Searching for optimal machine learning algorithm for network traffic classification in intrusion detection system

Alicja Gerka[1,*]

[1]Rzeszow University of Technology, The Faculty of Electrical and Computer Engineering, ul. Wincentego Pola 2 35-959 Rzeszów, Poland

Abstract. The main problem associated with the development of an effective network behaviour anomaly detection-based IDS model is the selection of the optimal network traffic classification method. This article presents the results of simulation research on the effectiveness of the use of machine learning algorithms in the network attacks detection. The research part of the work concerned finding the optimal method of network packets classification possible to implement in the intrusion detection system's attack detection module. During the research, the performance of three machine learning algorithms (Artificial Neural Network, Support Vector Machine and Naïve Bayes Classifier) has been compared using a dataset from the KDD Cup competition. Attention was also paid to the relationship between the values of algorithm parameters and their effectiveness. The work also contains an short analysis of the state of cybersecurity in Poland.

1 Introduction

Technological progress and increasing computerization of the society means that along with the development of technologies commonly used in computer networks and information systems, cyber attack and data breach techniques are also developed. Cybercriminals find new vulnerabilities and develop new techniques for their exploitation. According to a report published in 2017 by Panda Security, on average, about 280,000 new malware samples are produced every day [1].

Therefore, the protection of communication and information systems is becoming a growing challenge and generates ever-growing costs. According to estimates by Cybersecurity Ventures, the global IT security services and products spending will exceed $1 trillion over the next 5 years, from 2017, and the amount of losses related to the cybercriminal activities will increase from $ 3 trillion in 2015 to $ 6 trillion in 2021 [2].

In view of these facts, it is necessary to search for new, computationally efficient methods of detection of attacks in intrusion detection systems. One of the solutions to this problem is the use of unconventional methods such as machine learning algorithms.

[*] Corresponding author: alicja.gerka@op.pl

2 The state of cyber security in Poland

According to cybersecurity reports published by the Republic of Poland Governmental Computer Security Incident Response Team CERT.GOV.PL over the years 2014-2018, the number of reports of breach of security is constantly increasing as shown in Figure 1 [3-7].

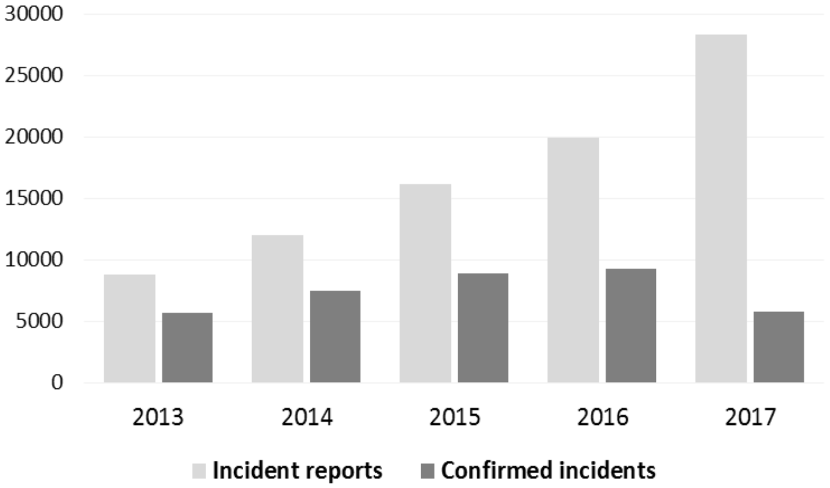

Fig. 1. The number of reports of potential security incidents and actual security incidents in 2013-2017 [3-7].

In 2017, an overall decrease in the number of cases of actual IT security breaches could be observed, including a significant decrease in the number of security incidents caused by incorrect configuration of devices. That tendency may be associated with an increase of public cyber security awareness. It is important that despite the use of newer methods of electronic systems protection and the increase of cyber threats awareness - the number of incidents caused by malware, as can be seen in Figure 2, increases each year, and in 2017 increased by over 70% in comparison to the previous year [3].

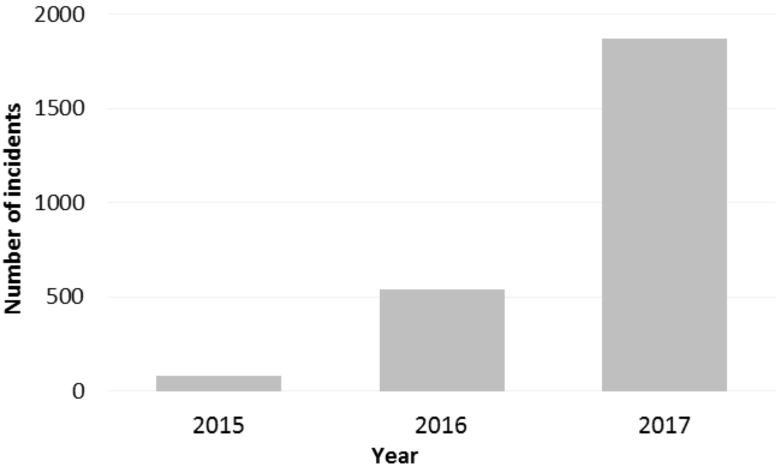

Fig. 2. Increase of number of security incidents caused by malware [3].

In turn, in 2018, cryptojacking became one of the most important global security challenges. This is a relatively new form of attack consisting of the installation of coin miner malware on the victim's device without his consent and awareness [8].

According to the McAfee Labs report, in the first quarter of 2018, the amount of coin miner malware increased by 629% (as shown in Figure 3) to more than 2.9 million known samples from nearly 400,000 samples in the fourth quarter of 2017. This suggests that cybercriminals have changed their tactics in relation to last year, when during global attacks using ransomware, they forced victims to make payments and are now more likely to choose less risky and simpler solutions that allow them to earn through the proliferation of malware. In this case, attackers can avoid brokers and the situations in which the victim refuses to transfer the ransom, as well as hide behind some of the misuse detection systems [8].

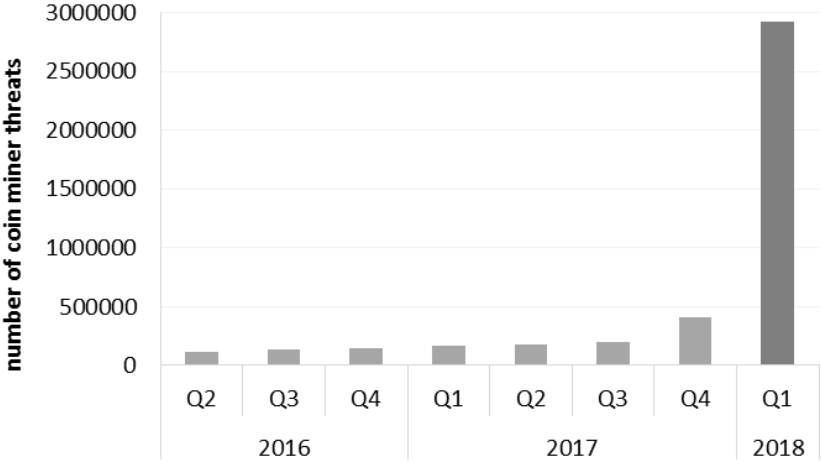

Fig. 3. Increase of samples of coin mining malware over the past two years [8].

3 Methods of Defense

The growing number of cyber threats and the development of techniques used by cybercriminals entail the necessity to develop new methods of information systems protection.

In accordance with the technical recommendations included in the report on the state of cybersecurity in Poland in 2017, in order to ensure the security of IT systems, it has to be implemented, inter alia, an intrusion detection and prevention system that will enable the detection of attacks using them network behavior anomaly detection [3].

Currently, the device protection strategy based on signatures is not sufficient to provide a basic level of security. Researchers and enterprises from the IT industry have long been aware of this, that's why they have been looking for new, more effective detection methods that will be able to detect attacks not only on the basis of system behavior, but also of the entire network. Therefore, anomaly detection methods based on the analysis of network behavior are increasingly used in Intrusion Detection Systems [9].

Systems that use network behavior anomaly detection techniques are particularly useful in situations where the analyzed network traffic is encrypted, and also enable the detection of zero-day attacks that are not detectable by signature-based detection systems. That is why enterprises and institutions are increasingly investing in such solutions [9].

Currently, machine learning methods are increasingly used in behavioral anomaly detection. Within this paper there has been examined the effectiveness of three o the most used supervised learning algorithms - ANN - artificial neural network, NBC - naive Bayes classifier and SVM - support vector machine. The chosen methods operate in different ways and they are characterized by a different computational complexity [10].

4 Research

The main goal of the research was to find the optimal method of attacks detection for IDS dedicated to small networks, which can work on host device. The research compares the effectiveness of attack detection using three different methods of network packets classification. The implementation of the goal consisted of several steps. First, three machine learning algorithms were implemented in Python language, then the algorithms' parameters were selected based on the conducted simulations, and finally, the accuracy and training time of the implemented algorithms were compared.

The work uses commonly used measures of evaluation of network traffic classifiers: length of training time, accuracy, F-measure, false negatives and false positives [11].

The study used a well-known dataset KDD Cup 1999 developed by MIT Lincoln Labs. According to a study conducted in 2015, KDD Cup 1999 is the most commonly used dataset in research related to the use of machine learning methods in intrusion detection systems. This dataset contains a standard set of audit data, which covers a wide range of intrusions simulated in the network environment. This set contains 22 types of attacks, whereby in this study, the classification model was binarized and all types of attacks were assigned to one "attack" class [12,13].

4.1 Stage I – searching for optimal values of algorithms' parameters

Before conducting a comparative analysis of the operation of the created classification models, it was necessary to select the optimal values of classifiers parameters. This is about: a regularization parameter C and the Gaussian function parameter gamma for the SVM based classifier and number of learning epochs and the batch size of neural network classifier.

During the test, the accuracy of the attack detection and the classifier learning time were chosen as quality criteria. The study consisted in teaching the classifier on a subset of 70% of the loaded dataset and then predicting class membership of samples from the rest of the set (constituting a classifier's test set).

For each configuration of parameters, 50 trials were performed, which in total gives over 5,000 simulations for both algorithms.

Classifier learning time results were rounded to two decimal places, and the results on the accuracy of the classification to five decimal places. The classifier's time of teaching was expressed in seconds, while the unit of accuracy of classification was a number between <0,1>, where 1 means 100% correctness of classification.

4.1.1 SVM

In order to maintain or increase the level of accuracy of classification, increasing the value of parameter C, the value of the gamma parameter should be simultaneously reduced, in example: increasing the fit of the classification model, is important to simultaneously increase the flexibility of the decision boundaries of the SVM model. In turn, Figure 4 presents the relationship between the classifier's training length and the value of parameter C.

For the discussed set, the classifier training time was the shortest for gamma value equal to 0.001 and C parameter value equal to 10,000.

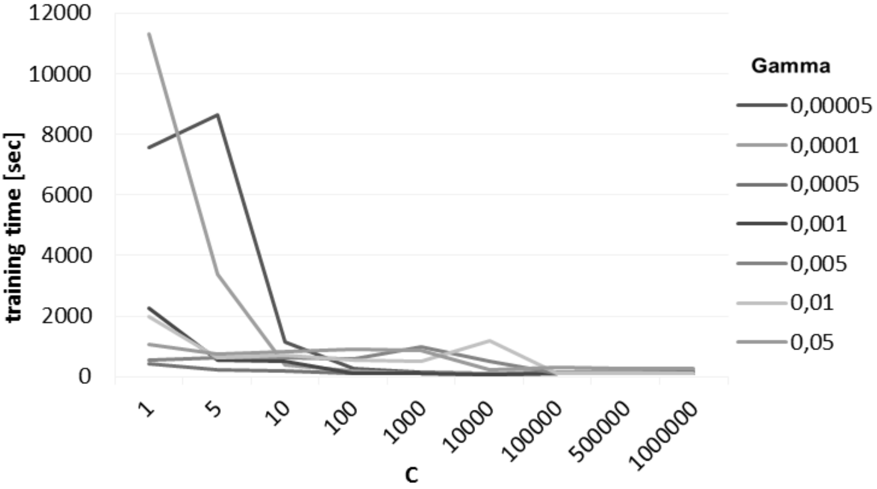

Fig. 4. The relationship between SVM classifier training time and value of C parameter for different gamma values.

As can be seen in the figure above, the change in the value of the gamma parameter has an increasingly lower influence on the length of training of the classifier along with the increase in the value of parameter C. The next figure (Figure 5) presents a graph of the relationship between the accuracy of classification and the value of parameter C.

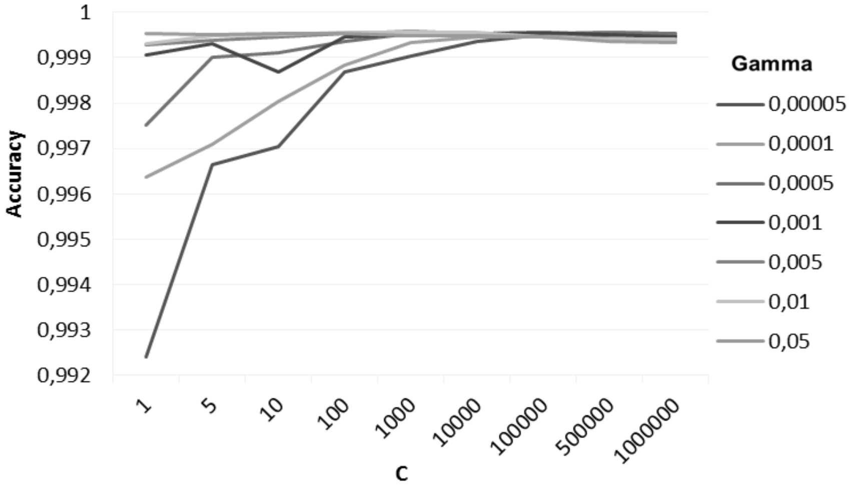

Fig. 5. The relationship between classification accuracy and C parameter value for different SVM classifier's gamma values.

At first glance, it may seem that the level of accuracy stabilizes after taking the appropriately high value of parameter C, but by zooming the graph (as shown in Figure 6), it can be observed that after reaching a certain maximum level, the accuracy of classification begins to decrease as the value of parameter C increases for each of the selected values of the gamma parameter.

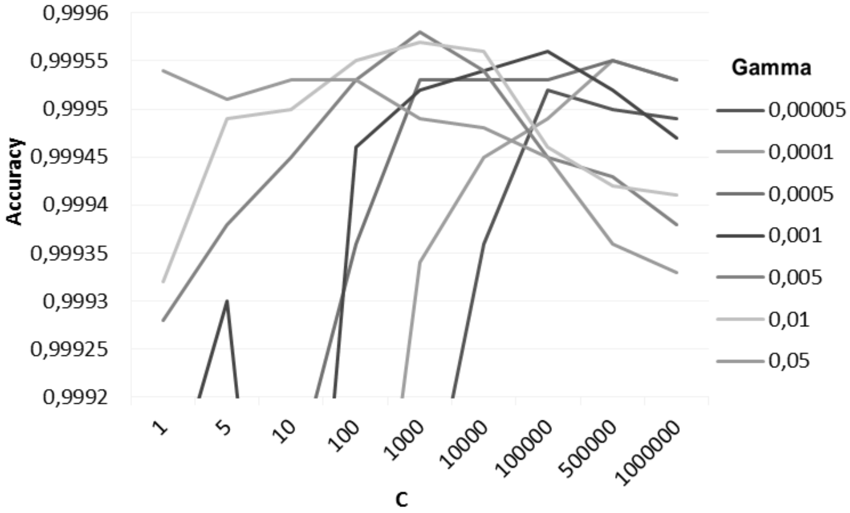

Fig. 6. The zoom of relationship between classification accuracy and C parameter value for different SVM classifier's gamma values.

Considering the dependences presented in Figures 4-6, it can be assumed that the selection of appropriate, optimal values of gamma and C coefficients consists in finding such values of these parameters for which the accuracy is as high as possible with the classifier training as short as possible.

In this case among the configurations with the shortest learning time, the highest accuracy value was obtained for the parameter gamma equal to 0.001 and C parameter equal to 100,000, therefore these parameter values were considered optimal.

4.1.2 ANN

Figure 7 presents a graph of relationship between training time and batch size of neural network for four different numbers of learning epochs. In turn, Figure 8 presents a graph of the relationship between the neural network batch size and the level of accuracy of classification.

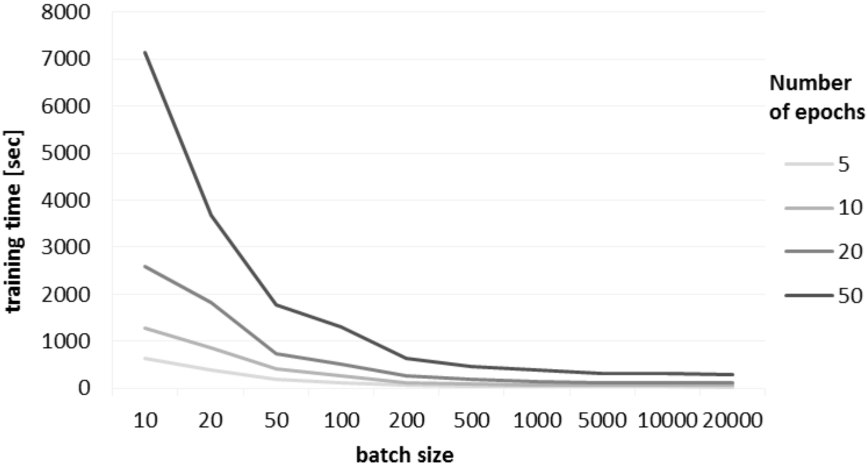

Fig. 7. The relationship between classifier's training time and neural network's batch size for different number of neural network lerning epochs.

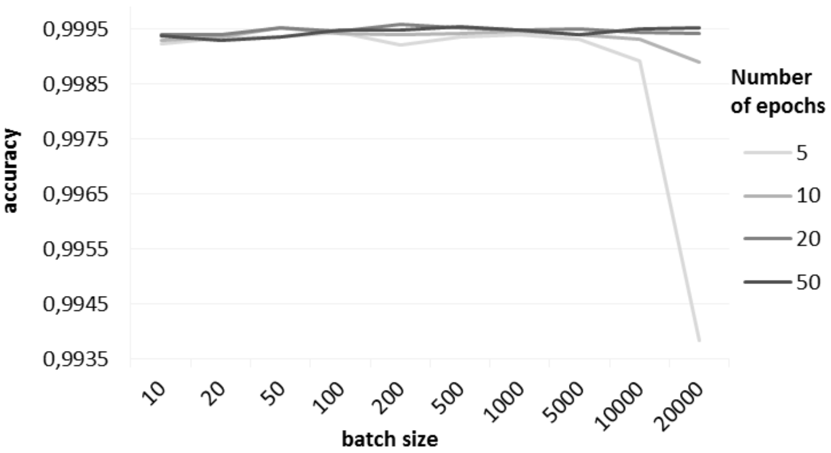

Fig. 8. The relationship between classification accuracy and artificial neural network's batch size for different numbers of learning epochs.

The worst results, both for learning time and accuracy, were obtained in the case when the network was taught only for several epochs with a large batch size. Unsatisfactory results were also obtained in the case when the network was taught for many epochs with a small batch size. In the first situation the classification model is underfitted, in turn in the second case, the classification model is overfitted.

This stage of research shows the relationship between the accuracy and the training time of the classification and the values of the parameters of both implemented algorithms.

Based on the results of the experiment, the following parameter values were adopted for further research:

a) for the neural network: number of epochs = 20 and batch size = 5,000,

b) for a classifier operating on the basis of SVM: C parameter = 100,000 and gamma parameter = 0.001.

4.2 Stage II – comparison

The second stage of the study was to compare the operation of three network attacks detecting methods. Each of them was based on a different network traffic classification algorithm and used another machine learning method.

100 trials were conducted for each algorithm. The average simulation results are shown in Table 1. The results were rounded to five decimal places.

The evaluation criteria while comparing the selected methods were: classification accuracy, classifier training time, F-measure and the error rates (false positives and false negatives).

Table 1. Comparison of three network attacks detection algorithms.

	ANN	**NBC**	**SVM**
Training time [sec]	125,72235	0,94814	95,59625
Accuracy	0,99947	0,94708	0,99956
False Positives [%]	0,0336%	6,5709%	0,0227%
False Negatives [%]	0,0318%	0,0228%	0,0319%
F-measure	0,998664	0,881651	0,998885

The obtained results confirm that the naive Bayesian classifier, due to the low computational complexity and simple structure, is characterized by a much shorter (more than hundred times) training time compared to other methods. Training the classifier based on the Bayes' theorem is very short but this is being done at the expense of the accuracy of the prediction and the increase in the number of false positive errors.

An interesting phenomenon is the fact that despite the general low accuracy of classification, this method achieved the lowest false negatives error rate in the study.

The study shows that the most accurate of the selected methods of attacks detection is the network traffic classification based on SVM. The advantage of detecting attacks using SVM over the use of ANN for this purpose has already been described in various articles, so the obtained results are in line with the results of other authors [14].

In addition, SVM-based detection method guarantees the best balance between the accuracy and sensitivity of the classification, as evidenced by the highest value of measure F. Although the SVM-based classifier achieved the worst result in terms of the false negatives rate, the result was only 0.0001% worse than in the case of classification using an artificial neural network, while the neural network being characterized by the longest training time of all discussed classification methods.

The presented results prove that even with a small computational effort, using machine learning methods, you can detect attacks in data collected as a result of network traffic sniffing. The low computational complexity of the presented classification methods (achieved by optimizing the values of parameters of the algorithms), means that they can be successfully used on the host devices.

5 Conclusion

A key element of each IDS is the threat detection subsystem. Finding an effective method of classifying network traffic is not a simple task. The analyzes show that attack detection modules based on binary classification using machine learning methods ensure effective detection of attacks with a small percentage of false alarms.

Considering the results of the conducted research, the author proposes to use the SVM algorithm as a method of classifying network packets, because this algorithm allows the most accurate detection of attacks with relatively low class time of the classifier. The solution gives optimal results in terms of the classifier's learning time as well as the accuracy of the detection. The advantages of using SVM in intrusion detection systems have also been presented in many other articles [15,16].

Due to the assumption that the system should work on the host device and be dedicated to small networks, this solution seems to be sufficient. However, for larger networks and networks with high security requirements, it is worth considering using several independently operating classifiers.

The conducted research is a contribution to the understanding of the relationship between the implementation of machine learning algorithms and the effectiveness of network attacks detection in behavioral based intrusion detection and intrusion prevention systems.

When analyzing the results of the conducted research, one should take into account the fact that the dataset used to test the implemented prediction models are not the perfect representation of network traffic. However, due to the lack of publicly available IDS learning data sets, these sets can be still good help in comparing different intrusion detection methods.

References

1. Panda Security, *PandaLabs' Annual Report 2017* (2017)
2. S. Morgan, Cybersecurity Ventures, *2017 Cybercrime Report* (2017)
3. Republic of Poland Governmental Computer Security Incident Response Team CERT.GOV.PL, *Report on the security status of the cyberspace of the Republic of Poland in 2017* (2018)
4. Republic of Poland Governmental Computer Security Incident Response Team CERT.GOV.PL, *Report on the security status of the cyberspace of the Republic of Poland in 2016* (2017)
5. Republic of Poland Governmental Computer Security Incident Response Team CERT.GOV.PL, *Report on the security status of the cyberspace of the Republic of Poland in 2015* (2016)
6. Republic of Poland Governmental Computer Security Incident Response Team CERT.GOV.PL, *Report on the security status of the cyberspace of the Republic of Poland in 2014* (2015)
7. Republic of Poland Governmental Computer Security Incident Response Team CERT.GOV.PL, *Report on the security status of the cyberspace of the Republic of Poland in 2013* (2014)
8. C. Beek, T. Dunton et al., McAfee, *McAfee Labs Threats Report June 2018* (2018)
9. A. Szmit, M. Szmit, *On the use of econometric models for forecasting network traffic.*, Scientific Notebooks Organization and Management of the Lodz University of Technology **55**, 1154, p.193-201 (2013)
10. J. Rusnacko, *Self-optimizing traffic classification framework*, Faculty of Informatics Masaryk University (2013)
11. D. M. S. Zekrifa, *Hybrid Intrusion Detection System*. Theses, School of Information Technology & Mathematical Sciences (2014)
12. N.F. Haq, M. Rafni, A.R. Onik et al., *Application of Machine Learning Approaches in Intrusion Detection System: A Survey*. International Journal of Advanced Research in Artificial Intelligence, **4**, 3, p. 9-18 (2015)
13. KDD Cup 1999, Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
14. S. Mukkamala, G. Janoski, A.H. Sung, *Intrusion Detection Using Neural Networks and Support Vector Machines*. IEEE International Joint Conference on Neural Networks 2002, IEEE Computer Society Press, p. 1702-1707 (2002)
15. J. Hussain, S. Lalmuanawma, L. Chhakchhuak, *A two-stage hybrid classification technique for network intrusion detection system*, International Journal of Computational Intelligence Systems, **9**, 5, p. 863-875 (2016)
16. Chen, W. H., S. H. Hsu, H. P. Shen, *Application of SVM and ANN for intrusion detection*, Computers & Operations Research, **32**, 10, p. 2617–2634 (2005)