An Effective Intrusion Detection Model Based on Pls-Logistic Regression with Feature Augmentation

Jie Gu^{1,2}(✉)

¹ Postdoctoral Research Station, Agricultural Bank of China, Beijing 100005, China
gujie@pku.edu.cn

² School of Electronics Engineering and Computer Science, Peking University,
Beijing 100871, China

Abstract. Computer network is playing a significantly important role in our society, including commerce, communication, consumption and entertainment. Therefore, network security has become increasingly important. Intrusion detection systems have received considerable attention, which not only can detect known attacks or intrusions, but also can detect unknown attacks. Among the various methods applied to intrusion detection, logistic regression is the most widely used, which can achieve good performances and have good interpretability at the same time. However, intrusion detection systems usually confront with data of large scale and high dimension. How to reduce the dimension and improve the data quality is significant to improve the detection performances. Therefore, in this paper, we propose an effective intrusion detection model based on pls-logistic regression with feature augmentation. More specifically, the feature augmentation technique is implemented on the original features with goal of obtaining high-qualified training data; and then, pls-logistic regression is applied on the newly transformed data to perform dimension reduction and detection model building. The NSL-KDD dataset is used to evaluate the proposed method, and the empirical results show that our proposed method can achieve good performances in terms of accuracy, detection rate and false alarm rate.

Keywords: Feature augmentation · Intrusion detection · Logistic regression · Partial least square · Network security

1 Introduction

With the rapid development of internet, networks are becoming more and more important in our daily life. Organizations rely heavily on networks to do on-line transactions, and also, individuals are dependent on networks to work, study and entertain. In a word, networks are an essentially indispensable part in modern society. However, this over-dependence on networks might have potential risk, because considerable information that relates to organization operation and individual activities is accumulated and stored. It would cause huge losses, when the networks are been invaded or attacked.

Intrusion detection systems are the most widely used tool to protect information from being compromised. Intrusion detection has been long considered as a classification problem [1, 2]. Various statistic-based and machine-learning-based methods have been applied to improve the performances of intrusion detection systems [3, 4]. However, machine learning-based methods for intrusion detection suffer criticisms [5]. Though many machine-learning-based detection methods, such as support vector (SVM) machine and artificial neural network (ANN), could achieve better detection performances, the detailed procedures of the detection process remain unknown. It is called the black-box which is not favorable for practical applications. Moreover, machine-learning-based detection methods are common time-consuming. For example, the training complexity of SVM cannot be tolerable when confront with large-scale and high dimension dataset. However, the statistic-based detection methods could cover these shortages to a large extent in terms of the model interpretation and training speed. Therefore, it can be inferred that when compared to machine-learning-based intrusion detection approaches, statistic-based intrusion detection method have some advantages, that is, good interpretability and fast training speed.

Among these statistic-based detection methods, logistic regression is the most widely used classification approach, which could achieve good detection performances [6–8]. It is worthy to noting that logistic regression could model the correlations among feature and take into account of the joint effects between features to produce a decision boundary to separate different classes effectively. Therefore, logistic regression can be considered as an effective detection method. However, we should also realize that to achieve further improvement in detection performance, it may not be sufficient to use logistic regression alone. Review of related work in intrusion detection indicates that data quality data quality has been considered as a critical determinant [9].

Therefore, in our study, we propose an effective intrusion detection framework based on pls-logistic regression with feature augmentation. Specifically, the feature augmentation technique is used to improve the data quality, and pls-logistic regression is chosen to reduce the dimension and build the intrusion detection model using the transformed data. The reminder of this paper is organized as follows. In Sect. 2, we give a brief overview of feature augmentation and pls-logistic regression. Section 3 describes the details of the proposed intrusion detection model. Section 4 presents the experiment settings, results and discussions. Finally, Sect. 5 comes to conclusion.

2 Methodology

To better illustrate the proposed detection model, firstly, we briefly review the main principles of the feature augmentation [10] in Sect. 2.1, as well as the pls-logistic regression classification model [11] in Sect. 2.2.

2.1 Feature Augmentation

Following Fan et al. (2016), suppose we have a pair of random variables $(\mathbf{X}, Y)$ with n observations, where $\mathbf{X} \in \mathbb{R}^p$ denotes the original features and $Y \in \{0, 1\}$ denotes the corresponding binary response. The logarithm marginal density ratio transformation is used

as the feature augmentation technique to transform the original features. Specifically, for $X_j, j = 1, 2, \dots, p$ in $\mathbf{X}$, denote by f_j, g_j the class conditional densities, respectively, for class 1 and class 0, that is, $(X_j|Y = 1) \sim f_j$ and $(X_j|Y = 0) \sim g_j$. Denote by $^1X_j = \{X_{ij}|Y_i = 1, i = 1, 2, \dots, n\}$ and $^0X_j = \{X_{ij}|Y_i = 0, i = 1, 2, \dots, n\}$. Then, f_j, g_j is obtained by kernel density estimation on 1X_j and 0X_j , and denote the estimates by $\hat{f}_j$ and $\hat{g}_j$, respectively. Thus, the feature augmentation for X_j using logarithm marginal density ratio transformation is shown as follows:

$$X'_j = \log \hat{f}_j(X_j) - \log \hat{g}_j(X_j), \quad (1)$$

where X'_j denotes the transformed feature for the j th feature X_j .

2.2 Pls-Logistic Regression Classification Model

Suppose we have a pair of random variables (X, Y) , where $X \in \mathbb{R}^p$ denotes the original features and $Y \in \{0, 1\}$ denotes the corresponding binary response. The procedures of pls-logistic regression is depicted as follows:

Step 1. Perform univariate logistic regression on each feature to obtain p coefficients denoted by $\omega^1 = (\omega_1, \omega_2, \dots, \omega_p)$. Denote the normalized ω^1 by $\bar{\omega}^1$.

Step 2. Extract the first pls component t_1 by $t_1 = \mathbf{X} \cdot \bar{\omega}^1$.

Step 3. Perform OLS regression of X against t_1 . Denote the residual of X by $\mathbf{X}^*$.

Step 4. Perform logistic regression on each feature of $\mathbf{X}^*$ against t_1 to obtain the p coefficients of features in $\mathbf{X}^*$, denoted by ω^2 , and then normalize ω^2 to $\bar{\omega}^2$.

Step 5. Extract the second pls component t_2 by $t_2 = \mathbf{X}^* \cdot \bar{\omega}^2$.

Step 6. Repeat Step 3, Step 4 and Step 5 until the stopping criteria are satisfied.

Step 7. Denote by $t_1, t_2, \dots, t_h$ the final extracted pls components. Perform the logistic regression on these pls components to build the classification model.

3 Proposed Intrusion Detection Model: Fa-Plslogistic

In this section, we present the main procedures of our proposed intrusion detection model based on pls-logistic with feature augmentation. By embedding the data quality improvement technique into pls-logistic, we can obtain an effective intrusion detection with good performances and less complexity. First, we perform feature transformations on the original features to obtain high-quality training data that can significantly improve the detection performances. Then, the pls-logistic regression is perform on the newly transformed data to conduct dimension reduction and build the intrusion detection model. For clarity, the detailed procedures are summarized as follows:

- **Step 1. Data transformation**

Perform feature transformations on the original data to obtain high-qualified training data.

- **Step 2. Detection model building**

Use the newly obtained data from Step 1 to train pls-logistic-based classifier and build the intrusion detection model.

• **Step 3. Intrusion detection**

For a new testing sample, it is first transformed by the logarithm marginal density ratio transformation illustrated in Sect. 2.1; then, the transformed data is fed into the built intrusion detection model to classify it as either an intrusion or a normal.

4 Experimental Setting

4.1 Dataset Description

In our study, the NSL-KDD dataset is used to evaluate the performance of the proposed intrusion detection model. The NSL-KDD dataset is a modified version of KDD 99 dataset which is considered as the benchmark dataset in intrusion detection domain. However, the KDD 99 dataset suffers from some drawbacks [12, 13]. For example, there are redundant and duplicate records which cause the classifier would be biased towards these more frequent records. The NSL-KDD dataset was proposed by [14] by removing all the redundant samples and reconstituting the dataset, making it more reasonable not only in data size, but also in data structure. The NSL-KDD dataset contains TCP connections that consist of 41 features and one labeling feature.

4.2 Experimental Results and Discussion

In order to prevent the dominance of features with large ranges, we normalize the data into a range of [0, 1] before conducting the experiments. To evaluate our proposed detection model, the 10-fold cross validation has been adopted and the performance is evaluated by the following measurements according to the confusion matrix presented in Table 1.

Table 1. Confusion matrix

		Predicted	
		Attack	Normal
Actual	Attack	TP	FN
	Normal	FP	TN

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \text{Detection rate (DR)} = \frac{TP}{TP+FN}, \text{False alarm rate (FAR)} = \frac{FP}{TN+FP}$$

To verify the effectiveness of our proposed intrusion detection model, we first compare the detection performance of Fa-plslogistic with that of the naïve-plslogistic detection model (pls-logistic regression on original data without feature transformation). The 10-fold cross validation results of these two detection models on NSL-KDD dataset with regard to accuracy, DR, FAR and training time are summarized in Table 2.

As the results shown in Table 2, our proposed intrusion detection model takes clear advantages over the naïve-plslogistic detection model, indicating that the data quality improvement technique can greatly boost the detection performance. More specifically,

Table 2. Performances of proposed methods

Metric	Pls-logistic (with feature augmentation)	Pls-logistic (without feature augmentation)
Accuracy (%)	97.39(0.33)	91.29(6.03)
DR (%)	96.95(0.32)	88.59(12.84)
FAR (%)	2.23(0.45)	6.35(3.05)
Training time (in sec)	98.66	137.51

the accuracy and detection rate of our proposed model both exceed 96%, while naïve-plslogistic only achieves 91.29% and 88.59%, respectively. Besides, in terms of false alarm rate, our proposed method is below 2.3%, while naïve-plslogistic is over 6%. Moreover, the performances of our proposed is also more robust than that of naïve-plslogistic.

To further demonstrate the advantages of our proposed method, the training time required by Fa-plslogistic and naïve-plslogistic is also compared in Table 2. As shown, the training time of our proposed method is superior to that of naïve-plslogistic. Specifically, naïve-plslogistic demands about 1.39 time as much training time as Fa-plslogistic does. Thus, it can be inferred that our proposed method is much more concise than naïve-plslogistic, which can reduce the training time.

Therefore, according to the comparison results, it can be concluded that our proposed intrusion detection model is more effective than naïve-plslogistic and can achieve better detection performances.

Standard errors are in the parentheses in percentage form.

In addition, we examine which features are influential on the intrusion detection. Here, for simplicity, the feature whose coefficient is greater than 1 after standardization is considered to be important. Thus, the influential features recognized during the 10-fold cross-validation are shown in Table 3.

According to the results in Table 3, the important features for intrusion detection are listed in descending order by frequency: land, su_attempted, num_failed_logins, src_bytes, urgent, hot, num_root, num_compromised, root_shell, is_guest_login and dst_bytes. These features are helpful in practice to efficiently detect network intrusion and attacks.

Furthermore, in order to better interpret the effectiveness of our proposed method in intrusion detection, performance comparisons between our proposed model and other existing methods in intrusion detection using NSL-KDD dataset are conducted. The comparison results are summarized in Table 4.

From the comparison results shown in Table 4, our proposed method outperforms other intrusion detection methods with regard to detection accuracy. However, it should be noted that Table 4 just provides a snapshot of performance comparison between our proposed method and other detection methods. Thus, it can be claimed that our proposed method always performs better when compared to any other methods. Nevertheless, from the results above, we can make a conclusion that our proposed method

Table 3. Influential features for intrusion detection

K-fold	Influential feature
1	src_bytes, land, hot, su_attempted, num_root
2	land, num_failed_logins
3	land, urgent, num_failed_logins
4	src_bytes, land, hot, root_shell, su_attempted, is_guest_login
5	dst_bytes, num_compromised
6	src_bytes, land, num_compromised, root_shell, num_root
7	land, urgent, num_failed_logins, num_compromised, su_attempted, num_root
8	src_bytes, urgent, hot, num_failed_logins, su_attempted
9	su_attempted
10	src_bytes, land, urgent, hot, num_failed_logins, su_attempted, is_guest_login

Table 4. Performance comparisons of proposed method and other detection methods

Method	Accuracy (%)
GHSOM [15]	96.02
A-GHSOM [16]	96.63
Naïve Bayes + N2B [17]	96.50
AdaBoost [17]	90.31
Proposed method	97.39

still possesses advantages in intrusion detection and can provide inspirations for the following researches.

5 Conclusion

Intrusion detection system is critical to network security. In this paper, we proposed an effective intrusion detection model based on pls-logistic with feature augmentation. Though the pls-logistic classifier might achieve a good performance, the detection capacity is much more dependent on the quality of the training data. Therefore, in order to increase the detection capacity, we use the logarithm marginal density ratio transformation on the original data to obtain high-quality training data for pls-logistic before building the intrusion detection model. Empirical results on NSL-KDD dataset show that our proposed intrusion detection model is effective and can achieve good and robust detection performances.

Acknowledgments. This research was financially supported by National Natural Science Foundation of China (Grant No. 72001222, 61832001, 61702016).

References

1. Kumar, G., Thakur, K., Ayyagari, M.R.: MLEsIDSs: machine learning-based ensembles for intrusion detection systems—a review. *J. Supercomput.* **76**(11), 8938–8971 (2020). <https://doi.org/10.1007/s11227-020-03196-z>
2. Bamakan, S.M.H., Wang, H., Yingjie, T., Shi, Y.: An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization. *Neurocomputing* **199**, 90–102 (2016)
3. Moustafa, N., Hu, J., Slay, J.: A holistic review of network anomaly detection systems: a comprehensive survey. *J. Netw. Comput. Appl.* **128**, 33–55 (2019)
4. Tsai, C.F., Hsu, Y.F., Lin, C.Y., Lin, W.Y.: Intrusion detection by machine learning: a review. *Expert Syst. Appl.* **36**(10), 11994–12000 (2009)
5. Sommer, R., Paxson, V.: Outside the closed world: on using machine learning for network intrusion detection. In: 2010 IEEE Symposium on Security and Privacy, pp. 305–316 (2010)
6. Wang, Y.: A multinomial logistic regression modeling approach for anomaly intrusion detection. *Comput. Secur.* **24**(8), 662–674 (2005)
7. Mok, M.S., Sohn, S.Y., Ju, Y.H.: Random effects logistic regression model for anomaly detection. *Expert Syst. Appl.* **37**(10), 7162–7166 (2005)
8. Ji, S.Y., Choi, S., Jeong, D.H.: Designing an internet traffic predictive model by applying a signal processing method. *J. Netw. Syst. Manag.* **23**(4), 998–1015 (2015)
9. Aburomman, A.A., Reaz, M.B.I.: A survey of intrusion detection systems based on ensemble and hybrid classifiers. *Comput. Secur.* **65**, 135–152 (2017)
10. Fan, J., Feng, Y., Jiang, J., Tong, X.: Feature augmentation via nonparametrics and selection (FANS) in high-dimensional classification. *J. Am. Stat. Assoc.* **111**(513), 275–287 (2016)
11. Bastien, P., Vinzi, V.E., Tenenhaus, M.: Pls generalised linear regression. *Comput. Stat. Data Anal.* **48**(1), 17–46 (2005)
12. Mahoney, M.V., Chan, P.K.: An analysis of the 1999 DARPA/lincoln laboratory evaluation data for network anomaly detection. In: Vigna, G., Kruegel, C., Jonsson, E. (eds.) RAID 2003. LNCS, vol. 2820, pp. 220–237. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45248-5_13
13. Bamakan, S.M.H., Wang, H., Yingjie, T., Shi, Y.: An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization. *Neurocomputing* **199**, 90–102 (2016)
14. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the KDD CUP 99 data set. In: Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications, pp. 1–6. IEEE (2009)
15. Yu, Z., Tsai, J.J., Weigert, T.: An adaptive automatically tuning intrusion detection system. *ACM Trans. Auton. Adapt. Syst.* **3**(3), 10 (2008)
16. Ippoliti, D., Zhou, X.: A-GHSOM: an adaptive growing hierarchical self-organizing map for network anomaly detection. *J. Parallel Distrib. Comput.* **72**(12), 1576–1590 (2012)
17. Panda, M., Abraham, A., Patra, M.R.: Discriminative multinomial naive bayes for network intrusion detection. In: Proceedings of 2010 Sixth International Conference on Information Assurance and Security, pp. 5–10. IEEE (2010)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.