Title: A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data

Authors:

- Markus Goldstein

- Seiichi Uchida

Affiliations:

1. Center for Co-Evolutional Social System Innovation, Kyushu University, Fukuoka, Japan

1. Department of Advanced Information Technology, Kyushu University, Fukuoka, Japan

Abstract:

This research article presents a comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. Anomaly detection involves identifying unexpected items or events in datasets that deviate from the norm. Unlike standard classification tasks, anomaly detection is applied to unlabeled data, considering only the internal structure of the dataset. This study addresses the challenge of unsupervised anomaly detection by evaluating 19 different algorithms on 10 datasets from various application domains. The paper provides the source code and datasets, aiming to establish a well-funded basis for unsupervised anomaly detection research. Additionally, the evaluation highlights the strengths, weaknesses, computational effort, and global/local anomaly detection behavior of the different approaches. The study concludes by providing guidance on algorithm selection for real-world tasks.

Introduction:

The paper discusses the significance of anomaly detection or outlier detection in machine learning. Anomalies refer to instances within datasets that deviate significantly from other members of the sample. The motivation for detecting anomalies has evolved over time, from removing outliers in data cleansing to recognizing interesting events or suspicious data records. Anomalies possess two important characteristics: they differ from the norm in terms of their features and are rare compared to normal instances. Anomaly detection algorithms have diverse applications, including network intrusion detection, fraud detection, data leakage prevention, medical applications, and life sciences. The paper explores various application domains and provides an overview of proposed algorithms in the field.

Anomaly Detection Applications:

The paper highlights several application domains where anomaly detection is employed. These include:

1. Intrusion Detection: Anomaly detection algorithms are utilized to identify potential intrusion attempts and exploits in network traffic and server applications. Both network-based and host-based intrusion detection systems are discussed.

1. Fraud Detection: Anomaly detection is applied to analyze log data and detect system misuse or suspicious events indicating fraud. Financial transactions and credit card payment logs are commonly examined for fraudulent activities.

1. Data Leakage Prevention (DLP): Anomaly detection is utilized for protecting sensitive information by detecting data loss at an early stage. Accesses to databases, file servers, and other information sources are monitored and analyzed for detecting uncommon access patterns.

1. Medical Applications and Life Sciences: Anomaly detection algorithms are employed in patient monitoring to detect critical situations using ECG signals or other body sensors. They are also used for analyzing medical images, such as CT scans, to detect abnormal cells or tumors.

Additionally, the paper mentions specialized applications such as surveillance camera data analysis, anomaly detection in smart buildings, monitoring mobile communication networks, forensic applications for detecting forged documents, and detecting critical states in complex systems.

Categorization of Anomaly Detection:

The paper discusses different setups for anomaly detection and categorizes them into three main types:

1. Supervised Anomaly Detection: This setup involves fully labeled training and test datasets, where an ordinary classifier can be trained and applied. However, this setup is not practical for many applications as anomalies are often unknown or unlabeled.

\[The truncated content ends here. If you need information from the remaining part of the paper, please provide specific queries.\]