A Novel Intrusion Detection System Based on Advanced Naive Bayesian Classification

Yunpeng Wang[1,4,5,6], Yuzhou Li[1,4], Daxin Tian[1,4](✉),
Congyu Wang[1,4], Wenyang Wang[2], Rong Hui[2], Peng Guo[2],
and Haijun Zhang[3]

[1] Beijing Advanced Innovation Center for Big Data and Brain Computing,
Beihang University, XueYuan Road No. 37, Beijing 100191, China
dtian@buaa.edu.cn
[2] China Automotive Technology and Research Center,
Automotive Engineering Research Institute,
East Xianfeng Road No. 68, Tianjin 300300, China
[3] School of Computer and Communication Engineering,
University of Science and Technology Beijing,
XueYuan Road No. 30, Beijing 100083, China
[4] Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems
and Safety Control, School of Transportation Science and Engineering,
Beihang University, XueYuan Road No. 37, Beijing 100191, China
[5] Jiangsu Province Collaborative Innovation Center of Modern Urban
Traffic Technologies, Si Pai Lou. 2, Nanjing 210096, China
[6] Key Lab of Urban ITS Technology Optimization and Integration,
The Ministry of Public Security of China, Hefei 230088, China

Abstract. Intrusion Detection System is a pattern recognition task whose aim is to detect and report the occurrence of abnormal or unknown network behaviors in a given network system being monitored. In this paper, we propose a machine learning model, advanced Naive Bayesian Classification (NBC-A) which is based on NBC and ReliefF algorithm, to be used in the novel IDS. We use ReliefF algorithm to give every attribute of network behavior in KDD'99 dataset a weight that reflects the relationship between attributes and final class for better classification results. The novel IDS has a higher True Positive (TP) rate and a lower False Positive (FP) rate in detection performance.

Keywords: IDS · Information security · NBC · ReliefF
Detection performance · KDD'99

1 Introduction

Network in a profound impact on people's lives and ways of working at the same time, it also brings a lot of security risks and threats. A variety of viruses, security vulnerabilities, attacks have caused the loss of users, enterprise, government, even national security. With increasing network security incidents in recent years, people have a strong sense of security and privacy protection. Therefore, the well-designed security system is a very

K. Long et al. (Eds.): 5GWN 2017, LNICST 211, pp. 581–588, 2018.
https://doi.org/10.1007/978-3-319-72823-0_53

important and urgent problem in the field of network information security especially in next generation networks (5G) [1, 2] with great security challenges.

At present, the network information security protection measures are divided into passive security and active security. Passive security includes data encryption [3, 4], security authentication [5], firewall [6] and other measures and these Active security is a technology represented by Intrusion Detection System (IDS) [7] which detect possible intrusion by collecting network datasets or information, and sending alerts and responding before an intrusion occurs, or before a hazard occurs. With the development of IDS, even that it could replace the traditional network security measures.

In recent years, with the rise of machine learning (ML)-related models [8], it is becoming a trend to apply machine learning methods into intrusion detection system. In the field of machine learning, Naive Bayesian Classification (NBC) [9] is widely used as the most classical learning algorithm with good classification accuracy. However, the NBC is based on the independence of event attributes, which is difficult to achieve in realistic network behaviors especially in future networks (such as 5G) with great complexity [10]. In response to this shortage, many scholars have put forth an improved method which is based on different attribute weights. Paper [11] propose a weighted BNC model based on Rough Set, it could performance well in small data sets and could do some changes in original information. Paper [12] use the value of every attribute to be as weights, but the attributes attribute are more, each weighted coefficients are small, it cannot play its role in real complex networks. Paper [13] propose a weighted NBC based on correlation coefficient, and it could improve the classification ability of the Bayesian, but the current measure formula is not described accurately for all conditions.

In this paper, we propose a novel IDS based on machine learning an advanced Naive Bayesian Classification (NBC-A) which we give every attributes a weight to reflect the relations between attributes and final classification results. We use the ReliefF [14] algorithm that is robust [15] and can deal with incomplete and noisy data to estimate weights and we get a higher True Positive (TP) rate and a lower False Positive (FP) rate [16] in detection performance that means it has better performance than NBC.

In this paper, we introduce some network information security related works and machine learning works in Sect. 2; we proposed the IDS based on NBC-A in Sect. 2.1; detection performance based on dataset KDD’99 and analysis is in Sect. 3; conclusion and outlook are in Sect. 4.

2 Advanced Naive Bayesian Classification (NBC-A) Model in Intrusion Detection System (IDS)

The method of intrusion detection is to design a network behavior classifier to distinguish the normal and abnormal data in the dataset, simulation or realistic network, so as to realize the alarm function of the attacking behavior. At the same time, intrusion detection by IDS is an uncertain behavior, and Naive Bayes theory is suitable for uncertain probabilistic events. Therefore, the introduction of intrusion detection technology based on NBC in IDS research design is completely reasonable.

2.1 Naive Bayesian Classification (NBC)

The Bayesian decision-making theory provides a probabilistic approach to reasoning. It assumes that the variables to be investigated follow certain probability distributions and can reason from these probabilities and observed data to make optimal decisions. Naive Bayesian Classification (NBC) model based on Bayesian decision-making theory [17], is a simplified Bayesian probability model. The classification model is simple in implementation, fast in classification and high in accuracy. It is one of the most widely used classification models in machine learning.

Given a data set of K attributes and assumed that the values of the K attributes are discrete, the purpose of classification is to predict the type of every case in the test set which is a part of dataset (the other part is train set whose task is to make the NBC's train). We can give a specific example, whose attributes are from a_1 to a_k. The probability of the example belonging to class C_i is $P(\mathrm{C} = c_i|A_1 = \mathrm{a}_1, \ldots, A_k = \mathrm{a}_k)$. Obviously, according to Bayesian decision-making theory:

$$P(\mathrm{C} = c_i|A_1 = \mathrm{a}_1, \ldots, A_k = \mathrm{a}_k) = \frac{P(A_1 = \mathrm{a}_1, \ldots, A_k = \mathrm{a}_k|\mathrm{C} = c_i)P(\mathrm{C} = c_i)}{P(A_1 = \mathrm{a}_1, \ldots, A_k = \mathrm{a}_k)} \tag{1}$$

Here, $P(\mathrm{C} = c_i)$ is a prior probability and can be easily calculated from train set. In data set, $P(A_1 = \mathrm{a}_1, \ldots, A_k = \mathrm{a}_k)$ is same to every class c_i and it assumes that the values of attributes are independent, we can know:

$$P(A_1 = \mathrm{a}_1, \ldots, A_k = \mathrm{a}_k) = 1 \tag{2}$$

$$P(A_1 = \mathrm{a}_1, \ldots, A_k = \mathrm{a}_k|\mathrm{C} = c_i) = P(A_1 = \mathrm{a}_1|\mathrm{C} = c_i)\ldots P(A_k = \mathrm{a}_k|\mathrm{C} = c_i) \tag{3}$$

Putting Formula (2) and (3) in to Formula (1), we can get the method used by Naive Bayesian Classification, that is:

$$V_{NBC}(x) = \arg max\, P(\mathrm{C} = c_i) \prod P\left(A_j = \mathrm{a}_j|\mathrm{C} = c_i\right) \tag{4}$$

Here $V_{NBC}(x)$ is indicated the target value out by NBC indicated that the output target. In theory, NBC has the minimum misclassification rate, compared with all the other classification algorithms and it is suitable to be used in IDS to find abnormal behaviors in network.

2.2 Attribute Weighted Naive Bayesian Classification

However, the independence assumption is difficult to meet in the real network behaviors, each network behavior has its own attributes, which have complex relationships and can directly affect the results of intrusion detection judgments. Paper [18] established an attribute weighted NBC, is assigned to give different weights to each attribute to make these relationships effect on NBC:

$$V_{wNBC}(x) = \arg max\, P(\mathrm{C} = c_i)\prod P\left(A_j = \mathrm{a}_j|\mathrm{C} = c_i\right)^{W_j} \tag{5}$$

Here, W_j is the weight of A_j. Different W_j has different inferences on NBC, great W_j makes great impacts on IDS. The key of NBC in IDS is how to determine the weights of different attributes.

2.3 The W_j Determined by ReliefF Algorithm

In the next generation network (such as 5G), the network behavior of the relationship will be far more complex than the current network. However, the algorithm for determining w_j are focusing on relations among on attributes instead of class C_i, these algorithms be very difficult to play a role because of the high complexity. Therefore, we propose to use ReliefF algorithm, which directly focuses on the relationship between attribute and final classification (class C_i) results rather than the relationship between attribute and attribute. The ReliefF algorithm as follows:

ReliefF algorithm is a multi-class attribute selection algorithm proposed by Kononenko. Its basic idea is to assign a weight value to each attribute in the attribute set, assign a higher weight to the attribute which could has direct and high relation to final classification (class C_i). For that purpose, given a randomly selected network behavior X_i (line 3), ReliefF searches for its two nearest neighbors: one from the same class, called nearest hit H, and the other from the different classes (class C_i, o $\neq$ i), called nearest miss M (line 4). Function $diff(A, I_1, I_2)$ (line 6) calculates the difference between the values of the attribute A for two network behaviors I_1 and I_2. The whole process is repeated for m (line 2) times, where m is a user-defined parameter. i, j, o and k are count constants.

ReliefF Algorithm for determining W_j:

Input: for each behavior X_i in train set attributes $A(A_1 = a_1, \ldots, A_j = a_j, \ldots, A_k = a_k)$ values and the class C_i.

Output: the vector $W(W_1 = w_1, \ldots, W_j = w_j, \ldots, W_k = w_k)$ of weights of the qualities of attributes A.

```
Step1 set all W as an initial value $W_j = 0$;
Step2 for $i := 1$ to $m$ do begin
Step3 randomly select a network behavior $X_i$;
Step4 find nearest $H_s(s = 1, 2, \ldots, q)$ in hit $H$ and nearest $M_s(s = 1, 2, \ldots, q)$ in miss $M$;
Step5 for $j := 1$ to $k$ do
Step6   $W_j := W_j - \sum_{s=1}^{q} \frac{diff(A, X_i, H_s)}{mq}$
            $+ \sum_{\substack{Y \neq class\, C_i \\ Y = class\, C_{io}}} \left[\frac{P(Y)}{1 - P(classC_i)} \sum_{s=1}^{q} diff(A, X_i, M_s)\right] / (mq)$;
Step7 end;
```

ReliefF Algorithm does not set the value range of W_j, it may be negative, in order to avoid this situation, the proposed standardized operation [19] of W_j, the formula is as follows:

$$W_j' = \frac{W_j - min_W}{max_W - min_W} \tag{6}$$

Here W_j' is the standard W_j, min_W is the minimum value of W and max_W is the maximum one.

2.4 The Processes of the Novel IDS Based on Model NBC-A

Combining 3.1–3.3, we get the advanced NBC (NBC-A) which be used in the novel IDS proposed in this paper is:

$$V_{NBC-A}(x) = \arg max\, P(\mathrm{C} = c_i) \prod P\left(A_j = \mathrm{a}_j | \mathrm{C} = c_i\right)^{W_j'} \tag{7}$$

$$W_j' = \frac{W_j - min_W}{max_W - min_W} \tag{8}$$

$$\begin{aligned} W_j &:= W_j - \textstyle\sum_{s=1}^{q} diff(A, X_i, H_s)/(mq) \\ &\quad + \sum_{\substack{Y \neq class\, C_i \\ Y = class\, C_{io}}} \left[\frac{P(Y)}{1-P\left(class C_i\right)} \textstyle\sum_{s=1}^{q} diff(A, X_i, M_s)\right]/(mq) \end{aligned} \tag{9}$$

We divide the IDS into two processes: in the train process, the Train Set includes the known network behavior data and the marked classes, and then Preprocesses: discretization and feature selection. Finally, we use the ReliefF algorithm to weight the feature to get NBC-A; in the test process, the Test Set includes unknown network behavior data, and then discretization, and finally the use of NBC-A to get behavior classification results. The processes of the novel IDS as follows (Fig. 1):

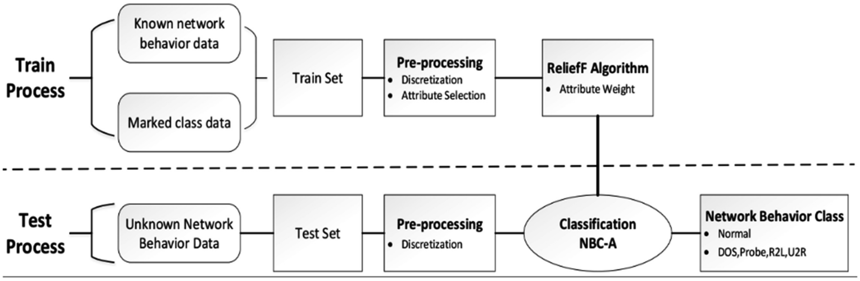

Fig. 1. The whole process is divided into 2 parts: Train Process and Test process, the model NBC-A based on NBC and ReliefF algorithm is used to get the network behavior class in Test Process.

3 Detection Performance and Analysis

3.1 Dataset KDD’99 for Detection Performance

We utilize dataset KDD Cup 1999 (KDD’99) [20] to be as the data for detection performance. KDD’99 is the standard dataset of intrusion detection and consists of two parts: 7 weeks of train data set, about 5,000,000 network connections; 2 weeks of test data set, about 2,000,000 network connections. Each network connection record is marked as normal (Normal) or abnormal (Anomaly), abnormal type is divided into 4 categories of 39 kinds of attack types. For time saving and computer performance, we utilize 10% KDD’99 to be the performance data. The distribution of data as follows (Table 1):

Table 1. The distribution of intrusion types in 10% KDD’99

Type	Train examples	Test examples	10% KDD’99 Train distribution	10% KDD’99 Test distribution
Normal	97,278	60,593	19.69%	19.48%
Probe	4,107	4,166	0.83%	1.34%
DOS	391458	229,853	79.24%	73.90%
U2R	52	228	0.01%	0.07%
R2L	1,126	16,189	0.23%	5.20%
Total	497,021	31,1029	100%	100%

3.2 Performance Analysis

In detection performance, the experimental platform environment is: Operation System: Windows 7 ultimate, CPU 3.00 GHz, RAM 8 GB, Hard Disk 500G; Programming tools: Spyder (Python 2.7), Dataset: 10% KDD’99 (80% Train Set and 20% Test Set), the detection performance results as follows:

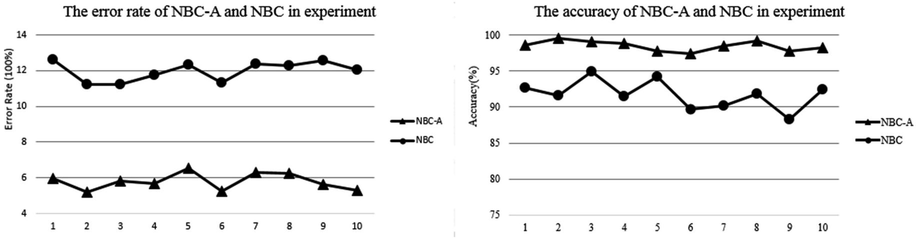

Fig. 2. Accuracy and error rate are the accuracy of NBC-A is higher than NBC and the error rate is lower than NBC, it means NBC-A has a better performance than NBC in detection performance.

(1) From Fig. 2, we could find that the accuracy of NBC-A is greatly higher than NBC one, and the error rate of NBC-A is lower than the NBC in intrusion detection performance. Actually, the average of NBC-A accuracy is 98.50% and NBC is 91.73%. The average of NBC-A error rate is 5.79% and the NBC is 11.98%. It means that model NBC-A can performance greatly better than in NBC in data mining by using ReliefF algorithm.

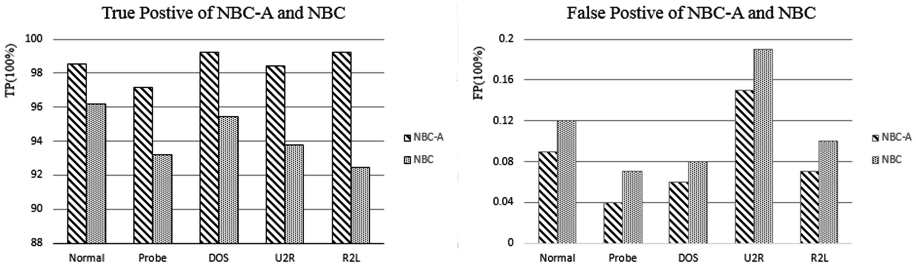

Fig. 3. TP of NBC-A is high than NBC and FP of NBC-A is lower, that means novel IDS based on NBC-A is more secure and useful the NBC one.

(2) From Fig. 3, for different types of intrusion attacks, the TP of NBC-A is generally higher than that of NBC and the FP of NBC-A is much lower than that of NBC. This means the novel IDS we propose in this paper, uses ReliefF algorithm to weight the class attributes to obtain a good effect, can effectively detect the intrusion behaviors in the networks, to ensure the safety of the system.

4 Conclusion and Outlook

Aiming at the massive and complex network attacks on the current Internet or next generation network 5G (the more massive and complex), it is reasonable to apply the NBC of Naive Bayes decision theory to IDS. We propose a novel IDS based on NBC-A which makes improvements on the NBC. The novel IDS we propose in this paper utilizes ReliefF algorithm to estimate attribute. Compared with other algorithm, Relief is more robust and efficient, it directly reflects the relationship between attributes and final class results. Model NBC-A is more suitable for large-scale and high-complexity networks. The detection performance shows that we get a higher TP rate and a lower FP rate in detection performance that means that NBC-A has better performance than NBC and practical application significance.

In this novel IDS, we have a lot of improvement in the performance of classifier space, in the future work, how to more effectively combine ReliefF algorithm and relationship of attributes between attributes, and other machine learning classification algorithm, which can further enhance the classifier's ability of intrusion detection for complex network.

Acknowledgments. This research was supported by the National Key Research and Development Program of China (2016YFB0100902).

References

1. Rappaport, T.S., Sun, S., Mayzus, R., Zhao, H.: Millimeter wave mobile communications for 5G cellular: it will work! IEEE Access **1**(1), 335–349 (2013)
2. Boccardi, F., Heath, R.W., Lozano, A., Marzetta, T.L.: Five disruptive technology directions for 5G. IEEE Commun. Mag. **52**(2), 74–80 (2013)
3. Matsui, M.: The first experimental cryptanalysis of the data encryption standard. In: Desmedt, Y.G. (ed.) CRYPTO 1994. LNCS, vol. 839, pp. 1–11. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-48658-5_1
4. Biryukov, A., Cannière, C.D.: Data encryption standard (DES) (2005)
5. Lowe, G.: An attack on the Needham-Schroeder public-key authentication protocol. Inf. Process. Lett. **56**(3), 131–133 (1995)
6. Manner, J., Karagiannis, G., Mcdonald, A.: NSIS signaling layer protocol (NSLP) for quality-of-service signaling. IETF **31**(2), 152–160 (2010)
7. Huang, M.Y., Jasper, R.J., Wicks, T.M.: A large scale distributed intrusion detection framework based on attack strategy analysis. Comput. Netw. **31**(23–24), 2465–2475 (1998)
8. Bache, K., Lichman, M.: UCI Machine Learning Repository (2013)
9. Flach, P.A., Lachiche, N.: Naive Bayesian classification of structured data. Mach. Learn. **57** (3), 233–269 (2004)
10. Muirhead, D., Imran, M.A., Arshad, K.: Insights and approaches for low-complexity 5G small-cell base-station design for indoor dense networks. IEEE Access **3**, 1562–1572 (2015)
11. Deng, W., Wang, G., Wang, Y.: Weighted Naive Bayes classification algorithm based on rough set. Comput. Sci. **34**(2), 204–205 (2007)
12. Cheng, K., Zhang, C.: Feature-based weighted Naive Bayesian classifier (2006)
13. Yao, S., Li, L.: Weighted Naïve Bayesian classification algorithm based on correlation coefficients. Int. J. Adv. Comput. Technol. **4**(20), 29–35 (2012)
14. Robnikšikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. **53**(1), 23–69 (2003)
15. Deng, Z., Chung, F.L., Wang, S.: Robust relief-feature weighting, margin maximization, and fuzzy optimization. IEEE Trans. Fuzzy Syst. **18**(4), 726–744 (2010)
16. Hamid, Y., Sugumaran, M., Journaux, L.: Machine learning techniques for intrusion detection: a comparative analysis. In: International Conference on Informatics and Analytics (2016)
17. Fei, Z., Guo, J., Wan, P., Yang, W.: Fast automatic image segmentation based on Bayesian decision-making theory. In: International Conference on Information and Automation (2009)
18. Zhang, C., Wang, J.: Attribute weighted Naive Bayesian classification algorithm. Microcomputer Information, pp. 27–30 (2010)
19. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 2nd edn. (The Morgan Kaufmann Series in Data Management Systems) (2006)
20. Pfahringer, B.: Winning the KDD99 classification cup: bagged boosting. ACM SIGKDD Explor. Newsl. **1**(2), 65–66 (2000)