Anomaly Detection and Data Labeling: Adaptive Model Scheduling using SVM, K-means, and Isolation Forest

Inuwa Amir Usman

June 26, 2023

Abstract

Data labeling is an important task for anomaly detection, which aims to identify abnormal behavior or events in a dataset. Anomaly detection is crucial in various applications, such as fraud detection, intrusion detection, and equipment monitoring. However, data labeling for anomaly detection can be challenging, as anomalies are often rare and difficult to identify. This paper reviews recent advances in data labeling for anomaly detection, including supervised, semi-supervised, and unsupervised approaches. The paper also discusses the gaps and limitations of current data labeling methods for anomaly detection and proposes a new method. Overall, the paper highlights the importance of data labeling for anomaly detection and the need for further research to improve the efficiency and effectiveness of data labeling methods for this important application.

1 Introduction

The task of data labeling is crucial in many machine learning applications, as it enables the training of accurate and robust models. However, data labeling can be time-consuming, expensive, and error-prone when done manually. To overcome these challenges, researchers have proposed various automated and semi-automated methods for data labeling. This paper review will focus on five recent papers that address different aspects of data labeling.

1.1 Problem Definition

The problem addressed in these papers is how to efficiently and accurately label data for machine learning tasks, including the application of semi-supervised techniques. The conventional manual labeling process can be time-consuming and expensive, particularly for large datasets. Therefore, there is a growing need to develop automated or semi-automated labeling methods that can effectively reduce the labeling effort while ensuring high labeling quality. These methods leverage the availability of unlabeled data in combination with a limited set of labeled instances to improve the labeling process. By harnessing the power of semi-supervised techniques, such as active learning, co-training, or self-training, these approaches aim to optimize the use of both labeled and unlabeled data to achieve more efficient and accurate data labeling for machine learning tasks.

1.2 Gap

While there have been many advances in automated and semi-automated data labeling, there are still some challenges that need to be addressed. For example, how to handle missing or incomplete labels, how to improve the efficiency and accuracy of labeling with limited resources, and how to deal with noisy or low-quality data. Each of the five papers selected for this review addresses one or more of these challenges, proposing innovative solutions to improve the efficiency and effectiveness of data labeling for machine learning.

1.3 Motivation

The increasing demand for labeled data in various domains, such as computer vision, natural language processing, and speech recognition. Labeled data is essential for training and evaluating machine learning models, and the quality of the labels directly affects the model's performance. However, manual data labeling can be a time-consuming and expensive process, especially for large datasets. Therefore, developing automated and semi-automated data labeling methods has become a crucial research area to reduce the labeling effort and improve the labeling quality. Furthermore, the availability of high-quality labeled data is essential for advancing the state-of-the-art in various domains and developing innovative applications that can improve people's lives. Therefore, improving the efficiency and effectiveness of data labeling methods is a crucial research area with broad practical implications for advancing machine learning and other related fields.

1.4 Baseline

The baseline for my project is the paper "Unsupervised deep learning and semi-automatic data labeling in weed Discrimination"[1]. It proposes a semi-automatic data labeling method for weed discrimination using unsupervised deep learning. The method achieves state-of-the-art performance on a large weed dataset by combining unsupervised deep learning and semi-automatic data labeling. My project aims improve this method for data labeling and make a more efficient data labeling solution.

1.5 Proposed Method

My proposed method, titled "Anomaly Detection and Data Labeling: Adaptive Model Scheduling using SVM, K-means, and Isolation Forest," addresses the challenge of efficiently and accurately labeling data for anomaly detection tasks. Conventional manual labeling processes can be time-consuming and expensive, particularly for large datasets. Therefore, we introduce an automated approach that combines adaptive model scheduling with the utilization of Support Vector Machines (SVM), K-means clustering, and Isolation Forest algorithms.

2 Related Work

In recent years, data labeling has become increasingly important for anomaly detection. To achieve accurate labeling efficiently, researchers have proposed several methods. One such method is adaptive model scheduling, which has been used for comprehensive and efficient data labeling[2]. Another method is self-supervised semi-supervised learning, which has been used for data labeling and quality evaluation[3]. Multi-label learning has been used to deal with missing labels via common and label-specific features[4], while adversarial data generation has been used for learning from incomplete labeled data[5]. Unsupervised deep learning and semi-automatic data labeling have been used for weed discrimination[1]. Automatic labeling of data has been used for transfer learning[6]. Reinforcement learning from partially labeled anomaly data has been proposed for deep supervised anomaly detection[7]. These methods have shown promising results in data labeling for anomaly detection, and they provide valuable insights for developing efficient and accurate labeling techniques.

3 Background

3.1 Anomaly Types

Anomaly detection is the task of identifying rare and unusual events or observations in a dataset. Anomalies can be classified into different types depending on their nature and characteristics[8]. Some of the common types of anomalies include:

1. Point anomalies: These are individual data points that are significantly different from the rest of the dataset. For example, in a temperature sensor dataset, a point anomaly could be a single reading that is much higher or lower than the other readings.

2. Contextual anomalies: These are data points that are unusual given their context or background. For example, in a credit card fraud detection system, a transaction made from a location that is inconsistent with the cardholder's past transactions could be a contextual anomaly.

3. Collective anomalies: These are a group of data points that are anomalous when considered together. For example, in a network intrusion detection system, a group of requests made to a server that are different from the usual pattern could be a collective anomaly.

Understanding the different types of anomalies is crucial for designing effective anomaly detection systems. Different types of anomalies may require different approaches and techniques for detection and classification.[9]

3.2 Anomaly Detection Techniques

There are various techniques for detecting anomalies in data. Some of the commonly used techniques include:

- Statistical techniques: These methods rely on statistical measures such as mean, standard deviation, and covariance to detect anomalies. One common approach is to assume that the data follows a normal distribution and detect data points that fall outside a certain number of standard deviations from the mean.
- Machine learning techniques: These methods use algorithms to learn patterns in the data and detect anomalies based on deviations from those patterns. Examples include clustering, decision trees, and support vector machines.
- Deep learning techniques: These methods use neural networks to learn complex patterns in the data and detect anomalies based on deviations from those patterns. Examples include autoencoders and recurrent neural networks.

Each technique has its own strengths and weaknesses, and the choice of technique depends on the characteristics of the data and the specific requirements of the application. In this work, we focus on using a combination of deep learning and semi-supervised learning techniques for efficient data labeling and anomaly detection.[10]

3.3 Time Series Anomaly Detection

Time series data refers to a sequence of data points that are collected over time intervals. Anomaly detection in time series data has attracted much attention due to its wide applications in various fields, including finance, healthcare, and manufacturing.[11]

There are several techniques that can be used for anomaly detection in time series data, including statistical methods, machine learning algorithms, and deep learning models. Statistical methods, such as moving average, exponential smoothing, and ARIMA, are commonly used for anomaly detection in time series data.

Machine learning algorithms, such as decision trees, random forests, and support vector machines (SVMs), have also been applied for anomaly detection in time series data. These algorithms typically require a set of labeled data to train the model and detect anomalies.

Recently, deep learning models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have shown promising results for anomaly detection in time series data. These models can capture complex patterns in the time series data and identify anomalies with high accuracy.[12]

3.4 Uni and Multivariate Time Series Anomaly Detection

Time series data is a common form of data in various fields such as finance, weather forecasting, and industry. Anomaly detection in time series data can be categorized into two main types: univariate and multivariate. Univariate anomaly detection is the process of identifying anomalies in a single time series data, while multivariate anomaly detection aims to detect anomalies in multiple time series data. There are various techniques used for both univariate and multivariate time series anomaly detection, including statistical methods, machine learning algorithms, and deep learning models. Each of these techniques has its advantages and limitations, depending on the characteristics of the time series data and the specific anomaly detection task. In this paper, we focus on exploring the use of machine learning and deep learning models for univariate and multivariate time series anomaly detection.[13]

4 Methodology

4.1 Proposed Method

In this section, we present our proposed method for anomaly detection and data labeling, which combines adaptive model scheduling[2], with the utilization of Support Vector Machines (SVM), K-means clustering, and Isolation Forest algorithms. The goal is to develop an efficient and effective approach to identify anomalies in datasets and label them for further analysis or classification tasks.

4.1.1 Adaptive Model Scheduling

One key aspect of our proposed method is adaptive model scheduling [2],. Traditional anomaly detection approaches often rely on static models that are trained on fixed datasets. However, in real-world scenarios, data distribution and characteristics can change over time, making static models less effective. To address this limitation, we propose an adaptive model scheduling strategy[14].

The adaptive model scheduling[2], technique involves training and updating the anomaly detection models at regular intervals or when significant changes in the data distribution occur. This ensures that the models remain up-to-date and capable of capturing the evolving patterns of anomalies in the dataset. By adaptively scheduling the model updates, we aim to improve the accuracy and robustness of anomaly detection.

4.1.2 Utilizing SVM, K-means, and Isolation Forest

To perform anomaly detection and data labeling, we employ three key algorithms: Support Vector Machines (SVM), K-means clustering, and Isolation Forest.

SVM: We utilize SVM as a supervised learning algorithm for binary classification, where the objective is to separate normal data instances from anomalous ones. SVM learns a decision boundary that maximizes the margin between the two classes, allowing us to identify and label anomalies with high accuracy.

K-means Clustering: K-means is an unsupervised clustering algorithm that groups data instances into distinct clusters based on their similarity. We employ K-means to partition the dataset into clusters and identify potentially anomalous clusters. This step helps in distinguishing normal clusters from anomalous ones, facilitating the labeling process.

Isolation Forest: Isolation Forest is an unsupervised algorithm specifically designed for anomaly detection. It constructs an ensemble of isolation trees, which isolate anomalies more effectively by recursively partitioning the data space. By utilizing the Isolation Forest algorithm, we can efficiently identify and label anomalies in the dataset.

4.1.3 Integration and Labeling

In our proposed method, we integrate the outputs of SVM, K-means clustering, and Isolation Forest to identify and label anomalies in the dataset. The SVM classifier assigns labels to individual instances, while K-means clustering helps identify potentially anomalous clusters. Additionally, the Isolation Forest algorithm provides anomaly scores for each data instance.

By combining these approaches, we can effectively detect and label anomalies in an unsupervised or semi-supervised manner. The labeled anomalies can then be further analyzed or used for subsequent classification tasks.

4.2 Evaluation

To evaluate the effectiveness of our proposed method, we conduct experiments on benchmark datasets and compare the results with existing state-of-the-art approaches for anomaly detection and data labeling. We measure the accuracy, precision, recall, and F1 score of our method to assess its performance. Additionally, we analyze the computational efficiency and scalability of the proposed approach to ensure its practical applicability.

Through these evaluations, we aim to demonstrate the superiority of our Adaptive Model Scheduling using SVM, K-means, and Isolation Forest for anomaly detection and data labeling, providing a robust and efficient solution for identifying anomalies and facilitating downstream tasks in various domains.

4.3 Dataset

We used the Secure Water Treatment (SWaT) testbed dataset[15], which is a real-world dataset collected from a water treatment plant located in Singapore. The dataset was collected on July 20, 2019, and includes measurements from 51 sensors and actuators over a period of approximately 8 hours. The SWaT testbed was specifically designed to simulate cyber-physical attacks on a water treatment plant, making it an ideal dataset for evaluating anomaly detection methods.

The system operated non-stop from its ”empty” state to fully operational state for a total of 11 days. During the first seven days, the system operated normally without any attacks or faults. In the remaining days, cyber and physical attacks were launched on the SWaT while data collection continued.

The dataset includes both normal and attack scenarios, where the attack scenarios simulate various cyber-physical attacks on the water treatment plant. The attacks were carefully designed to be realistic and to mimic the behavior of real-world attackers. The dataset is publicly available and has been used extensively in the research community for evaluating anomaly detection methods.

We preprocessed the data by removing any missing or erroneous values, and then partially labeled the dataset for attack and normal scenarios, after that we changed the data into time series data and then made sure the timing for the attacks and the dataset are in the same timezone.

5 Experiments

In this section, we present the experiments conducted to evaluate the performance of the baseline method, *”Unsupervised deep learning and semi-automatic data labeling in weed Discrimination using DeepCluster algorithm,”* and the proposed method, *”Anomaly Detection and Data Labeling: Adaptive Model Scheduling using SVM, K-means, and Isolation Forest”*.

5.1 Dataset

The experiments were performed on the SWAT dataset containing both normal and anomaly samples which we labelled. The dataset was preprocessed and scaled using standard scaling techniques.

5.2 Baseline Method

The baseline method utilized the DeepCluster algorithm for data labeling. The steps involved preprocessing and feature scaling, followed by dimensionality reduction using PCA. The deep features were obtained using an autoencoder model, and K-means clustering was applied to the deep features for labeling.

5.3 Proposed Method

The proposed method, *”Anomaly Detection and Data Labeling: Adaptive Model Scheduling using SVM, K-means, and Isolation Forest”*, aimed to improve the efficiency and effectiveness of data labeling for anomaly detection. The method incorporated adaptive model scheduling with SVM, K-means, and Isolation Forest algorithms.

The adaptive model scheduling technique monitored the performance of the current model and made adjustments if the performance dropped below a certain threshold. The model was retrained with new data, including a subset of normal training data, to enhance the model’s performance.

5.4 Evaluation Metrics

The performance of both the baseline and proposed methods was evaluated using several metrics, including accuracy, precision, recall, and F1 score. Additionally, the area under the precision-recall curve (PR AUC) was calculated. Confusion matrices were also analyzed to understand the classification results.

6 Results

The results obtained from the experiments are as follows:

6.1 Baseline Method

Evaluation Metrics:
F1 Score: 0.80 Accuracy: 0.94 Precision: 0.79 Recall: 0.80 PR AUC: 0.66

6.2 Proposed Method

Evaluation Metrics:
Accuracy: 0.81 Precision: 0.88 Recall: 0.77 F1 Score: 0.82 PR AUC: 0.87

7 Conclusion

Based on the experimental results, it can be observed that the proposed method, ”Anomaly Detection and Data Labeling: Adaptive Model Scheduling using SVM, K-means, and Isolation Forest,” outperformed the baseline method, ”Unsupervised deep learning and semi-automatic data labeling in weed Discrimination using DeepCluster algorithm.”

In conclusion, comparing the Baseline Method and the Proposed Method, it is evident that the proposed method outperforms the baseline in terms of several evaluation metrics. The proposed method achieves an accuracy of 0.81, precision of 0.88, recall of 0.77, F1 score of 0.82, and PR AUC of 0.87. These metrics indicate that the proposed method has improved performance across multiple aspects compared to the baseline method. The higher accuracy and precision values demonstrate better overall classification accuracy and the ability to correctly identify positive instances. Additionally, the higher recall score indicates improved sensitivity in detecting positive instances, while the higher F1 score represents a better balance between precision and recall. Moreover, the significantly increased PR AUC value suggests a substantial improvement in ranking the positive instances higher than the negative instances. These results highlight the effectiveness and superiority of the proposed method over the baseline, indicating its potential for enhancing the task's performance and achieving more accurate predictions.

These results suggest that incorporating SVM, K-means, and Isolation Forest algorithms in the data labeling process for anomaly detection can lead to improved performance and more accurate classification of anomalies.

Further research and experimentation can explore additional variations and enhancements to the proposed method to address specific challenges in anomaly detection and data labeling tasks.

References

[1] I. Lee, J. Lee, K. H. Kim, and S.-k. Kim, “Unsupervised deep learning and semi-automatic data labeling in weed discrimination,” *Computers and Electronics in Agriculture*, vol. 163, p. 104858, 2019.

[2] M. Yuan, L. Zhang, X.-Y. Li, and H. Xiong, “Comprehensive and efficient data labeling via adaptive model scheduling,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 4, pp. 741–755, 2018.

[3] H. Bai, M. Cao, P. Huang, and J. Shan, “Self-supervised semi-supervised learning for data labeling and quality evaluation,” in *Proceedings of the 2020 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1113–1123, ACM, 2020.

[4] M. Sun, P. Li, J. Li, and X. Hu, “Multi-label learning with missing labels via common and label-specific features,” *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[5] W. Wang, T. Derr, Y. Ma, S. Wang, H. Liu, Z. Liu, and J. Tang, “Learning from incomplete labeled data via adversarial data generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020.

[6] P. Dube, B. Bhattacharjee, S. Huo, P. Watson, and B. Belgoder, "Automatic labeling of data for transfer learning," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 3061–3068, IEEE, 2019.

[7] G. Pang, A. van den Hengel, C. Shen, and L. Cao, "Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6932–6939, 2019.

[8] IBM, "Data Labeling." https://www.ibm.com/topics/data-labeling, Accessed: 2023.

[9] A. Networks, "Anomaly detection." https://avinetworks.com/glossary/anomaly-detection/#:~:text=There%20are%20three%20main%20classes,available%20labels%20in%20the%20dataset., Accessed on April 18, 2023.

[10] C. Chuah and D. Ng, "A note about finding anomalies," *Towards Data Science*, 2019.

[11] Dataloop AI, "Data labeling challenges." https://dataloop.ai/blog/data-labeling-challenges/, 2021. Accessed: 2023-04-16.

[12] M. Kuchta, "Anomaly detection in time series," 2021.

[13] Z. Liu, Y. Yu, Y. Ma, W. Wang, and J. Tang, "Deep learning for multi-view multi-instance multi-label classification with incomplete labels," *Machine Learning*, pp. 1–37, 2022.

[14] X. Q. Xin Jin, "An adaptive anomaly detection method for cloud computing system," *IEEE*, 2021.

[15] iTrust Labs, "itrust dataset." https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/, 2021.