IJE TRANSACTIONS C: Aspects Vol. 29, No. 9, (September 2016) 1242-1246

International Journal of Engineering

Journal Homepage: www.ije.ir

A Hybrid Machine Learning Method for Intrusion Detection

H. R. Hemati[a], M. Ghasemzadeh*[a], C. Meinel[b]

[a] *Computer Department, Engineering Campus, Yazd University, Yazd, Iran*
[b] *Assoc. Prof. at Yazd University in Iran and Guest Researcher at HPI, Potsdam, Germany*
[c] *President and CEO of Hasso Plattner Institute (HPI), at Potsdam University, Potsdam, Germany*

PAPER INFO

Paper history:
Received 23 July 2016
Received in revised form 20 August 2016
Accepted 25 August2016

Keywords:
Intrusion Detection
Linear Discernment Analysis
Extreme Learning Machine

ABSTRACT

Data security is an important area of concern for every computer system owner. An intrusion detection system is a device or software application that monitors a network or systems for malicious activity or policy violations. Already various techniques of artificial intelligence have been used for intrusion detection. The main challenge in this area is the running speed of the available implementations. In this research work, we present a hybrid approach which is based on the "linear discernment analysis" and the "extreme learning machine" to build a tool for intrusion detection. In the proposed method, the linear discernment analysis is used to reduce the dimensions of data and the extreme learning machine neural network is used for data classification. This idea allowed us to benefit from the advantages of both methods. We implemented the proposed method on a microcomputer with core i5 1.6 GHz processor by using machine learning toolbox. In order to evaluate the performance of the proposed method, we run it on a comprehensive data set concerning intrusion detection. The data set is called KDD, which is a version of the data set DARPA presented by MIT Lincoln Labs. The experimental results were organized in related tables and charts. Analysis of the results show meaningful improvements in intrusion detection. In general, compared to the existing methods, the proposed approach works faster with higher accuracy.

doi: *10.5829/idosi.ije.2016.29.09c.09*

1. INTRODUCTION

Before organizations attempt to use computer systems seriously, they require enough security assurance for their data. According to former and recent surveys, the main challenges in this regard consist of security, performance and availability; among these, security always has been the most important one. Intrusion detection and prevention are the main mechanisms in providing network security.

Intrusion detection systems are responsible for identifying and detection of any unauthorized usage of the system and abuse or harm caused by any internal or external user.

Due to the importance of security, different studies have been conducted concerning this issue [1, 2]. The more remarkable have focused on applying artificial intelligence methods such as neural networks, pattern recognition and meta-heuristic algorithms. Next, we will introduce and discuss some of the relevant ones [3, 4].

*Corresponding Author's Email: m.ghasemzadeh@yazd.ac.ir (M. Ghasemzadeh)

2. LITERATURE REVIEW

Ramteke et al. [5] proposed a method based on using fuzzy clustering and neural networks. The volume of data is one of the main issues in intrusion detection systems; to overcome this challenge, they proposed a multithread intrusion detection system. Their system can process large volumes of data without much of data loss. They divide the training data into a number of subsets and then use a fuzzy clustering technique to train neural networks for each of the subsets. Although they have gained some valuable results, but their method suffer from several shortcomings like: it needs intensive human intervention, requires to set the network parameters by trial and error, it is slow in training and poor in learning scalability .

In another research, Kannan et al. [6] presented a new model for intrusion detection. They used a

Please cite this article as: H. R. Hemati, M. Ghasemzadeh, C. Meinel, A Hybrid Machine Learning Method for Intrusion Detection, International Journal of Engineering (IJE), TRANSACTIONSC: Aspects Vol. 29, No. 9, (September 2016) 1242-1246

combination of a genetic algorithm which was based on feature selection along with a fuzzy support vector machine which was introduced by Lin [7]. Their proposed model include four modules for 1-User Interface 2-Feature Selection 3-Classification and 4-Prevention. The user interface module, collects network data from the KDD dataset; the feature selection module selects necessary features using genetic algorithms. The classification module uses a fuzzy SVM to classify the selected data; while the prevention module determines whether the decision made by the classification module is valid or not. Since they reduce number of features, classification is accomplished faster.

In another attempt, Goztepe proposed a fuzzy expert system for cyber-attacks detection [8]. Construction of his fuzzy expert system consists of: defining the variables of the expected expert system, collection of the data which is concerned with cyber-attacks, system design and its imple-mentation.

The proposed system can successfully detect attacks of the already known types. The main drawback of the intrusion detection by suing such an expert system is that although it works very well in detection of already known forms of attacks, but it doesn't work properly when facing unknown and infrequently happening attacks.

Theseen and kumar [1] have used a combination of chi-square method for feature selection and the multi class SVM method for data classification. Chi-square is a numeric test that measures association between variables. It can also be used to test the association between one or more groups of variables. This methods consists of 7 steps: 1. Data normalization, 2. Generate feature subset using rank based Chi-square feature selection, 3. Train the SVM model on the validation set to obtain different kernel parameter gamma and overfitting constant C, 4. Select the optimal parameter pair C and gamma with best cross validation accuracy, 5. Train the SVM model on the training set with optimal parameter pair, 6. Predict the label for the test data set, 7. Evaluation of performance metrics.

In our proposed method, while we try to preserve the advantages of the existing systems, we have tried to avoid their disadvantages.

3. THE PROPOSED METHOD

The proposed solution is a hybrid method comprising of the linear discernment analysis and the extreme learning machine neural net-work. Since reduction of training time is an important factor in intrusion, we try to obtain this by using LDA and ELM neural network in a cooperating way. Figure 1 shows the training process in our proposed intrusion detection system.

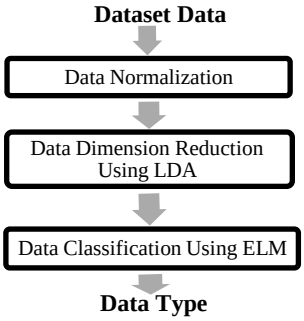

Figure 1. The proposed method.

3. 1. Data Normalization The first step of our proposed method, as shown in Figure 1, is dedicated to data normalization. It takes the dataset data as input and gives the normalized data at its output. The domain of the obtained normalized data would be between zero and one. Data normalization spreads the data according to its highest possible value and then maps them into a 0,1 interval. This phase would minimize the scaling side effects.

3. 2. Data Dimension Reduction The second phase of the proposed method is data dimension reduction. It is accomplished by using the linear discernment analysis method. Input of this section is the normalized data and its output would be the dimensionally reduced data with the corresponding scatter matrix. This matrix will be used later to reduce dimension of the test data.

Data Dimension is the number of variables measured in each observation. In the process of knowledge extraction, in many cases, only some of the data features can play a role. For this reason, in many cases, data dimension reduction is considered to be very important. In this regard, feature selection and feature extraction are the mostly used methods. In feature selection methods, a subset of the features, which are supposed to have a high impact, are selected. Genetic algorithm is one of the most important techniques being used for feature selection. The main drawback of using genetic algorithm for this purpose is its time-consuming characteristic [9].

The feature extraction methods, create fewer features by combining values of existing features in such a way that the obtained features include all (or most of) the initial features [10]. These methods are divided into linear and nonlinear methods. Linear methods are simpler and easier to understand. They try to map the data into a space with less dimension. Principle component analysis (PCA) [11] and linear discernment analysis (LDA) [12] are two of the main important of these linear feature extraction methods.

The LDA method is similar to PCA from this respect that they both try to find a linear combinations of the variables which could best explain the data. These methods also comprise a substantial difference. In LDA, difference between classes are modeled while in PCA difference between classes are ignored. In other words, LDA is a supervised method while PCA is an unsupervised method.

Figure 2 shows the differences between LDA and PCA methods. LDA looks for a line that separates two classes in the best way. This line is shown by 'LDA' label, while PCA looks for a line that has maximum variance associated with it. This line is shown by the 'PCA' label. LDA tries to find a linear space with minimum in class difference and maximum distance between two classes [13].

3. 3. Data Classification In the third part of the proposed method, we perform data classification using ELM [13] neural network. The ELM neural network was trained with the reduced data obtained from the former phase.

Output of ELM shows the class of the input data; normal or type of the attack. The system is trained with a set of n input vectors X along with their related output vectors T; where,

$$X = \{x_1, x_2, ..., x_n\}, T = \{t_1, t_2, ..., t_n\}$$

This lets us predict the output for unseen data vectors. In computational intelligence, the feed forward neural network and support vector machine have considered to be the major techniques for data classification. Both of these classification methods are involved with some challenging issues such as slow training, need of human intervene, lake of scalability and poor learning capability. The ELM model is used for single-hidden layer feed forward neural networks and it could overcome the mentioned challenges. In ELM the hidden layer doesn't need to be tuned and a hidden node in it can be a sub network of several nodes. ELM output is calculated by Equation (1) [14].

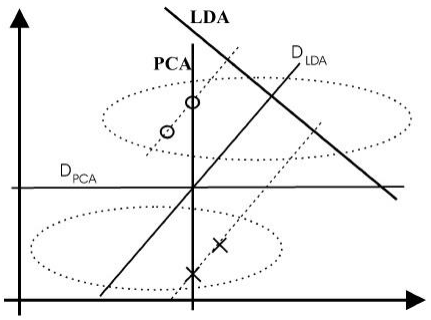

Figure 2. PCA and LDA methods in separation of two classes of data with Gaussian distribution [15]

$$f_L(x) = \sum_{i=1}^{L} \beta_i h_i(x) = h(x)\beta \quad (1)$$

where, $\beta = [\beta_1, ..., \beta_L]^T$ is the edge weight matrix which connects the hidden nodes to the output nodes and $h(x) = [g_1(x), ..., g_L(x)]$ are outputs of hidden nodes for input *x*, and $g_i(x)$ is output of *i*th node in hidden layer.

For the *N* given training samples $\{(x_i, t_i)\}_{i=1}^{N}$ the value of β is calculated by Equation (2).

$$\beta = \left(\frac{I}{C} + H^T H\right)^{-1} H^T T \quad (2)$$

where, $H = [h^T(x_1), ..., h^T(x_N)]^T$ and the target labels are: $T = [t_1, ..., t_N]^T$.

4. IMPLEMENTATION

We implemented the proposed method on a micro-computer with 4 GB RAM, and i5 CPU. In order to evaluate its performance, we run it on KDD dataset. This dataset has been created by Tavallaee et al. [15], and was widely used in evaluation of intrusion detection systems. The KDD was created from the DARPA intrusion detection system. DARPA consists of about 4 GB of compressed data obtained from tcp-dump of network traffic, during 7 weeks. In order to evaluate speed and accuracy of the proposed method, we implemented it along with SVM and MLP methods under the same conditions, then we compared the obtained results.

5. EXPERIMENTAL RESULTS

We run our method on 494366 samples of the KDD dataset which were classified in 21 classes. Table 1 shows the distribution of the samples.

TABLE 1. Distribution of samples in each KDD dataset class

Class number	Nr. of members	Class number	Nr. of members
1	97277	12	21
2	30	13	8
3	9	14	2203
4	3	15	12
5	107201	16	1589
6	281138	17	4
7	53	18	231
8	264	19	7
9	979	20	20
10	1040	21	1020
11	1247	22	10

Each instance of KDD dataset includes 40 features. These features are shown in Table 2.

In the first experiment, SVM, MLP and ELM are compared with each other without using data dimension reduction. Table 3 shows the results of this experiment.

TABLE 2. The Sample Features in KDD dataset

Feature Number	Feature Name
1	duration
2	protocol_type
3	service
4	src_byte
5	dst_byte
6	flag
7	land
8	wrong_fragment
9	urgent
10	hot
11	num_failed_logins
12	logged_in
13	num_compromised
14	root_shell
15	su_attempted
16	num_root
17	num_file_creations
18	num_shells
19	num_access_shells
20	num_outbound_cmds
21	is_guest_login
22	Count
23	serror_rate
24	rerror_rate
25	same_srv_rate
26	diff_srv_rate
27	srv_count
28	srv_serror_rate
29	srv_rerror_rate
30	srv_diff_host_rate
31	dst_host_count
32	dst_host_srv_count
33	dst_host_same_srv_count
34	dst_host_diff_srv_count
35	dst_host_same_src_port_rate
36	dst_host_srv_diff_host_rate
37	dst_host_serror_rate
38	dst_host_srv serror rate
39	dst_host_rerror_rate
40	dst_host_srv_rerror_rate

The results show average of using the K-Fold method with k=10 on the dataset. The Accuracy was obtained using Equation (3).

$$\text{Accuracy} = (TP + FN) / (TP+FP+FN+TN) \quad (3)$$

According to the results presented in Table 3, SVM runs faster than the other two methods. We expected this because SVM classification is accomplished in a binary manner; which is, data is divided into two categories; the category which include the items belonging to that class and the category which include the other items. Also, MLP neural network runs slower than ELM because it uses the back propagation learning algorithm.

In the second experiment, we run ELM against our proposed method which suggests to combine ELM with LDA. Results of this experiment are shown in Table 4.

The needed training time is much less for the combined method and for the accuracy, we can see a slight of decline. It's because of the mapping which is used by LDA to project the data to a line.

We may try to compare performance of our approach with the very recently published method proposed by Theseen and Kumar [1]. It should be noted that this comparison should be done with caution, because our proposed approach is based on data dimension reduction, while their method is based on feature selection. They have used a variant of the KDD dataset in their experiments. Performance of their proposed method is shown in Table 5.

Comparing Table 5 with Table 4, shows that in terms of accuracy their feature selection based method works slightly better than our LAD+ELM method, but our method is much faster than theirs.

TABLE 3. Accuracy / speed of SVM, MLP and ELM

Method Name	Accuracy (percent)	Train Time (s)
SVM	80	200
MLP	92	20
ELM	98	9

TABLE 4. Comparing our method vs. ELM

Method Name	Accuracy(percent)	Training Time (s)
ELM	98	9
LDA+ELM	97.5	5

TABLE 5. The result of Chi-square feature selection

Method Name	Accuracy(percent)	Train and test Time (s)
chi-squarefeature selection and multi class SVM	98	10235

6. CONCLUDING REMARKS

Nowadays, computers are being used extensively; therefore security is of more and more importance. This research showed how combining LDA and ELM could yield an efficient method for intrusion detection. We used LDA as an efficient method for data dimension reduction; while among different methods introduced for data classification, we used ELM because of its desired specifications. Experimental results show that this combination leads to a significant improvement in performance, comparing to the basic SVM and MLP methods.

7. REFERENCES

1. Thaseen, I.S. and Kumar, C.A., "Intrusion detection model using fusion of chi-square feature selection and multi class svm", ***Journal of King Saud University-Computer and Information Sciences***, (2016).
2. Cheng, F., Azodi, A., Jaeger, D. and Meinel, C., "Security event correlation supported by multi-core architecture", in IT Convergence and Security (ICITCS), IEEE., (2013), 1-5.
3. Jeyanthi, N., Shabeeb, H., Durai, M.S. and Thandeeswaran, R., "Rescue: Reputation based service for cloud user environment", ***International Journal of Engineering-Transactions B: Applications***, Vol. 27, No. 8, (2014), 1179.
4. Roschke, S., Cheng, F. and Meinel, C., "Using vulnerability information and attack graphs for intrusion detection", in Information Assurance and Security (IAS), IEEE., (2010), 68-73.
5. Ramteke, S., Dongare, R. and Ramteke, K., "Intrusion detection system for cloud network using fc-ann algorithm", ***International Journal of Advanced Research in Computer and Communication Engineering***, Vol. 2, No. 4, (2013).
6. Kannan, A., Maguire Jr, G.Q., Sharma, A. and Schoo, P., "Genetic algorithm based feature selection algorithm for effective intrusion detection in cloud networks", in 2012 IEEE 12th International Conference on Data Mining Workshops, IEEE., (2012), 416-423.
7. Lin, C.-F. and Wang, S.-D., "Fuzzy support vector machines", ***IEEE Transactions on Neural Networks***, Vol. 13, No. 2, (2002), 464-471.
8. Goztepe, K., "Designing fuzzy rule based expert system for cyber security", ***International Journal of Information Security Science***, Vol. 1, No. 1, (2012), 13-19.
9. Liu, H. and Motoda, H., "Feature selection for knowledge discovery and data mining, Springer Science & Business Media, Vol. 454, (2012).
10. Nikravesh, M., Guyon, I., Gunn, S. and Zadeh, L., "Feature extraction: Foundations and applications". 2006, Springer.
11. Bro, R. and Smilde, A.K., "Principal component analysis", ***Analytical Methods***, Vol. 6, No. 9, (2014), 2812-2831.
12. Izenman, A.J., "Linear discriminant analysis, in Modern multivariate statistical techniques."(2013), Springer.237-280.
13. Huang, G.-B., "An insight into extreme learning machines: Random neurons, random features and kernels", ***Cognitive Computation***, Vol. 6, No. 3, (2014), 376-390.
14. Cambria, E., Huang, G.-B., Kasun, L.L.C., Zhou, H., Vong, C.M., Lin, J., Yin, J., Cai, Z., Liu, Q. and Li, K., "Extreme learning machines [trends & controversies]", ***IEEE Intelligent Systems***, Vol. 28, No. 6, (2013), 30-59.
15. Tavallaee, M., Bagheri, E., Lu, W. and Ghorbani, A.-A., "A detailed analysis of the kdd cup 99 data set", in Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications, (2009).

A Hybrid Machine Learning Method for Intrusion Detection

H. R. Hemati[a], M. Ghasemzadeh[a], C. Meinel[b]

[a] *Computer Department, Engineering Campus, Yazd University, Yazd, Iran*
[b] *Assoc. Prof. at Yazd University in Iran and Guest Researcher at HPI, Potsdam, Germany*
[c] *President and CEO of Hasso Plattner Institute (HPI), at Potsdam University, Potsdam, Germany*

P A P E R I N F O

Paper history:
Received 23 July 2016
Received in revised form 20 August 2016
Accepted 25 August 2016

Keywords:
Intrusion Detection
Linear Discernment Analysis
Extreme Learning Machine

چکیده

امنیت داده‌ها یکی از مهم‌ترین نگرانی‌های هر دارنده یک سیستم کامپیوتری است. یک سیستم تشخیص نفوذ، دستگاه یا نرم افزاری است که مراقب سیستم کامپیوتری در قبال عملیات خرابکارانه می‌باشد. تشخیص و ممانعت از نفوذ غیر مجاز یکی از راهکارهای تأمین امنیت داده‌ها می‌باشد. در این رابطه قبلاً راهکارهای مفیدی ارائه شده است که با توجه به گسترش سیستم‌ها و افزایش حجم تبادل داده‌ها، روش‌هایی با سرعت بالاتر و دقت بیشتری را گوشزد می‌کنند. در این مقاله یک روش تلفیقی که الگوریتم «آنالیز تفکیک‌پذیری خطی» و شبکه عصبی «ماشین یادگیری سریع» را در تعامل با یکدیگر به کار می‌گیرد، ارائه می گردد. هدف از تلفیق یاد شده، بهره بردن از مزایای هر دو روش بوده است. در این رابطه الگوریتم آنالیز تفکیک پذیری خطی برای کاهش ابعاد داده و شبکه عصبی به منظور طبقه‌بندی، ایفای نقش می‌کنند. روش پیشنهادی روی یک میکروکامپیوتر با پردازنده پنج هسته‌ای و با به‌کارگیری جعبه ابزار یادگیری ماشین پیاده‌سازی گردید. به منظور ارزیابی عملکرد روش پیشنهادی، برنامه مورد نظر روی یک مجموعه‌ی داده مبنایی تشخیص نفوذ اجرا گردید. این مجموعه داده، توسط دانشگاه کالیفرنیا ارائه و در مسابقات تشخیص نفوذ بین‌المللی به‌عنوان داده استاندارد و محک، معرفی و بکار گرفته شده است. نتایج آزمایشی در جداول و نمودارهای مرتبط ساماندهی شدند. تحلیل نتایج، حکایت از بهبود در سرعت و دقت تشخیص نفوذ دارد.

***doi**: 10.5829/idosi.ije.2016.29.09c.09*