Evaluating the Performance of SVM, Isolation Forest, and DBSCAN for Anomaly Detection

Haowen Lu

WLSA Shanghai Academy, 200940 Shanghai, China

Abstract. With the advancement of computer technologies, various data models and algorithms have been integrated into industrial processes, significantly improving the efficiency of anomaly detection in datasets while reducing time and energy consumption. Identifying the most effective algorithm for anomaly detection is essential for enhancing industrial productivity. This study evaluates the suitability of three algorithms—Support Vector Machine (SVM), Isolation Forest, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) by comparing their accuracy and time efficiency in detecting outliers in different types of datasets. The algorithms are tested across various datasets, and their performance is systematically compared. The results are then analysed in relation to the structure of each algorithm to identify their respective advantages and disadvantages. The study finds that each algorithm performs differently depending on the dataset type. Specifically, SVM demonstrates superior performance in detecting point anomalies, while DBSCAN is more effective when the dataset is pre-processed. Additionally, Isolation Forests are most efficient at identifying collective anomalies within the dataset.

1 Introduction

With the industrial revolution of information technology and big data, networks and data techniques were applied in numerous fields, especially manufacturing. The assistance from computer systems and data models combined with industrial technologies, that grossly improves the profit of the enterprise by controlling the cost of the industrial equipment and shortening the production period with the improvement of productivity. Anomaly detection is one of the important techniques in manufacturing that improves the reduction of cost and security of production by data models. This paper will demonstrate and compare the algorithms and data models that are most appropriate and efficient for anomaly detection on industrial equipment.

With the development of data analysis models used in industrial production, some security problems of information technologies appear together. Some researchers find that some changes in values such as adding a pixel to an image may cause huge changes to output from data analysis models [1]. To solve these problems, Siegel suggests two unsupervised multivariate neural networks Recurrent Neural Network (RNN) Encoder-Decoder (RNN-ED)

Corresponding author: linsonglin@ldy.edu.rs

and One-Dimensional Convolutional Encoder-Decoder (1DCNN-RNN-ED) with the comparison of data analysis architectures. RNN Encoder-Decoder (E/D) is a combination of RNN architecture and Encoder-Decoder(E/D) architecture. RNN processes sequences by iterating through temporal data, and E/D uses the output from RNN to eliminate noise and unimportant features from the data. 1DCNN supports local anomalies, which makes RNN-E/D architecture more stable, has a lighter weight and faster network. There are also some data models that can resist malicious activities to protect industrial equipment [2]. Multifunction Information Distribution System (MIDS) is an anomaly detection system that use the data from the production of the Supervisory Control and Data Acquisition (SCADA) system which collect the data from ICS sensors. It can detect anomalies, especially stealthy attacks by investigating the measurement data. Intrusion-detection system (NIDS) is another system which monitors the input of data packets then discovering data that have the possibility to intrude the computer system [3].

In the training phase, the detection system can build a data model with the data collected by the sensors, which represents the behaviour of the machine in normal states. When the machine starts running, the system will compare the current data and the data model and then find the anomaly data. There are a few algorithms that can be applied to the detection system. The first is One class Support Vector Machine (SVM) At first, it finds a hyperplane to circle the positive examples in the sample. Then it uses the hyperplane to difference the normal data and the anomaly data. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an algorithm that uses the destiny of each cluster class to find noise because the density of noise is lower than the destiny of any cluster class. DBSCAN is a great algorithm for detecting noise in the dataset and it is suitable for a large proportion of data, but it is difficult to deal with large-scale datasets. One-Class SVM performs well in the sample-less case. However, as a binary algorithm, dealing with multi-class problems is difficult for this algorithm. Therefore, it is necessary to recognize the faults and advantages of each algorithm then find whether it is effective in anomaly detection of auto-engineering.

The primary objective of this study is to identify the most suitable algorithm for anomaly detection in automation engineering. Automatic anomaly detection plays a crucial role in industrial applications by efficiently and accurately identifying data faults and equipment errors. Additionally, the implementation of automated systems can significantly reduce labour costs. This research will evaluate various algorithms by testing them on datasets from automation engineering, analysing their performance in terms of computational time and resource utilization. The goal is to determine the most efficient algorithm that enhances the accuracy of anomaly detection systems while minimizing production costs. The findings of this study are expected to provide valuable insights for optimizing industrial processes and improving operational efficiency.

2 Methodology

2.1 Dataset

One of the datasets that can be used in this study is the University of California, Irvine (UCI) Steel Plates Faults Data Set, which is provided by researchers at Northeastern University. It is a dataset contains seven styles of steel plate drawbacks [4]. There's another dataset that is useful for this study which provides access to ball bearing test data for normal and faulty bearings. The bearing test rig accelerometer data of run-to-failure experiments in this dataset provide normal and anomalous industrial data, which is helpful for algorithms practicing discovering industrial abnormal data.

2.2 Algorithm

This chapter is used to introduce the algorithms which will be used in this study. It is proposed to analyse the theory and structure of these algorithms, in order to discover their advantages and disadvantages in anomalous detection. The analysis of these algorithms would include the features, theories, and advantages of these algorithms. The pipeline is shown in the Fig. 1.

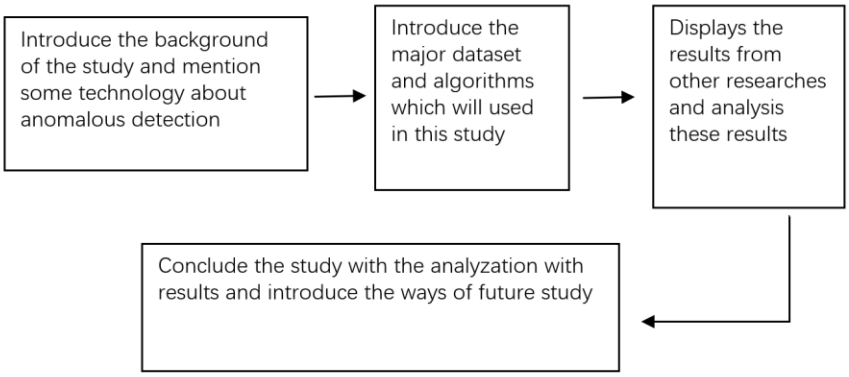

Fig. 1. The pipeline of this study (Picture credit: Original).

2.2.1 Support vector machine

SVM is a machine-learning algorithm based on binary research. The basic idea is to use a hyperplane to separate a d-dimensional dataset into two classes. Since the dataset is not linearly separable, the algorithms need to put the dataset in a higher-dimensional space and then separate the dataset directly (The algorithm doesn't need to deal with higher-dimensional space in this process). First, the dataset would need to be processed. To do this, the algorithm transforms the low dimensional dataset into a higher dimensional by the income of a vector. Next, the algorithm would find an optimal hyperplane to separate the dataset into two areas and maximize the distance to the nearest data points by the optimal curve. After separating the data, it will output the boundary of data which is not expected out-of-sample error [5]. In anomaly detection, the dataset collected by the machines will be turned into a hyperplane and separated into positive samples and negative samples. With the help of SVM, the data collected by the sensors can be classified into normal data and anomaly data. The generalization concept can help researchers to obtain better classification results. However, the SVM only supports solving the binary question, so it still needs to use other algorithms to combine with SVM for separating different kinds of anomalies.

2.2.2 DBSCAN

DBSCAN is the first density-based clustering algorithm. It is designed to use in high-dimensional database which contain noise in cluster data with any shapes. A vital idea in DBSACN each neighbour in a given radius must contain at least a minimum number of objects (MinPts) for each object of a cluster. DBSACN finding a cluster by calculating the number of neighbours around a object in the dataset. If the number of neighbourhoods of an object p is over MinPts, the algorithms will create a new cluster and set p as the core as this cluster. Then it will continue to search each object in this cluster and justify whether it can

become a core of a cluster. The process will not stop until no new object can be added to any cluster [6]. In anomaly detection, DBSCAN can label the data points as core points, border points, and outlier (anomalous) points. Core points is the core points of the clusters, border points are the points in the neighbourhood of the core points, and the outlier points are the points out of the neighbourhood of the core points. The outlier points in the data set indicate this data are anomalous are different from other points, which is probably anomalous [7].

2.2.3 Isolation forest

The design of Isolation Forest has two basic characteristics about the anomaly data (show in Table 1, Table 2 and Table 3): The first is the anomalous data only occupy a few parts of the entire data set. Secondly, there is a remarkable distinction between the exceptional data and the attribute value of the normal data. If a dataset that contains only data with the type of number, the algorithms will continue divide this dataset into two parts and stop when each data is separated into one part [8]. The algorithm randomly selects an attribute and a split value from the dataset, then divides each data in the dataset with this attribute. The data that is bigger than the split would be distributed into different subtrees of smaller ones. In the next step, the algorithm will continue this step into each subtree until every subtree only contains one value. This algorithm is very helpful to anomaly detection because this random partitioning produces noticeably shorter paths for anomalies, because the smaller number of anomalous data will lead to a reduction in partitions-which means a shorter path in tree structure, and instances with has a recognizable attribute value would be separated into partitions earlier than other data. Hence, the algorithms can justify whether the data is a anomalous data by finding if it has a shorter path in the randomly-generated forest [9].

Table 1. The efficiency of four algorithms in detecting point anomaly detection without time-series decomposition [10].

Model	F-score	Balanced accuracy	Precision	Recall
DBSCAN	0.895	0.935	0.968	0.871
Isolation forest	0.822	0.915	0.834	0.832
One-class SVM	0.916	0.983	0.923	0.970

Table 2. The efficiency of four algorithms in detecting point anomaly detection with time-series decomposition [10].

Model	F-score	Balanced accuracy	Precision	Model
DBSCAN	0.569	0.776	0.773	DBSCAN
Isolation forest	0.667	0.830	0.671	Isolation forest
One-class SVM	0.507	0.847	0.507	One-class SVM

Table 3. The efficiency of four algorithms in detecting collective anomaly detection with time-series decomposition [10].

Model	F-score	Balanced accuracy	Precision	Model
DBSCAN	0.905	0.942	0.960	DBSCAN
Isolation forest	0.816	0.940	0.784	Isolation forest
One-class SVM	0.810	0.930	0.811	One-class SVM

3 Result and Discussion

3.1 The performance of DBSCAN, SVM, isolation forest in anomalous detection

These tables represent the performance of three algorithms with different kinds of anomalies and various forms of data. Three algorithms all perform well in the detection of point anomalies in the data without time-series decomposition in Table 1. However, the performance of SVM is the best and the performance of isolation forest is the worst. When data is in time series, it is evident that DBSCAN performs best. In the detection of collective anomalies in the data with time-series decomposition, the isolation tree has the best performance. The SVM is good at distinguishing positive points and anomalous points based on a binary searching of the points. If data is pre-processed, DBSCAN can perform well. Since the data is already in a cluster, DBSCAN can save the time of clustering. Additionally, DBSCAN is sensitive to noise and outliers, therefore it is time-saved in this environment. Compared with Density-Based Spatial l Clustering of DBSCAN and SVM which discover outliers based on data points, isolation forests use trees to slice the dataset and find the anomalous data. Therefore, isolation can notice the local anomalous dataset with the depth of the tree, which enable the detection of collective data more efficient than another algorithm.

3.2 Analyzation of these algorithms

One-class SVM is a binary-based algorithm which present significant performance in classification of anomalous points and normal points. It can also deal with high-dimensional data. However, the SVM is limited by binary research, so it isn't efficient in training with large-scale samples. There are a few samples of advanced SVM algorithms nowadays, which are the approaches of more efficient application of SVM in anomaly detection.

DBSCAN is a clustering algorithm based on the density of data. It has a great sensation to the anomalous data in the dataset. However, the process of bringing data together into different clusters is time-consuming. If the data in the dataset is already in a series, DBSCAN would perform very significantly. Shorten the time of clustering would greatly improve the efficiency and accuracy of DBSCAN. Isolation forest is an outlier detection algorithm based on trees. It has a linear time complexity so that it has a low time consumption on detecting outliers in large-scale dataset. However, it is not sensitive to an anomalous point in a local area since it is focused on global anomaly detection. To improve this algorithm, enhance the ability of discovering a single anomalous point is a good idea, such as combine this algorithm with one-class SVM.

4 Conclusion

This study introduces and compares the performance of three algorithms in anomaly detection, highlighting their respective strengths and weaknesses in analyzing outliers within datasets. The second section provides a detailed analysis of the structure and theoretical underpinnings of these algorithms, while the third section examines their practical applications to determine the most suitable scenarios for each. The experimental results indicate that SVM algorithm is both time-efficient and accurate in detecting point anomalies. However, SVM has limitations, particularly in handling multi-class problems, as its efficiency is constrained by its binary classification approach. Moreover, DBSCAN demonstrates greater sensitivity to anomalous points, especially when the dataset is pre-processed, making it more effective for detecting such anomalies. The Isolation Forest algorithm excels in identifying collective anomalous points within a dataset by

simultaneously generating trees and quickly isolating the anomalous tree. However, this process is time-consuming, as it requires separating all data points into trees. Future research will focus on improving the performance of these algorithms in anomaly detection, with an emphasis on refining their structures and optimizing their specific applications. The next stage of research will aim to enhance the efficiency and accuracy of these algorithms, addressing the limitations identified in this study.

References

1. B. Siegel, Industrial anomaly detection: A comparison of unsupervised neural network architectures. IEEE Sensors Letters, 4(8), 1-4 (2020)
2. S. Jiawei, D. Vasconcellos Vargas, and K. Sakurai, One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23(5), 828-841 (2019)
3. S. Mokhtari, A. Abbaspour, K.K. Yen, & A. Sargolzaei, A machine learning approach for anomaly detection in industrial control systems based on measurement data. Electronics, 10(4), 407 (2021)
4. T. Yang, F. Mengyu, and W. Fang, Steel plates fault diagnosis on the basis of support vector machines. Neurocomputing, 151, 296-303 (2015)
5. D. Boswell, Introduction to support vector machines. Departement of Computer Science and Engineering University of California San Diego, 11, 16-17 (2002)
6. K. Kamran, et al. DBSCAN: Past, present and future. The fifth international conference on the applications of digital information and web technologies (2014)
7. M. Çelik, F. Dadaşer-Çelik, and A. Ş. Dokuz, Anomaly detection in temperature data using DBSCAN algorithm. international symposium on innovations in intelligent systems and applications (2011)
8. F.T. Liu, M.T. Kai, and Z. H. Zhou, Isolation forest. IEEE international conference on data mining (2008)
9. D. Xu, Y. Wang, Y. Meng, & Z. Zhang, An improved data anomaly detection method based on isolation forest. International symposium on computational intelligence and design, 2, 287-291 (2017)
10. A. L. Henriksson, Unsupervised Anomaly Detection on Time Series Data: An Implementation on Electricity Consumption Series (2021)