Enhanced Precision in Anomaly Detection: An Optimized k-Means Clustering Approach

Huining Huang
Department of Data Science
University of South Australia
Adelaide, South Australia, Australia
huahy057@mymail.unisa.edu.au

November 5, 2023

Abstract

Anomaly detection is an indispensable component in numerous applications ranging from fraud detection to system health monitoring. This paper presents a comprehensive investigation of anomaly detection algorithms, with a particular emphasis on a novel k-means clustering-based approach. We begin by delineating the theoretical underpinnings of various anomaly detection methods, including statistical, machine learning-based, and proximity-based techniques. Our literature review synthesizes advancements in clustering-based anomaly detection, highlighting enhanced DBSCAN, Gaussian Mixture Models, and refinements to k-means. We propose an optimized k-means algorithm that leverages centroid displacement and cluster density to identify outliers with heightened precision. Our empirical evaluation, conducted across diverse datasets, demonstrates the algorithm's superiority in accuracy and computational efficiency compared to conventional methods. The research findings suggest that the proposed k-means variant offers a robust alternative in the anomaly detection domain, capable of addressing the challenges posed by high-dimensional and complex data distributions.

1 Introduction

Anomaly detection is a critical process in data analytics that identifies deviations from the norm within a dataset, which may indicate errors, fraud, or new patterns. Its significance is recognized across various domains, including finance for fraud detection [**?**], industrial sectors for failure prediction [**?**], and cybersecurity for intrusion detection [**?**].

Key methodologies for anomaly detection are:

1. **Statistical Methods:** Utilizing statistical models to characterize normalcy and pinpointing data points that exhibit significant statistical deviations [**?**].

2. **Machine Learning-Based Methods:** Employing learning algorithms to discern anomalies, with supervised methods requiring pre-labeled data, while unsupervised methods such as K-means identify outliers without labels [**?**].

3. **Proximity-Based Methods:** Leveraging the distance between data points to detect anomalies, such as DBSCAN, which finds outliers in regions of low density [**?**].

While machine learning techniques are at the forefront due to their adaptability to complex datasets, their efficacy is contingent on the data's attributes, including dimensionality and distribution. Conversely, statistical methods, which are more transparent, might necessitate predefined assumptions about data distribution, and proximity-based approaches can be computationally intensive for large datasets.