Automatic Feature Selection and Ensemble Classifier for Intrusion Detection

Changjian Lin1*, Aiping Li 1 and Rong Jiang 1
1 National University of Defense Technology, Changsha, Hunan, 410000, China
*Corresponding author’s e-mail: lcj007@nudt.edu.cn

Abstract. Anomaly-based Intrusion Detection System (ADS) is one of the technologies widely used in network topology. Although many supervised and unsupervised learning methods in the field of machine learning have been used to improve the efficiency of ADS, achieving good performance is still a challenging problem for existing intrusion detection algorithms. Firstly, there are few public datasets available for evaluation. Secondly, a single classifier may not perform well in detecting each type of attack. Third, some of the existing schemes focus on feature subset selection, while ignoring the design of the classification decision algorithm, or focus on the classification decision algorithm. In order to address this issue, a new intrusion detection framework is proposed by comparing and studying various feature selection technologies and classification decision algorithms in this paper. An automatic parameter adjustment scheme is designed for feature selection and ensemble classification. It avoids the need to obtain the optimal parameters through manual experiments in advance, and can improve the robustness of the parameters and the model. We use the most classic NSL-KDD dataset and the latest CICIDS2018 dataset for comparative experiments. The experimental results demonstrate its efficiency in terms of Accuracy and False Positive Rate.

1. Introduction
In order to maintain the confidentiality, integrity and availability of network systems, intrusion detection systems (IDS) are deployed in various distributed systems. IDS can extract and understand features of data streams such as log data, IoT data and network traffic. Anomaly-based IDS (ADS) is a popular research direction, but they cannot avoid a high false positive rate.

Machine learning (ML) [2] is widely used in ADS. Artificial neural network is a machine learning algorithm that converts input into output through nonlinear latent processing of a group of artificial neurons. These methods are divided into shallow and deep neural networks.

There is a problem in the application of ML in ADS. A single classifier may not be able to deal with various and ever-changing attack patterns. Combination-based technologies have advantages because they have higher accuracy rate than single technology.

Feature selection has been widely proven to significantly improve the accuracy and speed of ADS when mining huge data. There are three main models for feature selection: wrappers, filters, and embedded methods. Models based on Mutual information, Chi-squared, ANOVA F-value are popular filter models. The three models will rank the features and give the values of the features. However, the feature selection model and parameters cannot be changed over time, which may have a great impact on the applicability of the model for the emerging new attack types.

In this paper, we propose a novel anomaly-based Intrusion Detection System (ADS). First, we use CICFlowMeter-V3 for feature extraction on log data and network traffic. Due to data redundancy,
unavailability, diversity and other factors, we need to converse and clean the extracted features. Second, we select three popular feature selection metrics, including Mutual information, Chi squared [7], ANOVA F-value and embed the faster-running random forest algorithm [1]. It can realize automatic parameters selection. Finally, we use three classification algorithms in parallel. Three classifiers include online unsupervised learning technology of SAE [4], offline supervised learning technology of CNN and Random Forest (RF) [1]. The parameters of the three classifiers are also adjusted through automatic parameter optimization.

Our first and core contribution is to realize the automatic adjustment and optimization of multiple parameters in the feature selection stage and classification decision stage in real time. It avoids the need to obtain the optimal parameters of the corresponding dataset through manual experiments in advance, and can improve the robustness of the parameters and the model. In the field of intrusion detection, automatic parameter adjustment and optimization have not yet been proposed.

Our second contribution is to propose a new framework of based on ADS. We combine feature selection and ensemble classification decision algorithms. The detection accuracy is greatly improved and the false positive rate of attacks is reduced. At the same time, the operating speed of the system is greatly improved.

Our third contribution is to use ten-fold cross-validation in parameter optimization, feature selection, and classification decision algorithms. It can reduce over-fitting and improve the true reliability of experimental results.

2. Related work

Public datasets may not be a problem for other fields. However, in this IDS field there are few public IDS datasets available for evaluation. By searching and comparing IDS datasets [5], We selected two datasets, namely NSL-KDD and CICIDS2018 dataset. The NSL-KDD is the classic and revised dataset, and the CICIDS2018 is the latest and most comprehensive dataset. [5] also introduces the datasets related to advanced persistent threat (APT). The datasets cover large-scale enterprise networks and cyber-physical systems and the Internet of Things. The only pity is that the APT datasets of the last three years have not been publicly available for download, which has brought difficulties to our subsequent verification of APT detection.

The research on feature selection methods in ADS is one of the hot spots. [3] uses wrapper-based feature selection. The algorithm speeds up the process of selecting important features. However, the ability to identify new attacks is relatively poor, and the detection accuracy is only 30%. Firefly algorithm deploys filter-based and wrapper-based methods to select features. The final result can reach 99.5% of DOS attack accuracy. However, the accuracy for other forms of attacks varies greatly, and the accuracy for U2R attack is only 17.5%.

In terms of classification algorithm research, Kitsune [4] uses an ensemble of autoencoders. The core algorithm of Kitsune is an anomaly detection algorithm called KitNET. KitNET uses small neural networks called autoencoders. Because it is a small neural network, it must be faster than deep neural networks in terms of training and execution speed. However, KitNET must be based on an assumption: all instances encountered in training mode are normal traffic, which is also a disadvantage of unsupervised learning. If unsupervised learning and supervised learning are integrated, this shortcoming can be compensated. Ensemble of Advanced Learners [1] uses 3 different methods to train ADS. The advantage is that it can produce acceptable performance under a zero-day attack. The classification method is an ensemble classifier that combines two deep neural networks and random forest. But like KitNET, it does not use feature selection technology to effectively reduce feature data.

TSE-IDS [6] proposes an improved system. The accuracy reaches 85.79% (using NSL-KDD dataset), but the false positive rate reaches 11.7%. For anomaly detection, every false alarm will result in a waste of computing resources and manpower. The performance of latest efficient ADS [8] is greatly improved than previous algorithms, but it has not been tested on the latest CICIDS2018 intrusion detection dataset. Therefore, there is a certain doubt about the scalability of the experimental results.
3. Proposed system design
Ensemble method is a prolific field in machine learning. As shown in figure 1, the model we propose is: first in data processing, data transformation and data cleaning, then automatically select the optimal parameters and number of features, and finally use three classifiers in parallel in the decision model. Three classifiers include unsupervised learning technology of SAE [4], supervised learning technology of CNN and RF. The parameters of the three classification algorithms are also adjusted through automatic parameter optimization.

3.1. Data preprocessing
To data transformation, the main steps are numericalization and normalization of the features. Numericalization embed categorical features into metric space. To data cleaning, the first is to correct and modify the data, and then use sub-sampling to generate a dataset with the required ratio.

3.2. Automatic parameter optimization
In the field of deep learning, the adjustment and optimization of hyperparameters should not be human work, especially when the system has been applied to practice. And it is best to leave it to the machine to do without human involvement at all. In the latest intrusion detection research [5], various parameters such as tree=60 are obtained through repeated manual testing and comparison. But it can only be said to be semi-automatic. As shown in Figure 2, we write the parameters calculated by the long-term machine to the memory file, and then read the new parameters into the system in real time. In this way, the self-learning ability and robustness of the intrusion detection system can be greatly improved.

3.3. Feature selection
Feature selection is regarded as an essential preprocessing step to significantly improve overall system performance when mining huge data. For different data, a single feature selection model and
parameters may affect the performance metrics of the entire system. Therefore, we select three popular feature selection metrics, including Mutual information, Chi-squared [7], ANOVA F-value, and use the faster-running random forest [1] as a classification strategy, as shown in Figure 3. It can also realize automatic selection of feature selection parameters and features.

![Figure 3. Feature selection model.](image)

### 3.3.1. Mutual information

The mutual information $I(X; Y)$ between the variable representing the dataset X and the class labels Y is defined as:

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

(1)

### 3.3.2. Chi-squared

Chi-square test is a numerical test, which is suitable for feature selection of intrusion detection set. The chi-square metric is defined as:

$$\text{chi square metric} = t(t_p, (t_p + f_p)p_{pos}) + t(f_n, (f_n + t_n)p_{pos})$$
$$+ t(f_p, (f_p + t_p)p_{neg}) + t(t_n, (t_n + f_n)p_{neg})$$

among them: $t(\text{count, expect}) = (\text{count} - \text{expect})^2/\text{expect}$

### 3.3.3. ANOVA F-value

In traditional statistics, the f value is used for analysis of variance. The formula for calculating f-value is as follows:

$$f = \frac{r_i^2}{1-r_i^2} * (n-2)$$

$$r_i = \frac{(x-\text{mean}(X))(y-\text{mean}(y))}{\sigma(x)\sigma(y)}$$

(4)

### 3.4. Ensemble Classifiers

Ensemble technologies have advantages because they have higher accuracy and detection rate than single technology. Therefore, when the current single method cannot make a breakthrough, we can optimize the performance of the existing classification algorithms and add the necessary modules to achieve better performance. We use support vector machines (SVM), k-nearest neighbors (KNN), random forest (RF), recurrent neural networks (RNN), convolutional neural networks (CNN), and deep autoencoders and simple autoencoder ensemble (SAE) for comparison experiments. We finally choose to use three classifiers in parallel, namely SAE, CNN and RF. We have also realized the automatic selection of parameters of classifiers, which are consistent with the parameters obtained by manual repeated experiments.

### 3.4.1. Simple autoencoder ensemble

SAE described in Kitsune [4] is to learn effective codes in an unsupervised way. The simplest architecture involves an input layer, more than one hidden layer, and an output layer that has the same number of neurons in the input layer for reconstruction. Since
unsupervised algorithms are more attractive when new attacks need to be detected, their main advantage is that they do not require labeled datasets for training. As shown in figure 4, it is the architecture of Kitsune’s anomaly detection algorithm [4].

![Figure 4. The architecture of Kitsune.](image)

3.4.2. **Convolutional neural network.** CNN has a good classification effect. Therefore, a CNN is embedded in the proposed system. We have improved the LeNet-5 version and modified the parameters. The specific structure is shown in figure 5.

![Figure 5. The specific LeNet-5 Structure.](image)

3.4.3. **Random forest.** Random forest (RF) overcomes the shortcomings of decision tree (DT) over-fitting and possible non-convergence through ensemble learning methods. Compared with SAE and CNN, RF has faster running speed and can be used as a useful supplement.

4. **Results comparison**

Based on the two datasets we selected, we achieve a comparative experiment on the real-time automatic optimization of parameters, feature selection and ensemble classifier that we proposed. The experiments are performed by Tensorflow v.2.0.0 and scikit-learn v0.21.3 on desktop PC with 3.4 GHz Intel Core i3-3240 processor and 16GB RAM.

4.1. **Datasets introduction**

In order to design and verify an effective ADS, a public intrusion detection dataset is needed. Table 1 lists the comparisons of the IDS datasets that can be downloaded and used. As a challenging and popular APT attack detection, this paper will not test and verify the 2013 APT datasets we have found. We take APT attack detection as one of our future research directions.

| Datasets          | Realistic network | Attack types | Input type | Labelled | Release time |
|-------------------|-------------------|--------------|------------|----------|--------------|
| Configuration     | Traffic           |              | Log\(^a\)  | Netw\(^b\)|              |
| KDD-99            | True              | False        | 4          | True     | True         | 1999        |
| NSL-KDD           | True              | False        | 4          | True     | True         | 2004        |
| UNSW-NB15         | True              | True         | 9          | False    | True         | 2015        |
| NGIDS-DS          | True              | True         | 7          | True     | True         | 2016        |
| TRAbID            | False             | True         | 2          | False    | True         | 2016        |
| CICIDS2017        | True              | True         | 7          | False    | True         | 2017        |
| CICIDS2018        | True              | True         | 7          | True     | True         | 2018        |

\(^a\) Log represents log data.

\(^b\) Netw represents network traffic data.
In view of the above available datasets, this experiment uses the classic and revised dataset of NSL-KDD, and the latest and most comprehensive dataset of CICIDS2018. The introduction of these two datasets will be described in the following.

4.1.1. The classic and revised dataset of NSL-KDD. NSL-KDD dataset solves the duplicate records and some statistical problems in the KDD99 dataset. It has three predefined datasets:

- Training dataset: “KDDTrain +”;
- Testing dataset: “KDDTest +”;
- The challenging testing dataset: “KDDTest-21”. It consists of all records misclassified by 21 designed machine learning algorithms in “KDDTest” dataset.

4.1.2. The latest and comprehensive dataset of CICIDS2018. The CICIDS2018 dataset contains benign and latest common attacks. There are seven different attack scenarios. The dataset includes captured network traffic and system logs. And has a total of 10 days of data.

4.2. Evaluation metrics for IDS

This paper uses k-fold cross-validation in metrics calculation. The result obtained in this way is the average of all tests, making the experimental results more universal. The evaluation metrics are defined as follows:

- Accuracy: The percentage of the total records correctly classified is defined as follows:
  \[ \text{Acc} = \frac{TN+TP}{TN+FN+FP+TP} \]  \hspace{1cm} (5)

- Detection rate: The percentage of correctly identified attack records is defined as follows:
  \[ DR = \frac{TP}{TP+FN} \]  \hspace{1cm} (6)

- False Positive Rate: FPR is the ratio of false attack alarms to all false identifications defined as follows:
  \[ FPR = \frac{FP}{FP+TN} \]  \hspace{1cm} (7)

4.3. Results and Discussion

Our experiment mainly includes three aspects, one is the automatic optimization of parameters, the second is feature selection, and the third is the ensemble of classifiers. In order to compare the performance of our proposed system, we used the performance data of related systems in previous papers. These results are shown in Table 2.

In Table 2, we compare the performance of the proposed classifier with previous studies (training with the dataset KDDTrain+ and testing with KDDTest+ and KDDTest-21). Based on the experimental results of KDDTest+, the proposed method achieves an accuracy of 85.5%, which is better than the latest anomaly-based IDS technology, namely SVM. Compared with Two-stage ensemble [6], the accuracy is only 0.29% behind. But the classifier we proposed is only 4.69% in FPR metric, which is much lower than 11.7% of Two-stage ensemble [6]. This greatly improves the efficiency of anomaly-based intrusion detection system. Because the time to analyse each alarm is time-consuming and expensive. According to the classification results applied to the more difficult KDDTest-21, the proposed method is significantly better than the classifiers in the current literature. Since most of the CICIDS2017 dataset are used, they are still comparable due to the similarity of attack types. We use the Thursday15-02-2018 dataset and the CICIDS2017 Wednesday dataset, both of which have DOS-GoldenEye and DOS-Slowloris attack styles. Here, we have considered other existing methods, such as [8]. Our proposed classifier yields a performance accuracy of 99.99%, which is much higher than other classifiers. We only have 0.00004% in terms of FPR.
Table 2. Comparison of baseline datasets (Publicly available).

| Methods       | Year | Feature selection | Accuracy (%) | FPR (%) | Dataset       |
|---------------|------|-------------------|--------------|---------|---------------|
| Proposed      | 2020 | Automatic         | 85.5         | 4.69    | KDDTest+      |
| Two-stage ensemble | 2019 | Hybrid            | 85.79        | 11.7    | KDDTest+      |
| SVM           | 2019 | MBGWO             | 81.58        | -       | KDDTest+      |
| Two-tier classifier | 2017 | LDA               | 83.24        | 4.8     | KDDTest+      |
| GAR-Forest    | 2016 | No                | 82.39        | 14.3    | KDDTest+      |
| Fuzzy classifier | 2011 | No                | 82.74        | 3.9     | KDDTest+      |
| SVM           | 2009 | No                | 69.52        | -       | KDDTest+      |
| Proposed      | 2020 | Automatic         | 72.98        | 6.63    | KDDTest-21    |
| Two-stage ensemble | 2019 | Hybrid            | 72.52        | 18      | KDDTest-21    |
| Decision      | 2012 | RCDFT             | 58.80        | 27.67   | KDDTest-21    |
| Two-tier classifier | 2009 | No                | 66.16        | -       | KDDTest-21    |
| Proposed      | 2020 | Automatic         | 99.99        | <0.001  | CICIDS2018(Wed.15.2) |
| Proposed      | 2020 | Automatic         | 99.99        | <0.001  | CICIDS2017(Wed.) |
| XGBoost-IDS   | 2018 | No                | 91.36        | 12      | CICIDS2017(Wed.) |
| Autoencoder   | 2019 | LDA               | 95.73        | 4.32    | CICIDS2017(Wed.) |

Best results with 10f validation

The comparative analysis in Table 2 shows that this method is highly competitive as an effective method for anomaly-based intrusion detection tasks. In the next three parts, we introduce the three experimental results we proposed. They are automatic parameter selection, feature selection and the ensemble of classifiers.

4.3.1. Automatic parameter optimization based on k-fold cross-validation. GridSearchCV in Sklearn is a very useful tool in parameter optimization and feature processing. The disadvantage is that when the amount of data is large, the calculation cost is very high. Therefore, in our proposed model, the automatic parameter adjustment is placed on the parallel program, and when the program calculates the parameters, the previous parameters are replaced. As showed in figure 6, the optimal parameters we obtained are basically the same as those obtained from repeated experiments, and also verify the correctness of the automatic parameter optimization.

At the same time, we can directly read the content of figure 6, realize zero participation of operators, and apply the latest and optimal parameters to the proposed system in real time.

![Figure 6](image-url)

Figure 6. The automatic parameter optimization results on RF (10-fold).

4.3.2. Results of feature selection. We use three feature selection techniques, and used the automatic selection and parameter optimization mentioned above. The results on NSL-KDD dataset for each selection technique are presented in figure 7. And we can clearly see that when n=10 and technology=Mutual information, the classification accuracy of feature selection is the best. For the CICIDS2017 and 2018 datasets, we only sample 13 features and achieved an accuracy of 99.99%.
4.3.3. Results of the proposed system. The system we proposed combines automatic feature selection and the ensemble of three classifiers. Figure 7 and figure 8 show the accuracy and FPR of the system without and with automatic feature selection (AFS). We can see that the accuracy with feature selection increased by 3.04% and the FPR decreased by 1.49%.

We also do an ablation study. As shown in Figure 8, we use SAE algorithm, CNN algorithm and RF algorithm separately for the same dataset, and the accuracy is 81.8%, 80.2% and 80.2% respectively. It can be seen that the significant advantages of ensemble learning. Through ensemble classifiers, three single classifiers have also played a better performance.

5. Conclusion and Future work
The download of seven intrusion detection datasets and two APT datasets has laid a public data foundation and benchmark for our research. Through experimental comparison, it can be concluded that our proposed intrusion detection system based on automatic feature selection and ensemble classifiers has superior performance. In particular, the FPR is significantly reduced, which will greatly improve the efficiency of the anomaly detection system. For future work, the advanced persistent threat (APT) is the latest and hottest intrusion detection research problem. We will conduct follow-up research based on the two downloaded APT datasets.

Acknowledgments
The work described in this paper is partially supported by the National Key R&D Program of China (No. 2016QY03D0603, 2019QY1406, 2017YFB0802204, 2017YFB0803303), the Key R&D
References

[1] Andalib A and Vakili V T 2020 An autonomous intrusion detection system using ensemble of advanced learners Preprint arXiv/2001.11936.

[2] Berman D S, Buczak A L, Chavis J S and Corbett C L 2019 A survey of deep learning methods for cyber security Information 10(4) 122.

[3] Li Y, Wang J L, Tian Z H, Lu T B and Young C 2009 Building lightweight intrusion detection system using wrapper-based feature selection mechanisms Computers & Security 28(6) 466–475.

[4] Mirsky Y, Doitshman T, Elovici Y and Shabtai A 2018 Kitsune: an ensemble of autoencoders for online network intrusion detection Preprint arXiv/1802.09089.

[5] Stojanović B, Hofer-Schmitz K and Kleb U 2020 Apt datasets and attack modeling for automated detection methods: A review Computers & Security 92 101734.

[6] Tama B A, Comuzzi M and Rhee K H 2019 Tse-ids: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system IEEE Access 7 94497–507.

[7] Thaseen I S and Kumar C A 2017 Intrusion detection model using fusion of chi-square feature selection and multi class svm Journal of King Saud University-Computer and Information Sciences 29(4) 462–72.

[8] Zhou Y, Cheng G, Jiang S and Dai M 2020 Building an efficient intrusion detection system based on feature selection and ensemble classifier Computer Networks p 107247.