Abnormal Traffic Prediction and Classification based on Information Big Data

Yuyang Tian*

Beijing Jiaotong University Beijing, China

*Corresponding author: 18711026@bjtu.edu.cn

Abstract. Intrusion Detection System (IDS) is a proactive security technique for detecting and alerting suspicious signals. However, the intrusion method developed as a fast and traditional method for detecting malicious traffic has a lot of shortcomings like low accuracy and low efficiency. To determine the different intrusion methods' features and promote the accuracy of malicious traffic detection, several Machine Learning models for classifying different intrusion methods such as KNN, Naive Bayes, SVM, LightGBM are compared. To further improve the accuracy of the model, ensemble models like Voting, Stacking for comparison are also introduced. Grid Search is used for the best parameters. The accuracy, precision, recall score and F1 score are used as metrics to evaluate the performances of different models. The experimental comparison and analysis show that the integrated learning algorithm based on Stacking has the highest accuracy for malicious traffic detection.

Keywords: Big Data, Intrusion Detection System, UNSW-NB15, multi-class classification.

1. Introduction

The wide spread of Internet usage has introduced many network security threats, and the recent trend of attack is to cause serious network performance degradation by introducing a large amount of malicious traffic into the network. Network abnormal traffic detection can monitor the network environment in real-time by extracting and analyzing network traffic characteristics, and plays an important role in network security protection. The network intrusion detection system can achieve good detection performance by using machine learning algorithm.

With the emergence of big data, massive and complex data make traditional machine learning algorithm affected by time complexity and space complexity, which reduces the accuracy and efficiency of the system. If every feature of the data is included in the classifier, the efficiency of machine learning will be greatly reduced.

Recent popular detection methods of IDS are the Signature-based Method[1] and the Anomaly-based Method[2]. To achieve a low error rate for benign traffic. Fengjie Hu[3] adopted a feature-based approach to design a Light GBM network intrusion detection system framework based on ADASYN data balancing and PCA feature dimension reduction. He Hongyan [4] generated data sets with typical data features as the data set of feature extraction, and used Kmeans clustering algorithm to conduct network intrusion detection experiments on the intrusion detection model. Smitha Rajagopal[5] proved the relevance of each feature class and the importance of various combinations of feature classes and uses two types of neural networks to conduct scalable machine learning research on data sets.

However, their researches all adopted a single Machine Learning method or linear fusion of simple model to study the data sets related to network security, leading to did not consider the limitations of a single machine learning algorithm. On the other hand, the method based on ensemble learning to solve the data mining work related to network security. The proposed method solves the problem of data jumbliness and improves the efficiency and further improves the accuracy via Ensemble Models.

In this paper, we tend to find a better method to detect anomaly from packet caught from the network and report potential threat by anomaly-based Method. We use the UNSW-NB15 dataset for training and testing our multi-class classification models. In addition, we select several ML models (LightGBM [6], LDA [7], NB [8], SVM [9], RF [10], CART [11], KNN [12], LR [13], Voting [14],

Weighted Voting [15], Stacking [16]) to classify the dataset. To further improve the accuracy and feasibility of the research, we give weight to the algorithm based on the correct rate of the model, and let them participate in the ensemble learning. By the training result, we determine the most accurate and efficient method for network abnormal traffic detection. The model and method can be used in further IDS system. Weighted Voting has the highest accuracy of the models, at 77.8 percent. The results clearly show that the ensemble learning method has higher accuracy than any base classifier and better prediction results for the dataset.

2. Method

This section describes the proposed method. The dataset used in this paper is first described (Sec. A). Then we show the method of preprocesses of the data (Sec. B).

2.1. Data Collection

The IXIA PerfectStorm tool in the Cyber Range Lab of UNSW Canberra produced the raw network packets for the UNSW-NB 15 dataset to provide a blend of real contemporary normal activities and synthetic current assault behaviors, containing 175341 items in the test set and 82333 items in the training set, totaling 99867 items.

According to the requirements of data and variables used in this study, the main variables are the data of abnormal traffic types and their related characteristics. Since the data comes from most abnormal traffic detection and prediction in modern normal life, UNSW-NB 15 can be regarded as a representative sample of abnormal traffic, which can be reflected in the attack behavior on the network.

The explained variable (dependent variable) in this study is the attack types of abnormal traffic. To distinguish different abnormal traffic, the abnormal traffic data are divided the abnormal traffic into nine discrete types of attacks: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. The independent variables of this study are measured by 197 related attributes in UNSW-NB 15. To reduce the workload in the modeling phase and increase the usability of the model, some of the attributes are selected by data preprocessing. According to the actual situation of abnormal traffic analysis and research needs, some continuous attributes are discretized. Maintaining the Integrity of the Specifications

2.2. Data pre-processing

The dataset contains a total of 196 abnormal traffic-related attributes, which contain discrete variables and continuous variables. To solve the problem that there are too many labels in categorical features, we disregard some attitudes. Since categorical features cannot be input into our model, we first use one hot coding to discrete the attributes. To solve attribute redundancy, the correlation of the attributes heatmap. For outlier analysis, we draw the violin. To solve the problem that the distance between features is different, we normalized the data.

2.2.1 Reducing the Labels in Categorical Features

For a total of 196 attributes, we drop ‘id’ and ‘label’ and define ‘attack_cat’ to be y column to obtain an of 42 features and start reducing. Variables with categorical data have label values rather than numerical values. In our research, the Categorical feature is ‘proto’, ‘state’, and ‘service’. Since we need one hot code here does not need too much redundant data, we only choose the top 5 frequent labels in each categorical feature and disregard others. Tab. 1 shows that the unique labels are reduced to 6 (which contains a N/A) and 5.

Table 1. lables before and after reducing

	Proto	**Service**	**State**
Count	257673	257673	257643
Top	TCP	-	FIN
Unique (Before)	6	5	6
Unique (After)	133	13	11

2.2.2. One Hot Coding

To convert categorical data variables, we use one hot encoding process, so they can be provided to machine learning algorithms to improve predictions. Many machine learning algorithms can't operate on label data directly. They require all input variables and output variables to be numeric. It converts each categorical value into a new categorical column and assigns a binary value of 1 or 0 to those columns. In this paper, we use using pd.get_dummies to one hot encode 'cat_cols'.

2.2.3. Feature Selection by Correlation Heatmap

To find the correlations of each feature and reduce very similar features, we use the correlation heatmap, which shows the correlation between various variables through a graphic representation of a correlation matrix. Among heatmap, we dropped 10 columns which features with a correlation greater than 90%.

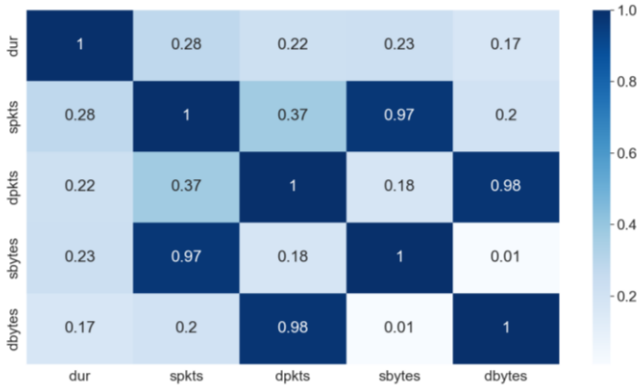

Figure 1. Heatmap of attributes

2.2.4. Density analysis

The violin diagram is used to make each attribute value of the data have a better visual display for density. Considering that the data is a combination of relative and absolute numbers, the data interpolation is too large. Therefore, we consider outliers deletion through parameter adjustment in the classification model.

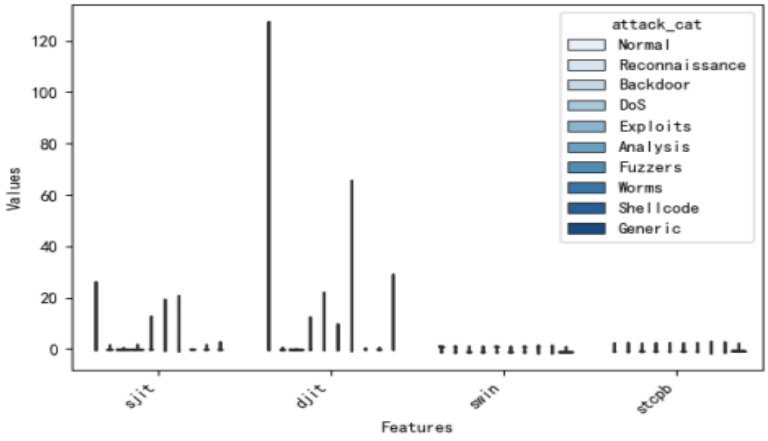

Figure 2. Density analysis was performed using a violin diagram

2.2.5. Normalization

To avoid distorting differences in the ranges of values or losing information. Normalization is the process of scaling all the columns in a dataset to the same value. Thus, we use the method of Min-Max Scaling which linearly rescales every feature to the [0,1] interval.

Table 2. Attributes after normalization

	dur	splts	dpkts	sbytes	bdytes	rate
mean	0.021	0.002	0.002	0.005	0.009	0.091
min	0	0	0	0	0	0
max	1	1	1	1	1	1

Therefore, this study solved the problem of data jumpiness and improved the efficiency through the data preprocessing steps for the UNSW-NB15 data set, and is expected to further improve the accuracy rate of the model for the UNSW-NB15 data set through Ensemble Models.

3. Experimental Settings

This section describes the experimental settings in our experiment. First, we introduce the basic ML models in the experiment (Part A). Then we integrate the basic ML models to build the ensemble models (Sec. B). Lastly, we have to evaluate the model (Sec. C).

3.1. Basic Machine Learning Models

To build the model of integrated learning, we build 8 basic machine learning models to find the most suitable classification model for this group of data through multi-model comparison and use multiple training models to realize a multi-model fusion of test data based on the method in the figure, to find a model with higher accuracy and feasibility. The basic ML models include the following 8 classification models (LightGBM) [6], Linear Discriminant Analysis (LDA) [7], Naive Bayes classifier (NB) [8], Support Vector Machines (SVM) [9], Random Forest (RF) [10], Classification and Regression Tree (CART) [11], K-nearest neighbors (KNN) [12], Linear regression (LR) [13]).

3.1.1. Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) [6] is a framework to implement the GBDT algorithm. Its main idea is to use decision tree to iteratively train to get the optimal model. The model also allows effective parallel training and benefits from shorter training times, cheaper memory requirements, improved accuracy, and support for distributed and quick processing of large amounts of data. To make the LightGBM model have better performance in the research, we conduct a Grid search on it in the implementation and get its best parameters by traversing, as shown in 0.

Table 3. Best parameter best_parameter after grid search

Feature	Values
bagging_freq	1
feature_fraction	0.8
lambda_l1	0.1
max_depth	5
n_estimators	50
num_leaves	16
bagging_fraction	0.7

3.1.2. Linear Discriminant Analysis

Linear discriminant analysis (LDA) [7] is a method used to find linear combinations of features to characterize or separate objects of two or more classes. The classifiers generated by LDA can be used

as linear classifiers and can help Attributes with dimensionality reduction operations to reduce algorithm complexity.

3.1.3. Naive Bayes classifier

Naive Bayes Classifier (NB) [8] is a supervised learning method. NB assumes that the value of an attribute has an effect on a given class independently of other attribute values. NB has the characteristics of high accuracy and high efficiency, low classification error rate, low time consumption, and low cost, and can well represent the causal dependence between attribute sets.

3.1.4. Support Vector Machines

Support Vector Machines (SVM) [9] is a binary classification model that associates some points in the space with the instance's feature vector. Drawing a line to "best" separate these two categories of points is the goal of SVM. Aiming at this research, we transform SVM to adapt to the problem of multiple classifications.

3.1.5. Random Forest

Random Forest (RF) [10] is an extended variant of Bagging. For this study, because there are many samples to be trained, RF can randomly select decision tree nodes to divide features, train weak classifiers for different sample sets, and the training can be highly parallelized to reduce the training model overhead.

3.1.6. Classification And Regression Tree

Classification And Regression Tree (CART) [11] is a kind of binary recursive segmentation technology. The binary Tree with a straightforward structure that the CART algorithm produces is a decision tree. For this experiment, since the decision of CART algorithm in each step can only be "1" or "0", it is easy to overfit the training set. Therefore, we avoid overfitting by controlling the termination condition of the tree structure by threshold in the experiment to avoid too thin branches.

3.1.7. K-nearest neighbors

K-nearest Neighbors (KNN) [12] is a classification method based on the similarity in the feature space. With this procedure, the category of the samples to be split simply depends on the category of the nearby sample or samples. Since the class domain with large sample size is used in this study, the sample points around the sample can be used to better identify the class to which the sample belongs.

3.1.8. Linear regression

Linear regression (LR) [13] is a machine learning method that maps consecutive values in attributes to the {0,1} space through Linear regression. For this research, LR can provide a concise method with less computation, faster speed and lower storage resources, which helps us to better find the linear decision boundary between different categories.

3.2. Ensemble Models

To obtain a classifier with higher accuracy and better interpretation, we use an ensemble model. The previous 8 machine learning are constructed and combined to further improve the research, and the classification model can also be compared. We use the following 3 ensemble learning methods: Voting, Weighted voting, and Stacking

3.2.1. Voting

Voting [14] is an integrated learning model that follows the rule of majority. The integration of multiple models reduces the variance and improves the robustness of the model. The hard voting method adopted in this study is classified voting. Voting discriminates the final classification results according to the clear category labels (Voting results) predicted by all base models, so that the research can get more accurate prediction results.

3.2.2. Weighted Voting

Weighted Voting [15] is a method of taking a Weighted sum of the probabilities of a predicted distribution and then taking the largest class of the probabilities. The non-negative weight weighting of each base classifier ensures that the ensemble performance is better than that of the single best individual learner. Weighted Voting is generally learned from training data.For our research, we give weight to the classifiers based on the accuracy of the model. Classification models with higher accuracy are given higher weight.

3.2.3. Stacking

Stacking [16] is a predictive ensemble machine learning algorithm that learns to best combine multiple basic machine learning algorithms. For this study, Stacking can utilize the ability of a series of models with good performance on classification or regression tasks to complete a good ensemble learning task.

3.3. Model Evaluation

We use accuracy, precision, recall and F1-score to evaluate the performances of the models we build.

3.3.1. Accuracy

Accuracy refers to the percentage of classes that are correctly divided into tuples, which reflects the classifier's ability to judge the sample space. Suppose that there are the following two types in the sample space:

Type 1: There are P samples with category i (i=1,2...,8) in total, assuming that category i is a positive example.

Type 2: There are P samples with category i (i,j=1,2...,8, i≠j) in total, assuming that category j is negative example.

For this multi-classification model, TP stands for tuples in category I are correctly classified into I category, FN stands for tuples in category I are incorrectly classified into J category, FP stands for prediction J category is actually I category, TN stands for j category which is actually j category.

$$\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{1}$$

3.3.2. Precision

Precision is the ratio of the abnormal traffic retrieved to the abnormal traffic of this type, which quantifies the retrieval system's accuracy.

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{2}$$

3.3.3. Recall

Recall rate refers to the ratio between the number of abnormal traffic of related categories and the number of abnormal traffic of all categories. It measures the recall rate of the retrieval system.

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} = \frac{\mathrm{TP}}{\mathrm{P}} \tag{3}$$

3.3.4. F1-Score

F1-score is introduced as a comprehensive metric to evaluate a classifier comprehensively in order to balance the influence of accuracy and recall. F1-score can balance the problem of excessive recall of classifiers by setting accuracy and recall to be the same in importance.

$$\mathrm{F}_{\propto} = \left(1 + \propto^2\right) \times \frac{\mathrm{precision} \times \mathrm{recall}}{\propto^2 \times \mathrm{precision} + \mathrm{recall}} \tag{4}$$

Results and Discussion

This section describes the model result and discussion. First, we compare the indicator of basic ML models (Sec. A). Then we carried on the research summary and the research prospect (Sec. B).

3.4. Model Comparison

To better verify the classification performance of each algorithm n multi-classification problems, the evaluation indexes of 11 model prediction results are output, including accuracy, recall, precision and F1-Score. Table Ⅴ shows the indicators of each classifier.

Table 4. The evaluation indicators of each classifier

Classifier	Accuracy	Precision	Recall	F1-Score
LightGBM	0.733	0.730	0.733	0.715
Linear Discriminant Analysis	0.661	0.701	0.660	0.650
Naive Bayes	0.480	0.690	0.480	0.536
Support Vector Machines	0.717	0.704	0.712	0.642
Random Forest	0.765	0.771	0.765	0.746
CART	0.748	0.744	0.748	0.733
KNN	0.720	0.705	0.720	0.697
Linear Regression	0.699	0.654	0.699	0.634
Voting	0.737	0.761	0.737	0.735
Weighted Voting	0.778	0.763	0.768	0.757
Stacking	0.768	0.763	0.768	0.753

According to the accuracy, we can see that the weighted voting model has the best performance and applicability among all models, and the random Forest is close to it. In ensemble learning, the accuracy of voting can be improved by adding weight, but stacking is more acceptable. Also, Recall, precision and F1-soc support this view.

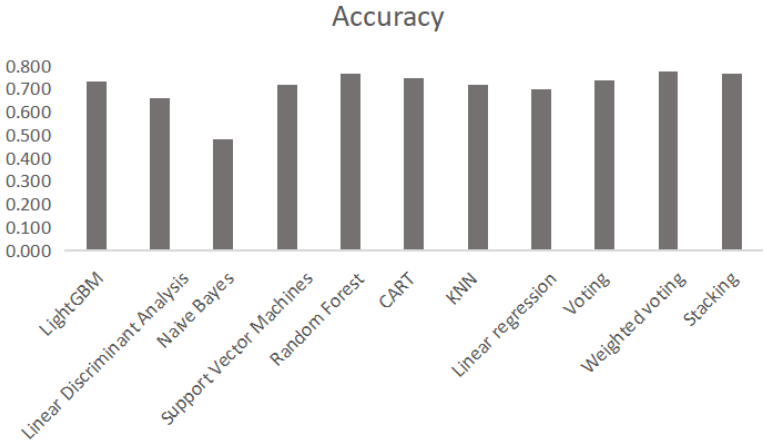

Figure 3. The accuracy of each classifier

To evaluate the performance of the classifier effectively, the confusion matrix is selected as a visual tool for the real recognition of different class tuples of the classifier. We output the confusion matrix of the weighted voting model. According to the confusion matrix, it can be seen that the model has a high accuracy for "Reconnaissance, Fuzzers, Analysis". In other classes, it can not be said that the correct rate is not high, because the tuple initially exists in the state of class imbalance.

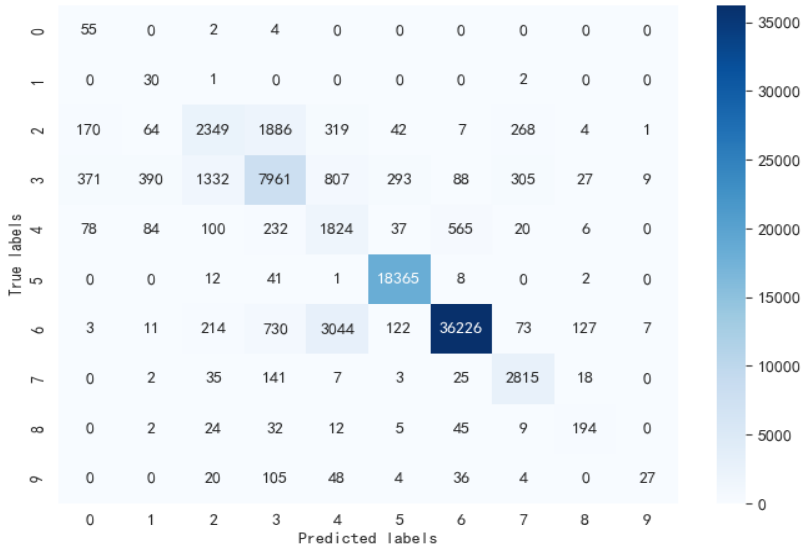

Figure 4. Confusion matrix analysis

4. Conclusion

This study presents the abnormal traffic based on UNSW data in many aspects. We establish a multi-method classification model. In addition, ensemble learning is carried out based on the established model, and the original ideal based on voting further improves the accuracy and availability of the model. In this paper, we suggest a method for assessing attack_CAT using information from several elements of anomalous traffic. And the accuracy of the model is acceptable in multi-classification models. Our experimental findings also support the existence of a model for classifying anomalous traffic using a variety of indications and categories. The research can follow ensemble learning, and the nodes of the model hold the predictive result in attack_cat. Meanwhile, our analyses illustrate that provide guidance for abnormal traffic definition. This study provides a multi-model classification study of abnormal traffic, comparing the accuracy and availability of different models.

In the future, we will study more high-dimensional unstructured data in abnormal traffic analysis, such as related pictures and videos.

References

[1] Ruiz-Correa, Salvador, Linda G. Shapiro, and Marina Melia. "A new signature-based method for efficient 3-d object recognition." Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Vol. 1. IEEE, 2001.

[2] Gurina, Anastasia, and Vladimir Eliseev. "Anomaly-based method for detecting multiple classes of network attacks." Information 10.3 (2019): 84.

[3] Hu Fengjie. Research on Network Intrusion Detection System Based on LightGBM[D].XidianUniversity,2020.DOI:10.27389/d.cnki.gxadu.2020.001824.

[4] He Hongyan, Huang Guoyan, Zhang Bing, and Chen Yu.Research on intrusion detection model based on multiple feature selection strategies[J].Information Security Research,2021,7(03):225-232.

[5] Rajagopal, Smitha, Poornima Panduranga Kundapur, and K. S. Hareesha. "Towards effective network intrusion detection: from concept to creation on Azure cloud." IEEE Access 9 (2021): 19723-19742.

[6] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).

[7] Izenman, Alan Julian. "Linear discriminant analysis." Modern multivariate statistical techniques. Springer, New York, NY, 2013. 237-280.

[8] Eitas, Timothy K., and Jeffery L. Dangl. "NB-LRR proteins: pairs, pieces, perception, partners, and pathways." Current opinion in plant biology 13.4 (2010): 472-477.

[9] Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

[10] Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25.2 (2016): 197-227.

[11] Lewis, Roger J. "An introduction to classification and regression tree (CART) analysis." Annual meeting of the society for academic emergency medicine in San Francisco, California. Vol. 14. Citeseer, 2000.

[12] Kramer, Oliver. "K-nearest neighbors." Dimensionality reduction with unsupervised nearest neighbors. Springer, Berlin, Heidelberg, 2013. 13-23.

[13] Seber, George AF, and Alan J. Lee. Linear regression analysis. John Wiley & Sons, 2012.

[14] Lau, Richard R., and David P. Redlawsk. "Voting correctly." American Political Science Review 91.3 (1997): 585-598.

[15] Banzhaf III, John F. "Weighted voting doesn't work: A mathematical analysis." Rutgers L. Rev. 19 (1964): 317.

[16] Ting, Kai Ming, and Ian H. Witten. "Stacking bagged and dagged models." (1997)