香港浸會大學 HONG KONG BAPTIST UNIVERSITY | 研發辦公室 Research Office | 北师香港浸会大学 BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY | 研究拓展及知识转移处 RESEARCH DEVELOPMENT AND KNOWLEDGE TRANSFER OFFICE

GUANGDONG AND HONG KONG UNIVERSITIES

"1+1+1" Joint Research Collaboration Scheme

粵港高校「1+1+1」聯合資助計劃

A Statistical Clustering-Based Method for Outlier Detection

Wei Dai, Qing-Guo Wang, Wentao Fan
Beijing Normal-Hong Kong Baptist University

1. INTRODUCTION & PROBLEM DEFINITION

Background: Outlier Detection is a critical task in data mining, with broad applications in cybersecurity, anti-money laundering, and healthcare.

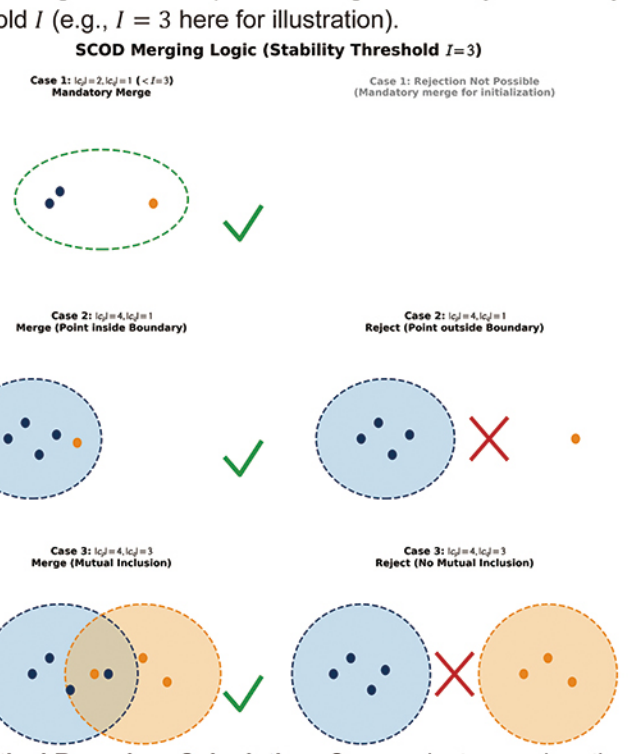
Research Status:

- **Parameter-free approaches** (e.g., COPOD, ECOD) often sacrifice detection accuracy.
- **Cluster-based methods** severely rely on the subjective, dataset-specific selection of hyperparameters (e.g., predefining the number of clusters).

Objective: To eliminate the need for predefining cluster numbers by proposing a statistically-guided bottom-up clustering approach that relies on stable internal statistics.

2. SCOD METHOD

SCOD: The SCOD algorithm utilizes a statistically-guided bottom-up clustering approach. Initially, every data point is treated as an independent cluster. Crucially, the merging judgment is only triggered when two clusters are mutually nearest neighbors. The process is governed by a stability threshold I (e.g., $I = 3$ here for illustration).



Statistical Boundary Calculation: Once a cluster reaches the stability threshold ($|c| \geq I$), a statistical boundary is dynamically calculated based on its internal statistics.

- **Case 1: Initialization Stage ($|c_p| < I, |c_q| < I$)**
  Both clusters lack sufficient samples to form boundaries. They merge unconditionally to assemble a stable core.
- **Case 2: Boundary Absorption ($|c_p| \geq I, |c_q| < I$)**
  A stable cluster c_p evaluates c_q. c_q is absorbed only if it falls strictly within c_q's boundary.
- **Case 3: Mutual Agreement ($|c_p| \geq I, |c_p| \geq I$)**
  Both clusters have boundaries. They merge only if each cluster's boundary contains at least one point from the other cluster.

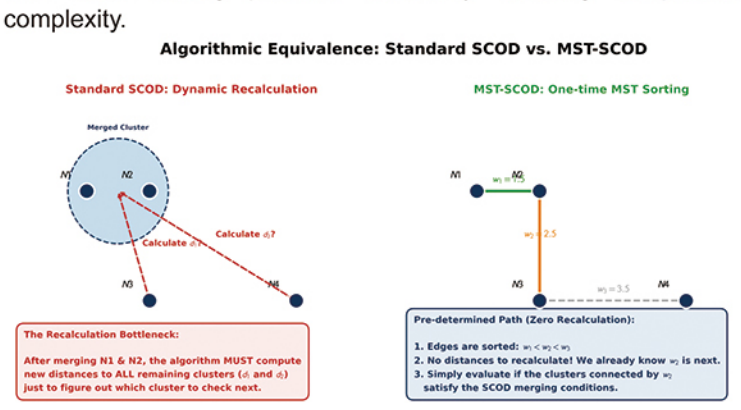
Termination & Outlier Scoring: This merging process repeats iteratively until no further clusters can be merged. Finally, similar to other clustering-based methods, an outlier score is computed for each data point based on the final structures to identify outliers.

3. MST-SCOD METHOD

Relationship with SCOD: While SCOD defines the statistical criteria for merging, MST-SCOD is its mathematically equivalent, highly optimized execution framework.

The Computational Bottleneck: In the standard SCOD approach, after merging the closest pair, the algorithm must dynamically recalculate the distances between the newly formed cluster and all remaining clusters to find the next nearest neighbor. This iterative recalculation consumes massive computational power.

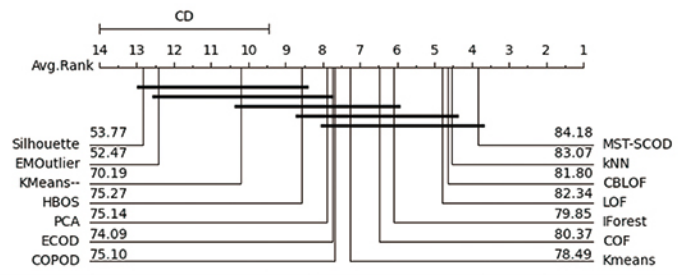
The MST-SCOD Strategy: We mathematically prove that pre-constructing a Minimum Spanning Tree (MST) and sequentially evaluating its edges in ascending order of weights to determine if the two clusters connected by the edge satisfy the merging conditions yields strictly identical clustering results to the original algorithm. This transformation replaces dynamic distance recalculations with a one-time static sorting process, drastically reducing computational complexity.



4. EXPERIMENTAL RESULTS

Superior Performance: The CD diagram shows MST-SCOD achieves the best overall average rank across 20 datasets. The lack of shared horizontal lines statistically confirms it significantly outperforms baselines like PCA, ECOD, and COPOD.

Computational Efficiency: The MST-SCOD framework achieves orders-of-magnitude acceleration in execution time, particularly on large-scale datasets, completely mitigating the traditional $O(N^3)$ bottleneck of dynamic hierarchical clustering.



5. CONCLUSION & FUTURE WORK

Conclusion: SCOD successfully bridges the gap between parameter-free simplicity and cluster-based accuracy. The introduction of the MST-SCOD framework makes this statistically-guided approach scalable and highly efficient.

Future Work: We aim to extend the dynamic statistical boundary calculation to high-dimensional streaming data and further optimize the MST update mechanisms.