Research on Intrusion Detection Method Based on Pearson Correlation Coefficient Feature Selection Algorithm

Pengtian Chen, Fei Li*, Chunwang Wu
School of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China

* lifei@cuit.edu.cn

Abstract. The current era is the era of big data and 5G. The network security data in the network is different from the past, and the network security data is growing exponentially. As an important line of defense for network security, intrusion detection technology can efficiently detect and process massive amounts of security data has become an important factor restricting its development. The feature selection method of intrusion detection data directly affects the efficiency of intrusion detection. Therefore, this paper proposes a feature selection algorithm based on pearson correlation coefficient, which performs feature specification on many features, which greatly reduces the amount of security data that needs to be processed, and effectively reduces the dimensionality of the data to increase the intrusion detection efficiency.

Keywords: Feature Selection; Intrusion Detection Technology; Pearson Correlation Coefficient

1. Introduction
With the rapid development of the Internet and computer systems, there are a lot of tricky security issues. Among these security issues, the most mentioned and the most harmful is a series of security issues caused by intrusion attacks. The number of intrusion attacks is also increasing year by year. Any intrusion or malicious attack directed at the loopholes of the network or computer system may have a devastating blow to the computer system. Therefore, the research on network threats and information security is still a hot spot at present and even in the future.

Intrusion detection is the process of detecting incidents that violate security policies by detecting computer networks and systems [1]. Intrusion detection systems are used to detect and analyze network behaviors and identify any deviations from normal events [2]. To classify intrusion detection systems based on detection analysis technology, it mainly includes abnormal intrusion detection systems and misused intrusion detection systems. The intrusion detection system based on misuse can detect existing attack behaviors well by constructing a rule base of normal behavior and attack behavior, with good accuracy and false alarm rate, but the disadvantage of this detection method is that it cannot be effectively new types of attacks are detected. The anomaly-based intrusion detection system establishes a detection model based on normal network behaviors in historical data, and attempts to detect deviations from normal behaviors. When the deviation exceeds a certain threshold, the system defines abnormal behaviors. This detection method can detect new and unknown abnormal behaviors, but the disadvantages are high false alarm rate and low accuracy rate.
Intrusion detection mainly takes security log behavior and information operation of audit data as the entry point to effectively detect related behaviors that break into the system. Through in-depth analysis and research on the intrusion detection system, it is found that the system is mainly based on computer applications. Basis, evaluate different loss situations and test results, provide more technical basis and support for later practice.

In order to improve the performance of intrusion detection systems, a variety of technologies have been applied to the intrusion detection research center, such as feature selection technology. This paper proposes a feature selection algorithm based on pearson correlation coefficient, which reduces the selection of features as much as possible while ensuring their independence. Under the premise of ensuring a certain detection rate, the data is effectively reduced in dimension to improve Intrusion detection efficiency.

2. Feature Selection

The data set used for analysis may contain hundreds of features (or attributes), most of which may be irrelevant or redundant with our intrusion detection system tasks. Feature selection is to remove irrelevant, weakly related or redundant features from the original features, or reduce the number of features by reorganizing the features, and then find the smallest feature set, which is the probability distribution of the data as much as possible Close to the original distribution obtained by using all features. The principle is to reduce the dimensionality of the feature vector as much as possible while retaining or even improving the original discrimination ability. The input of the feature selection algorithm is a set of features, and the output is a subset of the set of features. Feature selection generally includes 3 steps:

⑴ Search: Search for feature subsets in the feature space. Each subset is called a state and consists of selected features.

⑵ Evaluation: Input a state (subset), output an evaluation value through an evaluation function or a preset threshold, so that the evaluation value reaches the optimal value.

⑶ Classification: Use the final feature set to complete the classification algorithm
We usually use statistical significance tests to determine the "best" and "worst" features. The basic methods of feature selection are:

⑴ Step by step forward selection, that is, starting from an empty feature set, determine the "best" feature in the feature set, add it to the selection set, and iterate continuously, adding the remaining "best" features in the original feature set .

⑵ Deleting backwards step by step, that is, starting from the entire feature set, continue to delete the "worst" feature in the feature set, and iterate repeatedly.

⑶ The combined method, that is, a combination of stepwise forward selection and stepwise backward deletion methods. Each time selected a "best method", and each iteration deleted a "worst" feature from the remaining features.

⑷ Decision Tree Induction, that is, at each node, the algorithm will select the "best" feature and divide the data into categories. Each internal node represents a test on a feature, and each branch corresponds to a result of the test. Each external node represents a class prediction.

However, the prerequisite of the feature selection algorithm based on statistical significance test is to assume that the attributes are mutually independent, but in actual events, there is often a certain correlation between the attributes. For this reason, we propose a feature selection algorithm based on pearson correlation coefficient, and introduce pearson correlation coefficient to measure the degree of correlation between features.

3. Pearson Correlation Coefficient

The development of artificial intelligence, big data, the Internet of things and other professional fields is very rapid. If the teaching system cannot keep pace with The Times and evolve iteratively, it is not conducive to the cultivation of students' knowledge, ability and quality employment, the development
of teachers' ability and the development of scientific research, and ultimately, it is not conducive to the cultivation of high-level talents in universities. Therefore, the teaching system should be adjusted in time with the development of related professional fields.

On the basis of big data analysis platform, data mining analysis is regularly performed for teaching and scientific research, relevant policy, social services, teachers and students feedback, and evolution of teaching system optimization scheme is automatically generated, then through the analysis of the expert group, man-machine integration mechanism is adopted to define the teaching system adjustment scheme, teaching system adaptive evolution system logic architecture is shown in the following figure:

The Pearson correlation coefficient is used for measuring whether two data sets are on a line, and it is used to measure the linear relationship between the distance variables. The Pearson correlation coefficient is defined as the Pearson correlation coefficient between rank variables. For a sample with a sample size of n, n original data are converted into grade data, and the correlation coefficient r is

$$r_{xy} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2}\sum (y_i-\bar{y})^2}$$  \hspace{1cm} (1)

r describes the degree of linear correlation between two variables. The value of r is between -1 and +1. If r=0, it indicates that the two variables are positively correlated, that is, the greater the value of one variable, the greater the value of the other variable; if r<0, it indicates that the two variables This variable is negatively correlated, that is, the larger the value of one variable, the smaller the value of the other variable. The larger the absolute value of r, the stronger the correlation.

The greater the absolute value of the correlation coefficient, the stronger the correlation: the closer the correlation coefficient is to 1 or -1, the stronger the correlation, and the closer the correlation coefficient is to 0, the weaker the correlation.

Under normal circumstances, the correlation strength of the variable is judged by the following value range:

- Correlation coefficient 0.8-1.0 very strong correlation
- 0.6-0.8 strong correlation
- 0.4-0.6 moderate correlation
- 0.2-0.4 weak correlation
- 0.0-0.2 Very weak correlation or no correlation

For the correlation coefficient r of x and y:

- When r is greater than 0 and less than 1, it indicates a positive correlation between x and y
- When r is greater than -1 and less than 0, it indicates a negative correlation between x and y
- When r=1, it means that x and y are completely positively correlated, and r=-1 means that x and y are completely negatively correlated.
- When r=0 means x and y are not related

After obtaining the pearson correlation coefficient of each feature value, we will perform cluster analysis based on its value in order to achieve feature selection and complete data dimensionality reduction. Because the clustering basis we chose is the Pearson correlation coefficient, our clustering method is hierarchical clustering (systematic clustering), and because of its clustering result representation, it is also called tree clustering.

Hierarchical clustering is based on the distance between data, through a hierarchical structure, the data is repeatedly aggregated to create a level to decompose a given data set. Hierarchical clustering algorithms are often used for automatic grouping of one-dimensional data.

Hierarchical clustering algorithm is a very intuitive clustering algorithm. The basic idea is to reconnect nodes after sorting them from high to low based on the similarity between data. The whole process is to build a tree structure.

The steps are as follows:

1. Each data point is a separate class
② Calculate the distance between points (similarity)
③ Connect pairs according to the distance from small to large (from strong to weak similarity) (after the connection, use the mean value of the two points as the new class to continue the calculation), and get the tree structure

After completing the cluster analysis, we selected the most representative features in each category based on the value of the pearson correlation coefficient and included them in our new feature set to obtain a new feature subset. Efficient and reliable dimensionality reduction is performed on the original features.

4. Experiments and Analysis

4.1. Lab Environment
The host used in this experiment is Windows 10 Professional operating system, the processor is AMD Ryzen 5 2600X Six-Core Processor@4.10 GHz, and the memory is 16GB. The simulation experiment was performed on MATLAB R2018b.

4.2. Experimental Data
In order to ensure the timeliness and integrity of the data in intrusion detection, and avoid errors that lead to the collection of wrong data, the performance of our intrusion detection is reduced or even meaningless. Due to the limitations of the experiment, it is difficult to collect various attack data, so this experiment uses the UNSW-NB15 [3-4] dataset as our data source. Compared with KDD99, UNSW-NB15 includes contemporary covert attack methods, which can more completely reflect the real situation of contemporary networks.

The sample features in the UNSW-NB15 dataset have a total of 49 dimensions as shown in Table 1. Since the feature algorithm based on the pearson correlation coefficient in this experiment can only process numerical data, and the original data in our standard data set contains discrete features of nouns, certain preprocessing is required:

1) Convert discrete features into continuous numerical features. According to the composition of each noun in the same variable, the corresponding numerical value is assigned, and the numerical value replaces the noun.

2) Numerical normalization. Let max and min be the maximum and minimum values of a feature in the data set. Suppose the normalized value range is $[\text{min}',\text{max}']$, then the mapping relationship from the original value $u$ to the new value $v$ is:

$$v = \frac{\text{max}' - \text{min}'}{\text{max} - \text{min}} (u - \text{min}) + \text{min}'$$

In most cases, the normalized value range can be selected from [-1,+1] and [0,1]. In this experiment, we choose [0,1] as the normalized value range.
Tab. 1 Basic characteristics of UNSW-NB15 data set

Take the feature 5proto (Transaction protocol) as an example to show the data preprocessing of discrete features containing nouns. The composition of each noun in feature 5 is shown in Table 2.

| Proto      | Frequency | Percentage | Valid Percentage | Cumulative Percentage |
|------------|-----------|------------|------------------|-----------------------|
| Valid      | 62        | .6         | .6               | .6                    |
| arp        | 14        | .1         | .1               | .8                    |
| icmp       | 48        | .5         | .5               | 1.2                   |
| ospf       | 7512      | 75.1       | 75.1             | 76.4                  |
| tcp        | 2364      | 23.6       | 23.6             | 100.0                 |
| Total      | 10000     | 100.0      | 100.0            |                       |

Tab. 2 Features 5 Noun Composition

We assign 1 to represent arp, 2 to icmp, 3 to ospf, 4 to represent tcp, and 5 to represent udp, so as to preprocess the data containing discrete features of nouns and convert them from character to numeric.

4.3. Data processing
There are 9 kinds of attacks in the UNSW-NB15 data set, namely: Analysis, Backdoor, DoS, Exploits, Fuzzers, Genercill, Reconnaissance, Shellcode, Worms. Table 3 and Table 4 respectively show the number of 9 types of records in the UNSW-NB15 data set network events and their distribution in the Training-set and Testing-set of the UNSW-NB15 intrusion detection data set.
An efficient intrusion detection system requires not only a suitable feature selection algorithm but also a suitable feature classifier. In recent years, the development of machine learning has greatly improved the performance of the classifier. Therefore, in order to verify the feature algorithm proposed in this article, we choose a machine learning algorithm classifier for classification and recognition. In order to fully verify the classification accuracy of various machine learning algorithm pairs, this article uses the Classification Learner toolbox that comes with Matlab2018 to train the data in the Training-set dataset. The training results of each classifier are shown in Table 5. In order to reduce the network training time, the entire training process adopts parallel acceleration for training. In addition, to ensure that the randomness of operating environment factors is minimized, the running time and recognition accuracy of different machine learning models are taken as the average of the two runs.
Tab. 5 Comparison of recognition performance of different machine learning models

From Table 5, it can be learned that the correct recognition rate of different machine learning for intrusion detection is 66.1%-87.3%. The Bagged Trees classifier has an average recognition accuracy rate of 87.3% and the highest recognition accuracy. The average running time is 85.2905s. The test results are shown in Fig.1 Shown.

Fig. 1 Bagged Tree test results

The abscissa of the test evaluation matrix represents the type of intrusion predicted by the machine learning classification model. The ordinate represents the actual intrusion type of the data in the Training-set data set. The diagonal green element represents when the predicted type is consistent with the actual type, it means the recognition is correct. The color of the element represents the type of intrusion test. The amount of result data, the more the amount of data, the darker the color, and vice versa, the lighter the color. The remaining red elements represent the total number of misidentified
cases, and their color patterns are consistent with the green elements. Therefore, we chose Bagged Tree as our classification model for this experiment.

According to the Pearson correlation coefficient formula, the correlation coefficient between the basic features in the Training-set data set is calculated, and the cluster analysis is performed according to the correlation coefficient. Draw its cluster analysis diagram as shown in Fig.2.

![Cluster Analysis Diagram](image)

**Fig. 2 Cluster analysis diagram**

In order to select the most representative feature in the divided clusters, based on the pearson correlation coefficient of the feature and the Label feature, the feature with the high correlation coefficient in the cluster is added to the new feature set. From this we get a feature subset \{spkts, dpkts,
In order to compare with commonly used data dimensionality reduction methods, we performed principal component analysis on the data set, and obtained a subset of contrast features \{service, dbytes, sload, sloss, sjit, dwin, tcprtt, ct_src_dport_ltm, ct_ftp_cmds, ct_flw_http_mthd, is_sm_ips_port\}.\[6\]

4.4. Analysis of results
The detection effect of all features, the detection rate and false alarm rate of the selected feature subset and the main component feature selection feature set are compared. The results are shown in Table 6. Before and after feature selection, the detection rate and false alarm of the method proposed in this paper remain basically unchanged, which is consistent with the theoretical role of feature selection, while the detection rate of PCA feature selection has decreased and the false alarm rate has increased. And because of our effective dimensionality reduction of the data, the detection efficiency has been improved. In this experiment, the detection time was shortened by nearly 46%, and the detection efficiency was greatly improved.

| Feature selection scheme | Detection rate | False alarm rate | Detection time/s |
|--------------------------|----------------|------------------|------------------|
| All features + Bagged Tree | 87.3% | 1.24% | 85.2905 |
| PCA feature selection + Bagged Tree | 81.3% | 2.62% | 46.254 |
| Pearson feature selection + Bagged Tree | 87.3% | 1.25% | 45.451 |

Tab. 6 Comparison of detection results before and after feature selection

Fig.3 shows the detection rate of each intrusion category before and after feature selection. The results show that the detection rates of other categories are all except for the slightly reduced detection rates of Exploits, Fuzzers, Normal, Reconnaissance, and ShellcodeJ, while the total detection rate remains unchanged. Improve or remain unchanged. Especially for the Dos and Worms categories, the detection rate has been significantly improved.

5. Conclusion
In this era full of data, the surge in data volume has pros and cons. How to complete as much effective analysis as possible in a limited time requires us to process the data in detail. The feature selection algorithm based on pearson correlation coefficient proposed in this paper not only reduces the amount of data, but also reveals the relationship between the data. Under the premise of ensuring a certain recognition rate, the data is effectively reduced in dimension to improve the efficiency of intrusion detection.

Acknowledgments
This work was supported by Natural Science Foundation of Sichuan Province under grant No.2019YFG0201, No.2018GZDZX0011, and Chengdu Science and Technology Project No.2018-YF05-00707-SN.

References
[1] Jiawei D July 2018. NETWORK SECURITY SITUATION AWARENESS COLLECTING, PERCEIVING AND PREDICTING. (Beijing: Machinery Industry Press)
[2] Guang S, Ying H and Xiuchun L, et al 2019 Methods for obtaining mutually independent features in network intrusion detection systems J. Computer Engineering and Applications, 2 125-143
[3] Moustafa N and Slay J 2016, The evaluation of network anomaly detection systems: Statistical analysis of the UNSW-NB15 dataset and the comparison with the KDD99 data set J. Information Security Journal A Global Perspective, 25(1-3) 18-31
[4] Moustafa N and Slay J 2015 UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 net-work data set) Proc of 2015 Military Communications and Information Systems Conference (MiCIS) pp1-6
[5] Zhu K and Zhang Q 2017 Application of machine learning in network intrusion detection J. Data Acquisition and Processing 32(3) 479-488
[6] Xinqian L, Chun D and Jiadong R, et al. 2019 Intrusion detection method based on multi-dimensional optimization of traffic anomaly analysis J. Journal of Information Security, 4(1) 14-26