Exploring graph-based clustering and outlier detection algorithms

Li, J.

Chapter 1

Introduction

This thesis is an exploration of graph-based clustering and outlier detection machine learning algorithms. The research described in this thesis aims to find efficient and effective graph-based algorithms to process data mining problems.

1.1 Research Background

Data mining has become an essential tool for uncovering valuable insights and patterns within large and complex data sets. In machine learning, organizations in various industries generate and accumulate massive amounts of data daily, which require efficient and effective methods to analyze and interpret this information [57]. Clustering is one of the most fundamental and widely used techniques in data mining [136].

Clustering, also known as unsupervised learning in machine learning, involves grouping similar data points into clusters or groups without any prior knowledge of the class labels. The primary objective of clustering is to identify hidden patterns or structures within the data that may not be immediately apparent through simple observation or traditional analytical methods [133].

The importance of clustering in data mining lies in its ability to simplify complex data sets by reducing their dimensionality and highlighting meaningful relationships among data points. By grouping similar items together, clustering can help researchers and practitioners gain a deeper understanding of the under-

lying distribution of the data, identify outliers or anomalies, and uncover hidden subgroups within the population.

Several clustering algorithms have been developed over the years, each with its own strengths, weaknesses, and applicability to specific types of data. Traditional clustering algorithms such as k-means [80], hierarchical clustering [111], and DBSCAN [43] are widely used due to their simplicity, efficiency, and effectiveness in many real-world scenarios. However, with the advent of big data and the increasing complexity of modern data sets, researchers have been exploring new and advanced clustering techniques, including density-based clustering, grid-based clustering [150], spectral clustering [23], and ensemble clustering [53], among others.

Despite the significant progress made in clustering research, there are still numerous challenges and open questions that need to be addressed. One of the primary challenges lies in the inherent difficulty of defining what constitutes a "good" cluster, as this can vary significantly depending on the nature of the data and the objectives of the study. Furthermore, many real-world data sets are characterized by high dimensionality, noise, and non-linearity, which can make clustering a computationally intensive and challenging task.

In light of these challenges, there is a growing need for research that can develop more robust, scalable, and effective clustering algorithms capable of handling the complexities of modern data sets. This research is crucial not only for advancing the field of data mining, but also for enabling organizations to harness the full potential of their data to drive innovation, make informed decisions, and achieve their goals.

Outlier detection, also known as anomaly detection [45] or novelty detection [142], has emerged as a crucial tool for researchers across a broad spectrum of fields. From banking and marketing to social sciences and medical diagnosis [137], the identification of subjects that exhibit different and sometimes peculiar behaviors is paramount. Outliers are defined as observations that deviate markedly from the majority of the data, either due to mechanical faults, changes in system behavior, fraudulent activities, human or instrument errors, or simply through natural deviations in populations.

The process of detecting outliers within a data set is often challenging and requires special attention as these observations may indicate serious issues or critical

situations. For instance, they could represent intrusions in a network [84], cases of financial fraud [62] or money laundering [73], malicious individuals in online social networks [35], or pathologies in medical images [75]. The widespread applicability of outlier detection has driven the development of numerous algorithms designed to detect anomalies within a given data set.

However, despite the increasing popularity of these techniques, most of them are restricted to a single type of data, mainly continuous data. When researchers deal with data sets that consist of both discrete and continuous variables (mixed-type data), outlier detection presents unprecedented challenges. This has motivated the development of novel methods that can effectively identify outliers in mixed-type data settings, while reducing the required user interaction and providing general guidelines for selecting suitable hyperparameter values.

The current state of the art in outlier detection encompasses a diverse range of techniques [141], including statistical methods, clustering algorithms, and machine learning approaches [89]. Statistical methods, such as Histogram-based Outlier Score (HBOS) [54], rely on the assumption that the data dimensions are independent and perform interval (bin) partitioning to estimate the density of each dimension. Clustering algorithms, on the other hand, are often used to identify clusters of data points and treat outliers and noise as unvaluable data, although they are not specifically designed for outlier detection.

More advanced methods, like Local Outlier Factor (LOF) [65] and Isolation Forest (iForest) [97], provide more sophisticated ways to detect outliers. LOF measures the local density deviation of a point relative to its neighbors, while iForest utilizes an ensemble of isolation trees to isolate outliers in a linear time complexity. These methods have shown promising results in various applications, but they still face challenges when dealing with mixed-type data.

The research in outlier detection is not limited to traditional algorithms but also extends to emerging techniques that leverage the power of deep learning and generative models. For instance, the use of Generative Adversarial Networks (GANs) [128] and Diffusion Models [39] has been explored to generate synthetic outliers for training purposes, which can help alleviate the issue of scarce labeled outlier data.

In summary, outlier detection serves as a vital tool for researchers in various domains, and the development of novel methods that can effectively handle mixed-

type data is an active area of research. The challenges associated with detecting outliers in complex data sets motivate the exploration of new techniques that can improve the accuracy and efficiency of outlier detection algorithms.

1.2 Machine Learning Algorithms

In this section, several machine learning algorithms related to our research is introduced, including minimum spanning tree algorithms, spectral clustering algorithms, and clustering-based outlier detection algorithms.

1.2.1 Minimum Spanning Tree-based clustering algorithms

Minimum Spanning Tree (MST)-based clustering algorithms are a powerful tool in data mining and machine learning, particularly suited for detecting clusters with irregular boundaries [50]. These algorithms leverage the concept of minimum spanning trees, a fundamental graph theory concept, to partition data points into meaningful clusters.

In a nutshell, a minimum spanning tree of a connected graph is a subgraph that includes all vertices of the original graph and the minimum possible number of edges (n-1 edges for a graph with n vertices) while maintaining connectivity [26]. For a weighted graph, the minimum spanning tree is the one with the smallest total edge weight. MST-based clustering algorithms exploit this property to identify clusters in a data set.

Two notable MST-based clustering algorithms are proposed in the literature. The first algorithm aims to produce a k-partition of a set of points for a given k. It starts by constructing a minimum spanning tree of the point set and then iteratively removes edges that satisfy a predefined criterion, such as those connecting points that are dissimilar or far apart. This process is repeated until the desired number of clusters (k) is achieved.

The second algorithm, on the other hand, partitions a point set into clusters without requiring a predefined number of clusters (k). Instead, it partitions the points into clusters that maximize the overall standard deviation reduction. This approach allows the algorithm to adaptively determine the optimal number of clusters based on the data's inherent structure.

MST-based clustering algorithms have several advantages over traditional clustering methods like k-means [78]. For instance, they can effectively handle clusters with irregular shapes and varying densities, which can be challenging for algorithms that assume spherical or convex clusters. Additionally, MST-based clustering algorithms can be implemented efficiently using distributed computing frameworks like MapReduce, making them suitable for large-scale data sets.

In summary, Minimum Spanning Tree-based clustering algorithms offer a flexible and powerful approach to cluster analysis, capable of detecting complex cluster structures in diverse data sets. By leveraging the properties of minimum spanning trees, these algorithms can adaptively partition data points into meaningful clusters, making them a valuable tool for data mining and machine learning practitioners.

1.2.2 Spectral clustering algorithms

Spectral clustering is a powerful clustering technique that leverages graph theory and linear algebra concepts to partition data points into groups or clusters [37]. It is particularly useful for data sets with complex, non-convex cluster shapes, which can be challenging for traditional clustering algorithms like k-means.

The basic idea behind spectral clustering is to represent the data set as a graph, where data points are nodes, and edges connect similar nodes [149]. The similarity between nodes is often defined using a similarity matrix, which can be constructed based on the distances between data points. The next step is to compute the graph's Laplacian matrix, a fundamental concept in spectral graph theory.

The Laplacian matrix encapsulates important information about the graph's structure, including the connectivity and relative importance of nodes. Spectral clustering algorithms then use the eigenvalues and eigenvectors of the Laplacian matrix to identify clusters. Specifically, the algorithm focuses on the smallest non-zero eigenvalues and their corresponding eigenvectors, which often contain information about the graph's clusters.

To partition the graph into clusters, spectral clustering algorithms employ a dimensionality reduction technique known as the "spectral embedding." This involves projecting the graph's nodes onto a lower-dimensional space based on the selected eigenvectors. In this embedded space, data points that belong to the

same cluster tend to be closer together, making it easier to identify and separate the clusters.

Finally, spectral clustering algorithms apply a traditional clustering algorithm, such as k-means, to the embedded data points to obtain the final cluster assignments. However, since the data has already been transformed to reveal the underlying cluster structure, k-means (or any other clustering algorithm) is often able to find more accurate and meaningful clusters.

Spectral clustering algorithms have several advantages over traditional clustering methods. They can detect clusters with irregular shapes and varying densities, and they are robust to noise and outliers. Additionally, spectral clustering algorithms can be implemented efficiently for large-scale data sets using optimization techniques and parallel computing frameworks.

In summary, spectral clustering is a sophisticated clustering technique that leverages graph theory and linear algebra to partition data points into meaningful clusters. By transforming the data into a lower-dimensional space that reveals the underlying cluster structure, spectral clustering algorithms can detect complex cluster shapes and densities, making them a valuable tool for data mining and machine learning practitioners.

1.2.3 Clustering-based outlier detection algorithms

There are many kinds of outlier detection methods [68, 58]. Clustering-based outlier detection algorithms are a class of unsupervised learning techniques that leverage the concept of clustering to identify outliers or anomalies within a data set. These algorithms assume that the majority of the data points belong to a well-defined set of clusters, while outliers are those data points that do not fit well into any of these clusters.

The basic idea behind clustering-based outlier detection is to first partition the data into clusters using a clustering algorithm, such as k-means, DBSCAN [43], or hierarchical clustering [175]. Once the clusters are formed, outliers are then identified as data points that are significantly distant from their closest cluster centers or as points that do not belong to any cluster at all.

One of the most straightforward ways to identify outliers using clustering is by measuring the distance from each data point to its nearest cluster center. In this approach, the outliers are those data points that have a high distance

from their nearest cluster center, indicating that they do not belong to any of the identified clusters. The threshold for defining an outlier can be determined based on the distribution of these distances or by using statistical methods.

Another approach is to use clustering algorithms that inherently identify outliers as part of their clustering process. For example, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that can identify outliers as points that are not reachable from any other point in the data set [43]. These points are considered to be noise or outliers, as they do not belong to any of the dense regions of the data set.

Clustering-based outlier detection algorithms have several advantages. Firstly, they are unsupervised, meaning that they do not require labeled data to train a model. This makes them useful in scenarios where labeled data is scarce or unavailable. Secondly, they can detect outliers in high-dimensional data sets, where traditional statistical methods may struggle. Additionally, clustering-based approaches can provide insights into the underlying structure of the data by revealing patterns and groups within the data set.

However, clustering-based outlier detection algorithms also have some limitations. The performance of these algorithms can be sensitive to the choice of clustering parameters, such as the number of clusters or the distance metric used. Additionally, the identification of outliers can be influenced by the shape and distribution of the clusters, which may not always accurately reflect the true nature of the outliers.

In summary, clustering-based outlier detection algorithms are a powerful tool for identifying anomalies and outliers in unsupervised data sets. By leveraging the concept of clustering, these algorithms can provide insights into the underlying structure of the data and identify outliers that do not fit well into the identified clusters. However, careful consideration of the clustering parameters and the interpretation of the results is required to ensure accurate outlier detection.

1.3 Research Questions

The major addressed in this thesis is how to improve the efficiency and accuracy in graph-based clustering and outlier detection algorithms. Therefore, we have formulated seven research questions. That we present here.

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

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

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

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

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

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

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?

1.4 Structure of This Thesis

The structure of this thesis is as follows (Figure 1.1):

Chapter 2: "A Scaled Minimum Spanning Tree-Based Clustering Algorithm and Application on Image Segmentation" aims to analyze the suitability and effectiveness of different MST-based methods across various domains and data sets. To address these questions, we propose two enhanced MST-based clustering methods. The first method involves computing a scaled version of the MST to better distinguish clusters of different densities. The second method optimizes the process by merging MST construction and inconsistent edge detection into a single step, reducing complexity and potentially improving accuracy. Both methods are validated through image segmentation and integration tasks, demonstrating their effectiveness and practicality, particularly for small and high-dimensional data sets.

Chapter 3: "A Novel Approximate MST-Based Clustering Algorithm Based on Sampling" proposes a new method for approximate MST-based clustering that employs sampling to address the challenges of traditional MST-based clustering approaches when dealing with large, high-dimensional data

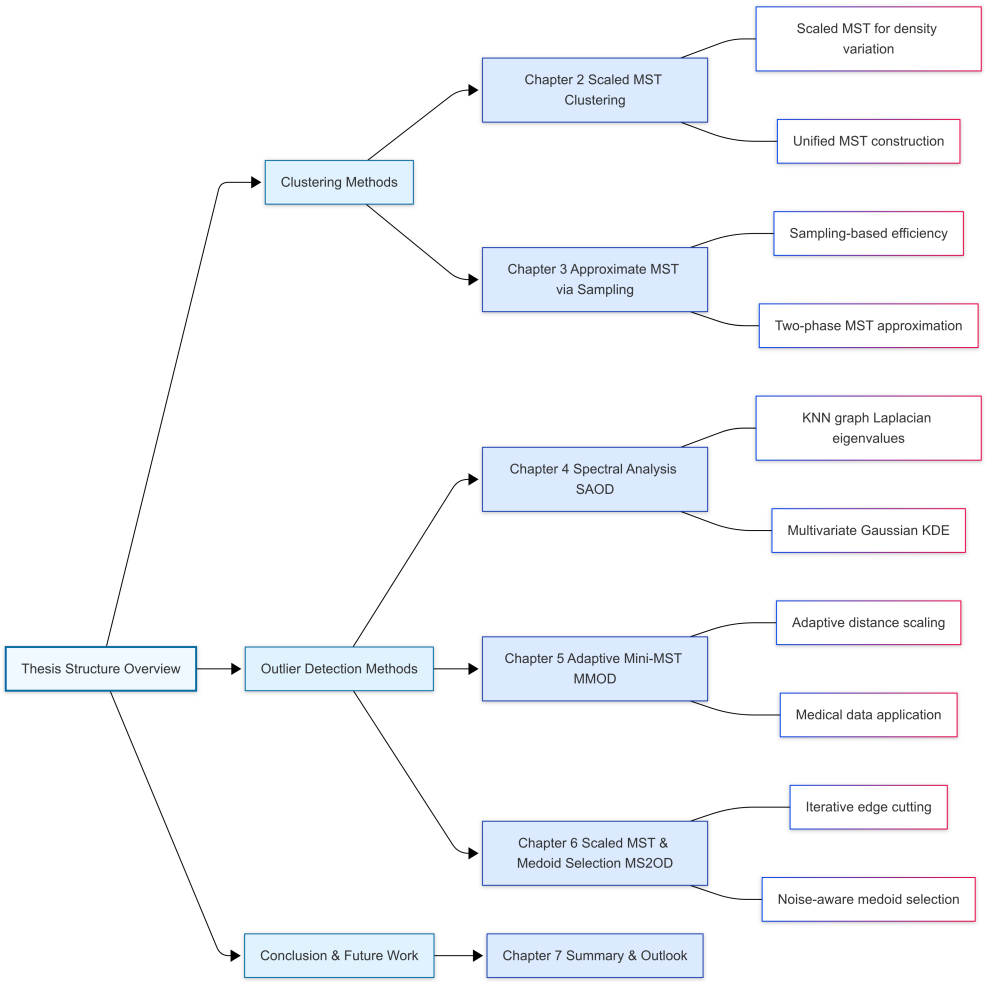

Figure 1.1: Thesis structure.

sets. The method first reduces the data set scale by sampling using nearest neighbors and then constructs an approximate MST using a two-phase algorithm. Experimental results show that the proposed method is effective in clustering both synthetic and real-world data sets, balancing high execution efficiency and accuracy.

Chapter 4: "Outlier Detection in Eigenvalue Spaces Based on Spectral Analysis of KNN Graphs" introduces a new spectral analysis-based outlier detection method called SAOD. The method constructs k-nearest neighbor graphs for each sample and computes the eigenvalues of their Laplacian matrices. Based on the different eigenvalue distributions of outliers and inliers, the method employs multivariate Gaussian kernel density estimation to empirically estimate the eigenvalue distribution of the data set. Experimental results demonstrate that the proposed SAOD method outperforms existing state-of-the-art outlier detection approaches.

Chapter 5: "Outlier Detection Using Iterative Adaptive Mini-Minimum Spanning Tree Generation with Applications on Medical Data" proposes an adaptive mini-minimum spanning tree-based outlier detection method called MMOD, designed to overcome the limitations of existing outlier detection techniques that require prior knowledge of outlier distributions. MMOD employs a novel distance measure that scales the Euclidean distance, enabling it to identify outliers in data sets with varying densities and shapes. Experimental results demonstrate the effectiveness of the proposed MMOD method in comparison to state-of-the-art outlier detection techniques.

Chapter 6: "MS2OD: Outlier Detection Using Minimum Spanning Tree and Medoid Selection" presents an advanced cluster-based outlier detection method that utilizes a scaled MST data structure and a novel medoid selection method. The method constructs a scaled MST, iteratively cuts the longest edge to form clusters, and employs a new medoid selection approach that considers noise effects to enhance the quality of outlier identification. Experimental results demonstrate the wide applicability and superior performance of the proposed MS2OD method compared to existing outlier detection techniques.

Chapter 7: Conclusion and Outlook summarizes our methods and results. As well as addressing their limitations, we propose solutions to overcome them.