Network Intrusion Detection Based on Naive Bayesian Model with Adjustable Weights

Yanjun DONG[1] *, Jingli LI[1], Guofu MA[2]

1Department of Information Technology, Baoding University, Baoding, 071000, China
2Department of Information Management, The Central Institute for Correctional Police, Baoding, 071000, China

***Abstract* — In view of the massive and complex network attack events in the Internet, intrusion detection must deal with a combined series of uncertain behavior processes. This makes the Naive Bayesian theorem the most suitable to solve the probabilistic events. In this paper we develop an improved Naive Bayesian classifier to the Network Intrusion Detection to deal with the deficiencies of the traditional Naive Bayesian Model. Our improved Naive Bayesian Classification Model is based on adjustable weight and is developed as follows: i) we introduce the Adjustable Weight parameter θ on the basis of the Traditional model, ii) we reduce the classifier's time complexity $O(n)$ by regulating the weight θ's value, iii) we calculate the weight θ's optimal value θ_{best} by experiment, and iv) obtain the classifier's best classification result by using the θ's optimal values. The simulation results show that the false alarm rate of the intrusion detection system using our proposed method is significantly reduced, and the results are better than those of other classification models.**

Keywords - network intrusion detection; adjustable weight; traditional naive bayesian model; improved model

I. INTRODUCTION

At present, the network security technology has become more and more important as the web-based services extend to every field of society, and the number of confidential information increases greatly. How to effectively detect and prevent network intrusion in order to protect the security of network data has become more and more attention from people of all walks of life.

In recent years, the number and severity of network attacks are increasing with the increase of Internet users and information content. According to Symantec the latest Internet security threat report[1] disclosed that from the beginning of 2004 to the end of 2012, the number of the network invasion records was increased from about 9000000 to about 6600 million, the growth rate exceeded 750%, results in economic losses worldwide each year up to an average of $11.4 billion. Therefore, the research and development of network intrusion detection is becoming more and more important. Intrusion detection system has become an essential infrastructure for most organizations.

Intrusion detection is a kind of security measure to find out a set of malicious actions that can harm the integrity, confidentiality and availability of information resources. The problem of intrusion detection is how to accurately distinguish between normal events and abnormal events, in order to achieve the purpose of filtering network attacks and reducing false alarm rate, at the same time to give consideration to the optimization of classification speed. According to the literature [2], the Intrusion Detection System based on data mining is divided into two categories: misuse detection and anomaly detection. Misuse detection is an algorithm that attempts to match the samples and features of known attack behaviors with the network traffic data. It has the advantages of low false detection rate and fast detection speed; the disadvantage of which is that the misuse detection can not detect a new type of detection behavior, but it can rely on a learning algorithm to make up for these shortcomings. In the literature [3], Guinde points out that this learning algorithm gets a data set through training, in which each event has been marked as a normal event or an intrusion event. Even if the algorithm can not detect the new attack types that are not included in the training set, but the new attack instances can be added to the training set, and the algorithm can automatically repeat training by updating the training set. For anomaly detection, Monowar believes that the technology should firstly establish a model based on normal network events, and then monitor network events in case of detachment from these models. The anomaly detection model can detect a new type of attack, because the model only depends on the known normal events. Although anomaly detection method has advantages, because of the normal events that may not be detected by the method in advance, it will lead to a high error detection rate and false negative rate, that will make it difficult to apply the method to practice[4]. Based on the characteristics of the above two models, in the research of intrusion detection, a hybrid intrusion detection model, which combines the two methods of misuse detection and anomaly detection, is proposed to improve the detection performance[5].

Therefore, the key to the development of efficient IDS is how to reduce the system false alarm rate, improve the classification accuracy. At present, many effective intrusion detection models are proposed. A new model combining network protocol analysis technique and decision tree mining technique is proposed in the literature[6], the model analyzes the data packet protocol type, and determines the best DT Algorithm for intrusion detection according to the protocol

type, this method has a good effect on Intrusion Detection in high speed network environment. There are other learning algorithms used in the field of intrusion detection, such as Support Vector Machine, Genetic Algorithm, Artificial Neural Network, etc.[7-10].

Based on the former research, this paper proposes an improved naive Bayesian model, introduced the adjustable weight parameter θ on the basis of the traditional model, obtained the best classification efficiency of the classifier by regulating the weight θ's value. Finally, compared with other methods in the simulation experiment, it is proved that the method proposed in this paper has obvious advantages in classification accuracy and error rate.

The paper's second section introduced the traditional naive Bayesian algorithm; the third section described and constructed the improved naive Bayesian model, and applied it to the network events' classification process in intrusion detection system; in the fourth section, we use the typical KDD '99 intrusion detection data set[11] in the simulation experiment to test the classification performance of the intrusion detection technology based on this algorithm, and experimental results are also given, at the same time, the results were compared and analyzed; finally, the full text is summarized.

II. CLASSIFICATION MODEL BASED ON NAIVE BAYESIAN THEORY

Bayesian decision theory is an important part of Bias's theory of induction. Bayesian decision uses subjective probability to estimate partial unknown state under incomplete information, then uses the Bayesian formula to correct the occurrence probability, and finally makes the best decision using the expected value and the correction probability. The core idea is to estimate the posterior probability by using a priori probability.

Naive Bias classification model is based on Bias's decision theory, and it is a simplified Bias probability model, the classification model which has the characteristics of simple implementation, fast classification speed and high accuracy, is one of the most widely used classification models. Its core algorithm is as follows:

The sample $A=(a_1, a_2, \cdots, a_n)$ is n-dimensional Boolean vector, used to represent whether the feature $\forall a_i$ in event A appears. Network event class $\forall C \in (C_1, C_2, \cdots, C_m, f)$ represents the classification problem of M classes. The mapping function $f: A_i \rightarrow C_j$ represents an arbitrary uncertain event instance A_i is classified as a tag $\exists C_j$ in M categories. Now there exist network event training samples $X_1, X_2, \cdots, X_N$, where $X=(x_1, x_2, \cdots, x_t)$ is the t dimension Boolean vector, training sample category $c_1, c_2, \cdots, c_k$ represent K events category label. Now consider the probability of a network event sample $\forall Y_i = (y_1, y_2, \cdots, y_n) \in (Y_1, Y_2, \cdots, Y_N, C_j)$ to be classified belonging to each category $C_j (\forall j = 1,2, \cdots, n)$. The calculation steps are as follows:

1) Estimate the probability $P(c_j)$ that the training sample is classified as a category c_j:

$$P(C = c_j) = \frac{\sum_{i=1}^{k} N(c_j)}{\sum_{s=1}^{m} T_s} \quad (1)$$

Where, $\sum_{i=1}^{k} N(c_j)$ represents the number of training samples which category is c_j, $\sum_{s=1}^{m} T_s$ represents the total number of training samples.

2) Estimate the relative probability $P(a_i|c_j)$ that the characteristic $\exists a_i$ of training sample appears in the event category $\forall c_j$:

$$P(A_i = a_i|c_j) = \frac{\sum_{i=1,j=1}^{m} S(\exists a_i \in \forall c_j)}{\sum_{i=1}^{k} N(c_j)} \quad (2)$$

Where, $\sum_{i=1,j=1}^{m} S(\exists a_i \in \forall c_j)$ stands for the number of training samples which category is $c_j (1 \le j \le m)$, and contain feature $\forall a_i$.

3) According to formula (1) 、 (2), calculate the probability $P(a_i)$ that the characteristic $\forall a_i$ of training sample appears:

$$P(a_i) = \sum_{i=1,j=1}^{k \cup c} P(a_i|c_j) P(c_j) = \sum_{i=1,j=1}^{k \cup c} \frac{\sum_{i=1,j=1}^{m} S(\exists a_i \in \forall c_j)}{\sum_{i=1}^{k} N(c_j)} P(c_j) \quad (3)$$

4) According to formula (3) and Bayesian formula, calculate the relative probability $P(c_j|a_i)$ that the sample belonging to category c_j when the characteristic $\exists a_i$ appears in the sample to be classified.

$$P(c_j|a_i) = \frac{P(c_j) \prod_{i=1}^{m} P(a_i|c_j)}{P(a_i)} \quad (4)$$

5) According to the naive Bayesian independent hypothesis, calculate the probability $P(k)$ that sample $\forall Y_i$ belongs to the category c_j.

$$P(k) = \prod_{i=1}^{m} P(c_j|a_i) P(Y = Y_i), \quad (\forall i, j, 1 \le k < \infty) \quad (5)$$

6) Using the same method, Calculate the probability $\exists P = (P_1, P_2, \cdots, P_k, f)$ that sample y_i belongs to other category $c_j (1 \le j \le k)$, use the mapping function $f: (y_1, y_2, \cdots, y_n) \rightarrow (c_1, c_2, \cdots, c_m), \ (\forall m \ne n)$ to normalize the $\forall k$ probability values. the similarity that sample y_i to be classified belongs to each category can be obtained after sorting, and the maximum a posteriori probability MAP is calculated:

$$MAP = \arg\max_{c \in M} P(c_j|y_1) = \arg\max_{c \in M} \frac{\prod_{i=1}^{n} P(y_1|c_j) P(c_j)}{P(y_1)} = \arg\max_{c \in M} \prod_{i=1}^{n} P(y_1|c_j) P(c_j) \quad (6)$$

7) According to the above conditions, the Naive Bayesian classifier (NBC) formula is defined as follows:

$$Classifier(y_1, y_2, \cdots, y_n)$$

$$= \arg\max_{c_j \in C} P(c_j) \prod_{i=1, j=1} P(y_i | c_j) \quad (7)$$

On the basis of the above mathematical model, the structure of NBC is constructed. The classifier has a node of category $\exists C \in (C_1, C_2, \cdots, C_m)$ and n nodes A_i to be classified. Each A_i is composed of n independent characteristic values a_i, the mapping function between nodes is:

$$f: \{C_1, (A_1, (a_1, a_2, \cdots, a_m))$$
$$C_2, (A_2, (a_1, a_2, \cdots, a_m)), \cdots,$$
$$C_k, (A_k, (a_1, a_2, \cdots, a_m))\}, \ (\forall m \neq n).$$

All the nodes A_i to be classified are subordinate to the common category node C, and the relationship between each node to be classified is independent of each other. The schematic diagram of NBC structure is shown in Figure 1.

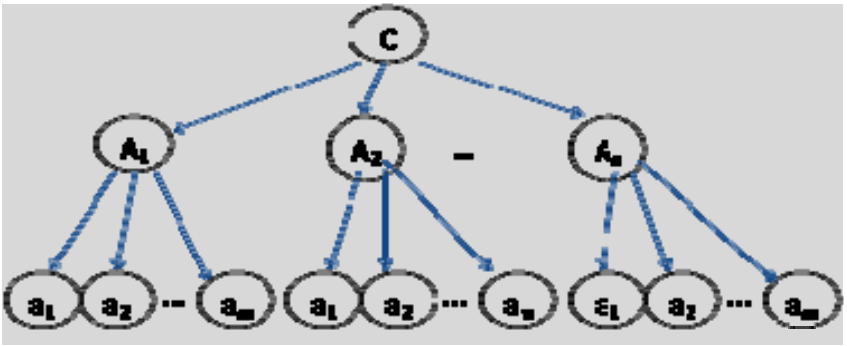

Figure 1. NB Classification Model Structure Framework

From Figure 1, we can see that the structure meaning of the model and the relation between each node in the relational model describe the Bayesian decision theorem's core idea , that the posterior probability is calculated by a priori probability, then the training learning process of sample set nodes is updated, finally, the classification of the maximum likelihood is calculated. Therefore, although the assumption of independence is often inaccurate , the experimental data show that some characteristics of NBC can still get satisfactory classification results in practice.

III. CONSTRUCTION AND APPLICATION OF THE IMPROVED NAIVE BAYESIAN CLASSIFIER

A. Construct The Improved Naive Bayesian Classifier with Adjustable Weight Parameter θ

In theory, the naive Bias classification algorithm determines the final classification node $c_j \ (1 \leq j \leq n)$ by calculating the maximum posteriori probability MAP that the node to be classified belongs to its category. But in practice, the influence of different factors, such as the different characteristics of the sample nodes selected and the different training samples set, will cause the NBC classification accuracy to a certain degree of decline[12]. Therefore, In view of the defects of the NBC, the adjustable weight parameter θ was introduced into the traditional model, the classification accuracy of NBC is improved and the error rate is reduced to the maximum extent by adjusting the weight θ's value. In network intrusion detection, the network event $A_i (\forall i, 1 \leq i \leq n)$ to be classified is classified as normal C_N or abnormal events $C_j (\forall j, 1 \leq j \leq m)$. Where full set $U_c = \{C_N, C_j | \exists P(C_N) + P(C_j) = 1\}$. Usually the posterior probability $P(C_N|A_i)$ and $P(C_j|A_i)(\forall i, j, 1 \leq i \leq j \leq n)$ is calculated separately according to the naive Bayesian formula. Therefore, Judge $P(C_j|A_i) > P(C_N|A_i)$ only according to the maximum posteriori probability MAP get by calculation, finally the network event to be classified is classified into a certain event category, that is not rigorous enough. This will not only result in higher false detection rate, but also create an error sample set. So then the predictive ability and robustness of the intrusion detection system are reduced to a certain extent.

Therefore, this paper introduces the adjustable weight parameter θ into the traditional model, improves classification accuracy by regulating the weight θ's value. and by properly choosing the parameter θ's value makes NBC achieve the best classification results. Further derived from the above formula: $\frac{P(C_j|A_i)}{P(C_N|A_i)} > 1$, after introducing the parameter θ, the formula is converted into: $\ln \frac{P(C_j|A_i)}{P(C_N|A_i)} > \theta$, finally , derived improved discriminant affected by the θ's value.

$$\frac{P(C_j|A_i)}{P(C_N|A_i)} > e^{\theta} \rightarrow \frac{P(C_j|A_i)}{1-P(C_j|A_i)} > e^{\theta} \rightarrow P(C_j|A_i) > \frac{e^{\theta}}{1+e^{\theta}} \quad (8)$$

Discriminant (8) shows that, the specific value θ get is substituted into the mapping function $f: \{U_{MAX}[\frac{e^{\psi}}{1+e^{\psi}}] \rightarrow \psi\}$ to calculate the corresponding discriminant value ψ . According to the clamping force theorem, given the $\lim P(C_j|A_i) = \psi$ and any positive α , where $\exists n > N$, $|P(C_j|A_i) - \alpha| < \psi$, that is $\alpha - \psi \leq P(C_j|A_i) \leq \alpha + \psi$, among which $\alpha - \psi > 0$ *and* $\alpha + \psi < 1$. Finally, the most suitable ψ's value can be estimated according to the

calculated limit $f(X_p)$ value, thus event A_i belongs to a category of C_j can be more accurately determined, where the best value θ gets is calculated based on the comparison of the experimental data.

B. *Application of The Improved Naive Bayesian Classifier In Intrusion Detection Process*

The implementation of intrusion detection method is to design a classifier to separate the normal and abnormal data from the data stream, so as to carry out the alarm to the attack behavior[13]. This paper applied the improved NBC to the classification module of intrusion detection model, through a series of intrusion detection module processing, and finally gets a reasonable classification set of network events. The intrusion detection process of the model is shown in Figure 2 below:

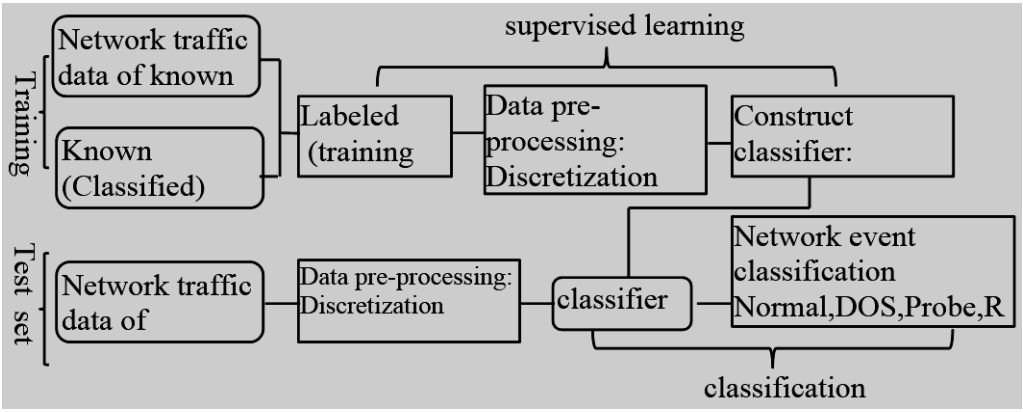

Figure 2. Intrusion detection process based on NBC.

According to the flow structure in Figure 2, we can know that, the first stage the known network traffic data T_k is combined with the known category C_k in the sample by the mapping set $U_{TC}=(\exists T_k, C_k | \forall T_k \in C_k)(1 \le k \le n)$, then join the training set (TRS) for training. Data which has been marked as complex should be pre -processed (discretization, feature selection) firstly, then the processed data which were effective and simple were counted and calculated. Finally, the event classifier based on NB algorithm is established based on the estimated prior probability $P(T_k|C_k)$, $(\exists k = 1,2,\cdots,n)$.

The whole detection process of this stage is a repetitive supervised learning process, in order to gradually improve the classification data of samples, so that the classifier has a better predictive effect. The second stage, firstly the unknown network traffic data T_u in test set (TES) should be pre-processed (discretization) before classification, then TES is classified according to the pre-established NBC: $\{(U_u = (T_u, C_k)) \sqsubseteq (U_k = (T_k, T_u, C_k))\}$ in the first stage, then the unknown event will be matched with marked event in the sample by mapping the relationship function $f_{NBC}: T_u \to C_k$, and then get the categories $\exists C \in \{(C_j \neq \emptyset,\ 1 \le j \le n) | Normal, DOS, Probe, R2L, U2R\}$, of network events $X = (X_1, X_2, \cdots, X_N)$ to be classified, at the same time, the sample set is updated in order to complete the detection performance of the model. The data of TRS and TES used in the whole process of the framework is from the classic KDD Cup1999（KDD'99）intrusion detection data set, the design pattern of the data set can test the generalization ability and prediction ability of the classifier model. In the pre-process of massive and complex network data before testing, this paper uses the idea of discretization and feature selection. The purpose of data processing is, (1) to make abstract network data concrete; (2) to remove redundant features and non-important features to simplify the data; (3) to reduce the time and space complexity, to improve the training speed and detection accuracy of the classifier.

IV. THE EXPERIMENT RESULT ANALYSIS

A. *Intrusion Detection Data Set*

In this paper, the experimental data using KDD Cup1999（KDD'99）intrusion detection data set, the data set includes 2 parts: 7 weeks of training data, including about 5000000 network connection records, the remaining 2 weeks of test data, which contains about 200000 network connection records. The category for each network connection is marked as normal or abnormal, exception types are subdivided into 4 categories, a total of 39 types of attacks, among them 22 attack types appear in the training set, the other 17 kinds of unknown attack types appear in the test set. Descriptions and examples of the 4 types of attack are shown

in Table 1, the data in Table 2 are the number of 5 types of network event records and the distribution of them in the TRS and TES of 10%KDD data set.

TABLE 1. EXCEPTION TYPE

Attack type	Type description	Typical examples
DOS	Denial of service	Ping-of-death、SYN flood
R2L	Remote-to-Login Attack	Guess password
U2R	User-to-Root Attack	Buffer overflow attack
Probe	Port monitoring or scanning	Port-scan

TABLE 2. DISTRIBUTION OF NETWORK EVENT RECORDERS IN 10%KDD DATA SET

Attack type	Training sample number	Test sample number	Distribution in the training set（TRS）	Distribution in the test set（TES）
Normal	97277	60592	19.69%	19.48%
DOS	391458	237594	79.24%	73.91%
R2L	1126	8606	0.23%	5.20%
U2R	52	70	0.01%	0.07%
Probe	4107	4166	0.83%	1.34%
Total	494020	311028	100%	100%

B. Experimental Results Analysis

The experimental environment of this study is based on the windows7 operating system, the hardware environment is Corei3-6100, 8G RAM. MATLAB7.13 is used as a simulation software.

The intrusion detection data sets which is used in the simulation experiment are the TRS and TES of the 10%KDD'99 intrusion detection data set that mentioned above. The experimental results are listed in table 3 and table 4.

TABLE 3. COMPARISON OF NB ALGORITHM'S DETECTION RATE BEFORE AND AFTER IMPROVEMENT

Method	Normal	DOS	R2L	U2R	Probe
Improved NB algorithm（TR%）	98.69	99.25	99.17	98.74	97.81
Improved NB algorithm（FP%）	0.07	0.05	0.08	0.14	0.03
Traditional NB algorithm（TR%）	96.28	95.87	92.12	93.85	94.74
Traditional NB algorithm（FP%）	0.09	0.06	0.13	0.18	0.08

TABLE 4. INFLUENCE OF θ'S VALUE ON DETECTION ACCURACY

θ's value	0.0	0.5	1.0	1.5	2.0	2.5	3.0
TR（%）	95.26	96.7	97.63	98.21	97.74	95.62	94.45
FP（%）	9.42	7.16	6.55	6.28	7.83	8.24	9.71

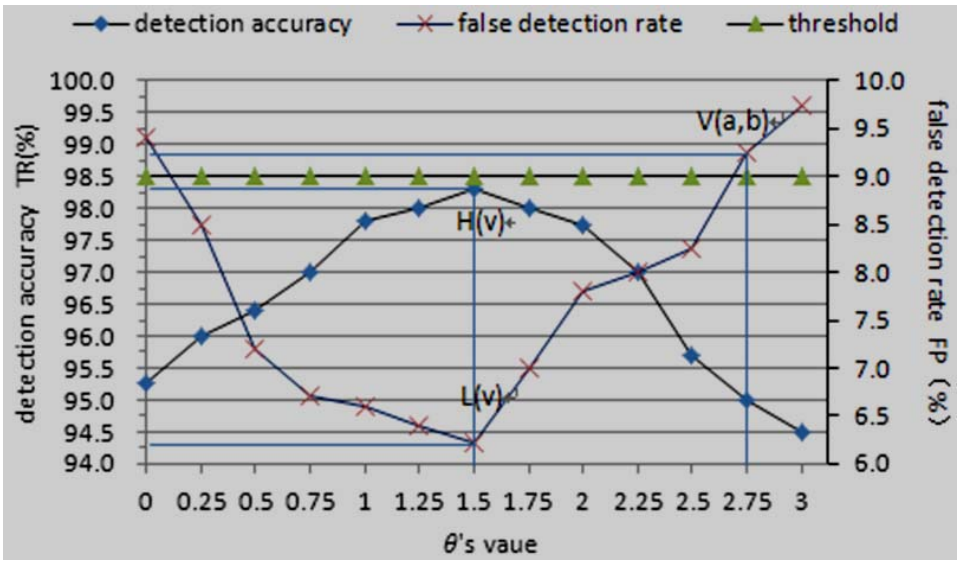

Figure 3. Distribution probability of θ's value.

(1) The experimental results of table 3 shows that, compared with the traditional NB model, the improved NB model has a significant improvement in classification accuracy (TR) and false detection rate (FP) in the process of judging the intrusion events, so it can be proved that, this paper introduced the adjustable weight parameter θ into the traditional NB model, the method that improve the classification accuracy by adjusting the parameter θ's value is feasible.

(2) The data in table 4 shows the effect of different values that θ gets on NBC's detection accuracy. With the gradual increase of θ's value $(\forall\theta \in [0,\infty))$, FP and TR will show a high peak value H (v) and a low value L (v) respectively, the θ's value corresponding to these two values is the best θ's value for the experiment, this can be illustrated by figure 3. Firstly, the reasonable range $(\forall\theta \in [0.5,2.5])$ of parameters is defined by setting threshold. Then, the weight parameter θ's optimal value $\theta_{test} = 1.5$ using the parameter sample set $U_u = \{V_\theta = (0.5,0.6,\cdots,2.5)|f_u\}$ which was already established is calculated according to the mapping function $f_u: \{[V_{HL} = (H(v), L(v),)] \to V_\theta\}$. Finally, the value θ_{test} will be substituted into the improved model, thus optimal detection performance of intrusion detection model based on the algorithm proposed in this paper is obtained.

(3) The improved NB algorithm proposed in this paper is compared with other classification algorithms through simulation experiment using the same experimental data, the experimental results are shown in table 5. From the experimental data in table 5, it can be seen that, the improved algorithm is superior to other classification algorithms in the 2 indexes of detection accuracy and error rate.

TABLE 5. PERFORMANCE COMPARISON OF VARIOUS EVENT CLASSIFICATION METHODS

Method	Accuracy（%）	Error rate（%）
Improved Naive Bayesian	98.21	6.28
Decision Tree	97.83	7.72
SVM	97.76	8.13
Naive Bayesian	96.95	9.85
Artificial Neural Network	94.98	9.47
Genetic Algorithm	91.64	8.82

V. CONCLUSION

It is a common method that the naive Bayesian algorithm is applied to network intrusion detection, but in practical applications, the classification performance of naive Bayesian algorithm is easy to be affected by various factors, such as the different characteristics of the sample nodes selected and the different training samples set, these factors can cause the classification accuracy of naive Bayesian model to decline in different degrees. In view of the defects of the NBC, the Adjustable Weight parameter θ is introduced into the traditional model in this paper, the classification performance of the classifier is improve by adjusting the weight θ's value, and the θ's optimal value is calculated by the experimental data, and the classifier can obtain the best classification performance by using the θ's optimal value. Finally, the simulation experiment results show that, the error rate of intrusion detection system based on the improved model proposed in this paper is obviously reduced, and compared with other classification models' classification results, the improved model proposed in this paper also shows obvious advantages.

ACKNOWLEDGMENTS

The authors thank the reviewers who gave a through and careful reading to the original manuscript. Their comments are greatly appreciated and have help to improve the quality of this paper. This work is supported by the Science Research Fund Project of Baoding University(2016).

REFERENCES

[1] 2013 Internet Security Threat Report (symantec.com [OL]. http://www.symantec.com/safety_response /Publications/threatreport.jsp).

[2] Wei Jie, Shi Hong-bo and Ji Su-qin, "Distributed Naive Bayesian text classification based on the Hadoop," Computer System & Application, vol. 21, pp. 210-213, 2012.

[3] Guinde N B and Ziavras, "Efficient hardware support for pattern matching in network intrusion detection," Computers & security , vol. 29, pp. 756-769, 2010.

[4] Chu C T, Kim S K and Lin Y A,et al, "Map-Reduce for machine learning on multicore," 20th Annual Conference on Neural information Processing Systems , vol. 16, pp. 281-288, 2007.

[5] Deng Su and Chang He, "Four kinds of Bayesian classifiers and their comparison," Journal of Shenyang Normal University (Natural Science Edition) , vol. 26, pp. 31-33, 2008.

[6] Yang Jie, Chen Xin and Wan Jian-xiong, "Research on Intrusion Detection Model Based on network protocol analysis and decision tree mining," Computer Applications and Software , vol. 27, pp. 19-55, 2010.

[7] Duan Dan-qing, Chen Song-qiao and Yang Wei-ping, "Network intrusion detection system based on SVM active learning algorithm," Computer Engineering and Science , vol. 28, pp. 33-36, 2006.

[8] J. L. Zhao, J. F. Zhao and J. J. Li, "Intrusion Detection Based on Clustering Genetic Algorithm," Machine learning and Cybernetics , vol. 6, pp. 3911-3914, 2005.

[9] Yang Jie, Zheng Ning and Liu Dong, "Optimization of Features with Weight and Model Parameters of SVM Based on Genetic Algorithm," Computer Simulation , vol. 25, pp. 115-118, 2008.

[10] Mao Jian, Zhao Hong-dong and Yao Jing-jing, "Application and Prospect of Artificial Neural Network," Electronic Design Engineering , vol. 19, pp. 62-65, 2011.
[11] http://kdd.ics.uci.edu/databases/kddcup99/task.html.
[12] Dong Li-yan, Sui Peng, Sun Peng and Li Yong-li, "Novel naive Bayesian classification algorithm Based on semi-supervised learning," Journal of Jilin University(Engineering and Technology Edition) , vol. 46, pp. 884-889, 2016.
[13] Zhang Kun, Cao Hong-xin, Yan Han and Liu Feng-yu, "Application of Support Vector Machines on Network Abnormal Intrusion Detection," Application Research of Computers , vol. 5, pp. 98-100, 2006.