The Caspian Sea Journal
ISSN: 1578-7899

Volume 10, Issue 1, Supplement 4 (2016) 456-461

A Study on Improvement of Intrusion Detection Systems in Computer Networks via GNMF Method

Milad Gholipoor Moghaddam
Department of computer engineering, south Tehran branch, Islamic Azad University, Tehran, Iran

ABSTRACT— Intrusion detection systems have been designed to provide security in computer networks. These systems collect the network traffic data from parts on the network or computer system and use this information to detect the attacks entered into a network. The main purpose of this inspection system is to identify and deal with the applications that are identified as unauthorized users. A major problem in machine learning algorithms is training time. In this study, graph regularized non-negative matrix factorization method has been used in order to select features for increasing the efficiency of support vector machine in intrusion detection; test and evaluation of the proposed model has been conducted on data collection NSL-KDD which is the modified version of KDD-CUP99. Experimental results showed that the proposed model has the significant performance in increasing the accuracy and rereading the attack detection and reducing the rate of error notifications and is able to provide more the accurate diagnosis than its base models.
KEYWORDS: Network security, Intrusion detection, Support vector machine, Feature selection, GNMF.

Introduction

Today, most critical infrastructures such as telecommunications, transportation, trade and banking are managed by computer networks; so the security of the systems against planned attacks is very important. Most attacks abuse the software errors and security gaps of the target system. Since it is impossible to completely eliminate the software errors, all softwares have security gaps which refer to software vulnerabilities. So, researchers have tried to find these vulnerabilities so that they can provide system protection by preventive methods or dealing with after identifying the intrusion ways to system. Each dictionary has proposed a meaning and concept for intrusion. In the computer world, there is much debate on meaning or meanings of intrusion. Many know the intrusion as unsuccessful attacks, while others consider the separate definitions for intrusion and attack. Intrusion can be defined as follows [1]: "Active series of events related to each other that their goal is unauthorized access to data, changing in data or damaging the system in a way that make the system unusable. This definition includes both successful and unsuccessful attempts". Intrusion detection systems have been designed to provide security in computer networks. Intrusion detection systems are systems that strive to detect attacks to a network. These systems collect the network traffic data from parts on the network or computer system and use this information to provide the network security. The main purpose of the inspection system is to identify and deal with the users that are known as unauthorized users. When a protected computer system or network is under attack, intrusion detection system generates warnings to attack, even if the system is not vulnerable to the reported attack. Therefore, the purpose of intrusion detection is to discover the unauthorized uses, misuse and damage to computer systems and networks by both categories of domestic users and foreign invaders. In a complete security system, the encryption and authentication methods of the intrusion detection systems are used as an auxiliary tool to enhance the system security alongside the use of firewalls.

The Proposed Method

Data classification has been always interested by researchers as one of the most common topics in the field of various sciences; so that many methods have been presented in this area. Each of them has been designed for using in various applications. The main purpose of data classification is to identify the different classes in a data set which can determine the new samples position through it in this collection. Classification is considered as one of the most important goals of pattern detection which has received a lot of attention today according to many applications of it. So far, many methods have been invented for the data classification that can mention the following cases: 1) Inference based on decision trees, 2) Bayesian theory, 3) Neural networks, and 4) Support vector machines. Nowadays, the use of classification is inevitable in applications with the volume of data. Applications such as biology, hyperspectral satellite images with large scale [40] and web pages on the Internet are significantly increasing. Development of classification methods performance for using in a field has been as a challenge today. In this area (from practical point in bulky and large-scale data), the nonparametric classification methods and without learning stage (without optimization) as SVM have the utmost importance. So according to importance of their applications, research on the invention of new methods is active in this category of applications. These problems caused to definition of large projects in recent years in

combination with areas having a wide variety of data such as malware detection. Applications [41-43] can be noted. Our aim of this proposal is to provide a classification method for intrusion detection on a large scale that improves the learning method of SVM in terms of time and accuracy and increases the detection of new and unknown types of intrusions based on the proposed combination. Feature of the proposed method is combination of feature selection methods (the use of graph regularized non-negative matrix factorization) with support vector machine for classification.

Details of the Proposed System

The proposed intrusion detection system in this study has five components including data collection, normalization, feature selection, support vector machine and the post-processing. The structure of proposed intrusion detection system has been shown in Figure (1). Then in addition to a closer examination of each of these components, we deal with the implementation of them.

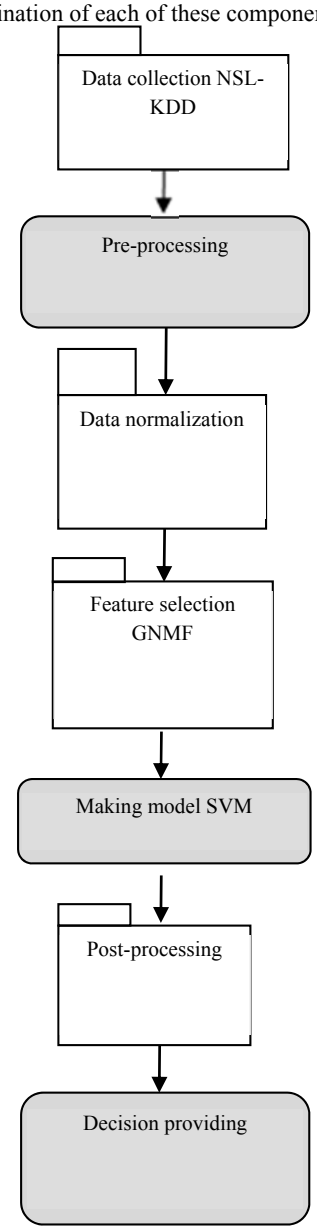

Figure 1. Components of the proposed system to implement the intrusion detection task

The Importance of Features Selecting For Intrusion Detection Systems

Analysis and classification is difficult in intrusion detection systems, because the amount of data that the system needs to check and monitor a network is too much. Thus, an intrusion detection system should reduce the amount of data for processing; so that the complexity of relationships that exist between some features to be deleted or less [50]. Since increasing number of features increases the computational cost of a system, system's design and implementation with the least number of properties seems to be necessary. On the other hand, attention to this issue is very important that the impressive subset of features should be selected that provides acceptable performance for the system. The main aim of feature selection is the reduction of feature vector dimension in classification so that an acceptable classification rate to be obtained. In these circumstances, features that are less discriminatory power are deleted and a number of features that includes proper information for discriminating the pattern classes will remain. Feature selection in intrusion detection systems causes to simplify the problem and faster and more accurate diagnosis of attacks due to removing the duplicate and unrelated data. Many solutions and algorithms have been proposed for problem of feature selection which some of them date back to thirty or forty years. The problem of some algorithms which were presented was their high computational load, although this problem is not visible today with the rapid rise of computers and large storage resources; but on the other hand, very large data sets for new issues has led that finding a fast algorithm for this purpose to be important as well.

Data Set

The data sets NSL-KDD [61] has been used to evaluate the proposed method. This dataset contains records from a selection of KDD-CUP99 [62]. The dataset KDD-CUP 99 has been the most widely used data set methods to assess the anomaly detection and network intrusion since 1999. This data set was made by Austolfo and colleagues [63] based on the evaluation program of intrusion detection systems DARPA'98 [64].

DARPA'98 is a set consists of nearly 5 million record from different connections which each of them has about 100 bytes. Training set of KDD consists of approximately 4900000 vectors from different connections that each vector consists of 41 properties and is classified as normal or attack. Each attack belongs to one of the following general groups:

1) Denial of service (DoS) attack: is a type of attack in which the attacker engages or fills a processing source or storage that cannot perform the authorized requests.

2) User to root (U2R) attack: is a class of attack in which the attacker has started its work with access to account a regular user on the system (which has been usually obtained through password sniffers, dictionary attack or social engineering) and gain the access of root to system by finding system vulnerabilities points.

3) Remote to local (R2L) attack: This attack occurs when the attacker can send packets over the network, but it has not an account on one of the machines on the network and access to system by investigating the system vulnerabilities points as system user.

4) Probing attacks: is an attempt to obtain information about computer network with the aim of network security threats.

An important point that should be mentioned is that the test set has not the same probability distribution with the training set and has specific types of attacks which do not exist in the training set that makes the assessments more realistic. In fact, some experts in intrusion field also believe that majority of new attacks originate from the old attacks, and signature of known attack will suffice for identifying these new attacks. More precisely, it should be also stated that the training set includes 24 types of attack, while the test set is also includes 14 other attacks that will never exist in the training set.

Attacks can be initially divided into different groups, but these groups are also included different categories. In Table 1, we can observe these categories.

Table 1- Types of attacks classification in data set NSL-KDD

Category	Subcategory
NORMAL	-
PRB	ipsweep, nmap, portsweep, satan
DOS	back, land, neptune, pod, smurf, teardrop
U2R	buffer_overflow, loadmodule, multihop, perl, rootkit
R2L	ftp_write, guess_passwd, imap, phf, spy, warezclient, warezmaster

As well as, the approximate distribution of different categories of data in the training and test sets is shown in Table 2. It should be noted that the total is not equal to 100% due to the rounding of percentages.

Table 2- Approximate distribution of training and test data in data set NSL-KDD

Category	Training	Testing
NORMAL	48%	19%
PRB	20%	1%
DOS	26%	73%
U2R	0.2%	0.07%
R2L	5%	5%

Preprocessing

Data normalization

One of the tasks of data normalization unit is to convert the values of text features into numerical values. In Table 3, characteristics of data set NSL-KDD has been shown. Some of the 41 feature values of data set NSL-KDD are in textual form, and as the support vector machines uses only the numerical data for training and testing, so we need to convert the values of text features into numerical values. Features with text values are (B) Protocol Type, (C) Service and (D) Flag which have numbers 2, 3 and 4 respectively in Table 3. For example, the feature values 1 for Protoclo_type, tcp, 2 for udp and 3 for icmp were considered. For coordination between the data limits SVM with the used kernel functions, it is necessary that the input data to it to be in one of two forms of binary or continuous depending on the type of kernel. So after converting the three above-mentioned characteristics from HTML format to numeric format, the main issue is to convert data into binary form or the normal continuous form in data normalization. The normal continuous form is used for features in the proposed intrusion detection system. To normalize features, a statistical analysis was initially carried out on any of the properties based on the data available in NSL-KDD and the maximum and minimum values were determined for each property. Then, normalization was performed in range [1, 0] according to equation (1-4).

$$Nf = \frac{f - MinF}{MaxF - minF} \qquad (1\text{-}4)$$

In which F the desired features, f feature value, max F maximum value of feature F, min F minimum value of feature F and the normalized value of F.

When we do not use the feature selection, normalization of parameters is done after converting the text values into numerical values and sent to support vector machine. But due to the use of feature selection option in the proposed model, the selection of important characteristics is initially done by graph regularized non-negative matrix factorization after converting text values into numerical values, and then normalization of parameters is performed.

Table 3- Characteristics of data set NSL-KDD

Label	Feature name	Label	Feature name
A	Duration	**V**	Is_guest_login
B	Protocol_type	**W**	Count
C	Service	**X**	Sev_count
D	Flag	**Y**	Serror_rate
E	Src_byte	**Z**	Sev_serror_rate
F	Dst_byte	**AA**	Rerror_rate
G	Land	**AB**	Srv_rerror_rate
H	Wrong_fragment	**AC**	Same_srv_rate
I	Urgent	**AD**	Diff_srv_rate
J	Hot	**AE**	Srv_diff_host_rate
K	Num_failed_login	**AF**	Dst_host_count
L	Logged_in	**AG**	Dst_host_srv_count
M	Num_comprised	**AH**	Dst_host_same_srv_rate
N	Root_shell	**AI**	Dst_host_diff_srv_rate
O	Su_attempted	**AJ**	Dst_host_same_src_port_rate
P	Num_root	**AK**	Dst_host_srv_diff_host_rate
Q	Num_file_creations	**AL**	Dst_host_server_rate
R	Num_shells	**AM**	Dst_host_srv_serror_rate
S	Num_access_files	**AN**	Dst_host_rerror_rate
T	Num_cutbounds_cmds	**AO**	Dst_host_srv_rerror_rate
U	Is_host_login		

Feature selection with GNMF

MATLAB software [65] has been used to implement the graph regularized non-negative matrix factorization algorithm in the proposed intrusion detection system. As it can be seen, we are looking for the optimal solution in a multidimensional space (the number of features). The NMF and GNMF algorithms implemented in [52] have been used in conducted tests that its parameters have been set in Table 4.

Table 4- Parameters used in the GNMF algorithm

Algorithm	nFactor	Weight	Alpha	maxIteration	K	WeightMode
GNMF	10	NCW	100	50	7	Cosine
NMF	10	NCW	0	-	-	-

Discussion and Conclusion

Different parameters have been measured in these experiments. Training and testing time significantly decreases which this approaches us to the speed of attack detection based on the objectives of intrusion detection system. For all cases where the data has the possibility of different views, the proposed method is a strong competitor for the other methods and shows its high performance. Although the proposed method has less efficiency than its base methods in some cases, it is the best method in many cases and is the best method in all cases. Classification strength and precision increases after selecting the important features and deleting the unrelated data, and the use of graph regularized non-negative matrix factorization reduces the incorrect classification rate, increases the detection precision, and accelerates ranking of the parameters. Today, intrusion detection systems have become the very important and functional tools to ensure the security of computer networks. Intrusion detection systems are security management systems which are used to detect unusual activities, unauthorized use and abuse in computers or networks. An intrusion detection system is a hardware or software tools that discover the attacks by monitoring the event stream. Intrusion detection allows organizations maintain their systems from threats that arise due to increasing the connection between networks and increase the reliability of their information systems. Intrusion detection methods are divided into two categories of misuse detection and anomaly detection. In misuse detection, the pre-made intrusion patterns are kept as rule. So that each pattern includes different types of a particular intrusion and the occurrence of intrusion will be announced if such a pattern in system takes place. By specifying a behavior as a normal behavior for the discussed subject (user, host or the entire system) in anomaly detection method, any deviation from this behavior is considered abnormal that this can be a possibility for occurrence of an attack. In this research, network-based intrusion detection system was evaluated in anomaly method which uses the graph regularized non-negative matrix factorization and support vector machine. In general, matrix factorization methods are widely used in information retrieval, machine vision and pattern recognition. Meanwhile, the graph regularized non-negative matrix factorization algorithm received special attention due to psychological and physiological interpretations that normally occurs in the data and has proportional representation to the human brain. On the other hand, data have usually sampled of a manifold with low dimensionality in a high-dimensional space from the geometric perspective. The use of all features available in network packets to evaluate and discover attack patterns increases the overhead and make a mistake for classifiers, because some of these features are irrelevant and redundant; and also the use of all features causes that the process of detection prolongs and performance of intrusion detection system reduces. The use of feature selection methods is beneficial because of managing the data and reducing the computation time. In this study, we dealt with the way of reducing data from a variety of attacks by this method in addition to reviewing some of the basic concepts of a graph regularized non-negative matrix factorization and expressing some of its applications. Graph regularized non-negative matrix factorization algorithms can lead to the reduction of redundant data in the desired data set in intrusion detection system with a compact representation method in different data while considering the intrinsic geometrical structure. Then, the different layers of this system which included data set, data preprocessing, feature selection, support vector machines and post-processing were separately evaluated. The data set NSL-KDD was used for training and testing the proposed model. One advantage of the proposed model is to reduce the computation cost and consequently reduce the computer resources such as memory and processor time which is necessary to attack detection. As well as, the level of precision, recall and F-Measure of the intrusion detection system increases by feature selection in the proposed model.

References

1. Endorf, Carl, Eugene Schultz, and Jim Mellander. Intrusion Detection & Prevention, McGraw-Hill, 2004.
2. I. Santos, B. Sanz, C. Laorden, F. Brezo, P. G. Bringas, "Opcode-Sequence-Based Semi-supervised Unknown Malware Detection", Computational Intelligence in Security for Information Systems, vol. 6694, 2011, pp 50-57.
3. P. Kane, S. Sezer, K. McLaughlin, Eul Gyu Im, "SVM Training Phase Reduction Using Dataset Feature Filtering for Malware Detection", IEEE Transactions on Information Forensics and Security, vol. 8, no. 3, 2013, pp. 500-509.
4. [4] P. Okane, S. Sezer, K. McLaughlin, Eul Gyu Im, "Malware detection: program run length against detection rate", IET Software, vol. 8, no. 1, 2014, pp.42-51.
5. N.Idika, A.P. Mathur, "A Survey of Malware Detection Techniques", Committee on Institutional Cooperation and General Electric, 2007.
6. A.E. Elhadi, M. Marrof, "Malware Detection Based on Hybrid Signature Behaviour Application Programming Interface Call Graph", American Journal of Applied Sciences 9, 2012, pp. 283-288.
7. P. Ning, and S. Jajodia, "Intrusion Detection Techniques", the Internet Encyclopedia: Wiley Online Library, 2004.
8. D. E. Denning, "An Intrusion-Detection Model", IEEE Transactions on Software Engineering, vol. SE-13, no. 2, 1987, pp. 222-232.
9. T. Verwoerd and R. Hunt. "Intrusion Detection Techniques and Approaches", Computer Communications, vol. 25, no. 15, pp. 1356-1365, 2002.
10. M. J. Ranum, Artificial Ignorance: How-to Guide. 1997. Available on: http://lists.insecure.org/firewall-wizards/1997/Sep/0096.html.

11. M. G. Kang, P. Poosankam, H. Yin, "Renovo: A hidden code extractor for packed executables", Proceedings of the 2007 ACM workshop on Recurring malcode, pp. 46-53, New York, USA, 2007.
12. M. J. Ranum, Intrusion Detection: Challenges and Myths. 1998. Available on: http://www.nfr.net/forum/publications/id-myths.html.
13. M. Egele, T. Scholte, E. Kirda, C. Kruegel, "A Survey on Automated Dynamic Malware-Analysis Techniques and Tools", ACM Computing Surveys (CSUR), vol. 44, 2012.
14. L. Nataraj, V. Yegneswaran, P. Porras, J. Zhang, "A Comparative Assessment of Malware Classification", in Proceedings of the 4th ACM workshop on Security and artificial intelligence, 2011, pp. 21-30.
15. K. A. Roundy, B. P. Miller, "Hybrid Analysis and Control of Malware", Recent Advances in Intrusion Detection (LNICS), Vol. 6307, 2010, pp 317-338.
16. S. Cesare, Y. Xiang, W. Zhou, "An effective and efficient classification system for packed and polymorphic malware", IEEE Transactions on Computers, vol. 62, no. 6, 2013.
17. S. W. Indratno, "Plug-in Classifier Dengan Bayesian Statistics Untuk Mendeteksi Situs Web Palsu", Prosiding Seminar Nasional Statika Universitas Diponegoro, 2013.
18. G. E. Dahl, J. W. Stokes, L. Deng, D. Yu, "Large-Scale Malware Classifiction Using Random Projections and Neural Networks", International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3422-3426, 2013.
19. Y. Reeves and D. Park, "Deriving common malware behavior through graph clustering", in Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security, New York, 2011.
20. L. Nataraj, V. Yegneswaran, P. Porras, J. Zhang, "A Comparative Assessment of Malware Classification", in Proceedings of the 4th ACM workshop on Security and artificial intelligence, 2011.
21. P. V. Hung, "An approach to fast malware classification with machine learning technique", Faculty of Environment and Information Studies Keio University, 2011.
22. G. Bronevetsky, "Detection of Control Flow Errors Survey of Hardware and Software Techniques", Lecture Slide, 2004.
23. M. N. Yusoff, A. Jantan, "Optimizing Decision Tree in Malware Classification System by using Genetic Algorithm", International Journal of New Computer Architectures and their Applications (IJNCAA), vol. 3, no. 1, 2011.
24. Chandrasekhar, A.M .; Raghuveer, K., "Intrusion detection technique by using k-means, fuzzy neural network and SVM classifiers," in The 2013 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-7, 4-6 Jan. 2013.
25. Chun-Wei Tsai, "Incremental particle swarm optimisation for intrusion detection," IET Networks, vol. 2, no. 3, pp. 124-130, Sept. 2013.
26. Abduvaliyev, A .; Pathan, A.-S.K .; Jianying Zhou; Roman, R .; Wai-Choong Wong, "On the Vital Areas of Intrusion Detection Systems in Wireless Sensor Networks," IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp.1223-1237, Third Quarter 2013.