Advances in Computer Science Research (ACSR), volume 72
IV International Research Conference "Information Technologies in Science, Management, Social Sphere and Medicine" (ITSMSSM 2017)

Network traffic clustering for intrusion detection

Nikishova Arina
Department of Information Security
Volgograd State University
Volgograd, Russia
nikishova.arina@volsu.ru

Ananina Irina
Department of Telecommunication System
Volgograd State University
Volgograd, Russia
matuny77@gmail.com

Ananin Evgeny
Department of Information Security
Volgograd State University
Volgograd, Russia
zananin@ya.ru

***Abstract*—The problem of network attacks detecting is considered. It is proposed to use clustering of network packets for anomaly detection in network traffic. Anomalies may indicate the implementation of network attacks. The used clustering algorithm is k-means method. It has a number of parameters, the choice of which affects the speed and accuracy of network attacks detection. Software package that implements different variants of values of k-means method's parameters is developed. With help of software package experimental studies are carried out. During experiments accuracy of simulated network attacks detection and speed of software package functioning is determined. Based on results the most effective set of k-means method's parameters for network attacks detection is offered.**

Key words—intrusion detection, network attack, clustering, k-means method, efficiency, errors of intrusion detection.

I. INTRODUCTION

According to statistics from InfoWatch, the network is the main channel for information leakage in organizations and amounted to 69.5% in 2016 (Fig. 1). If you take into account only intentional leakage through the network, then their amount in 2016 is 90.1%.

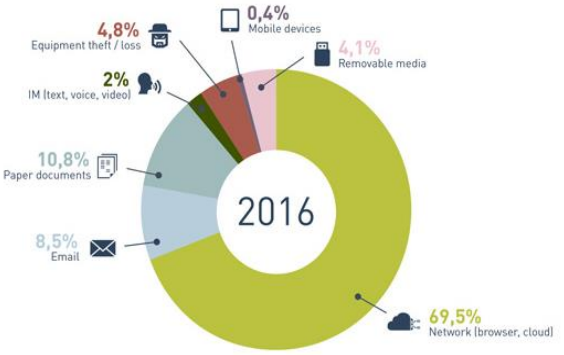

Fig. 1. The statistics of the channels for confidential information leakage in 2016

The resulted statistics allows to draw a conclusion that the main aim is to reveal signs of attacks in the network traffic.

In general, the functioning of organizations is characterized by a slight variation in the tasks to be solved, which makes it possible to characterize as stable functioning of the information system. In this case, anomaly detection methods can be used to detect attacks. Clustering is one of such techniques.

II. EXISTING APPROACHES TO INTRUSION DETECTION

Different approaches to intrusion detection process improvement are proposed in many papers. For example in article [1] the author proposes to use ensemble classifier based on naïve Bayes and ADTree. Usage of ADTrees complements naïve Bayes approach, which assumes that all analyzed features are independent. In cases where there is a complex dependency between analyzed features, the author proposes to combine both approaches to improve the classification accuracy. It is shown that the proposed classifier outperforms other classifiers in terms of accuracy. However, the work is not taken into account the time of classifier operation.

In paper [2] it is proposed to use outliers' detection with the help of data mining to provide a robust mechanism of distinguishing between normal and anomalous activities. The author claims that the proposed approach will reduce false alarm rate. The paper contributed enhanced mechanism of outlier detection to enhance accuracy in intrusion detection by introducing density based outlier detection into data mining using hamming densities of a data point. Hamming density is k-nearest neighbour divided by Hamming-distance. The author doesn't take into account the number of missed attacks by this method, which is more important than the number of false alarm in intrusion detection.

There are also papers which authors propose to use clustering to solve the problem of intrusion detection. For example, in paper [3] author proposes to use new intrusion detection technology suitable for cloud computing environment. It is based on silhouette coefficient and partitioned clustering subspace method to improve the bisecting K-means unsupervised learning method. Proposed approach is distributed. In article results of experimental studies of proposed method compare to basic clustering method on which it is based are introduced. The results of the experiments showed increase in speed and decrease in errors of both types amount for improved approach. The disadvantage is that the proposed method is specific to cloud infrastructure.

Author of paper [4] suggests using clustering of log lines to solve the intrusion detection problem and not for forensic purposes only. Author introduces semi-supervised concept for incremental clustering of log data. Its operation is independent

from the syntax and semantics of the processed log lines, which makes it generally applicable. It is claimed that the introduced approach has linear complexity. The disadvantage of this approach is that it can be used to detect a limited set of attack classes.

In paper [5] approach based on combination of unsupervised data mining techniques as intrusion detection system is introduced. Clustering of windowed incoming packets is applied. Proposed approach has been evaluated and compared with several existing intrusion detection approaches. Results indicates accuracy increase. However, speed of implemented approach operation is not assessed. The disadvantage of this approach is that it can be effectively applied only for the detection of DDoS attacks.

III. DATA CLUSTERING

The algorithm of the data clustering method consists of the following steps:

— Determination of the characteristics of the analyzed objects. It is necessary to determine the most significant properties of the analyzed object characterizing the clusters. It is also necessary to carry out the normalization of properties.
— Allocation of metrics for the formation of clusters. Each analyzed object is described by a vector of properties. To calculate the distance between vectors of properties, the metric or the distance function is used.
— Separation of objects from the training sample into clusters. It is carried out in accordance with the clustering algorithms.
— Assignment of the analyzed object to one of the clusters. After forming clusters based on the training sample, when receiving each analyzed object, the cluster is determined by the cluster distance which is the smallest among all clusters.

As the analyzed objects are network packets. To determine the characteristics, many of which will detect network attacks, the network packet headers for IP, TCP, UDP, and ICMP protocols were analyzed. Based on the results of the analysis, it is proposed to use the following set of characteristics (fields of packet headers) to detect attacks: (source address; destination address; source port; destination port; protocol; TCP flags; ICMP type) for IPv4 protocol and (source address; destination address; source port; destination port; next header; TCP flags; ICMP type) for IPv6 protocol.

It is necessary to normalize the selected characteristics. After that each network packet transmitted in the information system is described by the vector of normalized characteristics - properties. To calculate the distance between the vectors of the properties of packages, measures of similarity or similarity are used. They are also called metrics or distance functions:

- Euclidean distance.
- Squared Euclidean distance, which is used in cases when it is necessary to give greater weight to remote from each other points.
- Manhattan distance, for which the effect of individual emissions is less than using Euclidean distance, since the difference of corresponding coordinates is not squared.
- Chebyshev distance, which should be used, when it is necessary to classify some objects as distinct if they differ only in one of the properties.
- Power distance, which should be used if the task is to increase or decrease the influence of one of the properties whose values are very different in the vectors.

Separation of objects into clusters can be carried out in accordance with the algorithms of clustering. The following main algorithms are distinguished:

- hierarchical algorithms;
- k-means method;
- the method of the nearest neighbor;
- fuzzy clustering algorithms.

The most interesting is the k-means method, since it is rather simple, has a high speed of operation, it is easy to adapt, and it makes it possible to split objects not by a single property, but by a vector of properties.

The initial learning algorithm for the k-means method is shown in Fig. 2.

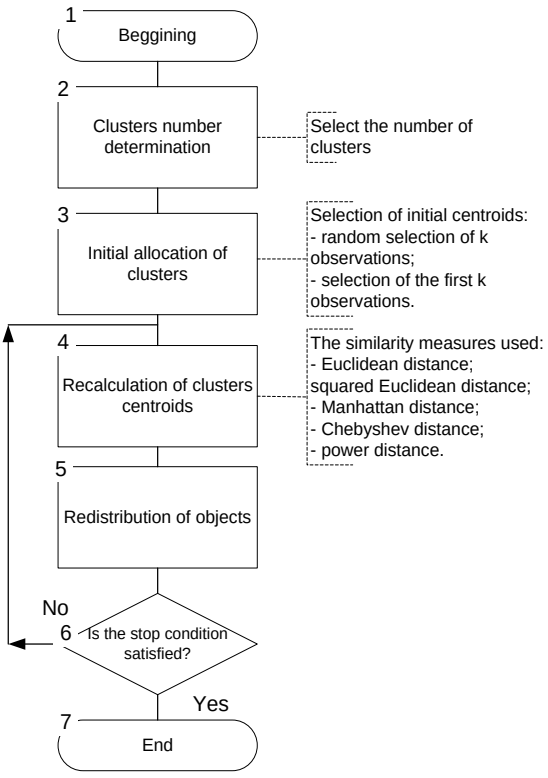

Fig. 2. Algorithm of initial learning of the k-means method

IV. K-MEAN METHOD PARAMETERS

The set of packets transmitted over the network is a multidimensional space and is difficult for description. In the process of clustering, it is divided into subspaces. It is proposed to determine experimentally the best number of subspaces from the point of view of clustering errors into which space is divided, which are clusters.

Other characteristics of the chosen clustering method that affect both the accuracy of network packet assignment to the generated clusters and the speed of learning and analysis with the method is similarity measure or distance function used during clustering and the method of initial selection of cluster centers.

During the initial allocation of objects to clusters the set of points in number equal to the number of clusters is chosen. In the first step of training these points are considered "centers" of the clusters. Each cluster corresponds to one center. The choice of initial centroids can be done in the following ways:

- select k points randomly;
- select the first k points of set.

The criterion for the quality of the functioning of the method is the number of errors. There are two kinds of errors:

- errors of the first kind (false positive), which include cases where there are no negative processes in information system, that would result from actions of intruder, but the network packet does not belong to the clusters of normal behavior;
- errors of the second kind (false negative), which include cases where there are negative processes in information system, that are the result of intruder actions, but the network package refers to the clusters of normal behavior.

To carry out experimental studies, a software package for network traffic clustering for detecting attacks was developed. For interaction with user and display of functioning results the user interface consisting of three tabs was developed.

The first tab of user interface is “Parameters selection”. It includes a group of radio buttons for selecting the number of clusters, a group of radio buttons to select the initial choice of centroids, a group of radio buttons for selecting the method of measures of similarity calculation and button for starting of training (Fig. 3).

In addition, the tab displays a list of network interfaces available on the computer, from which you must choose the analyzed interface. Clicking on the start button launches the collection of network packets for clustering.

The second tab of the user interface “Network packets” includes a table to display information on received network packets, Stop button to start the process of received network packets clustering and Result display button to output current clustered space.

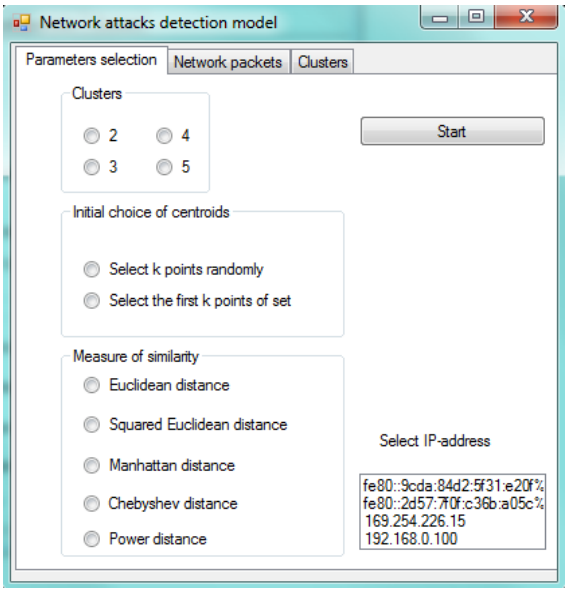

Fig. 3. Tab “Parameters selection” of user interface

The third tab of the user interface "Clusters" includes the area for clustered space projection, list of current clusters for selection attack included ones from them according to the results of cluster analysis, as well as field to display clustering time. After selecting the clusters that contain attacks, you can start the process of intrusion detection.

The task of experimental studies is to select the best values for the variable parameters of the k-means method of clustering, which were allocated during development of network traffic clustering model for intrusion detection. This task includes the following subtasks:

- collecting of network packets representing normal operation, i.e. absence of attacks;
- adding to the collected set the network attacks belonging to the main classes of network attacks: TCP-ports scanning, “Land” attack, “PUKE” attack, “UDP Bomb” attack, as well as samples of network packets with incorrect service data, the values of which exceed the allowed values. These data include the values of the source and destination ports, as well as the network packet identifier;
- clustering of the received set of network packets according to given parameters with calculation of false positive and false negative;
- the choice of such values of the parameters at which the effectiveness of intrusion detection will be greatest.

In the framework of experimental research is the initial formation of clusters on the collected set of network packets. Then, the resulting clusters are analyzed by the information security specialist. The cluster which includes the largest number of all the analyzed events is defined as the cluster of normal behavior of the system. If the cluster includes events

from the selection related to samples of network attacks, it is considered as missing attacks, i.e. error of the second kind. The clusters containing the analyzed events, which are attacks, are specified on the third tab of the user interface. After the initial splitting of learning sample into clusters, receiving packets are classified by assigning them to one of the formed clusters.

V. EXPERIMENTAL STUDIES

For clustering, 10,000 network packets were collected, which are considered normal. They were added 200 network packets - network attacks on the information system, which include the following packages:

1) TCP-ports scanning. This attack using the XMAS scan method. The attacker scans the TCP ports of the router using a program Zenmap to identify open ports. In this case, port scanning refers to the type of Stealth XMAS Scanning, that consists in sending TCP packets with the FIN, URG and PSH flags set. If is answered with RST packet then the port is considered closed, while no response means that the port is open or filtered. The port is marked filtered if is answered with ICMP unreachable error (type 3, code 1, 2, 3, 9, 10 or 13);

2) "Land" attack. Attacker sends to victim system a specially crafted TCP packet with the SYN flag set, in which the IP addresses, as well as the source and destination ports, are the same. As a result, the destination computer attempts to establish TCP session with itself. If successful, this attack can lead to loop of some TCP implementations and, as a consequence, disruption of computer operation. This attack was simulated by forming the input stream of TCP packets as described above;

3) "PUKE" attack. Attacker sends an ICMP unreachable error packet to the attacked host (unknown error of remote system), which causes the host to be disconnected from the server. In this case, the attacker replaces the source IP address with the IP address of the server. This attack was simulated by forming the input ICMP packet with field values as described above;

4) "UDP Bomb" attack. The attacker sends an incorrectly formed UDP packet to the attacking host (the source port value is exceeded). This attack was simulated by forming the input UDP packet with field value of the source port was unacceptable.

As an example screenshot of the program is presented.

The first example corresponds to the parameters: number of clusters — 2, the initial choice of centroids — select the first k points of set, the measure of similarity is Euclidean distance.

Next the collection of network packets that are transmitted over the network and packets for the simulated network attacks was launched.

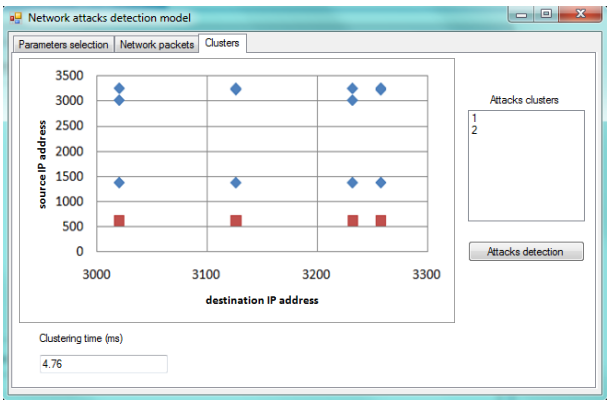

Fig. 4. Clustering results for 2 clusters

After collecting 10000 network packets corresponding to normal network operation and 200 packages, which correspond to the simulated network attacks, data collection was stopped and clustering was started. When you press the button Results display two formed clusters was displayed (Fig. 4). Cluster analysis showed that cluster 1 (blue diamonds) corresponds to the normal behavior of the system, while cluster 2 (red squares) — the attacker acts on the system, because the value of the parameter destination IP address is non-standard for network's IP addresses.

Next it was specified that cluster 2 is the cluster containing attacks and intrusion detection was launched. Collected packets are distributed into clusters. In case package is distributed to the 2nd cluster the message of possible attack appears.

The second example (Fig. 5) corresponds to the parameters: number of clusters — 3, the initial choice of centroids is select k points randomly, the measure of similarity is the Manhattan distance.

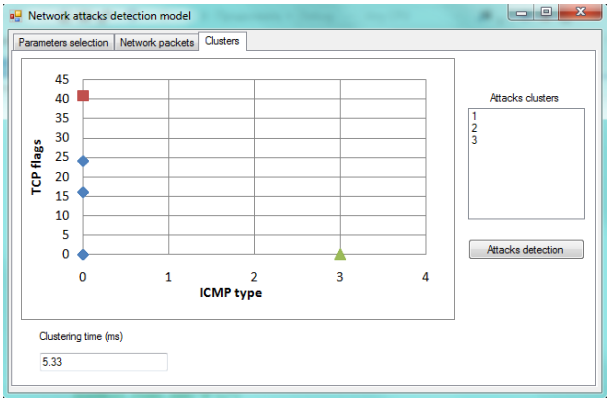

Fig. 5. Clustering results for 3 clusters

Of the three clusters: the 1st cluster (blue diamonds) contains normal network packets; the 2nd cluster (red squares) contains packets, TCP flags of which is not typical for network and in this case are evidence of TCP scanning; the 3rd cluster (green triangles) contains the packets of the ICMP, which ICMP type is not typical for network and in this case

indicates “PUKE” attack. The 2nd and the 3rd clusters can be considered as clusters of attacks for this experiment.

The third example (Fig. 6) corresponds to the parameters: number of clusters — 4, the initial choice of centroids is select k points randomly, the measure of similarity is Chebyshev distance.

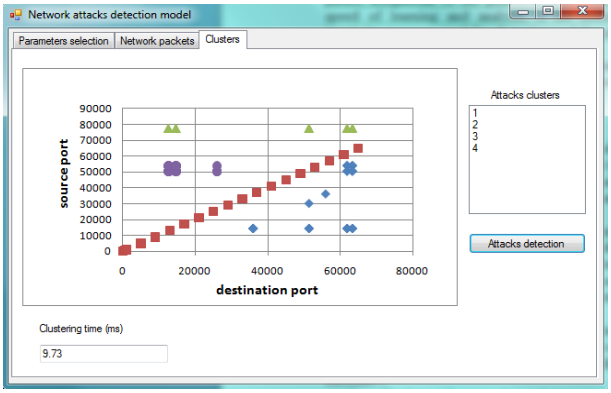

Fig. 6. Clustering results for 4 clusters

Of the four formed clusters: the 1st cluster (blue diamond) and the 4th cluster (lilac polygons) contain normal packets for the network; the 2nd cluster (red squares) contains packages for which the source port and the destination port are the same, which is one of the signs of “Land” attack; the 3rd cluster (green diamonds) contains packets for which the value of source port is exceeded, which may indicate “UDP Bomb” attack.

For each performed experiment errors of the first and second kind were counted, as well as the time in which clustering was performed. The obtained results were used to calculate the efficiency of operation of network traffic clustering software package for intrusion detection.

The best combination of parameters of k-means method is evaluated by the effectiveness of intrusion detection — ability to detect network attacks is correlated with spending resources.

Then, the efficiency index is calculated according to the formula:

$$E = \frac{A}{R}, \tag{1}$$

where E is the efficiency index; A is potential effect; R is resource consumption.

As potential effect is considered, the proportion of correctly clustered network packets, which is calculated by the formula:

$$A = \frac{N - Q_1 - Q_2}{N} \tag{2}$$

where N is the total number of analyzed packets; Q_1 is number of mistakenly detected attacks – false positive; Q_2 is number of missed attacks – false negative;

As resource consumption is accepted the time required to conduct the clustering.

VI. Conclusion

An analysis of the results of experimental studies has shown that the best combination of parameters for which the efficiency E = 58.29238 is combination of:

- number of clusters is 4;
- method of initial selection of centroids - select k points randomly;
- metric used is Chebyshev distance;

since it has the smallest number of false positive and false negative errors among all combinations: 17 false positive errors and 9 false negative errors from 10200 analyzed network packets.

Reference

[1] M.A. Jabbar, K. Srinivas, S. Sai Satyanarayana Reddy, “A novel intelligent ensemble classifier for network intrusion detection system” in 8th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2016); Vellore; India; 2016, pp. 490-497.

[2] N. Kumar, U. Kumar, “Anomaly-based network intrusion detection: An outlier detection techniques” in 8th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2016); Vellore; India; 2016, pp. 262-269.

[3] X. Zhao, W. Zhang, “Hybrid Intrusion Detection Method Based on Improved Bisecting K-Means in Cloud Computing” in 13th Web Information Systems and Applications Conference (WISA 2016), Wuhan, Hubei; China, 2016, pp. 225-230.

[4] M. Wurzenberger, F. Skopik, M. Landauer, P. Greitbauer, R. Fiedler, W. Kastner, “Incremental clustering for semi-supervised anomaly detection applied on log data” in 12th International Conference on Availability, Reliability and Security (ARES 2017), Reggio Calabria; Italy, 2017, pp. 1-6.

[5] W. Bhaya, M. Ebadymanaa, “DDoS attack detection approach using an efficient cluster analysis in large data scale” in 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), Baghdad; Iraq, 2017, pp. 168-173.