Intrusion detection method based on information gain and ReliefF feature selection

Yong Zhang
School of Computer and Information Technology
Liaoning Normal University
Dalian, China
zhyong@lnnu.edu.cn

Xuezhen Ren
School of Computer and Information Technology
Liaoning Normal University
Dalian, China
miamizhenbaolovely@qq.com

Jie Zhang
School of Computer and Information Technology
Liaoning Normal University
Dalian, China
1209112706@qq.com

***Abstract*—Traditional random forest has slow convergence in network intrusion detection and its learning performance is not perfect. In order to eliminate the redundant information in the original intrusion detection data, this paper proposes a random forest intrusion detection method based on the combination of information gain and ReliefF algorithm. The proposed method first uses the information gain to calculate the information gain value of each feature. Then, the ReliefF algorithm is used to calculate the weight of each feature. According to the information gain and the feature weight, the final feature subset is obtained. Finally, this paper uses a random forest classifier for classification. The experiment compares the three feature selection methods, including the proposed method, information gain based method and ReliefF-based method. The experimental results show that the precision, recall rate, and false positive rate of the proposed method are superior to those of the other two methods.**

Keywords—network intrusion detection, information gain, feature selection, random forest

I. INTRODUCTION

With the rapid development of network technology, network information security is particularly important. If the network environment is not safe, it will lead to many problems, such as privacy leakage, resource theft, etc., which will bring many losses to people's work and life. Therefore, it is imperative to improve the network information security of the Internet. The network intrusion detection system came into being. It is an effective means of dealing with the security of network information.

Network intrusions are usually divided into illegal access to information, modification of information, and destruction of user systems [1]. The safest way to protect computer from network attacks is to perform network intrusion detection. Network intrusion detection includes misuse detection and anomaly detection. Misuse detection can only detect known attacks. It cannot identify the variants of the attack and has no practical application value. The anomaly detection does not require prior knowledge, although the detection rate is slightly lower. However, some new or catastrophic intrusion attacks can be detected. Therefore, it has become the main research direction [2-4].

In addition, the data set for intrusion detection is too large. These data are often high-dimensional, with each feature having correlation and redundancy. This will have a great impact on the efficiency of intrusion detection, resulting in low detection accuracy. Therefore, most research work pre-optimizes the data before intrusion detection. Smith *et al.* [5] proposed a method combining Bayesian network and principal component analysis (PCA) to optimize the data, using Bayesian network to adjust the correlation of features, and then using PCA for dimensionality reduction. In [6], an intrusion detection method based on deep multi-layer extreme learning machine is proposed. Based on the extreme learning machine automatic encoder (ELM-AE), a multi-layer neural network is generated by stacking ELM-AE initialization. The hidden layer weights are used to reduce the dimension by singular value decomposition, so as to achieve the feature simplicity. Qian *et al*. [7] used an artificial bee colony algorithm to optimize the parameters of neural network, which can avoid the neural network falling into a local optimum and solve the problem of slow convergence speed of the neural network algorithm. Zhao *et al.* [8] proposed an improved feature selection algorithm to identify most appropriate subset of features for a certain attack, using Mahalanobis distance feature ranking and an improved exhaustive search to choose a better combination of features. Meng and Sun [9] introduced an improved ant colony clustering intrusion detection method. The convergence speed of the ant colony clustering algorithm is improved. In the optimization process, information entropy is introduced to prevent local optimization, so the method can automatically adjust the update pheromone and improve the clustering speed. AI-Jarrah *et al*. [10] proposed two feature selection methods, namely, random forest-forward selection ranking and random forest-backward elimination ranking, for large-scale network intrusion detection.

In this paper, a feature reduction method combining information gain and ReliefF algorithm is proposed and applied to the feature extraction stage of intrusion detection. The proposed method first uses the information gain technique to calculate the information gain value of each feature attribute. Then the proposed method uses ReliefF algorithm [11] to calculate the weight of each attribute. According to the information gain and the attribute weight, the final feature subset is obtained. This method can avoid the neglect of the minority class by ReliefF algorithm, which can not only retain the impact of attributes on majority classes, but also enhance the impact of attributes on minority classes. It also ensures that the attributes

978-1-7281-2009-6/$31.00 ©2019 IEEE paper N-19591.pdf

are independent of each other and will not fall into local optimum. It can better avoid excessive redundant attributes, effectively reduce information loss and speed up the convergence.

The remainder of this paper is organized as follows. Section II introduces preliminaries. Section III presents our proposed feature selection method based on information gain and ReliefF algorithm. Section IV presents our experimental results and discussions. Section V presents concluding remarks.

II. PRELIMINARIES

A. Entropy and Information Gain

Entropy is an important concept in information theory. Entropy represents whether the distribution of energy in space is uniform. The more uneven the energy distribution is, the smaller the entropy is, and vice versa, the larger the entropy is. Shannon first applied entropy to information processing research and proposed the concept of information entropy. Information entropy is actually quantifying information, which is the degree of uncertainty in measuring the value of a random variable [12].

Information Gain (IG) is an important concept in information theory and is widely used in the field of machine learning. For data classification, the information gain is calculated by counting the number of occurrences of each feature in each category to calculate the information gain for each category.

Let the vectors **X** and **Y** represent the sample features ($\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_m$) and the category features ($\mathbf{Y}_1$, $\mathbf{Y}_2$, ..., $\mathbf{Y}_n$), respectively. The information gain between a given feature **X** and the associated category feature **Y** is as follows.

$$IG(Y,X) = H(Y) - H(Y|X) \tag{1}$$

$$H(Y) = -\sum_{i=1}^{n} p(y_i) log_2 p(y_i) \tag{2}$$

$$H(Y|X) = \sum_{i=1}^{m} p(x_i) H(Y|X = x_i) \tag{3}$$

In equation (1), IG(**Y**, **X**) represents the information gain of the feature **X** for the category **Y**, H(**Y**) is the entropy of **Y**, and H(**Y**|**X**) is the conditional entropy. In this paper, **X** is the feature item of the data, and **Y** represents the category.

B. ReliefF Algorithm

The ReliefF algorithm is an extension of Relief algorithm [13], developed by Kira and Rendell that takes a filter-method approach to feature selection. Relief algorithm is a feature weighting algorithm, which is initially confined to two-class classification problems. Relief algorithm assigns different weights to features according to the correlation of each feature and category, and features whose weights are less than a certain threshold will be removed. The correlation between features and categories in Relief algorithm is based on the ability of distinguishing features from close samples.

In order to handle multiclass problems, Kononenko [14] proposed ReliefF algorithm. Given the class labels with l categories $C = \{c_1, c_2, \dots, c_l\}$. ReliefF algorithm randomly selects a sample R_i from the training set, then finds K near samples (called *near Hits*) of R_i from the same category, which is denoted by H_j $(j = 1,2, \dots, K)$, and also finds K near samples (called *near Misses*) of R_i from different categories, which is denoted by $M_j(c)$ $(j = 1,2, \dots, K)$. The ReliefF algorithm repeats the above procedure on each feature dimension, and gets the weight of each feature as follows:

$$W(A) = W(A) - \sum_{j=1}^{K} diff(A, R_i, H_j)/(m * K) + \sum_{c \notin class(R_i)} \left[\frac{p(c)}{1 - p(class(R_i))} \sum_{j=1}^{K} diff\left(A, R_i, M_j(c)\right) \right] / (m * K) \tag{4}$$

where m is the number of iterations. $p(c)$ is the probability of the class c. $diff(A, R_1, R_2)$ represents the difference between sample R_1 and R_2 about feature A, which is defined as

$$diff(A, R_1, R_2) = \begin{cases} \frac{R_1[A] - R_2[A]}{\max(A) - \min(A)}, & if\ A\ is\ continous \\ 0, & if\ A\ is\ discrete\ and\ R_1[A] = R_2[A] \\ 1, & if\ A\ is\ discrete\ and\ R_1[A] \neq R_2[A] \end{cases} \tag{5}$$

C. Random Forest

Random forest [15] is an ensemble learning method based on decision tree for classification and regression. It is constructed by a multitude of decision trees. The decision tree itself is a classifier, which is widely used in various fields. When a single decision tree is classified, it needs to be pruned to overcome over-fitting issue, then a decision tree classifier is established.

As shown in Figure 1, for network intrusion detection data, the final classification result is decided by the general majority voting of base classifiers in the ensemble. It is only to vote for each decision tree that participates in the classification. Therefore, the random forest classifier needs to decide the number of decision trees in advance, that is, the parameter k in Figure 1. However, the associate literature rarely provides how many decision trees should be used to construct a random forest classifier. In general, the number of decision trees is set based on constant trials and consideration of error rates.

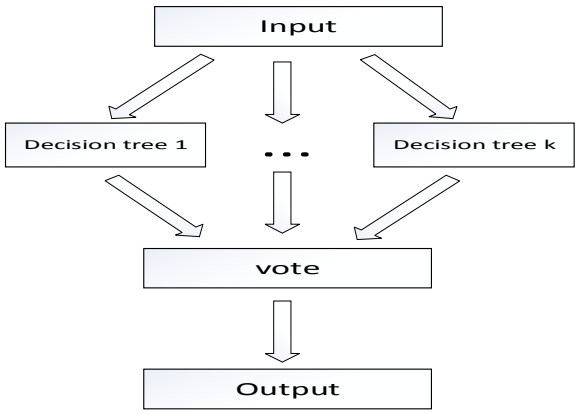

Fig. 1. Algorithmic diagram of a random forest.

III. Feature Selection Method Based on Information Gain and ReliefF

When network intrusion detection data is too large, too many redundant features will affect the classification performance, which will consume a lot of time, and affect the accuracy and the effectiveness of classification. Moreover, when ReliefF algorithm is used for feature selection of data, some categories are difficult to distinguish, which leads to the problem of inaccurate classification of minority classes. The information gain algorithm can improve the ReliefF algorithm for the differences between and within minority classes, and achieve the best classification effect.

Therefore, this paper proposes a feature selection method based on information gain and ReliefF, which can improve the ability of ReliefF to distinguish minority classes and improve the detection efficiency. The proposed method first uses the information gain technology to correlate the intrusion data features, calculates the information gain of each feature. For a given threshold T, the proposed method selects all the features whose information gain are larger than the threshold T, to obtain the feature subset M_1. Then, ReliefF algorithm is used to calculate the weight of each feature. Features whose weights are larger than the given threshold w are selected to obtain a feature subset M_2. Finally, the proposed method combines two subsets M_1 and M_2 to obtain the final feature subset.

The proposed feature selection method is described in Algorithm 1 as follows.

Algorithm 1. IG-ReliefF feature selection method

Input: Intrusion detection data set D, the number of iterations m, the information gain threshold T, and the weight threshold w.

Output: The feature subset M after feature selection.

Algorithm:

1. For each feature t in data set D do
   Calculate the information gain IG(t);
2. Select features with IG(t)>T to obtain subset M_1;
3. Set all weights $W[A]$=0 for each feature A;
4. For i = 1 to m do
5. Randomly select a sample R_i
6. For each class $c \in class(R_i)$ do
   Find *K near Hits* H_j;
7. For each class $c \notin class(R_i)$ do
   Find *K near Misses* $M_j(c)$;
8. Calculate the weight of each feature $W[A]$ using equation (4) and equation (5);
9. End For
10. Select features whose weights are larger than w to obtain subset M_2;
11. Output $M = M_1 \cup M_2$.

IV. Experiments

A. Data Set and Evaluation Criteria

This experiment uses the Matlab 7.12.0 (R2011a) simulation environment, the memory is 4 GB, the processor is Intel (R) Core (TM) i3-4160 CPU 3.60 GHz, and the system is Windows 7. The experimental dataset uses the NSL-KDD dataset. It contains 148,517 intrusion data in the network, each of which includes 41 features. The data is divided into five categories: Normal, DoS, Porbe, U2R, and R2L. Normal is normal data, and DoS, Porbe, U2R, and R2L are attack data. The distribution and identification type are shown in Table I.

TABLE I. Identification Types and Distribution in Intrusion Detection Data Set of NSL-KDD

Category	Meaning	Class label	Sample size	Percent (%)
Normal	Normal record	1	77054	51.88
DoS	Denial of service attack	2	53385	35.95
Probe	Probe attack	3	13081	8.81
R2L	Permission attack	4	3749	2.53
U2R	Permission access attack	5	1248	0.09

We divide the types of data sets into five categories as shown in Table I, The class label of Normal is 1, and the other four attack types are 2, 3, 4, and 5, respectively.

Before the experiments, the data in the data set need to be digitized and then standardized. The experiments first transform the string features into corresponding numerical features, and then the data sets are standardized by z-score as following:

$$x' = \frac{x - \bar{x}}{s} \tag{6}$$

where x is the original data, $\bar{x}$ is the mean, and S represents the standard deviation.

The evaluation criteria of this paper are confusion matrix, precision, recall rate, and false positive rate (FPR). The confusion matrix is shown in Table II, and other performance indicators can also be obtained through the confusion matrix as shown in Table III.

TABLE II. Confusion Matrix

		Predicted	
		Normal	Intrusion
Actual	Normal	True Positive (TP)	False Negative (FN)
	Intrusion	False Positive (FP)	True Negative (TN)

TABLE III. The Evaluation Criteria

Precision	$TP/(TP+FP)$
Recall	$TP/(TP+FN)$
FPR	$FP/(FP+TN)$

B. Experimental Results and Analysis

After normalizing the data, the experiments use the proposed IG-ReliefF method for feature selection, and finally the random forest is used for classification and detection. In order to better analyze the experimental results, we first need to determine the value of parameter k (the number of decision trees) in the random forest. The parameter k of the random forest is optimized and analyzed by the experiments. The accuracies and running time under different parameter k are shown in Figure 2 and Figure 3, respectively. As can be seen from Figure 2 and Figure 3, the classification performance is the best when the parameter k=10. Therefore, k=10 is chosen as the final experimental parameter in subsequent experiments.

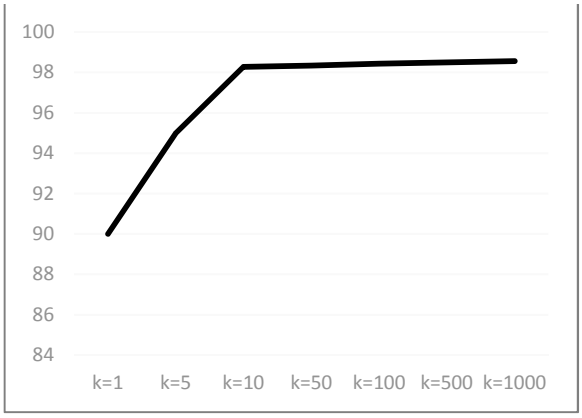

Fig. 2. The accruacies in differnet parameter k.

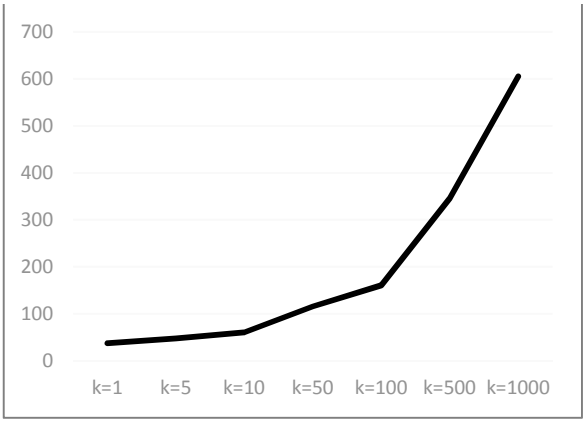

Fig. 3. The running time in differnet parameter k.

The following experiments adopt 10-fold cross-validation method to ensure the reliability of the evaluation. In 10-fold cross-validation, the NSL-KDD data set is randomly partitioned into 10 equal sized subsets. Of the 10 subsets, a single subset is retained as the validation data for testing the model, and the remaining 9 subsets are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsets used exactly once as the validation data. The 10 results can then be averaged to produce a single estimation.

In the process of feature selection, the information gain threshold T is set to be 0.9, the weight threshold w in ReliefF algorithm is 0.03, and the number of near samples K is 5.

The experiments first use the information gain to obtain a reduced feature subset. The information gain of each feature is calculated as shown in Table IV. We select those features whose information gain is greater than the threshold T=0.9 to form a new subset M_1. The subset M_1 includes 13 features, such as 3, 4, 9, 11, 12, 21, 22, 25, 26, 29, 30, 38, and 39.

TABLE IV. INFORMATION GAIN VALUES (IG) OF FEAUTES OF NSL-KDD DATA SET IN DESCENDING ORDER

Feature	IG	Feature	IG	Feature	IG
11	0.9614	**37**	0.8949	**13**	0.8788
26	0.9558	**5**	0.8932	**33**	0.8764
4	0.9553	**27**	0.8897	**36**	0.8761
12	0.9528	**18**	0.8895	**40**	0.8759
25	0.9473	**10**	0.8890	**17**	0.8759
39	0.9391	**28**	0.8871	**23**	0.8730
30	0.9297	**14**	0.8864	**16**	0.8718
38	0.9243	**2**	0.8844	**19**	0.8706
22	0.9203	**34**	0.8839	**32**	0.8686
21	0.9192	**8**	0.8830	**15**	0.8681
29	0.9069	**41**	0.8819	**24**	0.8619
3	0.9064	**31**	0.8815	**7**	0.8616
9	0.9033	**6**	0.8800	**20**	0.0000
35	0.8951	**1**	0.8789		

Then, the experiments uses ReliefF algorithm to feature selection. The weight of each feature is shown in Table V. For given the weight threshold w=0.03, the following 14 features are selected to form the subset M_2, including 2, 3, 4, 12, 25, 26, 30, 32, 33, 34, 35, 36, 38, and 39. According to our proposed IG-ReliefF feature selection method, the experiment finally obtains a subset $M = M_1 \cup M_2$ with 19 features, such as 2, 3, 4, 9, 11, 12, 21, 22, 25, 26, 29, 30, 32, 33, 34, 35, 36, 38, and 39.

TABLE V. THE WEIGHTS OF FEAUTES OF NSL-KDD DATA SET IN DESCENDING ORDER

Feature	Weight	Feature	Weight	Feature	Weight
26	0.1010	**40**	0.0287	**14**	3.429e-04
25	0.0945	**27**	0.0278	**21**	2.6525e-04
12	0.0843	**28**	0.0246	**18**	1.5716e-04
39	0.0785	**29**	0.0208	**9**	1.0318e-04
38	0.0742	**23**	0.0113	**19**	6.804e-05
36	0.0661	**31**	0.0101	**41**	6.3981e-05
2	0.0558	**8**	0.0095	**17**	5.9398e-05
4	0.0477	**22**	0.0072	**15**	2.7026e-05
35	0.0458	**24**	0.0047	**5**	5.3059e-06
34	0.0437	**37**	0.0037	**16**	1.2936e-06
3	0.0429	**10**	9.1672e-04	**13**	4.5225e-07
30	0.0428	**11**	8.1704e-04	**6**	4.3928e-07
33	0.0377	**7**	4.3774e-04	**20**	0
32	0.0330	**1**	3.497e-04		

The proposed method in this paper is compared with the traditional IG-based feature selection method and ReliefF-based feature selection method. The results are shown in Table VI. As can be seen from Table VI, the IG-based feature selection method has a poor classification effect for the minority classes, where the recall rates are 71.82% and 86.55% for R2L and U2R, and the accuracies are 85.48% and 86.07%, respectively. The RelifF-based feature selection method is slightly better than IG-

based method in classification of minority classes. The recall rates of R2L and U2R are 88.90% and 91.78%, and the accuracies are 83.11% and 88.71%, respectively. The proposed IG-ReliefF method in this paper improves the performance of minority classes. The recall rates of R2L and U2R are 97.91% and 96.76%, and the accuracies are 93.37% and 94.18%, respectively. Correspondingly, the FPR value of the proposed method is also better than the other two methods. Compared with the IG-based method and the ReliefF-based method, the proposed method can improve the classification effect of minority classes without affecting the classification effect of majority classes.

TABLE VI. PERFORMANCE COMPARISON

		IG-based method	ReliefF-based method	IG-ReliefF based method
Precision	Normal	94.02%	98.86%	98.84%
	DOS	99.19%	99.38%	99.36%
	R2L	85.48%	83.11%	93.37%
	PROBE	95.48%	98.96%	98.95%
	U2R	86.07%	88.71%	94.18%
Recall	Normal	98.42%	98.46%	98.74%
	DOS	96.40%	99.41%	99.51%
	R2L	71.82%	88.90%	97.91%
	PROBE	93.89%	98.44%	98.36%
	U2R	86.55%	91.78%	96.76%
FPR	Normal	6.79%	1.21%	1.07%
	DOS	0.43%	0.39%	0.25%
	R2L	0.18%	0.46%	0.31%
	PROBE	0.42%	0.11%	0.07%
	U2R	0.11%	0.09%	0.06%

V. CONCLUSION

In this paper, an effective feature selection method for intrusion detection is presented by using random forest classifier. In an intrusion detection system, because of the small number of U2R and R2L classes, the weight or information gain value of minority classes will be very low in the process of data analysis, which makes it difficult to select features that have a great impact on minority classes. This paper presents a feature selection method combining the information gain and ReliefF algorithm. Experimental results show that the proposed method has better performance and efficiency. Therefore, the proposed method can be effectively applied to the field of network intrusion detection, which provides a new method and idea for feature selection of network intrusion detection model.

ACKNOWLEDGMENT

This work is partly supported by National Natural Science Foundation of China (No. 61772252) and Program for Liaoning Innovative Talents in University (No. LR2017044).

REFEENCES

[1] J. P. Anderson, Computer Security Threat Monitoring and Surveillance, USA: PA 19034, 1980.

[2] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," IEEE Communications Surveys & Tutorials, 2018, doi: 10.1109/COMST.2018.2847722, in press.

[3] A. Sharma, I. Manzoor, and N. Kumar, "A feature reduced intrusion detection system using ANN classifier," Expert Systems with Applications, vol. 88, pp. 249–257, 2017.

[4] A. A. Aburomman and M. B. I. Reaz, "A novel weighted support vector machines multiclass classifier based on differential evolution for intrusion detection systems," Information Sciences, vol. 414, pp. 225–246, 2017.

[5] D. Smith, Q. Guan, and S. Fu, "An anomaly detection framework for autonomic management of compute cloud systems," in Proceedings of IEEE 34th Annual Computer Software and Applications Conference, 2010, pp. 376–381.

[6] X. Luo, Z. Zhu, C. Dong, and X. Zhang, "Applications of extreme learning machine in the intrusion detection," in Proceedings of 4th IEEE International Conference on Computer Science and Information Technology, 2011.

[7] Q. Qian, J. Cai, and R. Zhang, "Intrusion detection based on neural networks and Artificial Bee Colony algorithm," in Proceedings of 2014 IEEE/ACIS 13th International Conference on Computer and Information Science, 2014.

[8] Y. Zhao, Y. Zhang, W. Tong, and H. Chen, "An improved feature selection algorithm based on Mahalanobis distance for network intrusion detection," in Proceedings of 2013 International Conference on Sensor Network Security Technology and Privacy Communication System, 2013.

[9] L. Meng and G. Sun, "An improved ant colony clustering method for network intrusion detection," in Proceedings of IEEE Eighth International Conference on Networking, Architecture and Storage (NAS), 2013, pp. 312–316.

[10] O. Y. AI-Jarrah, A. Siddiqui, M. Elsalamouny, P. D. Yoo, S. Muhaidat, and K. Kim, "Machine-learning-based feature selection techniques for large-scale network intrusion detection," in Proceedings of IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW), 2014.

[11] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," Applied Intelligence, vol. 7, no. 1, pp. 39–55, 1997.

[12] M. Mitchell, "Machine Learning," Mc Graw Hill India, 2017.

[13] K. Kira and L. Rendell, "The feature selection problem: traditional methods and a new algorithm," in Proceedings of AAAI-92, 1992.

[14] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *European Conference on Machine Learning*, pp. 171–182, Springer, 1994.

[15] U. Yuma and N. Yasushi, "A proposal of regression hybrid modeling for combining random forest and x-means methods," Total Quality Science, vol. 3, no. 1, pp. 1–10, 2017.