A new Boosting algorithm used in intrusion detection

Zhixin Cai[1, a], Xiufen Fu[2, b]

[1] Faculty of Computer, Guangdong University of Technology, Guangzhou, 510006, China

[a]email: janson_1015@163.com

Keywords: intrusion detection; MDBoost; overfitting; Accuracy

Abstract. At this stage, the high dimension and large variety of network data have increased the difficulty of intrusion detection. In this paper, we discuss the advantages and disadvantages of the MDBoost algorithm. Subsequently to optimize it, we add a slack variable in the objective function, so that the algorithm can effectively prevent over fitting, and the accuracy of the prediction is also improved. Then, we propose a model, which uses the MDBoost-2 algorithm to generate a strong classifier, and we use this model for intrusion detection. Finally, we use the CUP KDD 1999 data set to carry out the experiment. The results show that the new approach outperforms MDBoost and other well-known methods.

Introduction

With the development of the computer network, the network has all kinds of viruses and malicious information. Intrusion detection system (IDS) for real-time detection of network transmission and the effective interception of malicious information have become an integral part of network security. At present, intrusion detection is mainly focused on improving the detection rate. Today some network attacks which are very similar with normal user behavior can evade the detection system, to achieve the purpose of malicious behavior. Thus many algorithms are used to solve these problems, e.g. Neural Network algorithm [1], Bayesian algorithm[2] and other Machine learning algorithms[3].

However the algorithms mentioned above are the single classifier whose capacity in identifying network data is limited. Thus, one view is that the plurality of weak classifiers trained by an algorithm to generate a strong classifier, e.g. [4,5,6].

Kearns and Valiant proved that a plurality of weak classifiers is equivalent to a strong classifier. Freund and Schapire proposed Adaboost algorithm in 1995. Compared to the Boosting algorithm, Adaboost algorithm has better performance. And Schapire proposed maximum interval theory. Based on this theory, Buhlmann in 2001 improved Adaboost algorithm and proposed L2Boost algorithm. And A.Demiriz put forward LPBoost algorithm. C.Shen and H.Li on improved Adaboost algorithm and proposed the Adaboost-CG algorithm[7] and MDBoost algorithm[8]. Guo[9] improve the L2Boost algorithm and proposed the MCBoost algorithm which has many advantages compared with L2Boost.

On the basis of MDBoost algorithm, this paper proposes the MDBoost-2 algorithm, which is used in the intrusion detection system. Finally, through the experimental comparison. Obtained MDBoost-2 algorithm has an advantage in terms of detecting intrusion data.

The paper is organized as follows: In section 2, we introduce the MDBoost algorithm and MDBoost-2 algorithm. In section 3, we propose MDBoost-2 model and use the model for intrusion detection .In Section 4 we put forward two experiment. At last, we give our conclusions in section 5.

ENSEMBLE LEARNING

A. MDBoost algorithm

Firstly, C.Shen and H.Li proposed Adaboost-CG algorithm, and proved that the mean interval and variance of Adaboost-CG reach the maximum at the same time when the training samples are

subject to the normal distribution. Secondly based on the above theory, they proposed the MDBoost algorithm. Suppose the number of training samples is N, and label is yi , H represents a weak classifier set(hi ∈H, i = 1⋯t).

Objective function of MDBoost is :

$$\text{Min} \quad \frac{D}{2(N-1)} \sum_{i>y} \left(mrg_i - mrg_j\right)^2 - \left\|mrg\right\|^1 \tag{1}$$

$$\text{s.t.} \quad w_i \geq 0, \left\|w\right\|^1 = 1 \qquad mrg_i = y_i \sum_{j=1}^{t} w_j h_j(x_i)$$

Define matrix B:

$$B = \begin{bmatrix} 1 & -\frac{1}{N-1} & L & -\frac{1}{N-1} \\ -\frac{1}{N-1} & 1 & L & -\frac{1}{M-1} \\ N & N & 0 & N \\ -\frac{1}{N-1} & -\frac{1}{N-1} & L & 1 \end{bmatrix}$$

Then the objective function of MDBoost is :

$$\text{Min} \quad mrg^T B\ mrg - \left\|mrg\right\|^1 \tag{2}$$

$$\text{s.t.} \quad w_i \geq 0, \left\|w\right\|^1 = 1$$

B. MDBoost-2 algorithm

In order to improve the robustness of the MDBoost algorithm. Firstly, We add a slack variable in the objective function.

$$\text{Min} \quad mrg^T B\ mrg - \left\|mrg\right\|^1 + C^T\varepsilon \tag{3}$$

$$\text{s.t.} \quad w_i \geq 0, \left\|w\right\|^1 = 1, \varepsilon_i \geq 0 \qquad mrg_i = y_i \sum_{j=a}^{t} w_j h_j(x_i)$$

The Lagrange function of the formula (3):

$$L\left(w, mrg, \varepsilon, r, \mu, q\right) = \frac{1}{2} mrg^T B\ mrg - \left\|mrg\right\|^1 + C^T\varepsilon + r(\left\|w\right\|^1 - 1) - q^T w \tag{4}$$

$$+ \sum_{i=1}^{M} \mu_i (mrg_i - y_i \sum_{j=1}^{t} w_j h_j(x_i) + \varepsilon_i)$$

Then Seeking the extreme value of w, mrg:

$$\inf_{w, mrg, \varepsilon} L\left(w, mrg, \varepsilon, r, \mu, q\right) = \inf_{mrg}\left[\frac{1}{2} mrg^T B\ mrg - (u-1)^T mrg\right] + \inf_{\varepsilon}\left[C^T\varepsilon + \mu^T\varepsilon\right] \tag{5}$$

$$- r\frac{1}{T} + \left[r * e - q^T + \sum_{i=1}^{M} \mu_i y_i H_i\right] w$$

Because W is a linear variable, its coefficient is 0. In equation 5, we calculate the derivative variable mrg, then after several steps we can get the dual problem from equation 3. By solving a convex optimization problem, finally get the right value for each weak classifier.

Institutions Optimization Design MDBOOST-2 MODEL

The framework of model consists of four parts, the listener module, data preprocessing module, data classification module, decision module and warning module. The whole framework of the model is illustrated in Fig

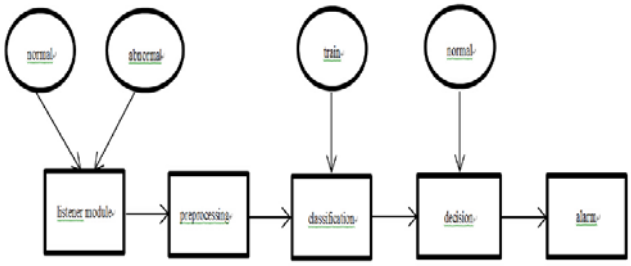

Figure 1. The framework of model

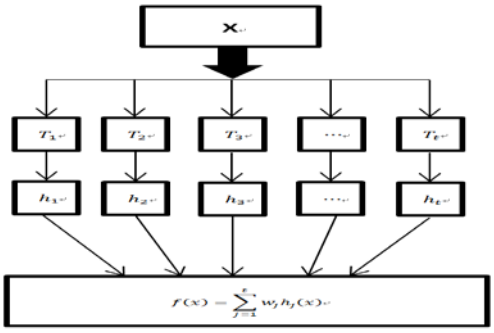

Figure 2. Data classification module

The monitor module is located at the front of the model, and it is the data input terminal. The key of this module is to extract the data feature. Because the network data is very large, the network protocol is very complex, the extracted features have a great influence on the model training and prediction. Therefore, the monitor module is the entrance and the most important module of the model.

The data preprocessing module is used to standardize the data frame. As the network data is very large, and the model is to detect the specific format of the data.

The data classification module is the main part of the model as shown in Fig 2. The module is composed of several weak classifiers.

The decision module makes the corresponding decision based on the data obtained from the classifier. The design of this module needs to be very careful. If the model is too sensitive, the model will treat the normal behavior as the invasion. If the detection ability of the model is weak, it will produce a lot of false negatives. So The design of this module needs to be very careful.

Alarm module, mainly to carry out a series of protective measures, such as closing the port or directly prohibit the user's rights and so on.

EXPERIMENT

We conducted experiments in MATLAB environment. Experiments using KddCup99 data set which contains 40 properties. All data are divided into 5 categories(e.g. normal,DOS,Probing,R2L and U2L). In order to reduce the correlation of weak classifiers, the training data of each weak classifier is random sampling and we select F features from KddCup99 data set as the feature of weak classifier.

First select normal and Dos two kinds of data to compare the MDBoost-2, MDBoost, MCBoost, AdaBoost, the accuracy of the four algorithms. We choose SVM as the weak classifier, and the number of weak classifier is 390. The size of the training and test data is (100，100) 、(1000，1000) 、(10000，10000).The result is shown in table 1.

Through experiments, the effect of different algorithms on different data is compared. In the case of the same data size, MDBoost-2 compared to other algorithms have higher accuracy. With the increase of the training data, the accuracy rate also increased. When the data is more and more big, the back of the three algorithms are almost the same. As the amount of data becomes larger and larger, the effect of the machine learning algorithm is very close.

Then, on the premise of the data scale is 1000, 100 sets of noise data are introduced to the above algorithm. The results of the experiment are shown in Table 2.

It can be seen from the experimental results that the accuracy of the algorithm is decreased after the introduction of the noise data. But MDBoost-2 compared to the other 3 algorithms, still can have a better effect. MDBoost-2 compared to the MDBoost algorithm can reduce the phenomenon of over fitting to a certain extent.

TABLE I. COMPARISON OF ALGORITHMS

Data size	Four algorithms			
	AdaBoost	MDBoost	MCBoost	MDBoost-2
100,100	99.01%	99.29%	99.31%	99.55%
1000,1000	99.23%	99.53%	99.55%	99.65%
10000,10000	99.50%	99.78%	99.77%	99.78%

TABLE II. AFTER INTRODUCING THE NOISE DATA,THE COMPARATION ALGORITHMS

Data size	Four Algorithms			
	AdaBoost	MDBoost	MCBoost	MDBoost-2
10000,10000	84.01%	94.27%	96.44%	98.72%

Conclusion

MDBboost-2 algorithm is proposed in this paper, which is the improvement of MDBoost algorithm. Then, a model is proposed, which uses the MDBoost-2 algorithm to generate a strong classifier, and it is used in the intrusion detection system. The results show that, in the case of limited samples, the strong classifier integrated by MDBoost-2 has better performance with respect to MDBoost. Therefore, it can be used in the intrusion detection, which can effectively predict whether the network data is the intrusion data.

Acknowledgement

This work is supported by the science and technology project of Guangdong Province (No. 9151008990).

References

[1] Gang Wang，Jinxing Hao，Jian Ma, Lihua Huang. A new approach to intrusion detection using Artificial Neural Networks and fuzzy clusterin[J]， Expert Systems with Applications, 2010, 37(9): 6225-6232.

[2] H.Altwaijry，S.Algarny. Bayesian based intrusion detection system[J], Computer and Information Sciences, 2012, 24(1): 1-6.

[3] Pavan Singhal，Gajendra Singh. Enhanced Intrusion Detection System using Hybrid Machine Learning Approach[J], Internarional Journal of Advanced Research in Computer Science and Electronics Engineering, 2014, 3(7): 383-388.

[4] Siva S. Sivatha Sindhu，S.Geetha, A.Kannan. Decision tree based light weight intrusion detection using a wrapper approach[J], Expert Systems with Applications, 2012, 39(1): 129-141.

[5] Weiming Hu，Jun GaoH, Yanguo Wang,Ou Wu. Online Adaboost-Based Parameterized Methods for Dynamic Distributed Network Intrusion Detection[J], IEEE TRANSACTIONS ON CYBERNETICS, 2013, 44(1): 66-82.

[6] Mrutyunjaya Panda，Ajith Abraham. Manas Ranjan Patra. Discriminative multinomial Naïve

Bayes for network intrusion detection[J], IEEE TRANSACTIONS ON CYBERNETICS，2010：5-10.

[7] C.Shen，H.Li. On the dual formulation of boosting algorithms[J], IEEE Trans.Pattern Anal.Mach.Intell, IEEE Trans.Pattern Anal.Mach.Intell, 2010, 32(12): 2216-2231.

[8] C.Shen，H.Li. Boosting through optimization of margin distributions[J] , IEEE Trans. Neural Networks, 2010, 21(4): 659-666.

[9] Guo Guangxu, and Chen Songcan. Research on Margin Distribution Based Boosting Algorithms[J], IEEE TRANSACTIONS ON CYBERNETICS, 2012, 24(1): 1-6