Universiteit
Leiden
The Netherlands

Exploring graph-based clustering and outlier detection algorithms

Li, J.

Citation
Li, J. (2025, November 12). *Exploring graph-based clustering and outlier detection algorithms*. Retrieved from https://hdl.handle.net/1887/4282945

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4282945

Note: To cite this publication please use the final published version (if applicable).

Chapter 7

Conclusion and Outlook

Throughout this thesis, we have addressed seven research questions regarding graph-based clustering and outlier detection. The current chapter summarizes the key findings obtained from our methodologies and results. Furthermore, we delve into the limitations of our approaches and propose potential solutions to overcome them. Finally, we highlight several avenues for future research.

7.1 Conclusions

This research has systematically explored the efficacy of graph-based clustering and outlier detection methods, addressing key challenges in handling complex, high-dimensional, and irregularly distributed datasets. The findings provide actionable insights for both theoretical advancements and practical applications, particularly in domains like medical data analysis and image segmentation. Below, we summarize the contributions in relation to the posed research questions:

RQ1: *Which of the proposed MST-based methods is most suitable for specific applications, such as medical data analysis and image segmentation?*

The proposed scaled MST-based clustering and spectral analysis outlier detection (SAOD) frameworks demonstrate superior adaptability for medical data analysis and image segmentation tasks. For medical data, the scaled MST clustering addresses density heterogeneity by leveraging distance scaling to identify irregular-shaped clusters (e.g., tumor regions in MRI/CT scans) without requiring prior knowledge of cluster densities. The SAOD framework, which analyzes eigenvalue distributions of k-nearest neighbor (kNN) graphs, effectively isolates outliers (e.g., rare genetic mutations or anomalous histopathological patterns) in high-dimensional medical datasets by exploiting spectral properties that deviate from normal data. In image segmentation, the approximate MST-based clustering with nearest-neighbor sampling reduces computational overhead while preserving structural integrity, enabling efficient segmentation of medical images.

RQ2: *How does scaling the Minimum Spanning Tree (MST) improve cluster identification in data sets with varying density distributions?*

Scaling the MST enhances cluster identification in datasets with varying densities by transforming Euclidean distances into density-aware metrics. The scaled MST clustering method addresses the limitation of traditional MST-based algo-

rithms, which tend to merge high-density clusters with sparse regions due to rigid distance thresholds. By scaling distances inversely proportional to local density (e.g., using k-nearest neighbor distances), the scaled MST prioritizes edges that span inter-cluster boundaries rather than intra-cluster variations. This ensures that the longest edges in the scaled MST correspond to natural separations between clusters of differing densities. Experimental results on synthetic and real-world datasets (e.g., overlapping Gaussian clusters or non-convex medical imaging data) confirm that scaled MST-based clustering achieves higher adjusted Rand index (ARI) and normalized mutual information (NMI) scores compared to baseline methods, particularly for datasets with skewed density distributions.

RQ3: *Can merging MST construction and inconsistent edge detection reduce computational complexity while maintaining or improving clustering accuracy?*

The proposed unified MST construction and inconsistent edge detection framework reduces computational complexity while maintaining or improving clustering accuracy. Traditional MST-based clustering involves two sequential steps: (1) constructing the MST and (2) pruning inconsistent edges (e.g., using edge length thresholds or density-based criteria). However, this separation can lead to redundant computations and suboptimal pruning decisions. By integrating these steps into a single-pass algorithm, the method dynamically adjusts edge weights during MST construction (e.g., by incorporating local density estimates or spectral properties) and prunes edges in real-time. This reduces the time complexity from $O(n^2 \log n)$ (for traditional MST construction) to $O(n \log n)$ for large datasets, while empirical evaluations on high-dimensional medical datasets (e.g., gene expression profiles) show no significant degradation in clustering purity or outlier detection recall.

RQ4: *How does sampling affect the performance of approximate MST-based clustering for large, high-dimensional data sets?*

Sampling-based approximate MST construction significantly improves scalability for large, high-dimensional datasets without sacrificing clustering performance. The two-phase approximate MST algorithm first samples a representative subset (e.g., using k-nearest neighbors) to construct an initial MST, then iteratively inserts the remaining points based on their nearest neighbors in the MST. This approach reduces the number of edge evaluations from $O(n^2)$ to

$O(\frac{n^2}{4}+\frac{n}{2}+(n+1)\log n)$, where n is the data size. The key insight is that sampling preserves the global topological structure of the MST, ensuring that inter-cluster edges (critical for clustering) are retained even with aggressive sampling rates .

RQ5: *What is the impact of using spectral analysis of k-nearest neighbor graphs on outlier detection accuracy?*

The SAOD framework leverages spectral properties of kNN graphs to improve outlier detection accuracy in high-dimensional data. By analyzing the Laplacian eigenvalue distribution of kNN graphs, the method exploits the observation that outliers exhibit atypical spectral signatures (e.g., lower algebraic connectivity or outlier eigenvalues). The multivariate Gaussian kernel density estimation of eigenvalue distributions enables robust outlier scoring without requiring prior knowledge of outlier proportions. On medical datasets (e.g., rare disease detection in genomic data), SAOD achieves higher F1-scores than state-of-the-art methods by distinguishing outliers based on structural deviations in the kNN graph rather than isolated density estimates. The framework' s effectiveness stems from its ability to capture both local and global data relationships, making it resilient to noise and density variations.

RQ6: *Can an adaptive distance measure based on mini-MSTs improve outlier detection in data sets with varying densities and shapes?*

The adaptive mini-MST outlier detection (MMOD) method improves outlier detection in datasets with varying densities and shapes by dynamically scaling distances using local mini-MSTs. Unlike global distance metrics, which fail in irregularly shaped clusters, MMOD constructs a mini-MST for each data point using its k nearest neighbors, then scales Euclidean distances based on the average edge length of the mini-MST. This ensures that outliers (e.g., rare medical anomalies or non-conforming image patches) are identified based on local connectivity patterns rather than absolute distance thresholds. The method' s parameter-free nature (no need to specify outlier proportions) and adaptability to non-convex data make it particularly suitable for high-stakes medical applications.

RQ7: *How can a scaled Minimum Spanning Tree (MST) data structure improve the accuracy and robustness of cluster-based outlier detection in high-dimensional medical datasets?*

The integration of a scaled MST data structure and a noise-aware medoid selection method significantly enhances cluster-based outlier detection in medical datasets by adaptive clustering, noise mitigation and generalizability. The scaled MST dynamically adjusts edge-cutting thresholds to partition data into clusters of varying densities, improving outlier identification in overlapping regions. The noise-aware medoid selection prioritizes cluster centers with high local density and low noise influence. Experimental results on 12 real-world datasets (including 5 medical corpora) demonstrate robust performance across heterogeneous data distributions, validating the method ' s applicability beyond medical domains (e.g., fraud detection in financial transactions).

7.2 Discussion

This thesis provides a comprehensive overview of the state-of-the-art in graph-based clustering and outlier detection methods. The four chapters discussed in this thesis demonstrate the versatility and effectiveness of these algorithms in various real-world applications, particularly in the medical field. While our initial explorations of clustering and outlier detection have achieved their objectives, there are still some limitations.

First, methodological assumptions introduce vulnerabilities: the reliance on k-nearest neighbor (kNN) parameters for density scaling (e.g., in scaled MST clustering) requires domain-specific tuning, limiting applicability in heterogeneous datasets (e.g., mixed tumor types in medical imaging). Additionally, the noise-aware medoid selection ' s sensitivity to noise influence thresholds may degrade performance in noisy environments (e.g., real-time ICU monitoring data). Second, computational scalability remains a challenge: the $O(n \log n)$ time complexity of approximate MST construction, while efficient for moderate-sized datasets, may prove infeasible for petabyte-scale medical repositories (e.g., whole-genome sequencing), where streaming or distributed algorithms (e.g., BIRCH, MapReduce-based MST) could offer superior scalability. Third, empirical validation is biased toward medical applications, with limited evaluation on non-medical high-dimensional data (e.g., cybersecurity logs, astronomy datasets), raising questions about cross-domain applicability. Furthermore, the absence of adversarial robustness testing (e.g., noise injection in medical images) exposes potential vulnera-

bilities in critical systems (e.g., automated cancer diagnosis). Fourth, theoretical guarantees are nascent: while the spectral analysis of kNN graphs (SAOD) improves outlier detection accuracy, it lacks probabilistic bounds on recall, unlike methods such as Isolation Forest. Finally, interpretability gaps persist—the proposed frameworks do not provide feature-level explanations for outliers, a critical requirement in high-stakes medical applications (e.g., identifying specific genetic mutations driving rare disease phenotypes).

7.3 Future Research Directions

While the proposed MST-based frameworks advance outlier detection and clustering in medical and high-dimensional data, their real-world impact hinges on addressing scalability bottlenecks, automation gaps, and cross-domain validation challenges.

First, to enable real-time processing of large-scale datasets (e.g., petabyte-scale genomic repositories, real-time satellite imagery for disaster response), future work must prioritize parallelized MST construction. By leveraging GPU acceleration (e.g., CUDA-optimized Borvka' s algorithm) and distributed computing frameworks (e.g., MapReduce-based Kruskal' s algorithm with edge-partitioning), the $O(n \log n)$ time complexity of approximate MST methods could be reduced to sub-linear or near-constant time for streaming data, enabling deployment in high-throughput medical pipelines (e.g., automated cancer screening from whole-slide histopathology images) and industrial IoT systems (e.g., anomaly detection in sensor networks).

Second, to eliminate manual parameter tuning (e.g., kNN density thresholds in scaled MST or mini-MST outlier detection), deep learning integration offers a promising avenue. By embedding adaptive distance measures (e.g., locally scaled Euclidean metrics) into neural network architectures, models could learn optimal scaling factors via meta-learning (e.g., gradient-based hyperparameter optimization) or self-supervised contrastive learning (e.g., aligning local density estimates with ground-truth cluster labels), thereby automating parameter selection for multi-modal medical data (e.g., MRI-PET fusion) or industrial datasets with non-stationary densities (e.g., manufacturing defect detection under varying lighting conditions).

Third, to validate generality, the frameworks must undergo rigorous domain-specific testing in genomics (e.g., detecting rare genetic variants in single-cell RNA-seq data with extreme density heterogeneity) and remote sensing (e.g., segmenting agricultural fields from satellite imagery with overlapping crop types). Here, deep clustering methods (e.g., autoencoder-enhanced MST with spectral regularization) and deep outlier detection (e.g., variational autoencoders for anomaly scoring in kNN graph Laplacians) could bridge traditional clustering with representation learning, improving robustness to out-of-distribution (OOD) samples (e.g., novel genetic mutations or unseen land-cover types). Additionally, applying traditional MST-based clustering to medical image data (e.g., tumor segmentation in 3D CT scans) could benefit from hybrid deep-MST approaches (e.g., using U-Net-generated feature embeddings for MST edge construction), balancing interpretability (from MST topology) with accuracy (from deep feature learning).

By pursuing these directions, the MST-based frameworks could transition from proof-of-concept prototypes to scalable, automated, and domain-agnostic tools for critical applications in precision medicine, climate monitoring, and industrial quality control.