QUID 2017, pp. 1918-1923, Special Issue N°1- ISSN: 1692-343X, Medellín-Colombia

DETECT NETWORK ANOMALIES USING A LINEAR SVM

(Recibido el 11-05-2017. Aprobado el 18-09-2017)

Mojgan Moradi
Department of Computer and Information, Faculty of Engineering, Islamic Azad University of Garmsar
mojgan1419@gmail.com

Amin Salehi
Network Manager of Rejal Petrochemical Co.
Email address:
aminsalehi1984@gmail.com

Khosrow Amiri Zadeh
Faculty member of Islamic Azad University of Garmsar
corresponding author :
khosrowamirizadeh@gmail.com

Resumen: De acuerdo con el advenimiento de la ciencia y el creciente volumen de comunicaciones, la necesidad de acceso seguro a la red ya la información se siente de antemano, las redes son una muy buena plataforma para atacar y transmitir datos mientras se transfieren o almacenan datos. La transferencia de datos causa diferentes anomalías en los datos. Por lo tanto, el análisis de datos es muy importante para detectar datos abrumadores y datos de problemas, por lo que los investigadores se han centrado mucho en descubrir la intrusión y detectar anomalías en la red. Los métodos más populares son el uso de la minería de datos para la detección de datos pertinentes. Se han desarrollado otros estudios sobre el uso de métodos de extracción de datos, métodos combinados para cubrir las debilidades de otros métodos. En esta investigación, el uso de un método híbrido para abordar las debilidades del método de descubrimiento anormal es utilizado por el algoritmo de soporte de aprendizaje motor profundo para mejorar la selección de características. Los datos utilizados en este estudio son los datos estándar KDD. Finalmente, el método propuesto se compara con otros métodos en términos de velocidad y precisión.

Palabras clave: detección de intrusos, detección anormal, máquina de copias de seguridad, red de aprendizaje profundo

Abstract: According to the advent of science and the increasing volume of communications, the need for secure access to the network and information is felt beforehand, networks are a very good platform for attacking and transmitting data while transferring or storing data. Data transfer causes different anomalies in the data. Therefore, data analysis is very important for detecting overwhelming data and problem data, hence the researchers have been focusing a lot on discovering the intrusion and detecting anomalies in the network. The most popular methods are the use of data mining for the detection of pertinent data. Further studies on the use of data mining methods, combined methods to cover the weaknesses of other methods Have been developed.
In this research, using a hybrid method to address the weaknesses of the abnormal discovery method is used by deep motor learning support algorithm to improve the selection of features. The data used in this study is KDD standard data. Finally, the proposed method is compared with other methods in terms of speed and accuracy.

Keywords: Intrusion Detection, Abnormal Detection, Backup Vector Machine, Deep Learning Network

1. INTRODUCTION

In this chapter, the general context of the research is examined, which examines the problem statement, problems, and definitions and the existing approaches. Then a general plan of the proposed method is used to diagnose this issue and the purpose of this presentation is to express the proposed method. Considering the problems that have been used in previous researches, the new and proposed method is trying to solve these problems. In the innovation section, the main difference of this research with other researches is expressed.

Each research is based on a question and objective, which in this chapter addresses the main issue of the research and the goal to be achieved. Finally, the method for doing this research is to answer this question and to reach this goal also in this chapter.

2. LITERATURE

Feng et al. (2014) investigated the intrusion detection using the combination of an SVM with an ant colony network. In this paper, a data classification method for intrusion detection in a machine learning network is presented. The main task is to classify the activities of a network (connection and logging records) as a natural or abnormal state to minimize the fall in classification. Although various models have been developed for classifying and detecting intrusions, each of them has weaknesses and strengths, including the most widely used method, using a backup vector machine Along with the Ant Colony Network. By studying the weaknesses of other methods, the new approach, with a combination of SVM and CSOACNs, has more advantages and less weaknesses. The proposed algorithm is evaluated using KDD99 Dataset Criteria and Standards. The results of the experiment indicate that the proposed method has increased the rate of classification and its implementation time and its efficiency. (Feng, W et al, 2014)

Abruchaman and Riyad (2017) reviewed penetration detection systems based on the group and composition of classifications. Due to the increased activity of malicious networks and violations of network policies, the system Intrusion Detection has emerged as a group of methods used to combat non-virtual uses in a network. Recent advances in information technology have produced a wide range of machine learning methods that can help with these systems. This research is a general review of penetration classification algorithms based on the methods used in the field of machine learning. Specifically, different techniques of group and combination are examined. This study focuses on both homogeneous and heterogeneous methods. There is also special attention paid to methods that are based on voting techniques. The recent literature review suggests that the hybrid method, in which the choice of attribute, or the reduction of a component of the property, and then the classification, has become commonplace. [2]

In this study, Kabir et al. (2017) investigated a new statistical method for intrusion detection system. In this study, a new method for intrusion detection systems based on minimum sampling in a backup vector machine is proposed. Decisions are made in two steps; in the first step, the total data is subdivided into predetermined subgroups. The proposed algorithm provides samples as representative of the subgroup so that it reflects the characteristics of the entire dataset from which it was selected. In the second step, using the least squares of the supporting vector machine algorithm to extract the penetration action. In the proposed algorithm, an optimized least squares error in backup vector machine is used for diagnostic systems. To demonstrate effectiveness, the proposed method is applied to the KDD99 data set. The result of the experiment shows that the proposed algorithm for the data set has been increased. (Kabir, W et al, 2017)]

In a study by De La Hades and his colleagues (2015), the use of PCA and SOM algorithms to detect potential penetration, the growth of the Internet as a result of the growth of the number of continuous computers provides a suitable platform for attackers and intruders. Fire walls are designed to detect predefined rules and to block potentially dangerous input traffic. However, completing the attack techniques will make it more difficult to detect traffic abnormalities. Different diagnostic methods have been proposed, including car learning methods based on neural network models such as self-organized mapping. In this paper, a hybrid classification method is proposed based on statistical methods and self-organized mapping to detect network anomalies. The principal component analysis and Fisher's separation ratio for character selection and noise elimination are considered. The main goal of self-organizing mapping is to enable the distinction between natural and abnormal communication. The proposed method can be activated without re-training the modified map and only by changing the unit's probabilities. These transactions are affected by the rapid implementation of intrusion detection systems and counterfeiting of false bandwidths. (De la Hoz et al, 2015)

Aqaral and Sharma (2015) explored KDD data analysis and a classification for intrusion detection, this data set is a well-known benchmark for intrusion detection

research. Many ongoing works is under way to improve the detection strategies of the intrusion detection system. In this study, KDD analysis was performed with respect to the general class of content, traffic and host, which in all data and features are categorized. This analysis is based on two prominent criteria for evaluation, the discovery of wax and the rate of false alarms. As a conclusion, it can be argued that in the data set, the contribution of each of the four classes of features in these two parameters indicates that it can be achieved by helping to raise the data appropriately to achieve Maximum correct diagnosis and minimization of misdiagnosis. (Sharma et al, 2015)]

During his research, Camacho et al. (2016) investigated multivariate PCAs for diagnosis of anomalies. In a decade ago, the use of multivariate method has attracted great attention on the basis of analysis of the main components for the diagnosis of anomalies. However, the limitations of this method were criticized by various authors. In this paper, the main steps of the MSPC method are introduced based on PCA. The literature of the network is also examined, highlighting the differences in MSPC and the drawbacks in its approaches. The features and challenges of using MSPC in network analysis are discussed. (Pérez et al, 2016)

Fimands et al. (2016) reviewed the analysis of the anonym colony optimization of the network by using IP and analyzing the network malformation. Today, active network management has grown from growing networks in size and with a variety of service complexes. In this approach, an approach based on identifying behavioral patterns of traffic is necessary that may damage the normal operation of the network. The purpose of automatic management to detect the prevention of potential problems is to combine two methods for the anomaly detection mechanism based on the method of the main statistical components and the supra-innovative ant-optimization method. In this way, a traffic profile, called the digital signature of the network segment using streaming analysis (BNFS), is used as a natural network behavior. Then, this signature is compared with actual network traffic using dynamic metric correction in order to recognize abnormal events. Therefore, a seven-dimensional analysis of the IP is performed. The system is examined using a real network environment and has promising results. Additionally, the correlation between false positive and false positive ratios indicates that the system's power is satisfactory to increase the detection of abnormal behavior by maintaining the warning level. (Carvalho et al, 2016)

Ahmed and his colleagues (2016) explored the technology of discovery of malformations in the network, ICT has had a huge impact on prosperity, economic growth and national security in the world. In general, ICT includes all computer devices, communication devices, mobile phones and networks. Information and communication technology, like other industries, is surrounded by a group of people who have malicious targets, such as network intruders, cybercriminals, and so on. Countering these harmful cybercrime activities is one of the international priorities in this research. Detection of anomalies and analysis of data to detect network penetration is very useful. In this paper, four categories of major methods for the detection of maladaptation, including classification, statistical, information theory and clustering, are analyzed and analyzed. Also, the existing challenges to present the datasets used for intrusion detection are discussed. (Mahmood et al, 2016)

3. METHOD

This research, using a linear SVM, attempts to identify abnormalities in a network. But most important of these are the features of network penetration, as has been said before, in intrusion detection systems, there are false alarms, which in some cases cause wrong alerts to operators The system becomes.

To overcome this problem, the method of extracting basic features is used to increase the accuracy of the network. In order to extract the essential features here, the deep-faith network algorithm is used.

Deep belief networks are highly trained because of the depth of the hidden layers, which is why they are used in extraction of basic features. The data used in this study is NSL-KDD.

The method of this research is to initially use the deep belief network to extract the basic characteristics of the data used, then using the SVM algorithm and the extracted features, attempts are made to cluster based on Normality and abnormality of the data.

The reason for using the SVM algorithm is the ability of this method to be the best separator. Also, unlike other neural networks in the backup machine, which considers modeling error as the objective function, operational risk is used as the target function. Also, using a nonlinear kernel function can also be decided nonlinearly.

Depending on the strengths of deep learning, it is used in a backup vector machine. In this method, a combination of deep trained learning is used to extract the attribute from among all the features, some of which may be irrelevant and change in inputs, so that the backup vector machine Able to efficiently isolate the data in the feature space.

The proposed profound learning approach is used as an algorithm to reduce data with high-dimensional dimensions to a low-dimensional set. From the set of results, the inputs of the instruction input to the supporting machine are given. The final structure of this research is a model of these two algorithms for detecting anomalies. The feature selection method flowchart with the help of the deep learning algorithm is shown in the following figure.

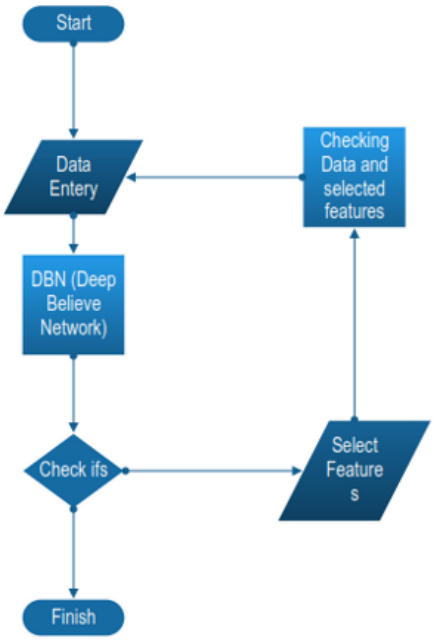


Figure1. Select Feature flow chart

In the first step, the whole data, along with all the features, is applied to the deep-faith network, and a deep prediction network performs a prediction based on the type of attacks, predictive accuracy as a reserve variable. To be in the second step, a feature is selected and again transferred to the deep belief network and its precision is stored. This prediction is performed for all permutations of the features. In the next step, two properties are selected among all the features and this prediction takes place.

The condition for stopping in this research is a prediction accuracy that should have the least difference when using all the features for prediction, and the second condition of the selected features must be the number of attributes The main ones are less. If the two conditions are present, the algorithm will stop. The deep belief network has the ability to examine relationships between features at different levels.

Deep network learning is divided into two stages: First, the discovery of the feature is done randomly and initialized, and in the second stage, using productive models, which include explicit layers and hidden layers, the network is taught. It will be. Deep belief networks have more than one secret layer and are composed of many parameters. This property allows the network model to model complex and nonlinear relationships between input and output.

The deep-faith belief network strategy for choosing the attribute is that it first computes the property-class-property matrix of the attribute-class and the attribute-property of the data set; in the second step, given that each of the attributes is input in A deep belief system is considered. Each of the features takes a weight randomly. In the deep belief network, a goal function is considered, which can be used to examine the significance of the features based on this function. The weight of each of the features changes to calculate the output based on the best weight, the output is calculated based on the best features that are arranged side by side respectively. With this in mind, the output of the feature extraction is first obtained through the Deep Intelligence Network. Then clustered data using SVM clustering. For training data, 80% of the data is used, to test 20% of the data used to obtain output results and the percentage of clustering error is calculated. This work is repeated to increase the accuracy of clustering.

To evaluate the proposed method for detecting abnormalities, we use the correct detection rate and the false alert rate:

1. False Positive (FP): The data rates that were secure but were mistakenly considered intrusive.
2. Negative Negative (FN): The data rates that were infiltrated but were mistakenly considered safe.
3. True Positive (TP): Data rates that were safe and properly considered.
4. Correct Negative (TN): Data rates that were permeated and correctly considered.

Accuracy is given by the following formula:

$$\text{Accuracy} = ((TN + TP) / (TN + TP + FN + FP)) * 100$$

This criterion is used to compare the proposed method with other methods.

4. DATA

The data used in this study is KDD data derived from a simulation. This set of data is used to test systems and methods around the network and is usually used as an input to intrusion detection problems. This dataset has 4 million records, each with a record of 41 characters.

Each of these features is a user's behavior that connects to the network, detecting malware by checking the critical behavior of users who connect to the network.

Due to the high volume of data, the analysis speed drops sharply. Therefore, the NSL-KDD data set is used. This data has an advantage over the main data set, which is:

• Do not have duplicate records.
• Test datasets do not have duplicate records and therefore have better performance.
• The number of divided tests and training is logical and the training is done correctly.

In this dataset, 4 attacks are detected, which include attacking a service ban, processing, routing an attacker, and remote user attack.

5. DESIGNING THE PROPOSED MODEL

To evaluate the proposed method, NSL-KDD data is used that has 38 attributes both numerically and in kind. In the first step, the data is converted into numbers so that the data are of the same sex. In the second step, using deep learning, we examine the behaviors of users connected to the network and examine the behaviors of the users and the unconventional ones. Deep learning is a branch of learning, in which one tries to increase the accuracy by learning at different levels and layers.

The first step of the proposed method is to use deep learning to determine the important behaviors of users. Deep learning is one of the new strategies for solving other problems of the methodology. The deep belief system has the ability to examine relationships between features at different levels.
The deep network learning is divided into two stages: first, the discovery is randomly and initialized, and in the second stage, using the productive models, which consists of obvious layers and hidden layers, the network is taught. Deep belief systems have more than one secret layer and have many parameters. This property enables the network to model complex and nonlinear relationships between input and output.

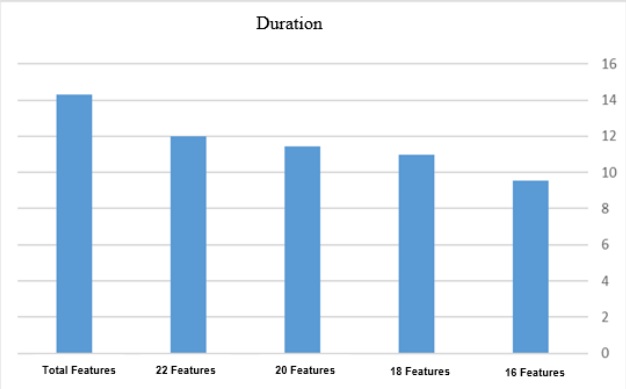
6. DISCUSSION

In this research, we have tried to increase the accuracy of the proposed model using a learning method. In the proposed method, we use the knowledge learning network to analyze the work. The main features of 38 specific user behavior are those that are first used and the second part uses deep learning of the 13 main features of the user's important behaviors. This extraction has resulted in the following output results to be obtained.

Table1. Result of SVM Algorithm

	Not mal	Pro be	DO S	R2 l	U2 R	Avera ge	Durati on
Feat ure 16	78.6	87. 19	88. 7	70. 1	72. 7	79.46	9.55
Futu re 18	80.1	88. 69	90. 2	71. 6	74. 2	80.96	11
Feat ure 20	82.1	90. 69	92. 2	73. 6	76. 2	82.96	11.45
Feat ure 22	82.8	91. 39	92. 9	74. 3	76. 9	83.66	12
Gran t featu re	88.6	97. 19	98. 7	80. 1	82. 7	89.46	14.3

indicates that the test result is very low with regard to the output of the volume of calculations and the accuracy level has been acceptable. In the table below, the comparison between the implementation time of the program is shown in two modes of user behavior and by extracting the behavior of the users.



As shown in the table above, the accuracy level in some cases has decreased, but the most important is to reduce the computational complexity and reduce the computational time using the feature extraction.

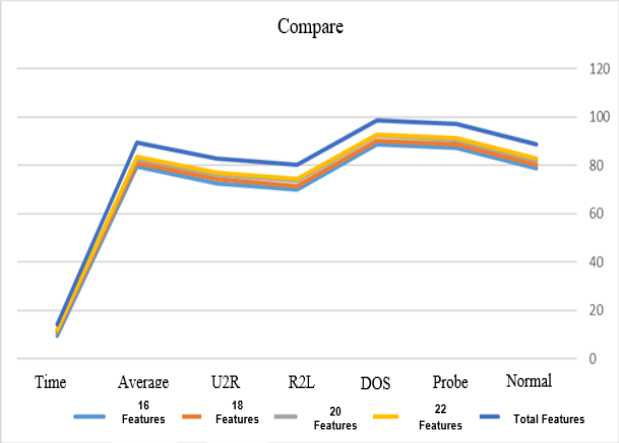


Figure 2. Compare of Selective Features

The comparison of the proposed method with respect to certain criteria

Table 2. Compare of Selective Features

Figure 3. Duration of executive Features

Regarding the output of the proposed model for analyzing user behaviors as maladaptation, the following table shows behaviors that are more important in detecting malware or malware attaching to a network.

Selection 22	Selection 20	Selection 18	Selection 16
dst_host_same_srv_rate	num_access_files	same_srv_rate	num_compromised
dst_host_srv_count	srv_count	rerror_rate	is_guest_login
serror_rate	wrong_fragment	serror_rate	root_shell
num_access_files	num_outbound_cmds	dst_host_srv_diff_host_rate	diff_srv_rate
srv_diff_host_rate	srv_diff_host_rate	dst_host_serror_rate	Countn
dst_host_same_srv_rate	num_failed_logins	hot2	dst_host_srv_diff_host_rate
is_guest_login	countn	is_host_login	num_failed_logins
protocol_type	dst_bytes	root_shell	su_attempted
srv_serror_rate	serror_rate	su_attempted	dst_host_srv_serror_rate
countn	logged_in	num_shells	srv_diff_host_rate
same_srv_rate	urgent	srv_serror_rate	Land
diff_srv_rate	num_compromised	dst_host_srv_rerror_rate	same_srv_rate
dst_host_count	dst_host_diff_srv_rate	num_access_files	difficulty_level
dst_host_same_src_port_rate	dst_host_srv_diff_host_rate	difficulty_level	dst_host_same_src_port_rate
num_failed_logins	dst_host_srv_count	srv_count	dst_host_rerror_rate
hot2	dst_host_same_srv_rate	protocol_type	num_shells
src_bytes	dst_host_count	dst_host_srv_serror_rate	
num_outbound_cmds	dst_host_same_srv_rate	durationn	
dst_host_serror_rate	dst_host_same_srv_rate		
service	durationn		
dst_host_srv_diff_host_rate			
rerror_rate			

This can identify the status of users as malicious or ordinary user, and according to the findings of this study and the table in chapter three, we can examine the proposed method in comparison with the method that Gimandes and his colleagues (2015), the duration of the program has decreased significantly, and compared with the method presented by Feng et al. (2014), the detection of abnormal behavior is more precise and, compared with Mirezai and Ashouri (1395), lesser behaviors it's been chosen.

7. SUGGESTIONS

At the end, it is suggested to use meta-innovative methods to reduce the amount of data to increase this diagnosis. Also, the use of a hyper bacterial method for intrusion detection and a method such as the genetic algorithm in the diagnosis can be very effective.

REFERENCES

Aburomman, A. A., & Reaz, M. B. I. (2017). *A survey of intrusion detection systems based on ensemble and hybrid classifiers*. Computers & Security, 65, 135-152.

Ahmed, M., Mahmood, A. N., & Hu, J. (2016). *A survey of network anomaly detection techniques.* Journal of Network and Computer Applications, 60, 19-31.

Camacho, J., Pérez-Villegas, A., García-Teodoro, P., & Maciá-Fernández, G. (2016). *PCA-based multivariate statistical network monitoring for anomaly detection*. Computers & Security, 59, 118-137.

De la Hoz, E., De La Hoz, E., Ortiz, A., Ortega, J., & Prieto, B. (2015). *PCA filtering and probabilistic SOM for network intrusion detection*. Neurocomputing, 164, 71-81.

Feng, W., Zhang, Q., Hu, G., & Huang, J. X. (2014). Mining network data for intrusion detection through combining SVMs with ant colony networks. Future Generation Computer Systems, 37,

Fernandes, G., Carvalho, L. F., Rodrigues, J. J., & Proença, M. L. (2016). *Network anomaly detection using IP flows with Principal Component Analysis and Ant Colony Optimization.* Journal of Network and Computer Applications, 64, 1-11.

Ggarwal, P., & Sharma, S. K. (2015). *Analysis of KDD Dataset Attributes-Class wise for Intrusion Detection*. Procedia Computer Science, 57, 842-851.

Kabir, E., Hu, J., Wang, H., & Zhuo, G. (2017). A novel statistical technique for intrusion detection systems. Future Generation Computer Systems.