On The Internet Method Of Identifying Threats Using Svm

Gafurov Sh.A.

Independent researcher of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

ABSTRACT

In this article a basic vector method for detecting threats implemented in web systems on the Internet is proposed, in which threats are detected using artificial intelligence, the introduction of the SVM method in the detection of threats, the process of detecting threats, as well as the construction of an IDS/IPS system based on the proposed method done

Keywords: artificial intelligence , IDS/IPS , support vector mashine(SVM) method , DDoS , web resources, national segment, threat

The system for detecting threats to web applications and resources located in the “.uz” domain of the Internet network is an important guarantee tool for network security. However, an arbitrary network threat detection system usually faces several problems in detecting threats due to the lack of knowledge in the knowledge base. This leads to the deterioration of the defense system. Today, the methods of detection of threats and attacks based on base vectors, even if the network threat and attack detection system does not have a sufficient level of knowledge base (that is, based on a small amount of data), web applications located in the ".uz" domain of the Internet network and allows identifying threats and attacks on resources with high accuracy. The main reason for this is the basic concept of network connectivity and the correct choice of a classifier based on support vectors.

That the “.uz” domain located on the Internet network has an objective state of vulnerability in the Internet and system security, constant network threats are made against the operating system, application software, and hardware devices. In traditional intrusion detection systems, it can detect unknown attack behavior after analyzing network data. Security professionals are increasingly in demand for systems that allow them to automatically detect behavior related to new types of network threats and attacks, as AI -based methods become available .

Based on the Vapnik-Chervonenkis measurement theory and the minimum principle of systematic risk of the statistical learning theory of existing threats and network attacks on the Internet, the support vector (SVM) method is used to determine the complexity of the model (i.e., the accuracy of learning the specified training samples) and learning ability (i.e., the ability to identify any sample without error) is the method with the best learning and generalization ability given limited sample data. Based on different internal parameter functions, the basis vector (SVM) method can be adapted to many existing learning algorithms, such as Polynomial Approximation, Bayesian Classifier, Radial Basis Function (RBF) method, and Multi-Layer

Perceptual Network. The application of the support vector (SVM) method helps the threat and attack detection system to achieve high accuracy.

The proposed method is primarily based on the basic concept of determining access to web applications and resources located in the ".uz" domain of the Internet network and the basic principle of the classifier based on support vectors. In this case, base vectors enable the formation of a network attack detection system that combines detection of threats and attacks, as well as detection of network anomaly attacks and detection of misuse.

The proposed method detects threats and attacks in Intrusion Detection System/Intrusion Prevention System (IDS/IPS), which are used to protect against network threats and attacks, which are web applications and resources located in the ".uz" domain of the Internet network. it is suggested to use the attack detection module. This includes detecting attacks aimed at compromising the confidentiality, integrity, and availability of IDS/IPS system resources, and detecting behavior that violates security strategies or threatens system security by examining operating system audit data or information in network data packets. . The IDS/IPS system is an additional means of protecting web resources located in the national segment.

The proposed method divides the detection system into three parts to fully identify a single threat in a national segment, and these three parts include: These are;

- data collection;
- data analysis;
- threat detection (final response).

The data collection module collects various types of data, such as network data packets and log logs or operating system application logs, including various attack signatures. It always pre-processes the collected data, including bringing the data into a format ready for analysis.

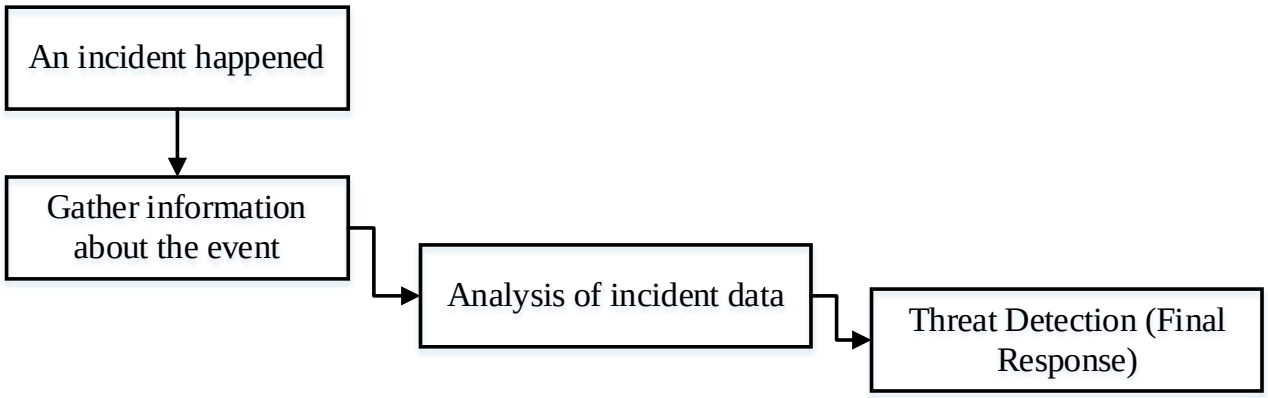

Figure 1. Threat identification process in the national segment

There are several types of IDS/IPS systems used to detect threats and attacks on web applications and resources located in the national segment of the Internet. These are the following IDS/IPS systems :

- according to the information source;
- according to the main component;
- according to the network level;
- according to application level;
- according to the signature ;
- according to the method of analysis of ax boron;
- according to statistical analysis ;
- according to integrity analysis;
- and according to other characteristics.

The proposed method mainly focuses on the IDS/IPS module according to the information analysis method.

To solve the problem of classification of threats and attacks on web applications and resources located in the ".uz" domain of the Internet network, the method of base vectors can be used as follows. In this case, it is assumed that $R_m$in linear separable problems, the method of basis vectors can be used as training data, and it can $\{(x_1, y_i), \dots, (x_i, y_i), \dots, (x_n, y_n), x_i \in R_m,\ y_i \in \{+1, -1\}\}$ be separated by an error-free hyperplane (-classification

hyperline), that is , if there is ma dimensional vector $w$and a constant quantity b, then

$$\begin{cases} (w * x_i) + b > 0 & y_i = 1 \\ (w * x_i) + b < 0 & y_i = -1 \end{cases} \quad (1)$$

In order to solve the problem of classification of threats and attacks on web applications and resources located in the ".uz" domain of the Internet network, optimal separation of two types of data is a matching search, (w, b)that is, optimal classification is a hyperlinear search. The classification hyperline that is farthest from the two types of sample points can have the optimal classification ability (that is, the ability to most reliably represent the boundary of the two types of samples). After normalizing the classifier hyperline, the optimal classifier is assumed to depend on a small number of sample points (basis vectors) that are closest to the hyperline and have nothing to do with other samples.

$$y_i\big((w * x_i) + b\big) - 1 \geq 0,\ i = 1, \ldots\ldots n \quad (2)$$

The range of threats and attacks against web applications and resources located in the national segment of the Internet network $\frac{2}{||W||}$is equal to . According to the basis vector method, the classification $\frac{1}{2}||w||^{2}$interval is maximized , which is equivalent to minimizing . In this case , a classification hyperplane that satisfies condition (1) and minimizes is called an optimal classification hyperplane.$\frac{1}{2}||w||^{2}$

In this case, the Lagrange function is expressed as follows.

$$L\,(w, b, \alpha) = \frac{1}{2}\big||w|\big|^2 - \sum_{i=1}^{n} \alpha_i (y_i (w * x_i) + b), \qquad \alpha_i \geq 0 \qquad (3)$$

$$\begin{cases} \dfrac{\partial}{\partial b} L\,(w, b, \alpha) = \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \dfrac{\partial}{\partial w} L\,(w, b, \alpha) = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \end{cases} \qquad (4)$$

According to the Kuhn-Tucker theorem,

$$\alpha_i\big(y_i\big((w * x_i) + b\big) - 1\big) = 0, \quad i = 1,2 \ldots n \qquad (5)$$

where $\alpha_i$is the Lagrange multiplier corresponding to each sample and is the $\alpha_i$sample $x_i$support vector corresponding to the nonzero . According to expressions (3) and (4), the maximum objective function is as follows.

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \big(x_i * x_j\big), \ \ \alpha_i \geq 0, \qquad \sum_{i=1}^{n} \alpha_i\, y_i = 0 \quad (6)$$

Let us assume that equation (6) $\alpha^* = (\alpha_1^*, \ldots \alpha_n^*)$is the optimal parameter, then the parameters of the hyperplane of optimal classification of threats and attacks to web applications and resources located in the ".uz" domain of the Internet network will be as follows.

$$w^* = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$b^* = y_i - w^* * x_i \qquad (7)$$

Then the optimal hyperline is as follows.

$$\sum_{i=1}^{n} \alpha^* y_i (x_i * x_j) + b^* = 0$$

Thus , the function of optimal classification of threats and attacks on web applications and resources located in the ".uz" domain of the Internet network can be expressed as follows:

$$f(x) \ = sgn\left\{\sum_{TV} \alpha^* y_i (x_i * x_j) + b^*\right\} \qquad (8)$$

Nonlinear problems can be transformed into linear problems in certain high-dimensional spaces by nonlinear transformations, and optimal classification hyperlines are solved in transformation spaces. It should be noted that the above dual problems, the optimal objective function (6) or the classification function (8) only include between training samples .$(x_i * x_j)$

If $K$the function is determined to complete the operation of the point network in the training space, the operation of the internal network is determined to be only in the high-dimensional space, and this function can be implemented in the national segment of the network. Therefore, linear classification after a certain nonlinear transformation $K\ (x_i * x_j)$can be performed by adopting the internal network state function in the optimal classification hyperline, but the computational complexity does not increase. Here, the objective function is transformed into expression (6).

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K\left(x_i * x_j\right) \qquad (9)$$

Thus, the optimal classification function of threats and attacks to web applications and resources located in the ".uz" domain can be expressed as follows.

$$f(x) = sgn\left\{\sum_{TV} \alpha^* y_i K(x_i * x_j) + b^*\right\} \qquad (10)$$

The principle of operation of the basis vectors is shown in the figure below, $x_1, \dots, x_j$the basis vectors and$X^1, \dots X^d$ based on input vectors.

Based on the above, the basis vector training algorithm is implemented following the following sequence.

First of all, it is necessary to determine the general format for presenting the collected information about threats and attacks on web applications and resources located in the ".uz" domain to the system. Having the same representation of the data in the input format and the data in the output value increases the efficiency of the support vector training system.

The information given above shows that the method of basis vectors is actually the selection of the optimal parameter ie $\alpha^* = (\alpha_1^*, \dots \alpha_n^*)$. The process of finding the optimal parameter is usually an iterative algorithm, which can be considered a learning process. In some literature, support vector optimization is also referred to as support vector training algorithm. This training can be mainly divided into two types.

Fragmentation. This is based on the fact that a training sample equal to zero in the above Lagrange multiplier is removed and does not affect the solution of the problem. The main goal of fragmentation is to solve the problem of one-by-one exclusion of those that do not belong to the base vectors through a certain iterative feature (selection).

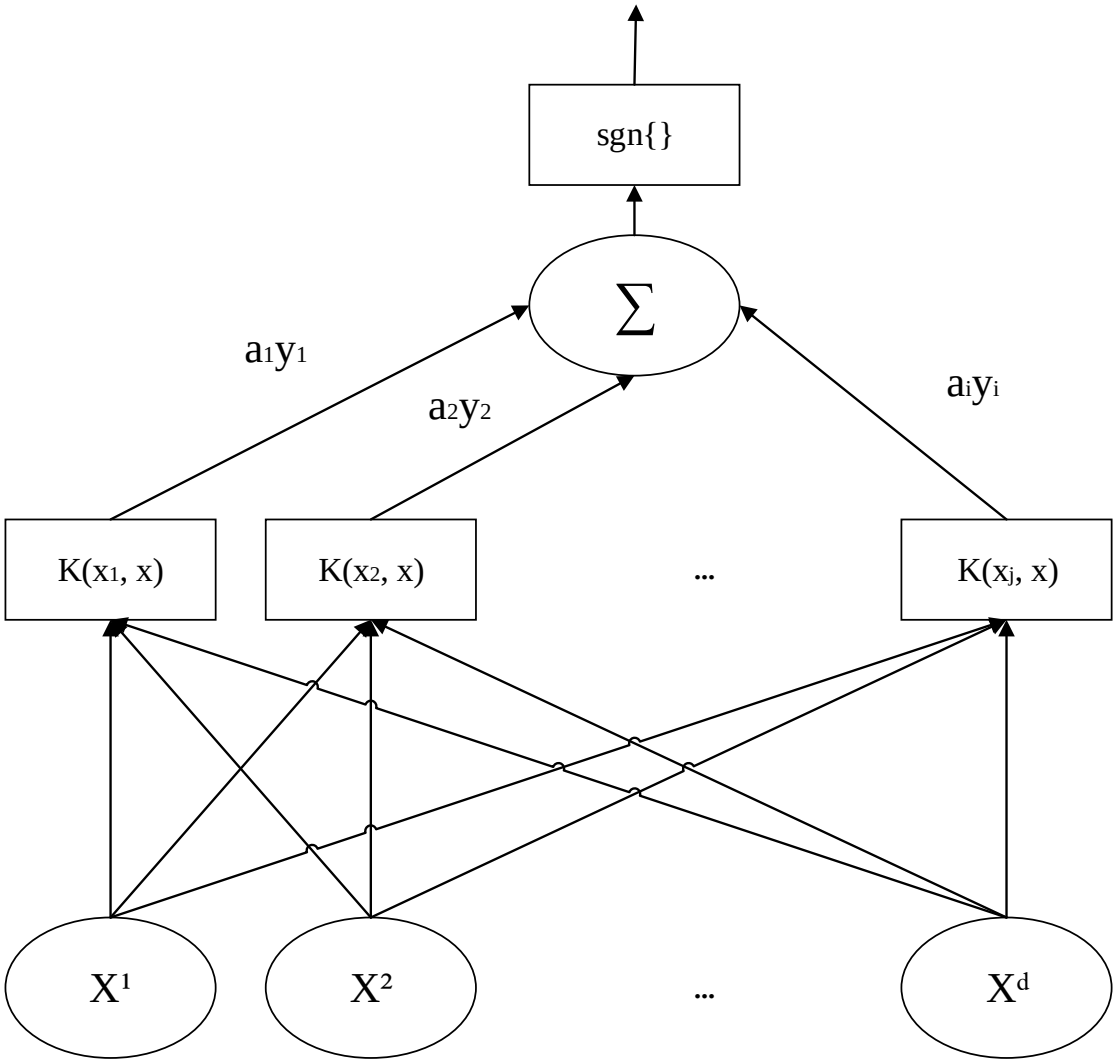

Figure 2. Principle of operation of base vectors

Secondly, the problem of identifying threats and attacks on web applications and resources located in the domain ".uz" is to divide them into small problems (problematic characters) with a fixed number of samples. In this case, the size of the working sample set is controlled within the allowable limit of the algorithmic speed, and the iterative process consists of exchanging equal amounts only between the "state-worst samples". Even if the number of basis vectors is larger than the size of the working sample set, the scale of the working sample set does not change and only a part of the basis vectors is optimized.

Differences between the proposed method and the current method:

- the objective function of the sample partitioning algorithm contains only the samples of the current working sample set;
- although the admissible working sample contains only the optimal variables, its objective function covers the entire training sample set;
- the Lagrange multipliers of the samples outside the working sample set are taken into the result of the previous iteration, but are not recorded as 0 (in current methods, the Lagrange multipliers are recorded as 0 in the fragmentation algorithm).

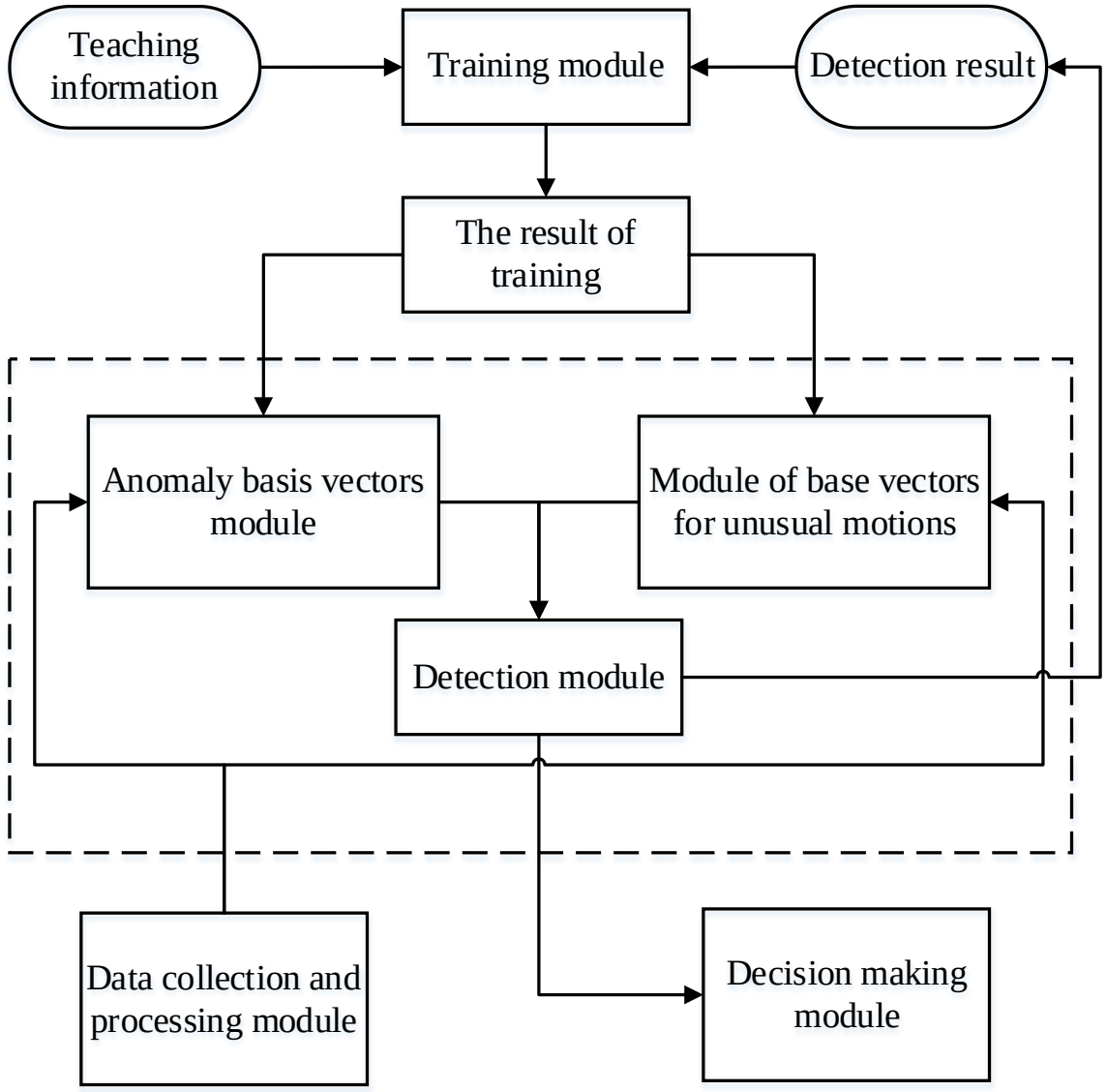

Figure 3. Module of the proposed method in the IDS/IPS system

This is also the case with the fixed working sample set method because replacement samples must be identified. Thus, the key to such an algorithm requires the development of a suitable iterative strategy that allows the algorithm to converge on the final result and quickly converge to the optimal result .

The module of the IDS/IPS system, i.e., the detection system of threats and attacks on web applications and resources located in the ".uz" domain, operating on the basis of the proposed method and algorithm, will be as follows. In this case, according to the network attack detection methodology, the processing object of the attack detection system is the packets in the network. It is necessary to monitor the various network packets being transmitted on network segments in real time and to analyze this data at the same time. If the events that do not have signatures in the database of signatures are detected, the protection system issues a signal (message) about this situation and considers this event as an anomaly.

Support vectors is a classification method with better learning ability for small samples, high training speed and decision-making speed in detecting threats and network attacks that share web applications and resources located in a national segment of the Internet network. , the size of the input data, offers advantages such as continuous adjustment of various parameters as the training data increases. In addition, support vectors allow solving many practical classification problems as well as small samples (symbols) and non-linear problems.

References:

1. M. Arafat, A. Jain, and Y. Wu, "Analysis of intrusion detection dataset NSL-KDD

using KNIME analytics" Proc. 13th Int. Conf. Cyber Warf. Secure. ICCWS 2018, vol. 2018-March, pp. 573–583, 2018.
2. V. Hajisalem, S. Babaie, A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection, Comput. Netw. 136 (2018) 37–50,
3. M.A. Ferrag, L. Maglaras, S. Moschoyiannis, H . Janicke, Deep learning for cyber security intrusion detection: approaches, datasets, and comparative studies, J. Inf. Secure. Appl. (2020) 50, https://d o i. org /10.1016/ j.jis a.2019.102419 .
4. Sharafaldin, A.H. Lashkari, A.A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, in: ICISSP 2018 - Proc. 4th Int. Conf. Inf. Syst. Secure. Priv., 2018, pp. 108–116.
5. R. Patil, H. Dudeja, C. Modi, Designing an efficient security framework for detecting intrusions in virtual network of cloud computing, Comput. Secure. 85 (2019) 402–422.