Classification of Sequences Generated by Compression and Encryption Algorithms

Alexander Kozachok, Spirin Andrey Andreevich

***Abstract*—The paper considers the possibility of using the method of testing the properties of bit sequences as one of the possible approaches to solving the problem of classifying pseudo-random sequences and the sequences formed by encryption and compression algorithms. The results of the analysis led to the conclusion that the proposed feature space could be used to identify ZIP, RAR compression algorithms and AES, 3DES encryption algorithms with an accuracy of more than 0.95.**

***Tóm tắt*—Bài báo này xem xét khả năng sử dụng phương pháp thử nghiệm các đặc trưng của dãy bít như là một trong các cách tiếp cận để giải quyết bài toán phân loại các dãy giả ngẫu nhiên và các dãy được tạo ra bởi các thuật toán nén và mã hóa. Các kết quả của việc đánh giá dẫn tới kết luận rằng không gian đặc trưng được đề xuất có thể được sử dụng để xác định các thuật toán nén ZIP, RAR và các thuật toán mã hóa AES, 3DES với độ chính xác lớn hơn 95%.**

Keywords—Identification of compression and encryption algorithms, statistical testing of information.

Từ khóa—Xác định các thuật toán nén và mã hóa, kiểm tra thông tin thống kê.

I. INTRODUCTION

According to reports from the Infowatch Analytics Center [1-3], the number of confidential data leaks is growing from year to year. Fig.1 provides statistics on the leaks occurred in 2011-2018. The total damage from leaks in 2013 amounted to more than US$7.5 billion. Equifax, the international credit reference bureau, has already spent US$1.4 billion to eliminate one information leakage that occurred in 2017 and the amount has continued to grow due to repetitive claims. The share of intentional leaks carried out through network channels in 2018 is 86.6 %.

One of the means to prevent leaks is using DLP (Data Leakage Prevention) systems. The methods used in them are divided into the methods of content analysis and the ones of analyzing the context of the information transmitted outside [4].

The methods of content analysis include the following:

1. Methods based on the predefined rules.
2. Hash function.
3. Statistical methods.

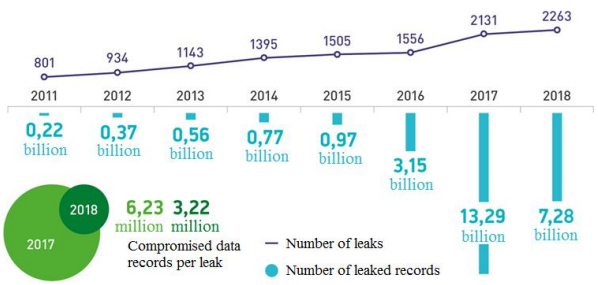

Fig.1. Statistics on the leakages occurred for 2011-2018

Methods based on the predefined rules possess high performance rate, but they badly cope with the problem if there is unknown data. Hash methods calculate the hash values of the data sent outside and compare them with the value of the prepared template, this method works only if the data has not been changed after the hash value template has been compiled. Statistical methods are based on statistical properties of data, for example, the frequency of words in the text. Encryption can be used to hide the meaning and content of the data [4].

Other means of preventing leaks are intrusion detection systems (Intrusion Detection System – IDS), which are able to analyze traffic and warn of a potential threat, but when

This manuscript is received April 22, 2019. It is commented on July 30, 2019 and is accepted on August 6, 2019 by the first reviewer. It is commented on August 20, 2019 and is accepted on August 27, 2019 by the second reviewer.

transmitting data in encrypted form, they perform only limited verification based on batch headers.

The complexity of analyzing the content of encrypted network traffic is one of the urgent problems for IDS.

II. A REVIEW OF RESEARCH

In a number of works [5-9] for traffic classification there have been applied machine learning methods using the characteristics extracted from the properties of the network streams (primarily from the packet headers). However, these methods have a significant drawback, their accuracy decreases when using encapsulation or encryption.

In early 2018, Cisco estimated that 60% of Internet traffic was encrypted. According to Gartner forecasts, by the end of 2019, 80% of the traffic will be the same. Encryption can also be used to hide interactions with malware command servers and solve other problems. According to the Ponemon Institute report for 2016, almost half (41%) attackers use encryption to bypass mechanisms of detecting their unauthorized activity. Security tools cannot inspect encrypted traffic (according to the Ponemon Institute, 64% of companies cannot detect malicious code in encrypted traffic) [10].

The existing security solutions are able to analyze the content of encrypted traffic, for example, by means of a man-in-the-middle attack, but due to the high cost of implementing the methods they are practically not applicable in real conditions [10]. However, there is at least one way to analyze the content of encrypted traffic without decrypting it – Cisco ETA (Encrypted Traffic Analytics), which allows, on the basis of network telemetry received from the network equipment and machine learning algorithms, classifying encrypted traffic, simultaneously separating the pure traffic in it from the malicious one. It is not the data field in the encrypted batch that is used for analysis, but rather its header, from which the extended telemetry is obtained [10]:

1. from Netflow – addresses and ports of the source and destination (SrcIP, DstIP, SrcPort, DstPort), information on the protocol, the number of transferred packets and bytes;

2. intra-stream – packet sizes & time parameters, byte distribution (occurrence in the stream) and their entropy (the higher it is, the higher the expectation to see the encrypted traffic);

3. from TLS metadata – extensions, cryptographic algorithm sets, SNI, certificate fields;

4. from DNS – domain names, query types, temporary query parameters;

5. from HTTP – headers and related fields, including other http requests from the same host.

It should be noted that in the case of data encapsulation in other protocols, the proposed method may give incorrect results.

The authors in [11] considered a method of detecting hidden encrypted partitions of a personal computer hard disk drive created by the TrueCrypt data encryption program. To classify files the authors used the NIST statistical test suite indicating encrypted data with an accuracy of more than 0,95.

In order to prevent information leakage, it is necessary to block the transmission of encrypted data, which determines the relevance of solving the problem of classifying encrypted, compressed and pseudo-random sequences (PRS).

III. PROBLEM STATEMENT

In general, the research problem is formulated as follows: it is necessary to map the original set of bit sequences X to the new set of classes Y by using the selected feature space. To be able to estimate the accuracy of the classification, the accuracy characteristic calculated by Formula 1 was used.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

where TP is the number of objects correctly assigned to class i, TN is the number of objects correctly assigned to class j, FP is the number of false positives *(type I error)*, FN is the number of false negatives *(type II error)*.

In a formalized form, the classification task can be defined by the following expression:

$$F : X \to Y, \quad (2)$$

where X is the initial set of bit sequences to be classified, Y is the set of classes.

The accuracy value in the classification must satisfy the condition presented in expression (3):

$$p(F(x_i) = y_j \mid i = j) \geq 0{,}95, \qquad (3)$$

where p is the probability, F(x) is the function of displaying the i-th file of the set X, y is the mark of the class from the set Y, i, j are the indices in the set of X and Y, respectively.

To develop and evaluate the classification method of encrypted, compressed and PRS the following particular tasks were set:

1. To check the possibility of binary classification of sequences generated by encryption and compression algorithms.

2. To check the possibility of multi-class classification of sequences generated by encryption, compression, and PRS algorithms.

IV. TESTING THE POSSIBILITY OF A BINARY CLASSIFICATION OF SENQUENCES GENERATED BY COMPRESSION AND ENCRYPTION ALGORITHMS

4.1 Binary classification for features generated on the basis of NIST test results

In the course of the research it was assumed that the results of NIST tests might be a feature space for constructing a classifier that allows distinguishing pseudo-random sequences by a source type.

To test the possibility of solving the classification problem, the NIST SP 800-22rev1a statistical test suite was used to evaluate random and pseudo-random sequence generators in cryptography [17].

To conduct the experiment the initial sample of 1000 files, 600 KB each, containing the text in Russian, was converted by encryption and compression algorithms. As a result, the resulting sample of 4000 files was divided into 2 classes:

1. Encrypted by AES, 3DES algorithms.

2. Compressed by RAR, ZIP algorithms.

Next, the files from the sample were processed with a package of statistical tests, as a result, 188 features were obtained.

To assess the applicability of the selected feature space, cross-validation (the number of subsamples equal to 10 were chosen) of machine learning algorithms with default parameters was used [18-23]: Decision Tree Classifier (DTC), Random Forest Classifier (RFC).

The results of the experiments are presented in Table 2.

TABLE 1. THE RESULTS OF CROSS-VALADATION OF MACHINE LEARNING ALGORITHMS IN THE NIST TEST FEATURE SPACE ON A SAMPLE COMPRESSED AND ENCRYPTED SEQUENCES

Algorithm	**Accuracy value**
Decision Tree	0,57508
Random Forest	0,642579

From the analysis of the results presented in Table 2, it is reasonable to conclude that the RFC has greater accuracy, but the resulting accuracy value of detecting sequence types equal to 0,642 does not allow constructing a classifier that meets the requirement presented in expression 3. Thus, it was concluded that the feature space based on the NIST test results would not allow solving the problem of binary classification taking into account the selected restrictions.

4.2 Binary classification on the feature space generated on the basis of the results of analyzing N length subsequence frequency.

Then it was assumed in the research that as a feature space for solving the problem of bit sequence binary classification the results of analyzing the frequency of independent bit subsequences of different length N (in bits) without taking into account the complete overlap of each subsequence can be used. For example, for the sequence S = "00011011" and N = 2 bits, the frequency occurrence of N bit length subsequences is presented in Table 3.

TABLE 2. EXAMPLE OF COUNTING THE FREQUENCIES OF BIT SUBSEQUENCES

Subsequence	**Amount**	**Frequency**
00	1	0,142857143
01	2	0,285714286
10	1	0,142857143
11	2	0,285714286

In the works of the authors it was shown that for binary sequences it was enough to analyze

half of all possible subsequences [24,25]. Using this assumption, the dimension of the feature space for N bit length subsequences has been halved, the number of possible feature on the assumption is presented in Table 4.

TABLE 3: THE DIMENSION OF THE FEATURE SPACE FOR N BIT LENGTH

Length of subsequence (N), bit	Amount of features	Amount of features, total	Cross-validation time on assumption, min
4	8	16	6
5	16	32	11
6	32	64	15
7	64	128	22
8	128	256	37
9	256	512	69

In the course of the experiments to assess the possibility of constructing binary classifiers,the RFC showed the highest accuracy, the results are presented in Table 5.

An average accuracy of more than 0,95 is achieved at N> = 8 bits.

The dependence of the average accuracy of classifying files on the basis of the RFC on the length of the subsequence and the time spent on the selection of features from the training sample is shown in Fig.2. A sufficient ratio of accuracy and time is determined at N = 9 bits length of the subsequence, in this case there is a significant increase in the accuracy of sequence classification at maintaining an acceptable amount of time spent on extracting features.

TABLE 4. ACCURACY OF DISTINGUISHING FILE TYPES WHEN USING RFC

File type	**Algorithm accuracy for length sequence**							
	N=4	*N*=5	*N*=6	*N*=7	*N*=8	*N*=9	*N*=10	*N*=11
	Time taken to retrieve features in minutes							
	6	11	15	22	37	69	135	268
AES/7-Z	0,821	0,843	0,834	0,835	0,938	**0,993**	**0,998**	**1,000**
AES/RAR	**0,986**	**0,992**	**0,994**	**0,992**	**0,991**	**0,997**	**0,993**	**0,993**
AES/ZIP	**0,986**	**0,992**	**0,988**	**0,990**	**0,993**	**0,998**	**0,999**	**0,999**
3DES/7-Z	0,834	0,846	0,865	0,865	0,947	**0,993**	**0,998**	**1,000**
3DES/RAR	**0,984**	**0,990**	**0,991**	**0,993**	**0,991**	**0,996**	**0,994**	**0,994**
3DES/ZIP	**0,991**	**0,988**	**0,991**	**0,989**	**0,991**	**0,997**	**0,998**	**0,999**
7-Z/ZIP	**0,968**	**0,974**	**0,974**	**0,973**	**0,971**	**0,972**	**0,976**	**0,978**
7-Z/RAR	0,948	**0,960**	**0,964**	**0,963**	**0,964**	**0,983**	**0,996**	**0,999**
RAR/ZIP	0,840	0,864	0,864	0,863	0,868	**0,977**	**0,999**	**1,000**
Mean value	0,929	0,939	0,941	0,940	**0,962**	**0,990**	**0,995**	**0,996**

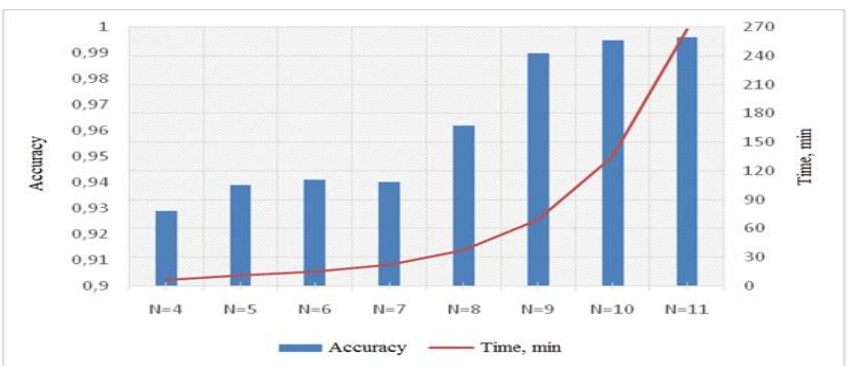

Fig.2. Dependence of the file type classification accuracy at using the RFC on the subsequence length

The AES/3DES pair was removed from the tested sample. Fig.3. shows the dependence of the accuracy of distinguishing sequences encrypted with the AES and 3DES algorithms on the length of the subsequence N.

Due to the ability of ciphers to dissipate the statistics of the source data, the accuracy of the classification of sequences encrypted with the AES and 3DES algorithms is on average 0,512, which makes it impossible to construct a classifier for sequences of this type.

In the course of the research, the possibility of applying the results of NIST statistical tests as a feature space for the classifying binary sequences generated by encryption and compression algorithms was verified. The accuracy of distinguishing the RFC sequences was 0,64, which does not satisfy the requirement specified in expression (3).

To solve the problem of binary sequence binary classification, a new feature space was proposed, it was formed by counting the frequency of different length bit subsequences. The classification accuracy of the RFC at the length of the subsequence N = 9 bits is 0,99, it satisfies the requirement specified in expression (3).

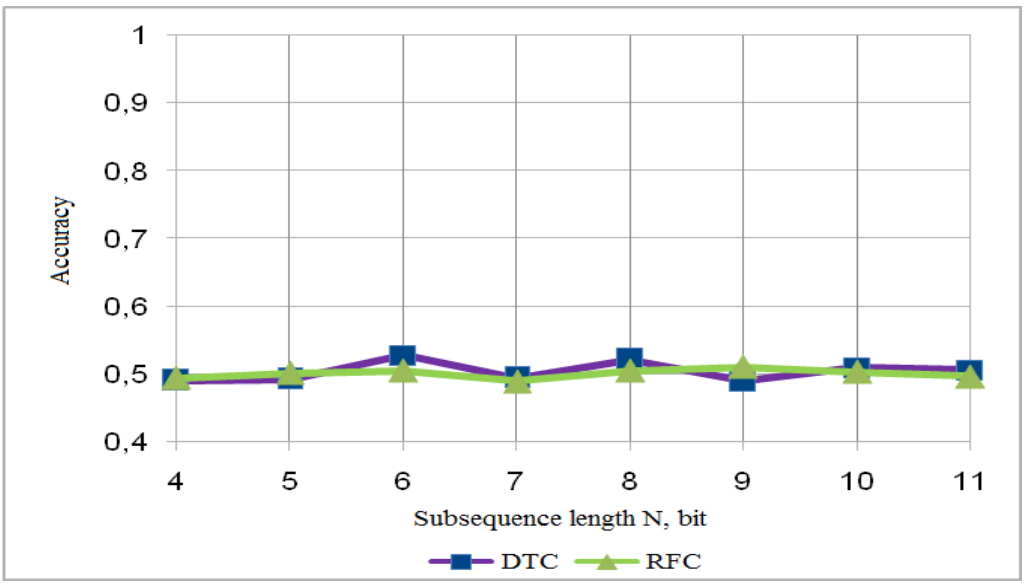

Fig.3. Dependence of the accuracy of distinguishing sequences encrypted with the AES and 3DES algorithms on the length of the subsequence

REFERENCES

[1]. INFOWATCH company group site. URL: https://www.infowatch.ru/analytics/reports.4.html (дата обращения: 30.05.2019).

[2]. INFOWATCH company group site. URL: https://www.infowatch.ru/sites/default/files/report/analytics/russ/infowatch_otchet_032014_smb_fin.pdf (дата обращения: 30.05.2019).

[3]. INFOWATCH company group site. URL: https://www.infowatch.ru/analytics/leaks_monitoring/15678 (дата обращения: 30.05.2019).

[4]. X. Huang, Y. Lu, D. Li, M. Ma. A novel mechanism for fast detection of transformed data leakage // IEEE Access. Special section on challenges and opportunities of big data against cyber crime. Vol. 6, 2018. pp. 35926-35936

[5]. Y. Miao, Z. Ruan, L. Pan, Y. Wang, J. Zhang, Y. Xiang. Automated Big Traffic Analytics for Cyber Security // Eprint arXiv:1804.09023, bibcode: 2018arXiv180409023M. 2018.

[6]. S. Miller, K. Curran, T. Lunney. Multilayer Perceptron Neural Network for Detection of Encrypted VPN Network Traffic // International Conference on Cyber Situational Awarness, Data Analytics and Assessment. 2018. ISBN: 978-1-5386-4565-9.

[7]. P. Wang, X. Chen, F. Ye, Z. Sun. A Survey of Techniques for Mobile Service Encrypted Traffic Classification Using Deep Learning // IEEE Access. Special section on challenges and opportunities of big data against cyber crime. Vol. 7, 2019. pp. 54024-54033 doi: 10.1109/ACCESS.2019.2912896

[8]. K. Demertzis, N. Tziritas, P. Kikiras, S.L. Sanchez, L. Iliadis. The Next Generation Cognitive Security Operations Center: Adaptive Analytic Lambda Architecture for Efficient Defense against Adversarial Attacks // Big Data and Cognitive Computing, 2019 3(6).

[9]. H. Zhang, C. Papadopoulos, D. Massey. Detecting encrypted botnet traffic // 16th IEEE Global Internet Symposium. 2013. p. 3453.

[10]. T. Radivilova, L. Kirichenko, D. Ageyev, M. Tawalbeh, V. Bulakh Decrypting SSL/TLS Traffic for Hidden Threats Detection // IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), 2018. ISBN: 978-1-5386-5903-8.

[11]. M. Piccinelli, P. Gubian. Detecting hidden encrypted volume files via statistical analysis // International Journal of Cyber-Security and Digital Forensics. Vol. 3(1). 2013 pp. 30-37.

[12]. NIST STS manual. URL: https://csrc.nist.gov/Projects/Random-Bit-Generation/ (дата обращения: 14.01.2019).

[13]. Toolkit for the transport layer security and secure sockets layer protocols.

URL: http://openssl.org (дата обращения: 14.01.2019)

[14]. Archive manager WinRAR. URL: http://rarlab.com (дата обращения: 14.01.2019).

[15]. Pedregosa F., et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research 12. 2011. pp. 2825-2830.

[16]. Breiman L., Friedman J., Olshen R., Stone C. Classification and Regression Trees // Wadsworth, Belmont, CA. 1984. 368 p. ISBN: 9781351460491.

[17]. Hastie T., Tibshirani R., Friedman J. Elements of Statistical Learning // Springer. 2009. pp. 587-601. ISBN: 978-0387848570.

[18]. L. Breiman, A. Cutler. Random Forests // URL:https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (дата обращения: 14.01.2019).

[19]. S. Raska. Python and machine learning // M .: DMK-Press. 2017. 418 p. ISBN: 978-5-97060-409-0.

[20]. L. Breiman. Random Forests // Journal Machine Learning 45(1). 2001. pp. 5-32.

[21]. M.Yu. Konyshev. Formation of probability distributions of binary vectors of the source of errors of a Markov discrete communication channel with memory using the method of "grouping probabilities" of error vectors. / M.Yu. Konyshev, A.Yu. Barabashov, K.E. Petrov, A.A. Dvilyansky // Industrial ACS and controllers. 2018. № 3. P. 42-52.

[22]. M.Yu. Konyshev. A compression algorithm for a series of distributions of binary multidimensional random variables. / M.Yu. Konyshev, A.A. Dvilyansky, K.E. Petrov, G.A. Ermishin // Industrial ACS and controllers. 2016. No. 8. P. 47-50.

ABOUT THE AUTHORS

D.S. Alexander Kozachok

Workplace: The Academy of Federal Guard Service of the Russian Federation.

Email: alex.totrin@gmail.com

The education process: received PhD. degree in Engineering Sciences in Academy of Federal Guard Service of the Russian Federation in Dec. 2012.

Research today: Information security; Unauthorized access protection; Mathematical cryptography; theoretical problems of computer.

Spirin Andrey Andreevich

Workplace: The Academy of Federal Guard Service of the Russian Federation.

Email: spirin_aa@bk.ru

The education process: graduated from the Academy of the Federal Guard Service of the Russian Federation in 2010.

Research today: Information security, information leakage prevention systems, statistical testing.