to the exclusion of certain features, whereas our study incorporates high-dimensional data; all features are included during the analysis and identification of outliers. INFS-DBSCAN only considered the top 10 features, while CBOS-GDBSCAN analyzes 22 features within the dataset. Experimental analysis shows that the proposed method performs better than INFS-DBSCAN.

Table 5. Result of Mushroom in High-dimensional data

	CatDBSCAN	CBOS-GDBSCAN (Proposed)	INFS-DBSCAN
ROCAUC	0.98	1	0.49
PR-AUC	0.98	1	0.26

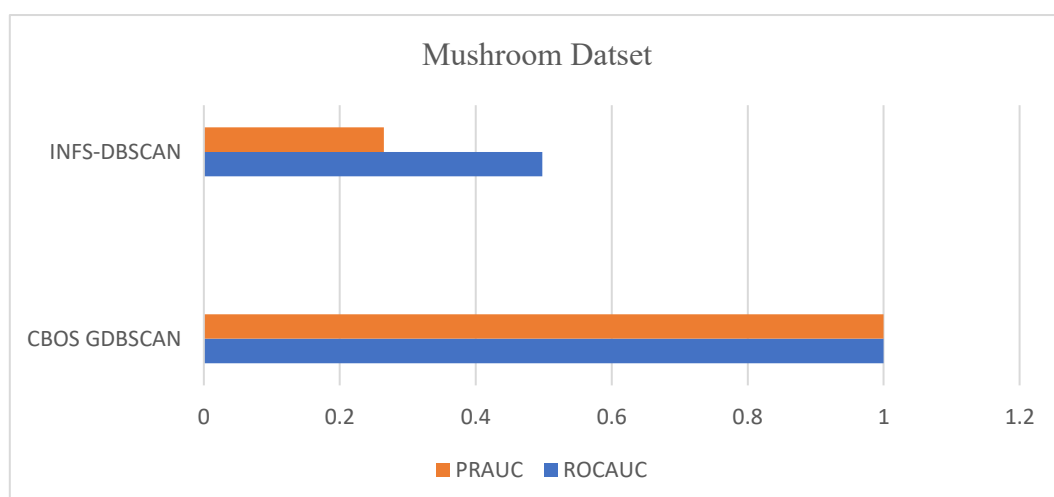
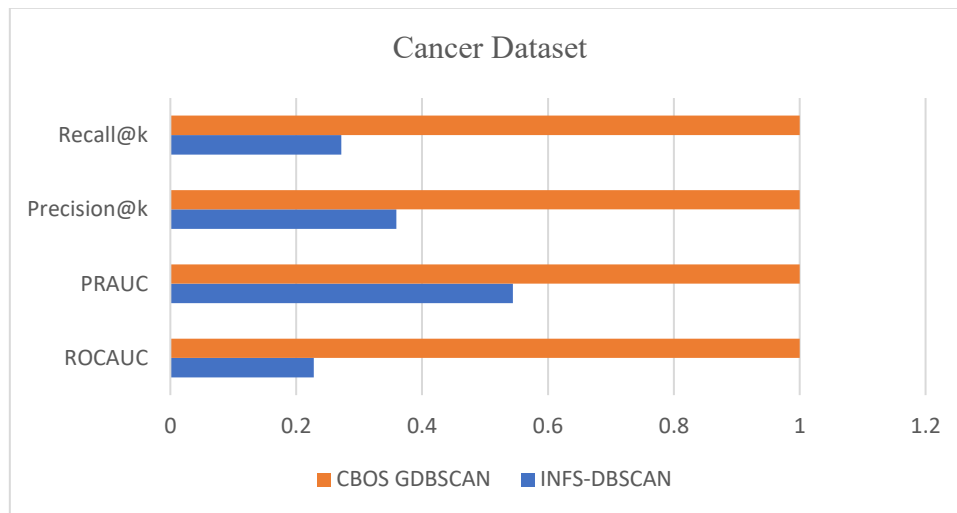


Figure 6. Analysis of results on high-dimensional data on the Mushroom Dataset

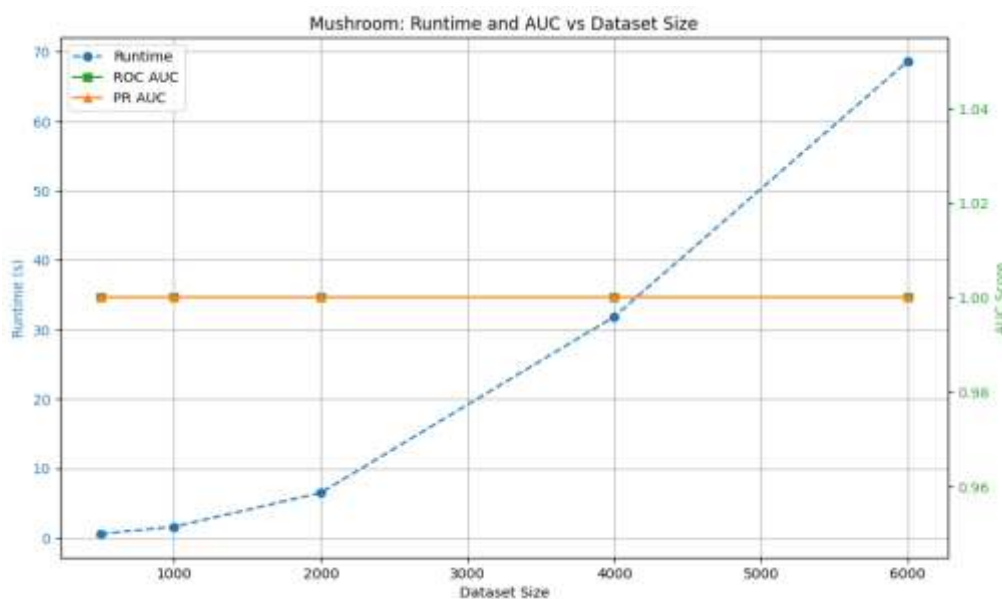
Table 6. Results of the Cancer dataset in High-dimensional data

	CBOS-GDBSCAN (Proposed)	INFS-DBSCAN
ROCAUC	1	0.22
PRAUC	1	0.54
Precision@k	1	0.35
Recall@k	1	0.27


Figure 7. Analysis results on high-dimensional data on the Breast Cancer Dataset

Scalability

Scalability analysis was conducted to investigate the potential and utility of the proposed method on larger categorical datasets. To measure the scalability of the method, it is executed on larger data by synthetically incrementing the size of the data. The mushroom dataset is increased progressively from 500 to 6000 observations. Figure 7. Represents the increase in execution with the growth in dataset size. The runtime is relatively low for smaller datasets but increases considerably as the number of observations becomes larger. In particular, the runtime increases from approximately 1.01 seconds for 500 observations to 56.39 seconds for 6000 observations. Despite the increase in computational cost, the ROC-AUC and PR-AUC remain approximately 1.00 across the tested dataset sizes. Thus, within the experimental range, increasing the number of observations did not produce a noticeable reduction in the measured outlier detection performance. The scalability experiment demonstrates that CBOS-GDBSCAN can be applied to progressively larger datasets while maintaining stable evaluation scores.


Figure 8. Scalability of the Proposed Method

Results and Discussion

The performance of the proposed method is compared against DBSCAN, CatDBSCAN, Reverse KNN, Adaptive KNN, K modes and ROCK using the ROCAUC, PRAUC, Precision@k and Recall@k. The evaluation measures are selected to evaluate the overall ranking of the anomalous observations and the quality of the highest-ranked outliers. Table 3 and Figure 3 present the results obtained for the mushroom dataset. Conventional DBSCAN achieved an ROC-AUC of 0.5000 and a PR-AUC of 0.0061, with Precision@k of 0.061 and Recall@k of 1.0000. Although its Recall@k indicates that the relevant anomalous observations were retrieved within the top-K set, the low Precision@k indicates that many non-anomalous observations were also included. CatDBSCAN substantially improved the results, achieving 0.9900 ROC-AUC and PR-AUC, with Precision@k of 1.0000 and Recall@k of 0.9900. The proposed CBOS-GDBSCAN achieved 1.0000 across all four-evaluation metrics, indicating consistent discrimination and top-K retrieval of the designated anomalous observations. The neighborhood-based approaches showed more variation. Reverse KNN obtained ROC-AUC and PR-AUC values of 0.62 and 0.17, respectively, while Precision@k and Recall@k were 0.1016 and 0.93. Adaptive KNN obtained ROC-AUC of 0.51, PR-AUC of 0.55, Precision@k of 0.58, and Recall@k of 0.1358. K-Modes produced ROCAUC of 0.546, PRAUC of 0.552, Precision@k of 0.549, and Recall@k of 0.81. ROCK achieved ROCAUC of 0.5115, PR-AUC of 0.61, Precision@k of 0.6667, and Recall@k of 0.52. These results show that the proposed method maintained consistent performance across both global ranking metrics and top-K metrics, whereas the competing approaches exhibited greater variation between precision and recall-based measures. The results also demonstrate the benefit of incorporating cluster-level information into anomaly scoring rather than relying solely on noise identification.

Results on the Breast Cancer Dataset

On the Breast Cancer dataset, DBSCAN obtained an ROCAUC of 0.5000, a PRAUC of 0.883, and a Precision@k of 0.7915. CatDBSCAN achieved 0.98 ROCAUC and PR-AUC, with Precision@k of 1.0000 and Recall@k of 0.96. CBOS-GDBSCAN again achieved 1.0000 for ROCAUC, PRAUC, Precision@k, and Recall@k. Reverse KNN obtained ROCAUC of 0.79, PRAUC of 0.18, Precision@k of 0.79, and Recall@k of 1.00. ROCK achieved ROCAUC of 0.50, PRAUC of 0.81, Precision@k of 0.627, and Recall@k of 0.224. K-Modes obtained an ROCAUC of 0.47, a PRAUC of 0.636, and a Precision@k of 0.708. The results demonstrate that CBOS-GDBSCAN provided consistent anomaly discrimination and top-K retrieval on both datasets. The method achieved perfect values of ROCAUC, PRAUC, Precision@k and Recall@k, which demonstrate that the ability of the method to correctly identify and rank the anomalous observations under the experimental condition.

Comparison with INFS-DBSCAN

Additional experiments were conducted using INFS-DBSCAN. The reported results show ROCAUC values of 0.49 and 0.22 and PR-AUC values of 0.26 and 0.54 across the evaluated experiments. Precision@k and Recall@k were 0.35 and 0.27, respectively, in the reported comparison. In contrast, CatDBSCAN achieved ROCAUC and PRAUC of approximately 0.98, while CBOS-GDBSCAN achieved 1.00 for both measures, together with 1.00 Precision@k and Recall@k. This difference indicates that the proposed cluster-aware scoring mechanism provided stronger anomaly ranking under the experimental conditions.

Discussion

The experimental results indicate that the choice of data representation, clustering strategy, and outlier scoring mechanism substantially affects categorical outlier detection performance. Traditional DBSCAN delivers limited performance, whereas CatDBSCAN substantially improved results by adapting a particular similarity measure to process categorical data. The proposed CBOS-GDBSCAN extends the approach by incorporating cluster characteristics such as cluster size and weight and small-cluster information into the identification of outliers. This enables the method to consider not only whether an observation belongs to a sparse region but also the characteristics of the cluster to which it belongs. The scoring awareness of clusters is relevant specifically for categorical datasets, where outlier and anomalous observations appear in the form of small groups rather than exclusively isolated points. The comparison with K modes and ROCK is performed as these are the relevant clustering algorithms specifically designed for categorical data; the comparison highlights the difference between categorical clustering and categorical outlier detection. The proposed method incorporates cluster information into the outlier scoring process. The method CBOS-GDBSCAN shows consistent performance across the evaluation measures on the datasets used in the study. The results show that the categorical density-based clustering and cluster-aware outlier scoring provide an effective method for identifying anomalous observations in categorical data. Despite this, the results are obtained and interpreted within the scope of the datasets and parameter configurations used in the experiments.

Conclusion

This study emphasizes the practical utility and effectiveness of CBOS-GDBSCAN for outlier detection on two entirely categorical datasets: the UCI Mushroom dataset and the structured health care dataset. While DBSCAN is traditionally applied to numerical datasets, our research demonstrates its effectiveness in identifying outliers within high-dimensional and large-scale datasets when appropriate preprocessing, such as one-hot encoding and a distance matrix, is computed using Hamming distance in categorical datasets. Outliers are particularly vital in

fields like toxicology or food safety; erroneous identification of a poisonous mushroom could lead to life-threatening situations. The outliers arises due to data entry errors, rare medical conditions, and fraudulent activities, emphasizing the need for effective outlier detection methods. Experiments are conducted to measure the outlier detection capability of the proposed method, and it is compared with other outlier detection methods including ROCK, K Modes, CatDBSCAN, INFS-DBSCAN, DBSCAN, Reverse KNN and Adaptive KNN. Experimental analysis demonstrates the effectiveness of the proposed method, and it is observed that CBOS-GDBSCAN achieved 1.0 for all evaluation measures on both datasets and outperformed other methods considered in this study. Furthermore, experiments with the INFS-DBSCAN show that the method can effectively identify outliers under high-dimensional settings. The scalability of the method is examined by progressively increasing the dataset size from 500 to 6000 observations. The experimental results show that there is an increase in runtime from 1.01 seconds to 56.39 seconds; despite the increase in execution duration, the method can effectively identify outliers. The scalability analysis shows that the method can handle large-scale datasets; however, the computational cost increases with data size. The results thus give a basis for future improvements and applications to large-scale categorical data. Moreover, Future work will be extended to use the proposed method in other application areas and data of mixed types.

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. D. M. Hawkins, *Identification of Outliers*. Springer Netherlands, 1980. DOI: 10.1007/978-94-015-3994-4.
3. E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications."
4. L. Duan, "5 Density-Based Clustering and Anomaly Detection." [Online]. Available: www.intechopen.com
5. A. D. Technique, F. K-means, E. K. Filter, and S. V. Machines, "Chapter 3 Ensemble-based Anomaly Detection Technique," pp. 64–83.
6. V. Chandola, "Outlier Detection : A Survey."
7. C. C. Aggarwal, "An introduction to outlier analysis. In *Outlier Analysis*," 2016. DOI: 10.1007/978-3-319-47578-3.
8. A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," 2000.
9. Q. Zhu, X. Tang, and A. Elahi, "Application of the novel harmony search optimization algorithm for DBSCAN clustering," *Expert Syst Appl*, vol. 178, Sep. 2021, DOI: 10.1016/j.eswa.2021.115054.
10. M. C. W. J. Gan Guojun, "Data Clustering: theory, algorithms, and applications," 2020.
11. 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
12. T. A. Messaoud, A. Smiti, and A. Louati, "A Novel Density-Based Clustering Approach for Outlier Detection in High-Dimensional Data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2019, pp. 322–331. DOI: 10.1007/978-3-030-29859-3_28.
13. Y. Wang, X. Cao, and Y. Li, "Unsupervised Outlier Detection for Mixed-Valued Dataset Based on the Adaptive k-Nearest Neighbor Global Network," *IEEE Access*, vol. 10, pp. 32093–32103, 2022, DOI: 10.1109/ACCESS.2022.3161481.
14. 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.
15. A. Smiti, "A critical overview of outlier detection methods," Nov. 01, 2020, Elsevier Ireland Ltd. DOI: 10.1016/j.cosrev.2020.100306.
16. 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.
17. 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.
18. 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.
19. C. C. Aggarwal and S. Sathe, *Outlier ensembles: An introduction*. 2017. DOI: 10.1007/978-3-319-54765-7.
20. S. Sharma and R. Jain, "Outlier detection in agriculture domain: Application and techniques," in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2018, pp. 283–296. DOI: 10.1007/978-981-10-6620-7_28.
21. IMS Engineering College, Institute of Electrical and Electronics Engineers Uttar Pradesh Section, International Conference on Advances in Computer Engineering and Applications 2 2015.03.19-20 Ghaziabad, and ICACEA 2 2015.03.19-20 Ghaziabad, International Conference on Advances in Computer Engineering and Applications (ICACEA), 2015 19-20 March 2015, Ghaziabad, India; conference proceedings.

22. "Mushroom - UCI Machine Learning Repository." Accessed: Jun. 20, 2025. [Online]. Available: <https://archive.ics.uci.edu/dataset/73/mushroom>
23. "Breast Cancer - UCI Machine Learning Repository." Accessed: Jun. 20, 2025. [Online]. Available: <https://archive.ics.uci.edu/dataset/14/breast+cancer>
24. 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>
25. Q. Xianting and W. Pan, "A density-based clustering algorithm for high-dimensional data with feature selection," 2016, DOI: 10.1109/ICIICIL.2016.83.