Implementation and Evaluation of an Adaptive Density-Based Outlier Detection Algorithm Using Mutual Nearest Neighbors

Type of this project: Implementation

Group No.3

Information of each member of your group：

Name	Student ID
Xu Dongliu	21208710
Xue Shaohua	21186118
Shi Yichen	21265914
Sun Dekai	21240639
Yun Hanxu	21286712

Declaration Statement：

This project, including all associated research, implementation, and documentation, is conducted exclusively within the scope and requirements of the course [Data Mining and Knowledge Discovery]. It is expressly understood that this work is not part of any other academic course, thesis, research project, or external endeavor. The content developed for this project will not be submitted, in whole or in part, for credit or use in any other context without prior formal authorization from the instructor of this course.

Project Description：

Background

Outlier detection plays a dual role in data analysis: it optimizes the performance of downstream tasks through "data cleansing" and facilitates early warning of risks in scenarios such as financial fraud, network intrusion, and medical diagnosis by identifying "rare valuable events." Among various techniques, proximity-based methods like kNN and LOF have garnered significant attention due to their independence from data distribution assumptions and high interpretability.

Nevertheless, these mainstream methods face two core challenges. Firstly, rigid parameter dependence. They typically rely on predefined static parameters (e.g., the k value in kNN, the neighborhood radius in LOF), failing to adapt to dynamic variations in intrinsic data densities. For instance, kNN's use of a uniform k value for all points often leads to misjudgment in data with complex density structures, while LOF still suffers from biased local density estimation and irregular decision boundaries near regions with significant density variations due to its static neighborhood definition. Secondly, performance limitations. Many methods struggle with high-dimensional and large-scale data. Some algorithms (e.g., KPCA, deep learning-based MO-GAAL) are prone to out-of-memory errors or timeouts due to high computational complexity, and generally exhibit insufficient generalization capability to unknown data.

Consequently, there is a pressing need within the research community for novel unsupervised outlier detection methods that can self-adapt to local density variations, reduce reliance on parameters, and simultaneously balance efficiency with generalization power.

Problem Addressed

This project focuses on the implementation and evaluation of ADOD (Adaptive Density Outlier Detection), a novel unsupervised algorithm designed to handle datasets with varying densities. The core problem tackled by ADOD is the inability of conventional proximity-based methods to adjust neighborhood boundaries dynamically, leading to inaccurate density estimates and outlier scores in heterogeneous data environments. ADOD introduces two key innovations:

Adaptive Local Scale Estimation: Using perplexity as a smoothing mechanism to determine the local scale σi for each point, enabling neighborhood boundaries to reflect local density variations.

Density Consistency Scoring: Leveraging a mutual neighbor graph to estimate local density and incorporating density differences between points and their neighbors to compute outlier scores.

By avoiding fixed parameters and incorporating adaptive mechanisms, ADOD aims to improve detection accuracy across diverse data distributions while maintaining scalability and interpretability.

Datasets

The evaluation will utilize the same datasets as in the original ADOD study to ensure a direct and reproducible comparison. These include:

Synthetic Dataset: The ThreeBlob Outlier dataset, consisting of 500 points generated from three Gaussian blobs with different standard deviations (σ=[0.6,1.2,0.3]) and 15% uniformly distributed outliers. This dataset allows for controlled analysis of density-varying regions.

Real-World Datasets: A total of 32 real datasets sourced from ODDS and ADBench repositories, spanning domains such as healthcare (e.g., hepatitis, arrhythmia), finance (e.g., credit fraud), astronomy (e.g., satellite), and network security (e.g., http, smtp). These datasets vary in size (80 to 567,498 samples), dimensionality (3 to 500 features), and outlier proportion (0.03% to 34.90%). Preprocessing steps include deduplication and standardization to ensure consistency.

The use of both synthetic and real datasets enables a comprehensive assessment of ADOD's performance under controlled and practical conditions, covering a wide range of data characteristics and outlier types.

Methods and Models to be Implemented

We will implement the Adaptive Density Outlier Detection (ADOD) algorithm and compare it against the model called Local Outlier Factor (LOF) as the baseline model.

LOF: A classical density-based algorithm that compares the density of a point with its neighbors. Its basic method is to compare the local density of a point with its neighbors. If the point's density is significantly lower than its neighbors, then it is considered as an outlier. However, LOF is based a fixed number of neighbors k, resulting in a possible neglect of outliers in sparse areas where the neighbors also have low local density, or confusing itself in boundary regions between clusters with different densities.

ADOD: ADOD is a newly proposed unsupervised anomaly detection method designed to deal with instability of traditional methods in a circumstance of significant density variation. Its main characteristic is introducing an adaptive neighborhood boundary for each data point by employing perplexity to dynamically estimate the local scale and constructing "mutual neighbor graph" to distinguish outliers and boundary points.

Both ADOD and LOF belong to the proximity-based method to detect outliers. By selecting the baseline method in the same category of algorithm, it can be ensured that the comparison is meaningful and highlights ADOD's strength.

Evaluation Metrics

To effectively and intuitively demonstrate the performance across different models, we will adopt the evaluation metrics same as the metrics in the original ADOD study:

ROC (Receiver Operating Characteristic curve):
It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across different thresholds. A higher ROC score indicates a better ability to distinguish positive (outlier) and negative (inlier) instances.

Precision@Rank N (P@N):
It measures the proportion of true outliers among the top N (the actual number of outliers in the dataset) instances ranked by the model's outlier scores. A higher P@N score indicates a better ability to detect real outliers.

Average Precision (AP):
It calculates the average precision across all recall levels. So a higher AP means both a better detection performance and a better ranking quality.

These metrics, which compare the ground truth with the predicted scores, indicate better performance with higher values.

Expected Outcomes

By comparing performance with LOF, we need to validate that ADOD indeed achieves a better performance on various datasets.

In the whole experiment and implementation process, some practical metrics like robustness to parameters, run time and adaptation to different datasets will be observed and summarized.

As is shown in the paper of the study of ADOD, it performs poorer than the baseline model in some specific datasets, so the reason for such situation will be studied and indicated. Possible improved method may also be proposed.

The final report will include detailed tables, graphs, and statistical analysis to support conclusions.

Implementation Schedule (tentative)
After our group discussion, the whole plan of the project is estimated to be scheduled and executed in four phases in four phases:

Phase 1: Paper Review and Early-stage preparation (Week 1, before 26th Sep)

- Study the ADOD paper in detail.
- Review LOF and other relevant algorithms.
- Evaluate and select proper datasets for the implementation.
- Complete project proposal.
- Explore the PyOD library for implementation support.

Phase 2: Data Preparation (Week 2, before 5th Oct)

- Download selected datasets from ODDS and ADBench.
- Preprocess data (unify to numpy, normalize features, and generate new feature columns.).
- Statistical Analysis across datasets. (number of samples, dimensionality, outlier ratio)
- Visualization. (histograms, scatter plots, and boxplots)

Phase 3: Model Implementation and Experiments(Week 3–4, before 20th Oct)

- Implement ADOD following the provided source code.
- Configure and run LOF using PyOD.
- Run each algorithm on selected datasets and collect metrics for comparison.

Phase 4: Analysis and Reporting (Week 5, before 27th Oct)

- Compare ADOD with baselines on all metrics.
- Visualize results (decision boundaries, score distributions).
- Write the final report summarizing findings and insights.

Moreover, if time permitted and our group has more interest to explore further on this project, we might conduct some more experiments like some ablation experiments, use more datasets (or even collect datasets by ourselves), or compare the performance with more models

List of Reference Papers:

[1] L. Qian, J. Qian, X. Sun, W. Guo, and C. Böhm, “ADOD: Adaptive Density Outlier Detection,” 2021 IEEE International Conference on Data Mining (ICDM), pp. 400–409, Dec.

2024, doi: https://doi.org/10.1109/icdm59182.2024.00047.

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF,” Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD ’00, 2000, doi: https://doi.org/10.1145/342009.335388.

[3] Y. Zhao, Zain Nasrullah, and Z. Li, “PyOD: A Python Toolbox for Scalable Outlier Detection,” Journal of Machine Learning Research, vol. 20, no. 96, pp. 1–7, 2019, Available: https://www.jmlr.org/papers/v20/19-011.html

[4] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, “ADBench: Anomaly Detection Benchmark,” Advances in Neural Information Processing Systems, vol. 35, pp. 32142–32159, Dec. 2022, Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/cf93972b116ca5268827d575f2cc226b-Abstract-Datasets_and_Benchmarks.html