Two-Level Feature Selection Method for Low Detection Rate Attacks in Intrusion Detection

Chundong Wang[1,2], Xin Ye[1,2(✉)], Xiaonan He[3], Yunkun Tian[3], and Liangyi Gong[1,2]

[1] Key Laboratory of Computer Vision and System, Ministry of Education, Tianjin University of Technology, Tianjin 300384, China
306187260@qq.com

[2] Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Ministry of Education, Tianjin University of Technology, Tianjin 300384, China

[3] Tianjin E-Hualu Information Technology Co., Ltd., Tianjin 300350, China

Abstract. In view of the fact that some attacks have low detection rates in intrusion detection dataset, a two-level feature selection method based on minimal-redundancy-maximal-relevance (mRMR) and information gain (IG) was proposed. In this method, irrelevant and redundant features were filtered preliminarily to reduce data dimension by using mRMR algorithm, and highly correlated features to low detection rate attacks were obtained based on the calculation of information gain, and finally these features were integrated together to get final feature subset. The experimental results showed that the classification result of the feature subset filtered by this method had a better classification performance than the current filtering methods and improved the testing results of some attacks with low detection rates effectively.

Keywords: Feature selection · Information gain · mRMR · Intrusion detection

1 Introduction

Network security is more and more prominent with the rapid expansion of Internet and computer technology. In recent years, intrusion detection system (IDS), which plays an increasingly important role in the network security engineering, has been widely studied. As a key technology of network security active defense system, intrusion detection can detect the malicious of network users without

Our work is supported by NSFC: The United Foundation of General Technology and Fundamental Research (No. U1536122), the General Project of Tianjin Municipal Science and Technology Commission under Grant (No. 15JCYBJC15600), the Major Project of Tianjin Municipal Science and Technology Commission under Grant (No. 15ZXDSGX00030).

J. Li et al. (Eds.): SPNCE 2019, LNICST 284, pp. 689–696, 2019.
https://doi.org/10.1007/978-3-030-21373-2_58

compromising the security of host and network [1]. It is a classifier designed to detect and classify network and host behaviors to identify whether they are normal or abnormal, and sent corresponding alerts. Therefore, improving the detection speed and accuracy is the key problem to be solved in IDS.

However, most of the results show that when there is a better whole detection rate, the detection rates of each classes may have great differences. This may cause the existence of low detection rate attack (LDRA). There are two reasons for LDRA: one is that the class distribution is nonuniform, the samples of lower detection rate classes are not enough and they are overwhelmed by other classes, so that these classes with a large sample size are dominant. This problem can be solved through increasing the number of their samples, but when there is a great disparity between samples sizes of each classes, this method may increase a great number of samples in the dataset, which can reduce the efficiency of the classifier; The other reason is that the selected features are not relevant to the low detection rate classes. To solve this problem, the most relevant feature subset should be found by improving the feature filtering method [2].

In order to prevent LDRA in intrusion detection from being ignored and having a tendency to give rise to security threat, a feature selection method for LDRA was proposed in this paper to improve their detection rates. In this method, irrelevant and redundant features were filtered preliminarily to reduce data dimension by using mRMR algorithm; in order to obtain the features most relevant to LDRA, we gathered them into a small dataset, then calculated the information gain in it and select some features greater than a given threshold. Finally these features were unioned together to get final feature subset. Experimental results show that our method can improve the detection rate of these classes effectively without affecting the overall and other class detection rates.

The rest of this article is divided into the following sections: Sect. 2 describes the related works; Sect. 3 introduces the feature selection method for LDRA; Sect. 4 reports the experiment and its results and Sect. 5 is the conclusion.

2 Related Work

In intrusion detection, reducing data dimension is an indispensable step in data preprocessing, so feature selection has become the focus of current research. Feature selection algorithms can be classified into 2 modes: filter and wrapper. The filter mode doesn't consider the learning algorithm and has a small computation. It can remove the noise and redundant features effectively. Information gain, mutual information, chi square distribution and mRMR [3] are the common filter method. The wrapper mode needs to determine the classification algorithm in advance, and then use the classifier to evaluate feature sets, which tends to a better classification performance, but has a higher computational cost. Aggarwal et al. [4] conducted a further research on the familiar intrusion detection dataset KDD 99, divided its features into 4 different classes according to the content and experimented with each combinations in classes to find the most influential combining class on detection rate and false alarm rate; Wu et al. [5] combined

the two modes, filtrated noise and irrelevant features by Fisher score and information gain respectively, then use the sequential backward selection algorithm to select feature subset; Cui et al. [6] proposed a feature selection method based on RS-PSO-SVM which can shorten time consumed greatly. However, the above methods just divided the dataset into 2 classes: normal or abnormal. They only thought about a better whole detection rate but made no comparison with each classes. If there are an obvious gap to the detection rates of each classes, it means that the whole detection rate cannot represent detection levels of each attack so that the whole detection rate should not be the main reference and evaluating standard.

In that case, Tang et al. [7] screened features through information gain and established an intrusion detection model with FCM clustering algorithm to improve its detectability; Huang et al. [8] proposed an intrusion detection method based on principal component analysis (PCA) increase its efficiency; Jia et al. [9] put forward a K-means based feature reduction method and reduced feature attributes by multiple clustering iterations; Mao et al. [10] integrated filter and wrapper methods. The filter method based on mutual information was firstly used to remove irrelevant attributes and the wrapper method based on improved adaptive genetic algorithm and improved evaluation function is used to select optimal attribute subset. For some situation that low detection rate classes are caused by the imbalance dataset, Feng et al. [11] combined SMOTE and GBDT algorithms, which were used to balance the dataset and make the classification respectively. These approaches took into account and improved the detection rates of each classes and had a better experimental results.

This paper makes some improvements based on the above researches, proposing a two-level selection method based on mRMR and information gain for the lower detection rate attacks. The experimental results show that the method has better detection rates than other methods whether in whole dataset or each class.

3 Two-Level Feature Selection Method

A good feature selection method can improve the performance of machine learning algorithm, simplify the model and increase the speed. A common feature selection method is to maximize the correlation between the feature and the classification variable, which is to select the first k variables with the highest correlation to the classification variables. However, in feature selection, the combination of these features does not improve the performance of the classifier, because it is possible that features are highly correlated with each other, and these features are redundancy features. Therefore, feature selection filters not only the unrelated features but also the redundant features [12]. The mRMR algorithm is used to remain the maximum relevance feature as well as filter the redundancy features.

3.1 Minimal-Redundancy-Maximal-Relevance

Minimal-Redundancy-Maximal-Relevance (mRMR) is a filter feature selection method based on mutual information, aiming at obtaining a subset contains features that are highly correlated with the class vector and uncorrelated with other features. Mutual information is a concept in information theory which is used to express the relationship and measure the correlation between features. Suppose there are two random variables x and y, their mutual information is defined as Eq. 1.

$$I(x;y) = \iint P(x,y)\, log \frac{P(x,y)}{P(x)\,P(y)} dxdy \tag{1}$$

Maximal-relevance is the measure of selecting features correlated with class and its computation is based on the mutual information between individual feature and class vector; Minimal-redundancy is the criterion of screen out redundancy features which is based on the mutual information between two features. Suppose the feature subset is S and C is the classification variable, the two formulas are Eqs. 2 and 3.

$$max\, D(S,c);\ D = \frac{1}{|S|} \sum_{x_i \epsilon S} I(x_i;c) \tag{2}$$

$$min\, R(S);\ R = \frac{1}{|S|^2} \sum_{x_i, x_j \epsilon S} I(x_i, x_j) \tag{3}$$

Define $\Phi(D,R)$ as the mRMR value of feature which is used to screen features, the formula is Eq. 4.

$$max\, \Phi(D,R);\ \Phi = D - R \tag{4}$$

3.2 Information Gain

Information gain is an important index of feature selection, which is defined as the amount of information that a feature can bring to the classification system. The more information it brings, the more important it is, the greater its information gain value is. In order to improve the accuracy of LDRA, we gather them into a new dataset and obtain the most important features to these classifications by calculating information gain of each features in this dataset.

Entropy represents the uncertainty of a feature. Suppose there are n classes in dataset, and class set $C = \{C_1, C_2, \cdots, C_i, \cdots, C_n\}$, $|C_i|$ is the number of samples of class Ci, $|D|$ implies sample size of dataset, $P(C_i)$ indicates the probability that class is C_i. We use H to represent entropy and the formula is Eq. 5.

$$H(C) = -\sum_{i=1}^{n} P(x_i)\, log_2 P(x_i) = -\sum_{i=1}^{n} \frac{|C_i|}{|D|} log_2 \frac{|C_i|}{|D|} \tag{5}$$

If feature T has m different values, divide the dataset into m subsets according to these values, that is $T = \{T_1, T_2, \cdots, T_k, \cdots, T_m\}$. Define $|T_k|$ as the number

of subsets T_k and $|T_{kc_i}|$ as the number of class C_i in T_k. The conditional entropy satisfies Eq. 6.

$$H(C|T) = -\sum_{k=1}^{m} \frac{|T_k|}{|D|} H(T_k) = -\sum_{k=1}^{m} \frac{|T_k|}{|D|} \sum_{i=1}^{n} \frac{|T_{kc_i}|}{|T_k|} log_2 \frac{|T_{kc_i}|}{|T_k|} \tag{6}$$

Make the subtraction to get the information gain value according to Eq. 7.

$$IG(T) = H(C) - H(C|T) \tag{7}$$

3.3 Two-Level Feature Selection Method

The flow diagram of the two-level feature selection method presented in this paper is shown in Fig. 1. The method is mainly divided into 4 stages, they are: preparation stage, execution stage, selection stage and integration stage. Detailed procedures for each stages are described as below:

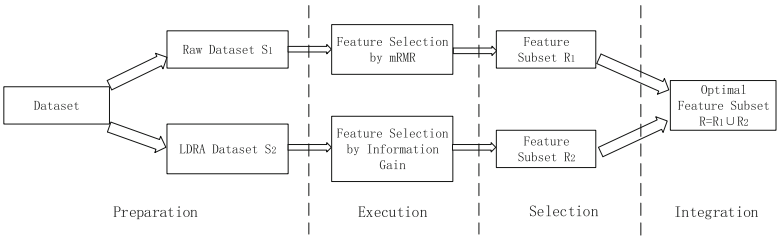

Fig. 1. The method flow diagram

- Preparation stage: define raw dataset as S_1 and duplicate LDRA classes from S_1 to a new dataset, namely the LDRA dataset S_2. Then quantize, normalize and discretize the data in dataset.
- Execution stage: mRMR feature selection method is used to select the features of the data in S_1 and calculate the mRMR values of each features; Information gain is used to select the features of the data in S_2 and calculate the entropy values of each features in LDRA. Both of the two values are sorted in descending order.
- Selection Stage: The results of the previous stage are further selected. Features whose mRMR value is greater than 0 are put into the feature subset R_1, which can guarantee the great reduction of data dimension; Features whose entropy value is greater than the specified threshold are put into the feature subset R_2, which can select out the features have high correlations with classification in LDRA.
- Integration stage: Integrate R_1 and R_2 as the final feature subset R. $R = R_1 \cup R_2$.

4 Simulated Experiment

4.1 Data Preprocessing

The dataset used in the experiment is the KDD 99 [13]. It contains 5 million network connection records and is composed of 41 features. Each connection is labeled as normal or abnormal. As for abnormal class, 39 types of attacks are summarized into four categories:

- DoS: Denial of Service.
- Probe: Monitor and other detection activities.
- R2L: Remote to Local. Illegal accesses from remote machines.
- U2R: User to Root. Unauthorized accesses to local superuser privilege by ordinary users.

The data distribution of dataset is shown in Table 1. From Table 1, the ratio of normal to abnormal is 1:4, but in abnormal class, DoS attack account for nearly 80% of the entire dataset and the remaining attack classes (especially R2L and U2R) account for a very small proportion so that this is an imbalanced dataset. To improve the classifier performance without drastic change about dataset sample size, we increase the sample of U2R to guarantee the basic classification. In order to reduce the time of experiment, the dataset is divided into two parts, 50000 samples are taken as training set and 30000 unduplicated samples as test set to verify the experimental results.

Table 1. Percentages of each classes in KDD 99

Class	Normal	DoS	Probe	R2L	U2R
Percentage	19.69	79.24	0.83	0.23	0.01

4.2 Results and Analysis

According to the classification results before feature selection, Probe, R2L, U2R has a relatively low detection rate, so we gather the three types into a new training set DR_Train. Then utilize mRMR to screen out the features into the feature subset mRMR_Sub and select the features in DR_Train through the calculated information gain values to the feature subset IG_Sub, at that time we can obtain the final optimal feature subset Final_Sub. Features contained in each subsets are shown in Table 2.

Our method is compared with the methods of literature [5,6,10] and the 41 attributes without any feature selection in the same experimental environment. The results will be considered not only in part but also in whole so the detection rates of each types of attacks and abnormal will be focused on. For the reason that the approaches we compare with are based on SVM, LibSVM was used

Table 2. Features contained in each subsets

Feature subset	Number of features	Features contained in subset
mRMR_Sub	15	3,24,12,32,31,37,6,23,1,2,40,38,5,39,36
IG_Sub	8	3,5,4,35,2,12,40,33
Final_Sub	19	1,2,3,4,5,6,10,12,23,24,31,32,33,35,36,37,38,39,40

Table 3. Comparison of feature selection methods

Methods	Normal	Dos	Probe	R2L	U2R	Abnormal
Literature [5]	99.8	99.9	86.6	78.1	50.8	99.60
Literature [6]	99.8	99.4	81.6	51.6	24.6	98.95
Literature [10]	99.7	100	73.9	50	70.5	99.51
All features	100	100	89.3	68.8	68.9	99.75
Our method	99.9	99.9	93.1	81.3	83.6	99.78

as the classification and testing algorithm and experiments were conducted on Weka. The results of the experiment are shown in Table 3.

Table 3 shows the validity and accuracy of our method. Compared with literature [5], literature [6] and literature [10], our method has good classification effect both from whole and part. Compared with all features, class probe, R2L and U2R classes have significant improvements. Therefore, our method wipes out the redundancy while preserves the correlation features, reduces the dimension and keeps the detection rate at the same time.

In order to test the stability of the method, define the abnormal detection rate, accuracy, false report rate (FNR), false alarm rate (FAR) and modeling time as evaluating criteria for further comparative testing. The experimental results are shown in Table 4. As can be seen from Table 4, though the methods of literature [5] and literature [10] have the smaller feature dimension and the shorter modeling time, their accuracy rate is low and FNR and FAR are also higher than our method, which shows that the number of features is not the less the better. Compared with all features, it maintains the detection rate of abnormal. Although the FAR has increased, the accuracy hold the level with all features in the case of shortening nearly 50% of the time, and it reduce the FNR, so our method is more stable.

Table 4. Comparison of stability of feature selection methods

Method	Number of features	Detection rate	Accuracy	FNR	FAR	Modeling time
Literature [5]	13	99.60	99.61	0.5	0.19	8.45
Literature [6]	17	98.95	99.08	1.04	0.22	16.22
Literature [10]	7	99.51	99.52	0.45	0.27	9.82
All features	41	99.75	99.76	0.27	0.03	17.5
Our method	19	99.78	99.77	0.2	0.1	10.1

5 Conclusion

Feature selection can effectively reduce the data dimension and improve the efficiencies of classifiers. However, the detection rates between categories may have great difference in a dataset. In this paper, a two-level feature selection method based on mRMR and information gain is proposed for LDRA in intrusion detection. The method uses mRMR to filter irrelevant features, reduces the data dimension, and improves some classes detection rates by information gain calculation, and finally combine these features together. Finally, the validity and expansibility of the method are proved.

References

1. Li, X., Yao, Y.: Master and use Snort tools for intrusion detection. Comput. Appl. Softw. **23**(3), 123–124 (2006)
2. Zhang, Y., Yang, A., Xiong, C., et al.: Feature selection using data envelopment analysis. Knowl. Based Syst. **64**, 70–80 (2014)
3. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
4. Aggarwal, P., Sharma, S.K.: Analysis of KDD dataset attributes - class wise for intrusion detection. Procedia Comput. Sci. **57**, 842–851 (2015)
5. Wu, X., Peng, X., Yang, Y.: Two-level feature selection method based on SVM for intrusion detection. J. Commun. **36**(4), 19–26 (2015)
6. Cui, W., Meng, X., Li, J.: Feature selection based on RS-PSO-SVM. Microelectron. Comput. **1**(3), 120–123 (2015)
7. Tang, C., Liu, P., Tang, S.: Anomaly intrusion behavior detection based on fuzzy clustering and features selection. J. Comput. Res. Dev. **52**(3), 718–728 (2015)
8. Huang, J., Chen, G., Ling, X.: Intrusion detection based on principal component analysis. J. China Jiliang Univ. **18**(3), 221–224 (2007)
9. Jia, F., Yan, Y., Zhang, J.: K-means based feature reduction for network anomaly detection. J. Tsinghua Univ. (Sci. Technol.) **58**(2), 137–142 (2018)
10. Mao, L., Yao, S., Hu, C.: A new hybrid attribute selection method and its application in intrusion detection. Trans. Beijing Inst. Technol. **28**(3), 218–221 (2008)
11. Feng, H., Li, M., Hou, X.: Study of network intrusion detection method based on SMOTE and GBDT. Appl. Res. Comput. **34**(12), 3745–3748 (2017)
12. Bolón-Canedo, V., Sanchez-Marono, N., Alonso-Betanzos, A.: Feature selection and classification in multiple class datasets: an application to KDD Cup 99 dataset. Expert Syst. Appl. **38**(5), 5947–5957 (2011)
13. Zhang, X., Cao, H., Jia, L.: Research of intrusion detection system dataset KDD CUP99. Comput. Eng. Des. **31**(22), 4809–4812 (2010)