Exploring graph-based clustering and outlier detection algorithms

Li, J.

Summary

In the era of big data, extracting meaningful insights from complex, interconnected datasets remains a critical challenge across application domains. In this thesis we study methods for clustering and outlier detection to support data analysis. Traditional clustering and outlier detection methods, such as k-means, DBSCAN, and statistical anomaly scoring, rely heavily on Euclidean distances, density assumptions, or global thresholds, often struggling to capture the intricate relationships and contextual anomalies present in high-dimensional, noisy, or dynamically evolving data. Graph-based approaches, such as minimum spanning tree (MST) based methods and spectral models, offer a powerful alternative by leveraging structural properties to uncover hidden patterns and deviations. Our research demonstrates, through a systematic approach, the superiority of graph-based techniques over some other methods, emphasizing their adaptability, efficiency, and real-world applicability without relying on deep learning or neural networks.

MST-based clustering excels in detecting irregularly shaped clusters and hierarchical structures that are missed by traditional methods. By constructing graphs where nodes represent data points and edges encode similarities (e.g., via kNN or Gaussian kernels), algorithms like spectral clustering partitionate the nodes by analyzing the Laplacian graph, a matrix that captures global connectivity patterns. Unlike k-means, which assumes spherical clusters, spectral clustering adapts to complex geometries, making it ideal for tasks like image segmentation, where boundaries are non-linear. These methods outperform density-based approaches such as DBSCAN in sparse or unevenly distributed data, where traditional methods fragment clusters or merge unrelated groups.

Outlier detection benefits equally from graph structures. Traditional statis-

tical methods (e.g., Z-score, IQR) flag anomalies based on global deviations from the mean, often overlooking contextual outliers that are normal within their local environment but abnormal in a global context. The Spectral Analysis Outlier Detection (SAOD) method introduces a novel graph-based approach to identifying outliers in high-dimensional spaces. By constructing $k$nearest neighbor graphs and analyzing the eigenvalue distributions of their Laplacian matrices, SAOD effectively distinguishes outliers from inliers based on their distinct spectral patterns. The method employs multivariate Gaussian kernel density estimation to quantify *outlierness*. Experiments with real-world data show that SAOD is particularly valuable for applications such as medical diagnosis and cybersecurity, where high-dimensional data is prevalent.

Many outlier detection methods fail to handle datasets with varying densities or irregular shapes and require prior knowledge of outlier proportions, limiting their practicality in medical applications. Adaptive Mini-MST Outlier Detection (MMOD) addresses the limitation of density-dependent approaches by developing a novel scaled distance measure and constructing local mini-MSTs, enabling effective outlier identification without prior knowledge of the outlier distribution. This novel distance metric of the algorithm dynamically scales according to local data density, enabling precise outlier identification in complex datasets with varying cluster shapes and densities. Validation on electronic health records demonstrates MMOD's superiority.

Cluster-based outlier detection often suffers from noise sensitivity and arbitrary medoid selection, leading to false positives in complex datasets. Scaled MST with Medoid Selection (MS2OD) combines the robustness of minimum spanning trees with an innovative noise-aware medoid selection process. The method iteratively partitions data through longest-edge cuts in a scaled MST, while its advanced medoid selection algorithm accounts for noise contamination in cluster cores.

The research presented in this thesis establishes graph-based clustering and outlier detection as indispensable tools for modern data analysis, offering unparalleled flexibility in handling complex, interconnected datasets. By advancing theoretical foundations, optimizing computational efficiency, and demonstrating real-world impact, the work bridges the gap between academic innovation and practical deployment. Future directions include extending these methods to streaming data

for real-time analytics and developing interpretable frameworks to enhance trust in graph-based decisions, ensuring their relevance in an increasingly data-driven world.

Summary