Research on network intrusion detection based on XGBoost algorithm and multiple machine learning algorithms

Zhihui Fan[1]**, Zhixuan You**[1, 2]

[1]Software institute, Nanchang University, Jiangxi, Nanchang, 330047, China

[2]lvuce1@163.com

Abstract. Network intrusion detection refers to monitoring and analysing network traffic, system logs and other information to identify abnormal behaviours and attacks in the network, and take timely and appropriate countermeasures to protect the security and stability of the network. This paper investigates the application of seven machine learning methods in network intrusion detection, and evaluates each model by indicators such as precision, accuracy, recall and F1 score. The results show that XGBoost, Random Forest and Decision Tree models have the best prediction results, while Support Vector Machines and Plain Bayes models have poor prediction results. XGBoost, Random Forest and Decision Tree models all belong to the category of integrated learning, which have strong generalisation ability and robustness, can handle high-dimensional and complex datasets and are not prone to overfitting. In addition, they are able to handle non-linear relationships and are suitable for complex classification problems. Catboost and logistic regression models have better prediction results, but their prediction results are also affected by feature engineering. They may under- or over-fit when dealing with high-dimensional, complex datasets. Support Vector Machines and Plain Bayesian Models have poorer prediction results, which is related to their limitations. Support vector machines may experience computational difficulties when dealing with high-dimensional, complex datasets and are weak in dealing with non-linear relationships. The plain Bayesian model assumes that the features are independent of each other, which may not hold true in practical applications, thus affecting the prediction results. The conclusions of this paper are instructive for research and application in the field of network security, and can provide reference and inspiration for research in related fields.

Keywords: Network security, Intrusion detection, Prediction accuracy.

1. Introduction

Network intrusion detection refers to the monitoring and analysis of network traffic, system logs and other information to identify abnormal behaviours and attacks in the network, and take corresponding countermeasures in a timely manner to protect the security and stability of the network. The research of network intrusion detection originated in the field of network security in the late 1980s, when it was mainly focused on traditional network security threats such as hacker attacks and virus attacks [1]. With the continuous development of network technology, network intrusion detection has also been developed and improved [2].

With the popularity and application of the Internet, the types and number of network attacks are increasing, and the means of attack are becoming more and more complex and covert [3]. Traditional rule- and feature-based intrusion detection methods can no longer meet the practical needs, so researchers have begun to explore intrusion detection methods based on artificial intelligence technologies such as machine learning and deep learning to improve detection accuracy and efficiency [4, 5].

In recent years, with the development of big data, cloud computing and other technologies, network intrusion detection faces new challenges and opportunities [6]. On the one hand, the scale and complexity of network traffic and log data are increasing, and traditional detection methods are no longer capable; on the other hand, big data and cloud computing technologies provide stronger computation and storage capabilities for intrusion detection, which provide strong support for researchers to develop and apply more efficient and accurate intrusion detection algorithms [7].

Some researchers have conducted intrusion detection based on the plain Bayesian algorithm, which can classify and identify abnormal traffic [8]; intrusion detection based on the support vector machine algorithm can classify network traffic; and some other researchers have conducted intrusion detection based on the deep learning algorithm can automatically extract the eigenvalues of the traffic data, and classify and identify the abnormal behaviours [9]. All these algorithms and models can be selected and adjusted according to specific application scenarios and needs to improve the accuracy and efficiency of intrusion detection. The application of machine learning in network intrusion detection provides strong support for network security and brings new development opportunities for research and application in the field of network security.

Network intrusion detection is one of the important research directions in the field of network security, and with the continuous development of network technology and the continuous updating of attack methods, intrusion detection technology also needs to be constantly innovated and improved in order to guarantee the security and stability of the network.

2. Source of data sets

The data used in this paper comes from kaggle’s network intrusion detection dataset, which consists of a variety of intrusions simulated in a military network environment. The process of acquiring the data began with creating an environment to acquire raw TCP/IP dump data of a network by simulating a typical Air Force LAN. The LAN was centralised like the real environment and was subject to multiple attacks. A connection is a series of TCP packets that begin and end at a certain duration, between which time data flows in and out of the source IP address to the destination IP address at some well-defined protocol. In addition, each connection is labelled as either normal or an attack with only one specific attack type. Each connection record consists of approximately 100 bytes.

The dataset contains a total of 25, 193 pieces of data, and for each TCP/IP connection, 41 quantitative and qualitative features (3 qualitative and 38 quantitative) were obtained from the normal and attack data, and the class variable has two categories, normal, which indicates a normal network, and abnormal, which indicates a network intrusion.

3. Pearson correlation analysis

Pearson correlation analysis is a commonly used statistical method to investigate the linear correlation between two variables. In network intrusion detection datasets, Pearson correlation analysis can be used to explore the correlation between individual data features. Specifically, the method derives the correlation coefficient between two variables by calculating the covariance and standard deviation between them. The correlation coefficient can take values ranging from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

When performing Pearson correlation analysis, the data were first pre-processed, including missing value processing, outlier processing and standardisation. Then, we selected several representative features to calculate the correlation coefficient matrix between each feature. Finally, we used a correlation heat map to display the correlation coefficient matrix in order to more intuitively analyse the correlation between individual data features. The correlation heat map is shown in Figure 1.

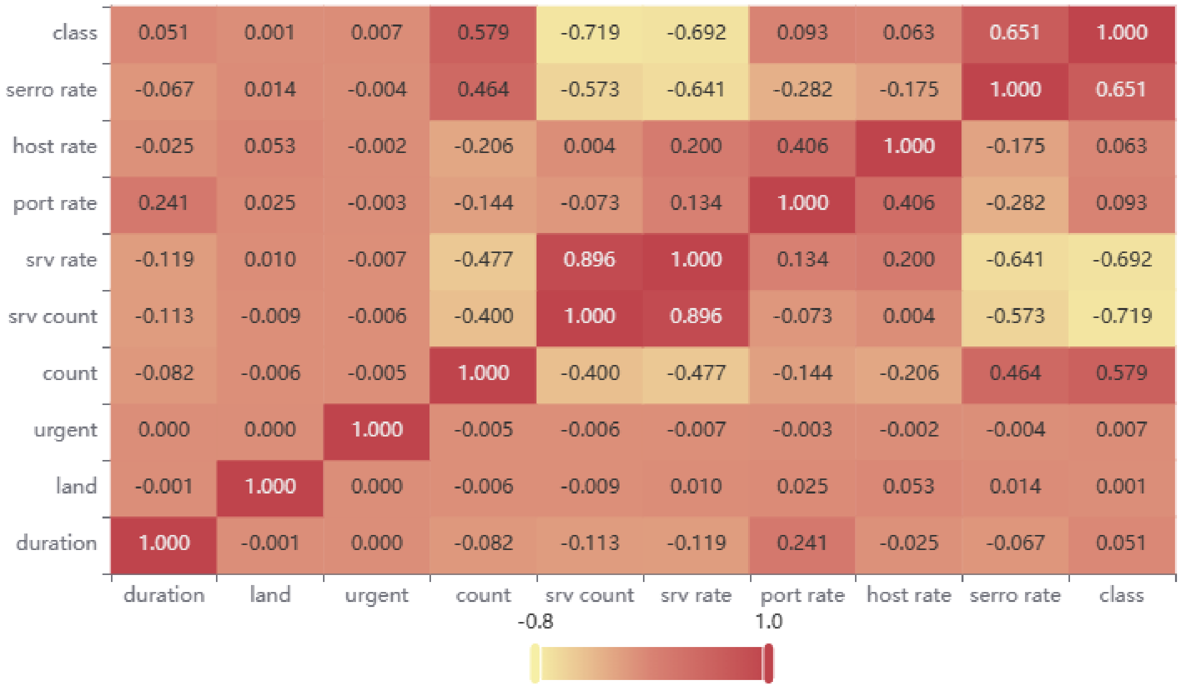


Figure 1. Correlation heat map. (Photo credit: Original)

From the correlation heat map, it can be seen that there is correlation between these network features we selected, and the positive correlation of a few features even reaches 89.6%, which shows that there are some correlations between the network features as well as between the network features and the final network intrusion detection decision category, and the internal laws between these network features can be further explored and analysed using machine learning methods.

4. Machine Learning Algorithms

4.1. Decision tree

A decision tree is a model for decision making based on a tree structure, where a tree is constructed for classification or regression by dividing the dataset into subsets. At each node, the direction of branching is decided by dividing the features and selecting the optimal ones. Decision tree models have the advantages of being highly interpretable, easy to understand and implement, but are prone to overfitting problems.

4.2. Random forest

Random forest is an integrated learning model based on decision tree, which constructs multiple decision trees by randomly selecting features and samples, and decides the final classification result by voting. The random forest model has high accuracy and robustness, and can effectively avoid the problem of overfitting.

4.3. Catboost

Catboost is a machine learning model based on gradient boosting decision trees, which gradually improves the accuracy of the model by optimising the loss function.The Catboost model has high accuracy and robustness, and is able to automatically handle both categorical and numerical features, as well as problems such as missing values and outliers.

4.4. Support vector machine

Support vector machine is a maximal interval based classifier that classifies data by mapping it to a high dimensional space. Support vector machine models have high accuracy and generalisation capabilities and can effectively handle high dimensional data and non-linear classification problems.

4.5. XGBoost

XGBoost is a machine learning model based on gradient boosting decision trees, which gradually improves the accuracy of the model by optimising the loss function.The XGBoost model has high accuracy and robustness, and is able to effectively deal with high-dimensional data and nonlinear classification problems, as well as problems such as missing values and outliers.

4.6. Bayesian model

Bayesian is a classifier based on Bayes' theorem that classifies features by counting the conditional probabilities between them. The plain Bayesian model has high speed and interpretability and can effectively handle high dimensional data and large scale datasets.

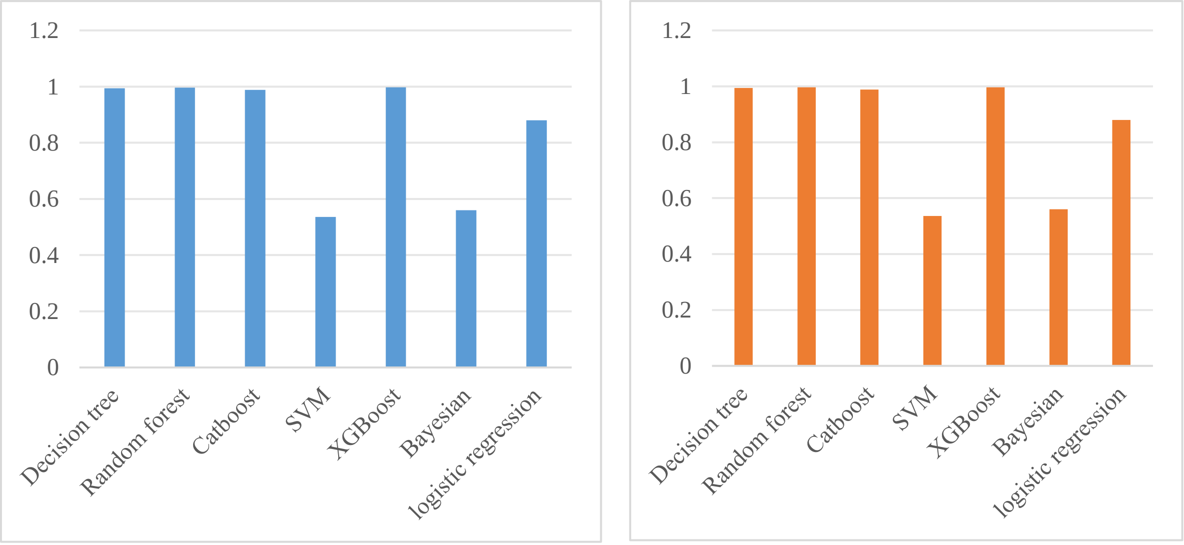
4.7. Logistic regression

Logistic regression is a probability-based classifier that classifies data by mapping it to a probability space. Logistic regression models have high speed and interpretability and can effectively deal with binary classification problems. Logistic regression models are mainly used for classification problems.

5. Experiments and results

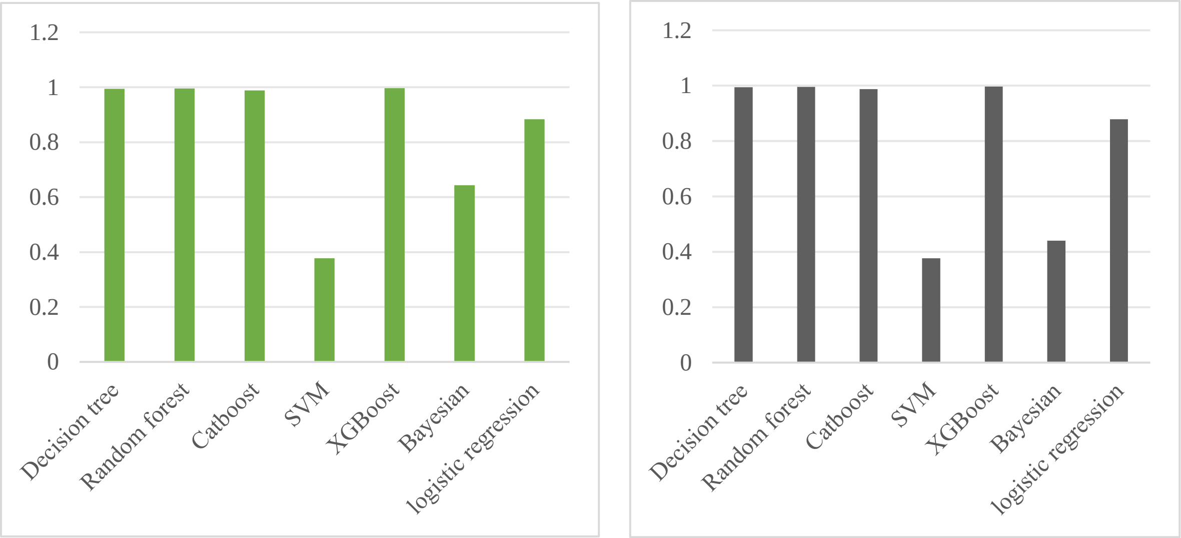
The training set, validation set and test set are divided according to the ratio of 6:2:2, the training set is used for model training, the validation set is used to validate the results of the training set, and the test set is used to evaluate the model results.

Seven machine learning methods, including Decision Tree, Random Forest, Catboost, Support Vector Machine, XGBoost, Plain Bayes and Logistic Regression, are used to detect network intrusion and judge the normal and abnormal states of the network, compare the advantages and disadvantages of each model, and evaluate the models in terms of precision, covariance, recall and F1 score. The results are shown in Table 1 and Figure 2.



(a)

(b)



(c) (d)

Figure 2. (a) Accuracy, (b) Recall, (c) Precision, (d) F1. (Photo credit: Original)

Table 1. XGBoost predicts results.

Model	Accuracy	Recall	Precision	F1
Decision tree	0.994	0.994	0.994	0.994
Random forest	0.996	0.996	0.996	0.996
Catboost	0.988	0.988	0.988	0.988
SVM	0.536	0.536	0.377	0.377
XGBoost	0.997	0.997	0.997	0.997
Bayesian	0.56	0.56	0.643	0.44
logistic regression	0.88	0.88	0.884	0.879

From the precision, accuracy, recall and F1 score of each model, it can be seen that the best prediction results are obtained from XGBoost model, Random Forest and Decision Tree model, with their respective accuracies of 99.7%, 99.6% and 99.4%, and their prediction accuracies are all more than 90%; Catboost model and Logistic Regression model also have better prediction results, with accuracies of 98.8% and 88%; while the support vector machine model and the plain Bayesian model have poor prediction results, with binary prediction results of only 53.6% and 56%.

6. Conclusion

For the three models XGBoost, Random Forest and Decision Tree, all three models belong to the category of integrated learning, i.e., combining multiple weak classifiers into a strong classifier. Decision tree is a classification algorithm based on a tree structure, which classifies the data set by recursively dividing it into smaller subsets and eventually generating a decision tree. Random forest is an integrated model consisting of multiple decision trees, where multiple decision trees are constructed by randomly selecting features and samples, and their predictions are voted for classification. XGBoost is an integrated model based on gradient boosting trees, where multiple weak classifiers are trained iteratively, and sample weights are adjusted in each iteration to give more attention to samples that were misclassified in the previous classification, and finally multiple weak classifiers are combined into strong classifiers. multiple weak classifiers into a strong classifier.

From the principle point of view, these three models have strong generalisation ability and robustness, and can handle high-dimensional and complex datasets, and are not prone to overfitting. In addition, they are able to handle non-linear relationships and are suitable for complex classification problems.

Next, for the two models Catboost and Logistic Regression, Catboost is an integrated model based on gradient boosting trees, similar to XGBoost, but it has some advantages in dealing with categorical features. Catboost is able to deal with categorical features automatically without the need for processing such as solo thermal coding, thus reducing the workload of feature engineering. Logistic regression is a generalised linear model that performs binary classification by linearly combining input features with weights, and then mapping the result to between 0 and 1 via a sigmoid function.

In principle, both models are more basic and suitable for more fundamental classification problems, but their predictive effectiveness is also affected by feature engineering. When dealing with high-dimensional, complex datasets, they may under- or over-fit.

Finally, for Support Vector Machines and Plain Bayesian Models, Support Vector Machines are a model based on Maximum Interval Classification, which classifies data by mapping it to a high-dimensional space and finding the optimal hyperplane in it. The Plain Bayesian model is a classification model based on Bayes' theorem, which performs classification by calculating prior probabilities and conditional probabilities.

Taken together, XGBoost, Random Forest and Decision Tree models have the best prediction effect in network intrusion detection, which is closely related to their integrated learning principle and ability to handle complex data. While support vector machine and plain Bayesian models have poorer prediction effect, which is related to their limitations. In practical applications, we need to select suitable models according to the characteristics of specific problems, and carry out work such as feature engineering and model tuning to improve the prediction effect.

Acknowledgement

Provincial Project of 2023 Student Innovation and Entrepreneurship Training Programme of Nanchang University, Jiangxi Province (Project No. S202310403001X)

References

[1] Daly L M, Siamak L, Weng W L, et al.FlowTransformer: A transformer framework for flow-based network intrusion detection systems[J].Expert Systems With Applications, 2024, 241.

[2] Elif D, Yunus K S, İlker Ö, et al. ROSIDS23: Network intrusion detection dataset for robot operating system[J]. Data in Brief, 2023, 51109739-109739.

[3] Ghani H, Salekzamankhani S, Virdee B. A Hybrid Dimensionality Reduction for Network Intrusion Detection[J].Journal of Cybersecurity and Privacy, 2023, 3(4):830-843.

[4] Tarek G, Bamidele J A, Mohamed T, et al.Metaverse-IDS: Deep learning-based intrusion detection system for Metaverse-IoT networks[J].Internet of Things, 2023, 24.

[5] Abiodun A, Amrit K, Anit K, et al.Network intrusion detection using feature fusion with deep learning[J].Journal of Big Data, 2023, 10(1):

[6] Manderna A, Kumar S, Dohare U, et al.Vehicular Network Intrusion Detection Using a Cascaded Deep Learning Approach with Multi-Variant Metaheuristic[J].Sensors, 2023, 23(21):

[7] Ivandro L O, Deqing Z, H. I A, et al.Network intrusion detection based on the temporal convolutional model[J].Computers Security, 2023, 135.

[8] Zhang J, Zhang X, Liu Z, et al.A Network Intrusion Detection Model Based on BiLSTM with Multi-Head Attention Mechanism[J].Electronics, 2023, 12(19):

[9] Goh, K. L., and Singh, A. K 2015 Procedia Computer Science, 70, 434-441.