Research of multiclass fuzzy classification of traffic for attacks identification in the networks

D Parfenov, L Zabrodina, A Zhigalov and I Bolodurina

Department of Applied Mathematics, Orenburg State University, Orenburg, 460018, 13 Prospekt Pobedy, Russia

E-mail: parfenovdi@mail.ru

Abstract. Currently, data mining methods such as neural networks, decision trees, genetic algorithms, restricted search algorithms, evolutionary programming, reasoning systems based on similar cases, rule induction, analysis with selective action, logical regression, algorithms for determining associations and sequences, data visualization, combined methods are actively used in various specialized areas. The introduction of analytical methods of data mining is primarily aimed at adapting existing solutions to solve new problems related to the informatization of business processes. One of the actively developing areas that use data mining and artificial intelligence methods is network security. To identify and detect anomalies in networks, it is most effective to create profiles of data flow behavior depending on current conditions. In this work, we developed a method that allows us to identify Exploits, Fuzzers, and Generic attacks based on multiclass fuzzy classification. In the experimental part of the study, it was found that the proposed solution is comparable in accuracy to the Naïve Bayes, SVM, and KNN methods, but it has higher performance and less resource consumption for large data flows. This is quite effective for networks with many devices.

1. Introduction
Currently, the issues of security of global data transmission networks, as well as ensuring the required level of confidentiality and integrity of data during their transmission through communication channels, have more relevant. At the same time, existing intrusion detection systems have to analyze network traffic under conditions of uncertainty [1, 2], which are formed due to the reduced quality of the source data. Note that the quality of data is determined by objective indicators of uncertainty [3] (omitted values, anomalies, noise, outliers, etc.) and linguistic uncertainty [4], which occurs due to the subjective assessment of an expert or a group of experts. At the same time, information about the studied objects and their States can be different and can be expressed in both quantitative and qualitative features. In this case, it is not possible to use any machine learning algorithms for classification, since the source data must be pre-processed and represented by a representative sample.

To exclude objective uncertainty in the training data, various algorithms are used for processing and filtering outliers, anomalies, noise, and so on. The most difficult stage of preprocessing is the analysis and reduction of linguistic uncertainty. In this area, algorithms for constructing fuzzy inference systems have shown good efficiency [5], which allows replacing the subjectivity of expert evaluation to some extent. At the same time, the use of neuro-fuzzy classification systems allows you to simultaneously build an intelligent fuzzy model of logical inference mechanisms, fully automating the process of classifying cybersecurity incidents. Subjectivity in the results of the neural grid system remains only at
the stage of constructing the corresponding linguistic variables and their accessory functions while increasing the accuracy of the constructed models [6]. The great importance of the stage of forming the base of fuzzy rules is based on the need not only to select realistic term sets for lexical variables but also to construct adequate accessory functions. There are many studies in this area that describe the capabilities of fuzzy classification algorithms. This work is aimed at developing a fuzzy inference system that allows you to classify abnormal network traffic and identify current attacks by type.

2. Related works

One of the directions that determine the modern development of information technologies for monitoring and diagnostics of the technical condition of local computer networks can be considered the intellectualization of monitoring information processing processes and their classification using the technology of expert systems. Research on methods and algorithms for constructing fuzzy inference systems for solving such problems is actively underway around the world.

The authors of [7] studied in detail a special case of the problem of managing a complex object in conditions of fuzzy initial information and developed a hybrid method for automatic classification of object States based on the fuzzy k-means clustering algorithm and the situational inference algorithm.

An approach based on a linear combination of fuzzy membership functions in fuzzy inference systems is demonstrated in [8]. This method allowed converting not only numeric but also categorical attributes into logical rules.

The study [9] presents a comparative analysis of single-element fuzzy inference systems and proves that such systems can approximate any given function that is continuous on a compact set. This approach allowed us to implement models with sufficient accuracy on noisy data.

To identify a limited set of network attacks Smurf, Mail-Bomb and Ping-of-Death in [10], a fuzzy decision tree algorithm was constructed. As a result of monitoring real data, the authors confirmed the effectiveness of the method by increasing the speed of identification and increasing the overall performance of the system.

In the article [11], the authors presented an approach based on comparing each type of threat with a characteristic vector of fuzzy values. The efficiency and effectiveness of this approach was assessed by calculating the integral indicator of the search for threats that have passed within the network based on a fuzzy rule base.

Thus, a review of existing methods, algorithms and odd inference systems for analyzing network traffic under conditions of uncertainty showed that modern systems do not allow considering all relevant types of attacks, and also leave room for modification and improvement of the accuracy of identification results, since they depend on expert judgment.

3. Mathematical formulation of the neural mesh classification problem

Consider a traditional network of telecommunications service providers that allows end users to access information and Analytics platforms. It is necessary to detect and classify malicious fragments of continuous network traffic. In other words, we will consider the problem of network security as a multi-class classification of network traffic for detecting network attacks.

For the pilot study will use a dataset UNSW-NB15, which contains data about the normal traffic and data 9 types of actual attacks: Worms, DoS, Analysis, Fuzzers, Shellcode, Generic, Reconnaissance, Exploits, and Backdoor [12].

Let network traffic data is acquired continuously and recorded as a set of records \( \{r_1, r_2, ..., r_m\} \), where each record \( r_i = \{r_{i1}, r_{i2}, ..., r_{ik}\} \) is a set remove from the device network characteristics: bandwidth, number of lost packets, access violation, etc.

Let us describe each object of the network X by a subset of the parameters \( x \in X \) and traffic characteristics \( r = \{r_1, r_2, ..., r_a\} \) fixed on it with the corresponding labels of the classes of various types of attacks, including the normal behavior of network devices. Note that an attack in the network can spread to several devices, however, within the framework of this approach, objects are analyzed in
parallel and the identification of attacking influences is performed as many times as there are devices in the network. Also, we will assume that each object for some given characteristics is associated with exactly one class from the set $Y$.

Let's assume that there is an objective function $f: X \rightarrow Y$ whose values are known on a finite set of objects $x_1, x_2, \ldots, x_n \in X$ and associating each object with a class label from the set $Y = \{1, \ldots, K\}$. Let's consider some initial data divided into test and training sets, which are a set of pairs $(x_j, y_j)$, $j = 1, n$, for which you need to build a map such that $y_j = f(x_j)$.

The task of training with a teacher is to build such a function $f(x)$ based on the training sample, which performs the mapping $X \rightarrow Y$, not only on the objects of the training sample, but also on the entire set $X$.

In this regard, it is necessary to build a multi-class classifier $f_c(R): R \rightarrow Y$ that matches each element of the set of network traffic records from some network objects $R = \{r_1, r_2, \ldots, r_m\}$ with label $y_l$, $l = 1, K$ that determines the type of attack.

Thus, we need to build a multi-class classifier $f_c(R): R \rightarrow Y$ that assigns a label to each element of the set of network node $R = \{r_1, r_2, \ldots, r_m\}$ matches a label $y_l$, $l = 1, K$ is a network traffic class. By class we mean the type of attack.

4. Data Pre-processing

Because the source data set under study has a large number of fixed characteristics of network objects over a fairly long period, it is more efficient to pre-process data to reduce the dimension of the feature space and increase the representativeness and informativeness of the sample.

Sequential data preprocessing should start with analyzing the impact of input features on the result – a class of attacks, because uninformative features reduce overall performance, but do not affect the accuracy of the obtained models. At the same time, the features presented in the data set should be able to be physically interpreted, since the constructed system of fuzzy inference as a result of the study should form fuzzy logical rules with specified lexical variables.

![Figure 1. The correlation matrix of input features with label field class.](image)

Correlation analysis of the data on the training set showed (figure 1) that out of the entire UNSW-NB15 feature space for constructing a label of an attacking impact class, the most significant features with a cutoff threshold of more than 0.2 are: $d_{\text{mean}}, ct_{\text{srv}_sirc}, ct_{\text{state}_ttl}, ct_{\text{dst}_\text{lim}}, ct_{\text{slimc}_d}, ct_{\text{src_sport}_\text{lim}}, ct_{\text{dst}_\text{src}_\text{lim}}, ct_{\text{src}_\text{class}_\text{lim}}, is_{\text{sm_ips_ports}}, dpkts, \text{sttl}, \text{sload}, \text{dload}, dloss, \text{sinpkt}, \text{swin}, \text{stepb}, \text{dstepb}, \text{dwin}$ [13].
5. Data analyzing
To build a fuzzy inference system for classifying anomalous traffic and identifying various types of network attacks, we define the following concepts.

The fuzzy variable is a tuple \((N, X, A)\), where \(N\) – this is the name of the variable; \(X\) – universal set (feature definition domain); \(A\) – given fuzzy set on \(X\).

The linguistic variable is a tuple \((N, T, X, A, G, P)\), where \(T\)– base term set; \(G\) – syntax rule (generation of new terms); \(P\) – semantic rule (mappings for \(A\) in some area \(X\)).

For the selected network traffic analysis features, we will form a base member of the set and define their linguistic variables to divide the specified network feature definition area. To do this, we will build the corresponding auxiliary functions in the FuzzyTech system based on the analysis of the mathematical expectation of the studied characteristics, as well as the values of their variance by attack classes. Examples of accessory functions are shown in figure 2.

![Figure 2. The examples of accessory function graphs for term-sets of network characteristics.](image)

To create a fuzzy logical system that allows you to classify attacks in network traffic, you need to build a set of fuzzy rules.

The most effective algorithms for constructing a fuzzy inference system are machine learning and data mining methods that allow us to evaluate the distribution of values of the studied characteristics and the relationship between different parameters.

In this work, fuzzy rules will be extracted from a decision tree based on the UNSW-NB15 training dataset (175341 unique records). Input features for building a decision tree are the characteristics of network traffic identified at the preprocessing stage. The output object is a field with the label of the attacking effect class. We used the C4.5 algorithm to build a decision tree.

The result of using the C4.5 algorithm to build a decision tree is shown in figure 3.

![Figure 3. The fragment of accessory function graph for term-sets of network characteristics.](image)

The next stage of data processing was extracting fuzzy rules from the constructed decision tree. As a result, 550 rules were obtained.

For example, the built fuzzy rules for exploits and universal attacks can be represented as the following rules:

\[
R_{\text{Exploits}} = \text{IF } \text{ct_src_sport_ltm} = \text{medium} \& \text{is_sm_ips_ports} = \text{low} \& \text{swin} = \text{low} \& \text{ct_src_dport_ltm} = \text{medium} \& \text{sload} = \text{medium} \& \text{sttl} = \text{low} \text{ THEN } \text{Exploits}
\]

\[
R_{\text{Generic}} = \text{IF } \text{ct_src_sport_ltm} = \text{medium} \& \text{is_sm_ips_ports} = \text{low} \& \text{swin} = \text{low} \& \text{ct_src_dport_ltm} = \text{medium} \& \text{sload} = \text{medium} \& \text{sttl} = \text{high} \& \text{sinpkt} = \text{low} \& \text{ct_dst_ltm} = \text{high} \& \text{sload} = \text{medium} \& \text{sttl} = \text{low} \text{ THEN } \text{Generic}
\]
Note that the cutoff threshold for nodes in the tree is set by the confidence level of the constructed fuzzy rule and is set to 20%. Fuzzy rules are constructed in such a way that they use all the studied linguistic variables of characteristics that describe network traffic. Also, each unique set of input fuzzy variable values has a single fuzzy rule associated with it.

6. Data analyzing
To conduct a computational experiment to identify attacking actions, the developed solution algorithm, and the proposed fuzzy inference technique were implemented as an independent module in Python.

The effectiveness of the built fuzzy inference system for determining the class of attacks was evaluated based on the analysis of network traffic on the UNSW-NB15 data set (more than 82 thousand unique records), and the classification results are presented in table 1.

Table 1. Results of attack identification on the test data set.

| Class          | Actually | Normal | Analysis | Backdoors | DoS  | Exploits | Fuzzers | Generic | Reconnaissance | Shellcode | Worms | Total |
|----------------|----------|--------|----------|-----------|------|----------|---------|---------|----------------|-----------|-------|-------|
| Normal         | 23581    | 401    | 5        | 136       | 3695 | 8227     | 66      | 888     | 1              | 37000     |       |       |
| Analysis       | 56       | 2      | 2        | 378       | 39   | 155      | 7       | 2       | 583            | 4089      |       |       |
| Backdoors      | 2        | 2      | 1        | 129       | 3380 | 170      | 232     | 156     | 5              | 677       |       |       |
| DoS            | 19       | 2      | 643      | 20        | 154  | 2        | 7       | 2       | 583            | 4089      |       |       |
| Exploits       | 70       | 14     | 6        | 110       | 1417 | 3500     | 319     | 154     | 3              | 11132     |       |       |
| Fuzzers        | 621      | 51     | 1417     | 3500      | 319  | 154      | 7       | 2       | 583            | 4089      |       |       |
| Generic        | 376      | 2      | 68       | 2869      | 386  | 15043    | 122     | 5       | 1              | 18871     |       |       |
| Reconnaissance | 26       | 13     | 1707     | 541       | 12   | 1196     | 1       | 1       | 3              | 3496      |       |       |
| Shellcode      | 2        | 3      | 143      | 73        | 157  | 7        | 8       | 7       | 44             | 378       |       |       |
| Worms          | 24       | 5      | 8        | 2972      | 1    | 16       | 1       | 5       | 1              | 82332     |       |       |
| Total          | 24751    | 417    | 16       | 514       | 23772| 13514    | 16359   | 2972    | 1              | 16        |       |       |

The results of the study showed that the fuzzy inference system allows you to identify the type of network attack with an accuracy of 64.58% (74.94% for the training set) and indicate the effectiveness of the proposed approach. The proposed approach most accurately defines Exploits, Fuzzers, and Generic attacks.

An experiment to evaluate the performance of the proposed solution was conducted on real network traffic using an experimental stand. The performance analysis was performed in a virtual network where 100 virtual machines were formed, connected, and having the roles of an attacking device, a connected device, or an analyzing device containing a developed module of the fuzzy output system for detecting attacking influences.

Initially, the experimental stand collected characteristics of legitimate traffic, which were combined in streams in the form of pcap files.

The next step was to simulate attacks such as Exploits, Fuzzers, and Generic using a customized Kali Linux distribution. Traffic received as part of the attacks was also collected as pcap files.

To check and compare the effectiveness of the proposed solution, we calculated the accuracy of detecting attacks using various machine learning methods: naive Bayes, SVM, and KNN. The comparison is based on the same data set, and the results are presented in table 2.
Table 2. Experimental results of attack identification.

| Attack / Method | AUC          |
|-----------------|--------------|
|                 | Multiclass Fuzzy Classification | Naïve Bayes | SVM | KNN |
| Normal          | 0.89         | 0.82       | 0.92 | 0.9 |
| Exploits        | 0.82         | 0.7        | 0.85 | 0.81 |
| Fuzzers         | 0.85         | 0.8        | 0.82 | 0.83 |
| Generic         | 0.9          | 0.84       | 0.9  | 0.89 |

7. Conclusion

As a result of this research, a fuzzy inference system has been developed that allows you to classify abnormal network traffic and identify current attacks by type. The effectiveness of the built system was confirmed by the example of determining network traffic attacks of the UNSW-NB15 data set, also, the system performance was evaluated on real network traffic using an experimental stand. The results showed that the developed fuzzy inference system can be used as an effective tool for detecting threats in networks with dynamic topology.

However, it is worth noting that the effectiveness of fuzzy modeling and control methods can be significantly improved if they are used together and in interaction with methods based on artificial neural networks and genetic algorithms. In future work, it is planned to investigate the use of the ANFIS neural network for forming fuzzy rules for identifying attacks according to the method proposed in this paper. Also, it is planned to evaluate the performance of a neural network solution based on graphics accelerators and create a low-level library for devices for software-defined networks.

Acknowledgment

The study was carried out with the financial support of the RFBR in the framework of scientific projects No. 20-07-01065, as well as a grant from the President of the Russian Federation for state support of leading scientific schools of the Russian Federation (NSh-2502.2020.9) and a grant from the President of the Russian Federation Federations for state support of young Russian scientists - candidates of sciences (MK-860.2019.9).

References

[1] Bolodurina I, Parfenov D, Torchin V and Legashev L 2018 Development and Investigation of Multi-Cloud Platform Network Security Algorithms Based on the Technology of Virtualization Network Functions 2018 International Scientific and Technical Conference Modern Computer Network Technologies (MoNeTeC) Moscow 1-7

[2] Bolodurina I and Parfenov D 2018 The development and study of the methods and algorithms for the classification of data flows of cloud applications in the network of the virtual data center International Journal of Computer Networks and Communications 10(2) 15-22

[3] Krasnov A E, Nikol'skii D N, Repin D S, Galyaev V S and Zykova E A 2018 Detecting DDoS Attacks Using the Analysis of Network Traffic as Dynamical System 2018 International Scientific and Technical Conference Modern Computer Network Technologies (MoNeTeC) Moscow 1-6

[4] Jin S, Jiang Y and Peng J 2018 Intrusion Detection System Enhanced by Hierarchical Bidirectional Fuzzy Rule Interpolation 2018 IEEE International Conference on Systems Man and Cybernetics (SMC) Miyazaki 6-10

[5] Rustam Z and Talita A S 2018 Fuzzy Kernel Robust Clustering for Anomaly based Intrusion Detection 2018 Third International Conference on Informatics and Computing (ICIC) Palembang 1-4

[6] Pradeepthi K V and Kannan A 2018 Detection of Botnet traffic by using Neuro-fuzzy based
Intrusion Detection 2018 Tenth International Conference on Advanced Computing (ICoAC) Chennai 118-23

[7] Alsirhani A, Sampalli S and Bodorik P 2019 DDoS Detection System: Using a Set of Classification Algorithms Controlled by Fuzzy Logic System in Apache Spark IEEE Transactions on Network and Service Management 16(3) 936-49

[8] Mendel J M 1995 Fuzzy logic systems for engineering: a tutorial Proceedings of the IEEE 83(3) pp 345-77

[9] Mouzouris, George C and Jerry M Mendel 1997 Nonsingleton fuzzy logic systems: theory and application IEEE Transactions on Fuzzy Systems 5.1 56-71

[10] Tabatabaei S F, Salleh M, Abbasy M R and Torkaman M R N 2011 Denial of Service (DoS) Attack Detection by Using Fuzzy Logic over Network Flows Proceedings of the International Conference on Security and Management 1

[11] Levonevskiy D K, Fatkueva R R and Ryzhkov S R 2015 Network attacks detection using fuzzy logic 2015 XVIII International Conference on Soft Computing and Measurements (SCM) St. Petersburg 243-4

[12] Moustafa N and Slay J 2016 The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set Information Security Journal: A Global Perspective 25(1-3) 18-31