AARMS Vol. 15, No. 2 (2016) 187–196.

Data Mining in Cyber Threat Analysis – Neural Networks for Intrusion Detection

Eszter Katalin BOGNÁR[1]

The most important features and constraints of the commercial intrusion detection (IDS) and prevention (IPS) systems and the possibility of application of artificial intelligence and neural networks such as IDS or IPS were investigated. A neural network was trained using the Levenberg-Marquardt backpropagation algorithm and applied on the Knowledge Discovery and Data Mining (KDD)'99 [14] reference dataset. A very high (99.9985%) accuracy and rather low (3.006%) false alert rate was achieved, but only at the expense of high memory consumption and low computation speed. To overcome these limitations, the selection of training data size was investigated. Result shows that a neural network trained on ca. 50,000 data is enough to achieve a detection accuracy of 99.82%.
Keywords: *IT security, intrusion detection, neural networks*

Introduction

At present, all the most important companies use a complex, highly interactive IT system to support their operation. The security of the IT system has a high-priority, as an increasing amount of transmitted and stored data is generated.

There are several widely used methods to prevent the unauthorized intrusions to networks, like the application of user identification, firewall and antiviral softwares. However, these methods due to their vulnerability are unable to provide a maximal security. To ensure the confidentiality, integrity and the accessibility of transmitted or stored highly relevant information for authorized users are great challenges, therefore a complex, multi-layer security system is required.

The application of IDS and IPS systems is one of the most efficient solutions to improve the network security, where the unauthorized intrusions into the network are detected and the harmful effects and attacks are prevented using a proper security system (e.g. packet filter firewall).

In the first part of the paper the role and relevance of IDS and IPS systems in the network security are described. In the subsequent part the application's efficiency of neural networks in the detection of network intrusion is investigated using a reference dataset. The analysis of this investigation proves the relevance of the new generation of IDS system; the anomaly-based detection.

1 Ph.D. student, National University of Public Service, Faculty of Military Sciences and Officer Training, e-mail: bgeszti@email.com

IDS and IPS Systems

For the protection of an information system a multi-layer approach is applied. The key element of this system is the user identification on the basis of biometrical, key and knowledge features. The different firewalls represent higher level protection, however, they act only as filters, which are unable to identify the intruders that have already penetrated the filter and protect against internal attacks.

Those highly-secure IT systems require further protections, whose important components are the IDS and IPS systems. The term of intrusion detection/prevention system refers to their major task; the detection and prevention of the internal and external attacks affecting the host or the network. They act as alarm systems, monitor permanently the network traffic and the host events, and report to an authorized person (administrator), if any unusual activity is detected. The administrator analyses the provided information and activates the system to prevent successfully the attacks.

Signature and Anomaly-Based Systems

At present that signature based detection is applied mostly which stores well-known attack patterns as rule in the database, and the unauthorized intrusion trials are tested and removed, if they match with previously stored rules. This solution has a disadvantage of storing large amount of data and identifying new types of attacks only after the refreshment of the database. The recent IDS and IPS systems require permanent user control, the reports should be monitored and the database should be extended with the new rules matching with the pattern.

To improve the accuracy of IDS and IPS systems a new approach, the anomaly-based detection has been recently developed. The main feature of the anomaly-based detection system is that normal user activity is modeled using statistical methods and every deviation from this behavior is classified as an anomaly. This method is in sharp contrast to signature based detection, where specifically the attack patterns are investigated. This approach has an advantage of detecting even such intrusions, which have not been identified and stored yet in the database. Although the anomaly-based detection method seems to be promising, but it has not been introduced yet due to its complexity and high resource requirements. [1]

Figure 1 shows the structure of an anomaly network-based IDS.

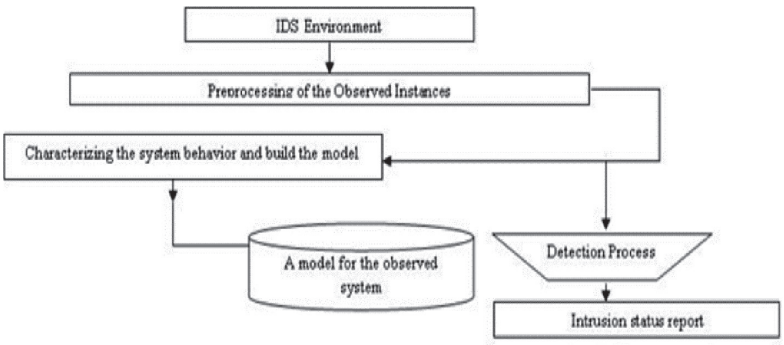


Figure 1. The structure of anomaly network-based IDS. [2: 28]

Overview about Relevant Publications

There are several methods to implement the anomaly-based IDS and to describe normal user behavior. In this section machine learning is described in more detail. Although machine learning is a very popular research field today and a lot of papers have reported about very high (98%) detection ratio and very low (1%) false alert, yet the signature based IDS are still much more preferred. Although in some applications the machine learning systems proved highly efficient, their implementation is rather difficult, consequently the method did not bring a decisive breakthrough. Robin Sommer and Vern Paxson analyzed in their paper Outside the Closed World: *On Using Machine Learning For Network Intrusion Detection* [3] the pitfalls of the implementation: there is not any dataset for teaching the system, the cost of the problem management, the deviations between the reported results and real operation, and the large variability of input data.

Despite these difficulties several attempts were made to implement the anomaly-based IDS using machine learning methods. Denning [4] published first in 1987 about the implementation of a host based IDS system using statistical metrics and profiles.

Other techniques based on machine learning, like genetic algorithm and application of neural network are also widely used and a lot of results were published in this topic. Susan M. Bridges and Rayford B. Vaughn [5] implemented their IDS using fuzzy data mining and genetic algorithm, consequently they proved that the application of genetic algorithm is highly useful to select the most important features characteristic to the attack. Some notable machine learning based systems are the netGA [6] and the system of the University of Minnesota called MINDS. [7] In this topic another work should be mentioned: Hua Tang, Shaolin Cao [8] and Chunmin Qiu, Jie Shan [9] implemented a system using neural networks and the application of the backpropagation algorithm were investigated in [10] and [11].

Method

For the analysis the MathWorks MATLAB [13] was applied with internal neural network components, while for teaching and testing the KDD'99 dataset was used. The algorithm was running on 10% of the KDD'99 dataset during the teaching period, and subsequently 500,000 elements from the test dataset was used to test the efficiency of the neural network and evaluate the results.

Neural Networks

The main goal of neural network research is to understand the information processing and the learning process of the human brain to model the behavior of a human.

The human brain is capable of achieving outstanding performance, to process the information parallel via multiple connections of neurons and to adapt to the changing environment, thus to learn, organize and collect information. It can both generalize and specialize, which is highly advantageous for thinking. The neural networks are designed to the similarity to the human brain. Although the model of human thinking is still too simple, the application of neural networks solved several problems, where the statistical methods failed. The neural

networks are capable of processing the information in parallel, and adapt and learn. If incorrect, noisy input data are provided, the neural network is still capable of achieving correct results. The application of neural networks is an important area of data mining. In case of a large amount of variable data this technique provides the best result. [12] As the network traffic generates a large amount of noisy data, therefore the application of neural networks proves an ideal solution.

Working with the KDD'99 Dataset

How successfully the neural network is taught, depends mostly on the dataset. It is very difficult to find a standardized and freely available dataset. After a proper analysis of the literature the KDD'99 was chosen.

The dataset was generated with a typical US Air Force LAN simulation, TCPDUMP [15] data were collected during seven weeks and several attacks were generated in this period. At final status the dataset contains roughly 5 million records. Each TCP/IP connection was associated with 42 different numerical and non-numerical variables, and in the dataset for each record the 42nd value indicates, whether it is a normal connection or an attack. The name of the attack is indicated, as well.

Table 1 shows the number of attacks and normal connections in the dataset.

Table 1. The number of different records (attack or normal connection) in the training dataset. [Made by the author.]

Attacks	3,925,650
Normal connections	972,781
Total	4,898,431

The values in the dataset can be categorized into three different groups: basic features related to the TCP connections, content features related to the connection and the values referring to the data transmission in the 2 sec time window. The dataset is divided into two groups: one part for teaching and the rest for testing the efficiency of the neural network. The records used for the test contain unknown attacks, as well.

The intrusions can be categorized into four groups:

- *Denial of Service* (DoS) attacks: The intruder blocks/inhibits the accessibility of the system by overuse of the processor/memory, therefore the authorized users cannot access the resources.
- *User to root* (U2R) attacks: The intruder has a user right, but abuses the system to get a root access due to the vulnerability of the system.
- *Remote to user* (R2L) attacks: The intruder monitors the system externally and gets user right by abusing the security of the system.
- *Probing* (PROBE): The intruder collects information about the weak points of the system by scanning (e. g. port scanning).

Coding and Preparing the Dataset

Data standardization is required in the dataset to handle properly the values by the neural network. There are both numerical and non-numerical values among the 42 attributes. [16] As the MATLAB operates with numerical matrices, it was necessary to modify the dataset, that each discrete output value is associated with numerical values.

For simplicity the attack types in the last column are not handled separately, but all attacks are coded with 1, and all normal behaviors are coded with 0. This type of data reduction simplifies the learning process significantly, because the output vector has a single value, which is either 1 or 0.

Backpropagation Algorithm

The Levenberg-Marquardt backpropagation algorithm [17] was used for the analysis.

The learning process is supervised during the run of the backpropagation algorithm, which means that input patterns are provided for the network and the expected output value associated with the input patterns is defined, as well. The network compares the output value for the actual weights with the expected output value and modifies the weights so that the difference between the actual and expected output values becomes as small as possible.

The backpropagation model is a multiple layer model therefore it contains several hidden layers with further junction points between the input and output layers.

Running the Algorithm

The algorithm requires input and expected output values. The dataset used for the teaching is 10% of the total KDD dataset and is coded as described above. The first 41 attributes in the dataset corresponds to the input vector, while the value in the 42nd column coding the attack and normal behavior represents the expected output for the network.

The number of hidden layers is an important parameter. The more layers are in the network the more precisely the network can learn the input dataset. However, there is a risk of overtraining, which affects the network as well. The network recognizes only the values, which are very similar to the input data, consequently it cannot generalize and adapt anymore. These lost features are the major advantage of the neural networks. The large number of hidden layers prolongs the learning time too. The most important advantage of the anomaly-based IDS is the capability to recognize new types of attacks, therefore it is very important to avoid overtraining. Additionally the analysis of the network traffic should be done optimally in real time, so the time for teaching should be decreased as low as possible. By taking into account these facts the number of layers was chosen to be 10 and this value was tested and proved a proper selection.

Efficiency Test of the Neural Network

The operation of the network is tested on 500,000 data selected randomly from the first 697,513 rows of the test dataset.

The output value is a floating point number between 0 and 1, an approximate value provided by the network. The output values are coded with 1 and 0, for the attacks and normal behavior, respectively. Furthermore the upper and lower threshold output values should be defined as well: above or below this value the output is classified as an attack (1) or normal behavior (0). The selection of the optimal threshold values was investigated and found to be highly crucial.

For the analysis of the network efficiency 5 criteria were taken into account: true positive, false positive, true negative, false negative and the accuracy.

The true positive value indicates the ratio of correctly interpreted attacks – when the network correctly identifies the attack – relative to the total number of attacks in the test dataset.

The false positive value is the opposite of the true positive value. In this case even the normal behavior is classified as an attack.

The true negative value similarly indicates the ratio of correctly interpreted normal behavior data in the whole dataset. The false negative measures the ratio of unidentified attacks, when the network does not detect the attack and code as a normal behavior.

Accuracy was calculated taking the sum of true positive and true negative values, divided by the number of data in the dataset.

Evaluation of the Results

The Receiver Operating Characteristic (ROC) analysis [18] was used for the evaluation and illustration of the results and the evaluation criteria introduced in the previous chapter were plotted as a ROC curve. The test was carried out on a dataset with 500,000 elements selected from the KDD'99 dataset. On the basis of statistical analyses the dataset seemed to be suitable for the test, because all types of the attacks are included and the normal and the attack type behavior are represented equally.

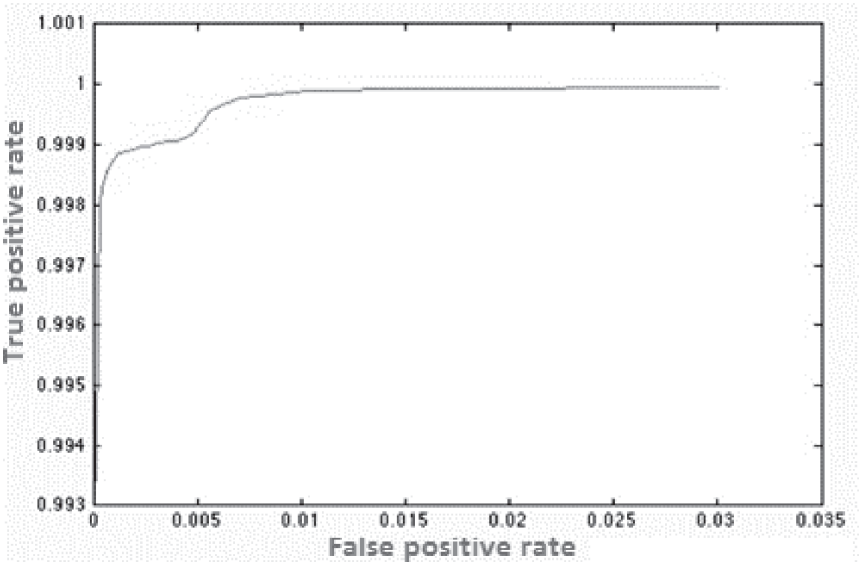
Table 2. Ratio of normal and attack type rows in the test dataset.
[Made by the author.]

Attacks	304,499
Normal connections	392,014
Total	697,513

First it was investigated how the network can distinguish after intensive training with a high number of data between normal behavior and attacks on unknown dataset.

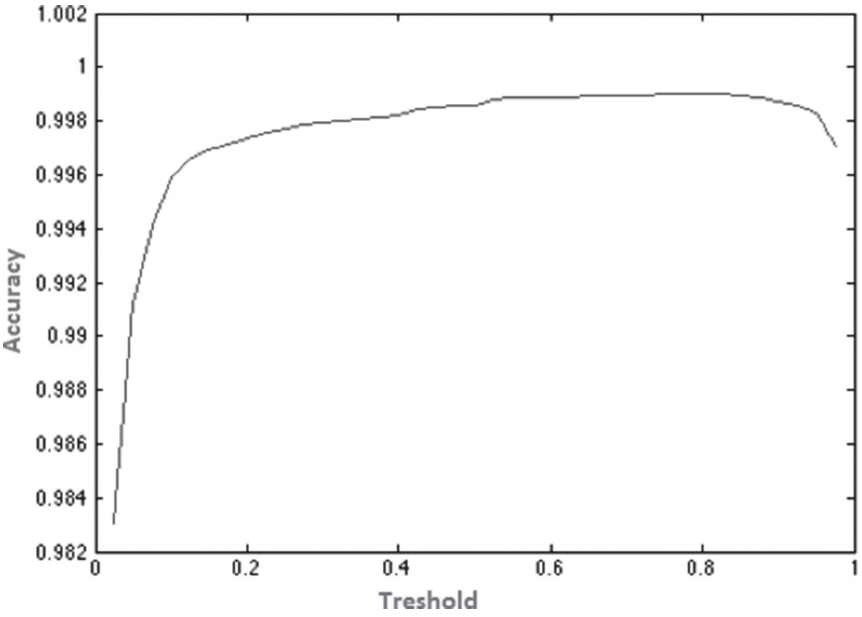
The following results were obtained after the training on 100% of the training dataset:

It can be clearly seen on Graph 1 that more precise detection of attacks accompanies a higher number of false alerts. 99.994% true positive rates can be achieved at false alert value of 3.006%. The optimal ratio depends on the demands of the company operating the IDS.



Graph 1. The ratio of correctly detected attacks and false alerts.
[Made by the author.]

An important parameter, the connection between the threshold value and the accuracy was investigated. The maximum accuracy can be achieved at the threshold value of 0.8, but roughly at around 0.5 and 0.6 a proper accuracy can be expected. (Graph 2)

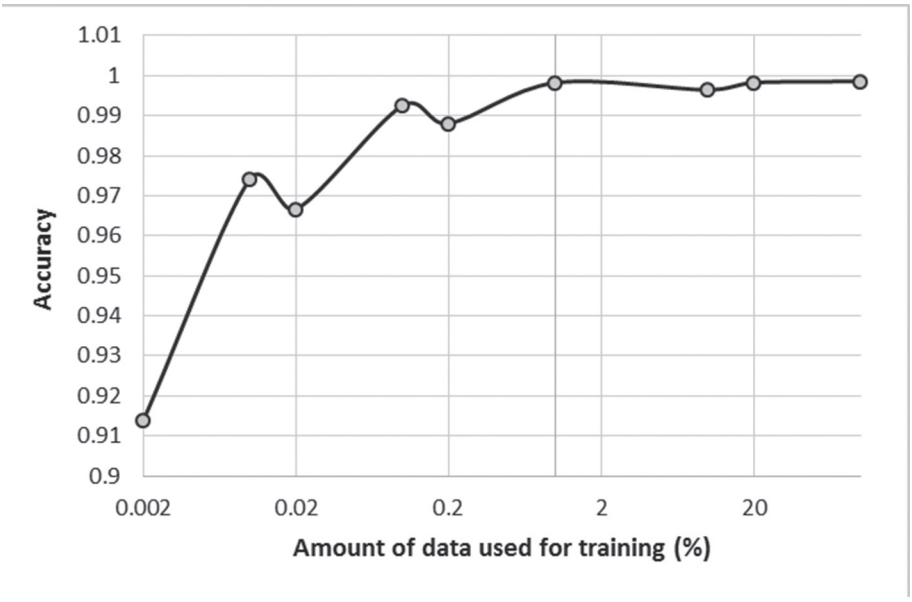


Graph 2. The dependence of the accuracy on the threshold values.
[Made by the author.]

Another interesting question was investigated, namely, how the size of the dataset used for the training influences the performance of the neural network. Several test runs were performed on different percentage of the dataset (0.002, 0.01, 0.02, 0.1, 0.2, 1, 20 and 100%) and Table 3 and Graph 3 show the corresponding accuracy.

Table 3. The accuracy of the neural network and the data used for training.
[Made by the author.]

Training %	Accuracy
0.002	0.913828
0.01	0.974154
0.02	0.966624
0.1	0.99249
0.2	0.988058
1	0.998218
10	0.996454
20	0.998246
100	0.998584



Graph 3. The accuracy of the neural network as a function of the amount of data used for training.
[Made by the author.]

In order to reach the maximum value, it is enough to teach the 1% of the complete dataset. If the training was performed on few data, the accuracy value fluctuates a lot, while the training on high number of data provides a stable accuracy close to the maximum value. This feature is very important for the IDS systems, because the capability of the system to learn on few data helps to monitor effectively the network traffic and to adapt successfully to the permanent change of the user behavior on the network.

According to the obtained data the neural network was able to learn effectively the user behavior, therefore it can distinguish between the normal and attack type behavior very precisely even on unknown data.

Summary, Future Work

This article introduced the possibility of application of artificial intelligence and neural network as IDS or IPS. A very high (99.9985%) accuracy and rather low (3.006%) false alert rate was achieved but several errors were detected. Although the Levenberg-Marquardt backpropagation algorithm used for teaching is highly efficient, it requires a lot of memory. It would be useful to test other less memory demanding teaching algorithms to optimize the use of the resources.

All the 41 features of the KDD dataset were used, and further investigations can reveal which elements influence significantly the output value and reduce the amount of input data to make the training more efficient and to reduce the use of the resources. Several publications focus on this topic, as the learning rate and limited use of resources are very important for the IDS system.

References

[1] GYURÁK G.: Kritikus infrastruktúrák védelme hálózati behatolás jelző rendszerekkel. *Hadmérnök*, X 2 (2015), 223–233.

[2] JYOTHSNA, V., RAMA, V., PRASAD, V.: A Review of Anomaly based Intrusion Detection Systems. *International Journal of Computer Applications*, 28 7 (2011), 26–35.

[3] SOMMER, R., PAXSON, V.: Outside the closed world: On using machine learning for network intrusion detection. In. *Security and Privacy – 2010 IEEE Symposium*, 305–316. California, 2010.

[4] DENNING, D. E.: An Intrusion Detection Model. In. *Proceedings of the Seventh IEEE Symposium on Security and Privacy*, 119–131. Oakland, 1986.

[5] BRIDGES, S. M., VAUGHN, R. M.: Fuzzy Data Mining and Genetic Algorithms Applied to Intrusion Detection. *Proceedings of the Twenty-third National Information Systems Security Conference.* Baltimore, 2000.

[6] LAVENDER, B. E.: *Implementation of Genetic Algorithms into a Network Intrusion Detection System (netGA), and Integration into nProbe.* Sacramento: California State University, 2010.

[7] CHANDOLA, V., EILERTSON, E., ERTOZ, L., SIMON Gy., KUMAR, P.: *Data Mining for Cyber Security.* New York, Philadelphia: Springer, 2006.

[8] HUA TANG, D., CAO, Z.: Machine Learning-based Intrusion Detection Algorithms. *Journal of Computer Information Systems*, 5 6 (2009), 1825–1831.

[9] QIU, C., SHAN, J.: Research on Intrusion Detection Algorithm Based on BP Neural Network. *International Journal of Security and its Applications*, 9 4 (2015) 247–258.
[10] ANBALAGAN, E., PUTTAMADAPPA, C., MOHAN, E., JAYARAMAN B., MADANE, S.: Datamining and Intrusion Detection Using Back-Propagation Algorithm for Intrusion Detection. *International Journal of Soft Computing*, 3 (2008), 64–270.
[11] CHANG, R-I., LAI, L-B., SU, W-D., WANG, J-C., KOUH, J-S.: Intrusion Detection by Back propagation Neural Networks with Sample-Query and Attribute-Query. *International Journal of Computational Intelligence Research*, 3 1 (2007), 6–10.
[12] ALTRICHTER M., HORVÁTH G., PATAKI B., STRAUSZ Gy., TAKÁCS G., VALYON J.: *Neurális hálózatok.* Budapest: Panem kiadó, 2007.
[13] *MathWorks MATLAB.* www.mathworks.com/products/matlab/ (Downloaded: 28 02 2016)
[14] *KDD'99 Dataset.* http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (Downloaded: 01 03 2016)
[15] *TCPDUMP.* www.tcpdump.org/ (Downloaded: 28 02 2016)
[16] *Attributes in the KDD'99 Dataset.* http://kdd.ics.uci.edu/databases/kddcup99/kddcup.names (Downloaded: 28 02 2016)
[17] *Levenberg-Marquardt backpropagation algorithm.* www.mathworks.com/help/nnet/ref/trainlm.html (Downloaded: 15 03 2016)
[18] *Receiver operating characteristic (ROC) or ROC curve.* https://en.wikipedia.org/wiki/Receiver_operating_characteristic (Downloaded: 02 04 2016)