DOI: 10.5604/01.3001.0012.8022

# BUILDING INTRUSION DETECTION SYSTEMS BASED ON THE BASIS OF METHODS OF INTELLECTUAL ANALYSIS OF DATA

**Serhii Toliupa, Mykola Brailovskyi, Ivan Parkhomenko**
Taras Shevchenko Kyiv National University, Faculty of Infirmation Security

***Abstract.*** *Nowadays, with the rapid development of network technologies and with global informatization of society problems come to the fore ensuring a high level of information system security. With the increase in the number of computer security incidents, intrusion detection systems (IDS) started to be developed rapidly.Nowadays the intrusion detection systems usually represent software or hardware-software solutions, that automate the event control process, occurring in an information system or network, as well as independently analyze these events in search of signs of security problems. A modern approach to building intrusion detection systems is full of flaws and vulnerabilities, which allows, unfortunately, harmful influences successfully overcome information security systems. The application of methods for analyzing data makes it possible identification of previously unknown, non-trivial, practically useful and accessible interpretations of knowledge necessary for making decisions in various spheres of human activity. The combination of these methods along with an integrated decision support system makes it possible to build an effective system for detecting and counteracting attacks, which is confirmed by the results of imitation modeling.*

**Keywords**: intrusion detection systems, attacks, fuzzy logic, neural networks

## BUDOWA SYSTEMÓW WYKRYWANIA ATAKÓW NA PODSTAWIE METOD INTELIGENTNEJ ANALIZY DANYCH

***Streszczenie.*** *W chwili obecnej szybki rozwój technologii sieciowych i globalnej informatyzacji społeczeństwa uwypukla problemy związane z zapewnieniem wysokiego poziomu bezpieczeństwa systemów informacyjnych. Wraz ze wzrostem liczby incydentów komputerowych związanych z bezpieczeństwem nastąpił dynamiczny rozwój systemów wykrywania ataków. Obecnie systemy wykrywania włamań i ataków to zazwyczaj oprogramowanie lub sprzętowo-programowe rozwiązania automatyzujące proces monitorowania zdarzeń występujących w systemie informatycznym lub sieci, a także samodzielnie analizujące te zdarzenia w poszukiwaniu oznak problemów bezpieczeństwa. Nowoczesne podejście do budowy systemów wykrywania ataków na systemy informacyjne jest pełne wad i słabych punktów, które niestety pozwalają szkodliwym wpływom na skuteczne pokonanie systemów zabezpieczania informacji. Zastosowanie metod inteligentnej analizy danych pozwala wykryć w danych nieznane wcześniej, nietrywialne, praktycznie użyteczne i dostępne interpretacje wiedzy niezbędnej do podejmowania decyzji w różnych sferach ludzkiej działalności. Połączenie tych metod wraz ze zintegrowanym systemem wspomagania decyzji umożliwia zbudowanie skutecznego systemu wykrywania i przeciwdziałania atakom, co potwierdzają wyniki modelowania.*

**Słowa kluczowe:** systemy wykrywania włamań, ataki, logika rozmyta, sieci neuronowe

## Introduction

With the rapid development of network technologies and global informatization of society to the fore problems of ensuring a high level of information system security. With the increase in the number of computer security incidents, started to be developed rapidly intrusion detection systems (IDS).

Traditionally, intrusion detection systems (IDS) are classified according to two characteristics: detection method and system level at which the protection is performed. All developers of intrusion detection systems and organizations that use IDS should understand and study their classification in order to choose the best solutions for information security systems. In the study of various aspects of taxonomy and the application of various options we can achieve a higher level of information system security [8].

## 1. Main part

Today intrusion detection systems are usually software or hardware-software solutions, that automate the process of control occurring in the information system or network (IDS) also independently analyze these events in search of signs of security issues [4, 7, 11]. Since the number of different types and methods of organizing unauthorized intrusions into foreign networks has increased significantly in recent years and intrusion detection systems have become a necessary component of the security infrastructure of most organizations.

Detection of intrusions has been an area of active research for several decades. There are a large number of different methods and approaches for identifying remote network attacks. To protect the information system such common means and methods are used: corporate network security policy; firewalls; router level protection; network audit; intrusion detection systems; procedure for responding to identified attacks, etc. (Fig. 1).

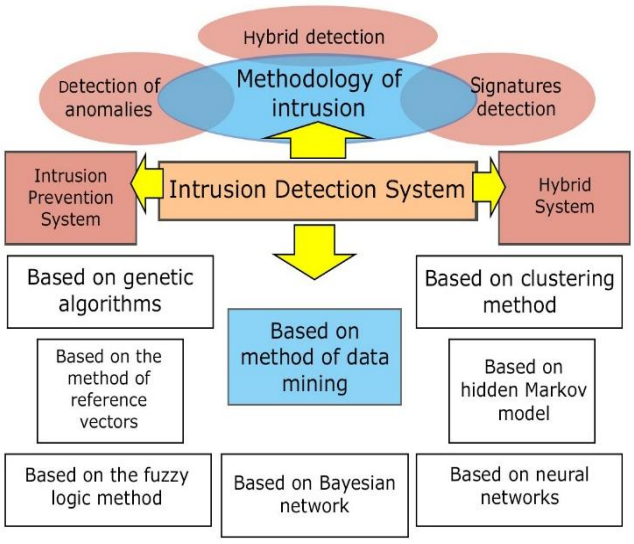


*Fig. 1. Classification features intrusion detection systems*

In their work, intrusion detection systems are guided not only by network traffic and *a* variety of rules, but also by auditing the system, different audit logs, operating system performance indicators, etc. There are also intrusion detection systems that allow not only to detect the fact of the intrusion into the system, but also to minimize the consequences by breaking the network connection, blocking suspicious user activity or even the administrator.

The most effective way to prevent unauthorized use of information systems and network resources is to support multi-level protection, when firewalls, intrusion detection systems, auditing systems, security policies, and other types of protection are used together.

The most general structure of an intrusion detection system, developed by a group of researchers CIDF (Common Intrusion Detection Framework) [14] is presented in Fig. 2.

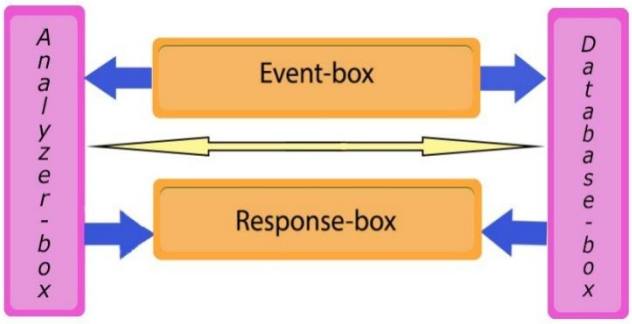


*Fig. 2. General structure of an intrusion detection system*

Event-box (Sensor) – analyzes data for processing and decision making by the analyzer [12]. The Data may contain names of controlled parameters, their features and values. The sensor can convert data to the required format or to reduce the amount of data transmitted.

Analyzer-box – makes a decision about the presence or absence of signs of an attack or anomaly based on data from the sensors. As part of the analysis, this block can perform the functions of filtering, normalizing, transforming, and correlating data. When an attack is detected, the analyzer block can add a description of the detected attack to the source data. The analyzer unit may have a multi-level system.

Database-box – contains a set of decisive rules and a semantic description of attacks, as well as accumulation of information from sensors. The data may be in text files, a database, etc.

Response-box – informs the administrator about a fixed attack, and in the case of an intrusion detection system creates an active response. Intrusion detection systems track activity in real time and quickly implement actions to prevent attacks. Possible measures are: blocking traffic flows in the network, resetting connections, issuing signals to the operator. Also, intrusion detection systems can perform packet defragmentation, TCP packet sequencing to protect against packets with modified sequence numbers and confirmation.

Network attack detection systems collect information from network traffic packets, system logs and system performance indicators. Traditional systems of detecting network attacks are based on the signature approach: with the help of a set of rules or signatures, which are formed by experts and placed in the database of decision rules, all possible scenarios and features of attacks are described. In this approach, there are many known shortcomings. Using signature analysis it is impossible to detect new types of attacks, because the base of decision rules does not contain information about the corresponding attack. The process of analyzing signatures for distributed attacks is extremely difficult. In addition, the databases of decision rules of popular intrusion detection systems are practically publicly available, so the perpetrator can test the possibility of hiding the attack.

The listed problems of the signature search compel experts to look for alternative ways to organize protection against network attacks. One of the popular areas of research is the use of various data mining methods in network attack detection systems. One of the popular areas of research is the application of various methods of intelligent data analysis in systems for detecting network attacks [2, 13].

Data mining – a set of methods for detecting previously unknown, non-trivial, practically useful and accessible interpretations of knowledge necessary for decision-making in various spheres of human activity. The basis of these methods is the assumption that all legitimate activity in the system can be represented as a mathematical model. Mining methods used to detect network attacks have one of the following aims: detection of violations; detection of anomalies.

The first method is simulate attacks and use classification tools. The other method is about simulating normal behavior and searching for exceptions.

Using data mining technique to detect network attacks, the following problems can be identified: considered data by intrusion detection systems have a high dimension and volume the requirement of data processing in real time; a large amount of noise and inconsistencies in the processed data and those which cause inadequate response of data mining methods.

Analyze IDS based on data mining techniques. One such method is to detect an attack using the Hidden Markov Model (HMM) [10]. The Hidden Markov Model (HMM) is a statistical model [6] where the system is modeled as a Markov process with unknown parameters. The objective of the method is to evaluate the hidden parameters that are based on the observed parameters. Event sequences collected from normal operating systems used as a training sample to evaluate the parameters of the Hidden Markov Model (HMM). After investigating the hidden Markov model, probabilistic estimates are used as threshold values for identifying network anomalies in test data.

Detection of attacks using Bayesian networks. The Bayesian network is a model that encodes probabilistic relationships between variables. The main method of using Baessian networks implies independence among the attributes. Several variants of the use of Bayesian networks have been proposed to detect network anomalies [15]. Most of the methods are aimed at forming conditional\to form conditional dependencies between attributes using complex Bayesian networks. Bayesian methods are often used in the procedure for the classification and localization of false positives. Bayesian networks may be effective in some cases to detect intrusions or predict intruder behavior, but in the general case, the accuracy of this method depends on the assumptions related to the behavior of the target system model. Thus, any significant deviation from the assumptions will reduce the accuracy of the detection.

Detection of attacks using clustering methods. The methods of clustering group data into clusters based on the similarity of objects. Most clustering methods begin with the choice of a central point for each cluster, and the set of elements is distributed among the clusters. After that, the centers are adjusted, and the elements are redistributed. Clustering allows you to study and identify anomalies without requiring a plurality of classes or types of anomalies. That is, to identify anomalies using clustering methods, there is no need for a training set. Clustering is widely used to detect network anomalies [16]. Detection of unknown network attacks is most often based on clustering methods. Homogeneous groups with similar characteristics or clusters are formed by splitting a set of elements without any marks. It is extremely important for a system to correctly identify the clusters in order to keep them from emissions as far as possible. The ultimate goal of these methods is to determine the degree of deviation of emissions from clusters. By simply comparing with a threshold values, emissions with a high degree of deviation from clusters are marked as anomalies.

Detection of attacks using the support vector machine. This method is one of the most popular classification methods. The method constructs an optimal hyperplane in the characteristic space: $w \times x - b = 0$, that separates normal and abnormal elements [3, 5]. As a result, the problem can be reduced to quadratic programming: $min \parallel \omega \parallel \frac{2}{H}/2 + c\sum_{i=1}^{N}\xi_i$, when $c_i(wx_i - b) \geq 1 - \xi_i$, $1 \leq i \leq N$, where $\xi_i$ is the magnitude of the error in the objects $x_i$. The parameter $c$ is a compromise between the accuracy of the description of the model, is determined by the magnitude of the error $\sum_{i=1}^{N}\xi_i$ and the possibilities of the model to generalization, that is, the value of the limit $1/\parallel \omega \parallel \frac{2}{H}$ .

Detection of attacks using neural networks. The interest in artificial neural networks is caused by the fact that the human brain produces computing operations in a completely different way than a conventional digital computer. Neural networks provide a variety of tools for a variety of applications: clustering of data, drawing of signs, diminishing of dimension, etc [1].

Detection of attacks using genetic algorithms. Genetic algorithms represent a computational model based on the principles of evolution and natural selection. With this approach, the problem that needs to be solved is transformed into an environment that uses the chromosomal data structure. Chromosomes have evolved over many generations, using operations such as choice, recombination, and mutation [9]. In the task of detecting network anomalies, the record chromosome contains genes that correspond to attributes such as services, flags, number of hits, etc.

Detecting attacks using rules of fuzzy logic. Fuzzy network intrusion detection systems use a variety of fuzzy rules to determine the probability of specific or general network attacks. A fuzzy set can be formed to describe the traffic on a particular network. Work [9] describes a method for constructing classifiers that use fuzzy associative rules that are used to detect an intrusion into a network. Fuzzy sets of association rules are used to describe normal and anomalous classes. Belonging of record to particular class is determined by the appropriate metric. Fuzzy association rules are formed based on the usual training samples. A test sample is classified as normal if the indicator generated by the set of rules is above a certain threshold value. Samples with a lower rate are considered abnormal.

Consider the possibility of using a group of data mining methods for designing a network attack detection system.

According to the previously presented subtasks associated with the identification of network attacks, we can distinguish several groups of data mining methods. At the core of most IDS is the classification process, which forms a conclusion about the fixation of an attack or anomalous behavior. Currently, there are numerous studies on the identification of network attacks. The basis of these studies are such techniques as neural networks, decision trees, association rules, genetic algorithms, and many others.

As a result of the analysis of a large number of studies, the support vector machine has been selected as a classifier. This method shows one of the best indicators of attack detection and has ample opportunities for internal configuration.

The support vector machine finds patterns that are on the borders between two classes, which are called support vectors. In the case of linear inseparability, the kernel mechanism is used, which translates the initial data into a large-dimensional space, in which there is a linear boundary between the classes. The following types of nuclei were most widely used: polynomial, radial-base and sigmoidal nuclei. Due to the high computational complexity, most studies are limited to using radial-base kernel.

Other functional tasks solved by the network attack detection system are mainly aimed at improving the quality of detection, speed of the system, unification and performance of other auxiliary tasks.

Learning principal component analysis is a complex computational task. In addition, the quality of the classification essentially depends on the internal settings of this method, individual for various types of attacks and training sets. In this regard, there is a need to simplify the procedure for preparing a classification model. The solution to this problem is to reduce the size, including the rejection of noise and emissions, as well as the breakdown of the training set into parts – the implementation of the clustering procedure.

The main purpose of the methods of reducing the size is to search for a smaller space in which the internal properties of the source data are stored. These methods allow you to define the set of most important parameters to detect a specific attack. The choice of a particular size reduction method is strongly depends on training data. For the problem that is solved, the method of main components is the most promising.

The purpose of the cluster analysis is to partition data sets into groups in such a way as to minimize the differences between the elements of the same group and maximize the differences between the elements of different groups. The main methods of cluster analysis are divided into hierarchical and non-hierarchical.

Hierarchical methods allow us to construct an optimal cluster structure, but have an exponential dependence on the number of records. Due to the large size of the training data arrays, the use of hierarchical clustering is impossible for all data arrays. But when considering fragments of training data containing records of individual attacks, the amount of information allows the use of hierarchical clustering.

Of the non-hierarchical methods of cluster analysis, iterative methods, which are more universal than hierarchical, but have one serious drawback, are the most commonly used – the need for a priori knowledge of the number of clusters, which greatly complicates their automation.

For the problem of forming the composition of the detection modules, both hierarchical methods and iterations are selected. The hierarchical ones are applied to a subset of attacks from the training set, iterative allows clustering the entire training set, the initial number of clusters and their centers are determined based on the hierarchical clustering subset of attacks.

Fuzzy logic is the generalization of classical formal logic and set theory. Instead of the values of lie and truth, the membership function of an element in the fuzzy set whose values are in the interval [0; 1]. For unclear logic, there are general logical operations based on operations with fuzzy sets. The current base of fuzzy rules allows you to compare the detected anomaly with a set of known attacks. The modular architecture allows you to build an over-the-top model, increasing the likelihood of an attack, and the fuzzy conclusion allows you to reduce the number of false positives.

Fuzzy logic allows you to solve the problem of having the same records with different marks in the training data: "questionable" network packets are a set of anomalous with some probability.

A side application of fuzzy logic is is the construction of clusters that intersect, and an extension of the support vector method. Of course it is ineffective to resist invasion and attacks based on only one of the data mining methods, so it is necessary to approach this issue comprehensively and to build an intellectual system to counteract the invasion. When constructing such an expert system it is suggested to choose a fuzzy model. This is due to the fact that much of the information about the causes and source of attacks can be obtained only by expert or in the form of heuristic descriptions of processes. To determine the sources of attacks, the security system must be represented by a model of the information network on which it is oriented. This model divides the problem of moving information between computers through a network environment into a number of levels of smaller and easier to solve subtasks. Each of these subtasks is solved using a single network level. Therefore, the primary task of a security specialist can be represented by the decomposition of security tasks at individual levels of the network.

Imagine a separate level of security in the form of a nonlinear object with a set of input variables $\{x_i\}, i = \overline{1,n}$ and one output variable *y*: $y = f_y(x_1, x_2, ..., x_n)$.

As input variables are selected signs of sources of attacks. The output variable y is an indicator of the degree of network-level state capability.

The model uses the following assumptions and limitations:

- input variables $\{x_i\}$ within the same level are independent,
- at each level of the network, separate network functions are isolated.

Building an integrated intelligent decision support systems to determine intrusions based on data mining methods should contain a set of functional components that allow to maximize automation and speed up the development of managing actions when chaning the security situation.

## References


[1] Bankovic Z., Stepanovich D., Bojanic S., Nieto-Taladris O.: Improving network security using genetic algorithm approach, Computers and Electrical Engineering, 33(5-6)/2007, 438–451.

[2] Barsegyan A. A., Kupriyanov M. S., Stepanenko V. V., Kholod I. I.: Technologies of data analysis: Data Mining, Visual Mining, Text Mining, OLAP, SPb. BHV, Petersburg 2007.

[3] Bhattacharyya D. K., Kalita J. K.: Network Anomaly Detection. A Machine Learning Perspective, CRC Press, 2014.

[4] Brailovskyi M. M., Pogrebna T. V., Ptakhok O. V.: Essential requirements for the construction and safety of next-generation networks. Telecommunication and Information Technologies 2/2014, 41–49.

[5] Brailovskyi N. N., Ivanchenko E. V., Khoroshko V. A.: Diagnostics of information space protection systems" Information protection. Special issue 2014, 59–67.

[6] Ghahramani Z.: An Introduction to hidden Markov models and Bayesian networks. International Journal of Pattern Recognition and Artificial Intelligence 15/2001, 9–42.

[7] Koboseva A. A., Machalin I. O., Khoroshko V. O.: Analysis of the security of information systems. DUIKT, Kiev 2010.

[8] Pavlov I. M., Toliupa S. V., Nishchenko V. I.: Analysis of Taxonomy of Attack Detection Systems in the Context of the Current Level of Information Systems Development. Modern Protection of Information 4/2014, 44–52.

[9] Tajbakhsh A., Rahmati M., Mirzaei A.: Intrusion detection using fuzzy association rules. Applied Soft Computing 9(2)/2009, 462–469.

[10] Tereikovskiy I., Toliupa S., Parkhomenko I., Tereikovska L.: Markov Model of Normal Conduct Template of Computer Systems Network Objects. 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering TCSET-2018.

[11] Toliupa S. V, Borisov I. V.: Methodology of evaluation of the complex system of information security at the object of information activity. Scientific and Technical Journal "Modern Information Protection" 2/2013, 43–49.

[12] Toliupa S. V, Parkhomenko I. I., Konovalenko A. D.: Analysis of vulnerabilities of local wireless networks and ways to protect them from possible attacks. Journal of the Engineering Academy of Ukraine 3/2017, 72–76.

[13] Toliupa S. V., Parkhomenko I. I.: Multilevel hierarchical models of information security systems. Proceedings of the II International scientific and practical conference Trends in the development of corvergent networks: decision of the post: NGN, 4G, 5G. Kyiv 2016, 111–114.

[14] Valdes A., Skinner K.: Adaptive model-based monitoring for cyber attack detection. Proc. of the Recent Advances in Intrusion Detection, Toulouse, France, 2000, 80–92.

[15] Valdes A., Skinner K.: Adaptive model-based monitoring for cyber attack detection. Proc. of the Recent Advances in Intrusion Detection. Toulouse 2000, 80–92.

[16] Yang H., Xie F., Lu Y.: Clustering and classification based anomaly detection. Fuzzy Systems and Knowledge Discovery 4223/2006, 1082–1091.



**Prof. SerhiiToliupa**
e-mail: tolupa@i.ua

Scientific and practical interests are related to such areas as intelligent control systems, the direction of improving the efficiency of information technology, information security systems, cybersecurity and cyber defense. He is the author of 5 monographs and over 150 scientific and methodological works, 16 textbooks and manuals.

ORCID ID: 0000-0002-1919-9174



**Prof. Mykola Brailovskyi**
e-mail: bk1972@ukr.net

Scientific and practical interests are related to such areas as information and communication systems and networks, systems of technical protection of information, cyber defense. He is the author of over 40 scientific and methodological works.

ORCID ID: 0000-0002-3031-4049



**Prof. Ivan Parkhomenko**
e-mail :parkh08@ukr.net, parkh08@gmail.com

Scientific and practical interests are related to such areas as information and communication systems and networks, systems of technical protection of information, cyber defense. He is the author of over 80 scientific and methodological works.

ORCID ID: 0000-0002-9197-2600


*otrzymano/received: 1.10.2018* *przyjęto do druku/accepted: 15.12.2018*