UDC 004

Dziadek M.I. student of group 125м-24-1
Research supervisor: Olishevskyi I.H., Associate Professor of the department of information security and telecommunications
(*Dnipro University of technology, Dnipro, Ukraine*)

ANALYSIS OF THE RESULTS OF APPLICATION OF MACHINE LEARNING ALGORITHMS IN NETWORK TRAFFIC ANOMALIES DETECTION SYSTEMS

As of today, the detection of network traffic anomalies is an urgent task of ensuring the information security of any enterprise. The value of the information that is processed and stored in the ICS of enterprises increases, thereby increasing the motivation of criminals to obtain NSD for this information. Therefore, there is a need to implement a solution that will minimize the risk of attacks on information via the Internet. A modern and effective solution for an enterprise against threats of this type is the implementation of network traffic monitoring based on a machine learning algorithm[1-3].

There are many machine learning algorithms, so the information security specialist is faced with the task of deciding on an algorithm that will demonstrate the highest results of correct response to threats. As part of the study, a comparative analysis of the results of the application of seven machine learning algorithms in SBA - Naive Bayes, QDA, Random Forest, Decision Trees, AdaBoost, MLP, KNN was performed.

Training and testing of the SBA model was performed on the basis of the CICIDS-2017 dataset. The use of this sample in the research is due to the presence of records of modern attacks such as Brute Force, DoS, HeartBleed, WebAttack, Infiltration, Botnet, DDoS, which makes it more relevant to the current research.

To train the model, the Python programming language and the libraries available in it were used, such as Sklearn - a library for implementing machine learning; Pandas is a library for data manipulation and analysis; Matplotlib - a library for visualizing data in the form of two-dimensional graphics; NumPy is a library for performing operations on multidimensional arrays of data[4-5].

The evaluation of the effectiveness of the application of machine learning algorithms in the framework of the study was carried out according to four indicators - Accuracy (an indicator of the ratio of the number of successfully classified attacks to the total amount of data); Recall (an indicator of the ratio of data classified as an attack to all attack data); Precision (indicator of the ratio of the number of successfully classified attacks to all classified records); F-measure (indicator of the harmonic mean value of sensitivity and accuracy). It was the F-measure that was used as a result of the overall success of the algorithm. The evaluation also includes an indicator of the time required for training the model, but this indicator is subjective and directly proportional to the technical characteristics of the PC on which the model is trained.

Two approaches were used for modeling, in the first one, the four most important features for each of the attacks were used. This approach aims to investigate the effectiveness of applying a certain machine learning algorithm to certain types of attacks and its performance, which it demonstrates during simulation. The second approach is based on applying the algorithm to the features that occur most often in the set of records of all attacks. The results of the application of machine learning algorithms according to different approaches will be compared and analyzed. In this way, two combinations of features were formed for training the system - a combination of eighteen features and a combination of seven features. The aforementioned combinations are shown in Tables 1 and 2.

Table 1 – Combination of eighteen features for attack types

Bwd Packet Length Max F	Flow IAT Mean	Fwd Packet Length Min
Bwd Packet Length Mean	Flow IAT Min	Fwd Packet Length Std
Bwd Packet Length Std F	Flow IAT Std	Total Backward Packets
Flow Bytes/s	Fwd IAT Total	Total Fwd Packets
Flow Duration	Fwd Packet Length Max	Total Length of Bwd Packets
Flow IAT Max	Fwd Packet Length Mean	Total Length of Fwd Packets

Table 2 – Combination of seven features for attack types

The name of the feature	Importance indicator	Frequency of appearance of the symptom, %
Bwd Packet Length Std	0,246620	38,9%
Flow Bytes/s	0,178786	28,27%
Total Length of Fwd Packets	0,102427	16,19%
Fwd Packet Length Std	0.063894	10,11%
Flow IAT Std	0,009896	1,55%
Flow IAT Min	0,006940	1,09%
Fwd IAT Total	0,005117	0,7%

After determining the combinations of features for the implementation of machine learning, you can directly start training the model. A program code written in Python was applied and results were obtained. The values of the above criteria for evaluating the application of machine learning algorithms for eighteen and seven features are shown in Tables 3 and 4, respectively.

Table 3 – Results of application of machine learning algorithms for eighteen features

The name of the machine learning algorithm	The value of the efficiency evaluation criterion				
	F-measure	Precision	Recall	Accuracy	Час, хв
Naive Bayes	0,63	0,63	0,64	0,78	3,2
QDA	0,31	0,58	0,58	0,31	4,7
Random Forest	0,88	0,96	0,83	0,94	22
Decision Trees	0,90	0,97	0,86	0,95	29,66
AdaBoost	0,91	0,95	0,87	0,95	378,2
MLP	0,53	0,75	0,54	0,84	188,8
KNN	0,95	0,94	0,95	0,97	2088,6

Table 4 – Results of application of machine learning algorithms for seven features

The name of the machine learning algorithm	The value of the efficiency evaluation criterion				
	F-measure	Precision	Recall	Accuracy	Час, хв
Naive Bayes	0,65	0,66	0,64	0,82	1,9
QDA	0,38	0,58	0,61	0,38	2,3
Random Forest	0,88	0,96	0,83	0,94	20,37
Decision Trees	0,91	0,93	0,89	0,95	12,56
AdaBoost	0,88	0,93	0,85	0,94	168,51
MLP	0,53	0,75	0,54	0,84	164,54
KNN	0,94	0,94	0,95	0,97	214,41

CONCLUSION

The models were trained on the basis of the above-mentioned two approaches. From the data shown in Tables 3 and 4, according to the main criterion for the success of model training

By F-measure, two best algorithms can be distinguished - Decision Trees and KNN. But, according to the above data, in addition to the four evaluation criteria, an evaluation of the time required to train the model was also performed. The KNN algorithm performs very well on the four algorithm evaluation criteria, but requires significantly more time and computational resources to train the model.

So, after analyzing all the obtained values, it can be concluded that the Decision Trees algorithm shows very good results according to the metrics for both eighteen and seven features, and also requires an adequate expenditure of time and, accordingly, computing resources, which are necessary for training the model. Based on the results of the research, the use of the Decision Trees algorithm in network traffic anomaly detection systems is the most optimal and balanced solution, which is based on its high performance indicators.

REFERENCE

1. OLISHEVSKYI I.H. Substantiation of energy efficiency of automated heating technology at HPS / OLISHEVSKYI I.H., // Електротехніка та електроенергетика. / Запорізький нац. ун-т «Запорізька політехніка». – Запоріжжя, 2024. – № 2. – С. 36-43 https://doi.org/10.15588/1607-6761-2024-2-4

2. Khabarlak, K. S. (2022). FASTER OPTIMIZATION-BASED META-LEARNING ADAPTATION PHASE. Radio Electronics, Computer Science, Control, (1), 82. https://doi.org/10.15588/1607-3274-2022-1-10

3. K. Khabarlak, "Post-Train Adaptive U-Net for Image Segmentation," Information Technology: Computer Science, Software Engineering and Cyber Security, no. 2, pp. 73--78, 2022, https://doi.org/10.32782/IT/2022-2-8

4. Lewis, T. G., & Denning, P. J. (2018). Learning machine learning. Communications of the ACM, 61(12), 24–27. https://doi.org/10.1145/3286868

5. Siddique, S. (2020). Machine Learning and Cryptography. Journal of Advanced Research in Dynamical and Control Systems, 12(SP7), 2540–2545. https://doi.org/10.5373/jardcs/v12sp7/20202387