Poisoning attack of the training dataset for on-line signature authentication using the perceptron

Evgeny Kostyuchenko¹, Balgit Lodonova¹, Yakov Usoltsev¹, Artem Konovalov¹

¹Tomsk State University of Control Systems and Radioelectronics, 40 Lenina St., Tomsk, 634050, Russia Federation

E-mail key@fb.tusur.ru

Abstract. The study of the dependences of the effectiveness of a poisoning attack for the task of on-line signature authentication using a perceptron to make a decision. The dependences on the structure of the neural network, the volume of poisoning sets and the noise of the training sample were studied. Conclusions are drawn on the lack of the ability to detect such attacks by changing the overall accuracy of the system.

Keywords: Training sample poisoning, Biometric authentication, On-line signature, Neural network

1. Introduction

Attacks on machine learning algorithms, and in particular on neural networks, can lead to various vulnerabilities in systems using these algorithms, in particular, an attacker can deceive the recognition algorithms of certain images [1, 2] or system developers there will remain the possibility of unauthorized access to the system [3]. Considering how important the component of the modern information space is machine learning [4], the tasks associated with the study of attacks on machine learning algorithms are now more than ever an urgent topic. Watching over development trends of modern information technologies we can observe significant increase areas of application to machine learning systems. Now system of speech recognition [5] or complex system of security systems [6] are created on neural networks basis.

As examples of such attacks, adversarial attacks [7, 8] and attacks at the training stage [9] can be distinguished.

The process of preparing a training sample is one of the key points of machine learning that determines its outcome. For this reason, providing counteraction to attacks while ensuring the security of the system based on machine learning methods is relevant and counteracting poisoning attacks of a training dataset [10] is important.

The poisoning attack to dataset is analyzed [11]. The essence of this attack is that when training a neural network, an attacker injects his data into the training set. Within the framework of the system under study, an attacker can embed his data sets in the data of legal users in order to increase the percentage of statistical errors of the second kind and gain potential access to the system under the guise of a legal user [12].
As part of a review aimed at identifying typical set poisoning attacks and evaluating their effectiveness, we can distinguish such works as [13-18]. Consider them, the methods used within the system and the results obtained in more detail.

In [13], a study is proposed of counteracting the methods of selecting features for training. Selection methods such as LASSO, ridge regression and the elastic net are considered. The general methods of machine learning are analyzed, it is shown that poisoning of 5% of the training dataset of the class leads to a complete disruption of the system.

The work [14] explores the stability of federated learning systems against training set poisoning attacks. The possibility of an attacker generating new training examples and using them to poison a training sample is shown. It is shown that this approach to attack allows achieving accuracy up to 80% for the attacker.

The vulnerability of the least squares method for the algorithm uLSIF, which is used to calculate the importance ratio for the adaptation of a covariant shear region to poisoning attacks, was shown in [15]. The results allow us to talk about the applicability of training dataset poisoning for the method under consideration.

The work [16] shows the possibility of the combined effect of the poisoning attack of the training sample regarding the integrity and accessibility of information. Attack efficiency for logistic and linear regression methods is shown.

In [17], the effectiveness of an attack for an intrusion detection system is considered. The Edge Pattern Detection (EPD) algorithm is analyzed to develop a new poisoning method that attacks several machine learning algorithms used in IDS.

In [18], the effectiveness of pollution attacks of training sets for intrusion detection systems, in particular, for mobile applications, is shown. A solution has been proposed to increase the detection efficiency of such attacks by at least 15%.

The analysis of the work allows us to talk about the relevance of the study of poisoning attacks of the training set based on new research in this direction over the past 2 years. It is seen that these attacks can be applied to various systems, in particular, intrusion detection systems and identification systems. Various characteristics, such as voice [2, 19], keyboard handwriting [20], off-line [21] and on-line [22] signature can be used as input for authentication systems using biometric characteristics. This work is devoted to the study of the effect of a poisoning attack on a training dataset on one of the systems not represented in existing works, namely, an identification system based on signature dynamics.

2. Description of the existing on-line signature authentication system

The system without attack is a script on a mathematical package MATLAB [23] that includes:

1) extraction of data sets of signatures from the database;
2) creation of a neural network with specified parameters;
3) splitting the signature datasets into three parts – for neural networks training, for selection decision threshold and for primary network testing with retraining;
4) random generation of new sets based on existing sets with a deviation of 1% for increasing the number of input data for training;
5) neural networks training;
6) selection decision threshold;
7) testing a trained neural network;
8) obtaining and correct presentation of various statistical errors [24].

Input data for neural networks training is signature dataset, several dozen for each user. Signature data sets are presented in the form of 144 signature parameters [22] that provide complete information about the dynamics of signature writing.

The study was conducted on dataset of 2445 parameterized signatures from 22 users. User is identified in system by a unique number for each user from 1 to 22. 13 users are legitimate on the basis of which the network will be trained. The rest 9 users are not registered in the system, and they will be used for testing system.
For all users, a process is provided for a tenfold increase in the number of signatures by generating nine signatures based on one with making a random change for each of the 144 signature parameters, not exceeding 1% of the parameter value.

By default, created neural network takes 144 parameters to input layer, have 30 neurons in intermediate layer and return vector of result. Number of elements in this vector is equal to the number of legitimate users in the system and displays to whom the entered signature belongs (Figure 1).

Signatures from legitimate users are divided into three sets: training set, set for selection decision threshold and set for primary training, in the ratio of 60%, 20% and 20%.

Based on the results of testing the trained neural network, data is formed on the sum of the errors of the first and second kind, the system error and the total error of the first and second kind with a tenfold priority of the error of the second kind.

3. Variants of the investigated attack parameters

As a part of a research on machine learning with biometric authentication and attacks, we conducted studies with gradual increase of various parameters of the neural network learning.

Parameters were increased within predetermined intervals.

In particular, we studied the effect on the output of the following parameters:

- number of neurons in intermediate layer;
- the percentage of inclusion of alien set under the guise of a legitimate user;
- number of users in system;
- coefficient of noise.

3.1. Dependence on the number of neurons in the intermediate layer

In result, we changed the number of neurons in intermediate layer and received plots of dependences of the total error of the first and second kind at tenfold priority when attacking the system (Figure 3) and without attack (Figure 2).
Figure 3. Dependence of the sum of errors sum at tenfold priority of second kind error to number of neurons in intermediate layer

Also error of the second kind that occurs on the attacked user value was received. The value reflects the probability of an intruder penetration in system as legal user.

The following graph shows dependence statistical error of second kind value of the number of neurons in intermediate layer (figure 4).

As seen from the graphs, gradual increase of number of neurons in intermediate layer doesn’t influence poisoning attack efficiency. The extremum position is blurred for legal user, but extremums value is similar in all studies.

Figure 4. Dependence of changes in the statistical error of the second kind to number of neurons in intermediate layer

3.2. Dependence on the percentage of inclusion

Similar in methodology study was conducted for estimates of the effect of the percentage of introduced by the attacker sets on system functioning and on the probability of an intruder penetration in system as legal user.

Probabilities are shown on figure 5.

Percentage values introduced sets located on the abscissa axis, value of errors located on ordinate axis.
As graphs show the percentage of inclusion doesn’t influence on systems efficiency. It means poisoning attack is difficult to detect by system performance analysis and by user data. We received value of dependence of the error on attacked user. The error value reflects the probability of an intruder penetration in system as legal user.

The dependence of the effectiveness of the attack on the percentage of poisoning is shown in the figure 6.

As seen from the dependency graphs percentage of poisoning sets influences user errors. Therefore, by using a set poisoning attack, an attacker is theoretically able to gain access with minimal risk of being detected.

3.3. Influence of the number of users on the system operation

We changed the number of users in training set for attacked system and received graphs of dependence of errors sum at tenfold priority of second kind error to the probability of a successful attack.
As seen from the error graphs, increase in the number of users with a fixed number of examples from the user doesn't change errors value. Also it becomes clear attack on system is more effective with a small number of users.

3.4. *Influence of the percentage of noise on the operation of the system*

Increase the training set by generation of new sets with a noise on the basis of available is a typical method for enhance learning sample representativeness and the quality of the system [25].

We conducted a study of influence of multiplicity of noising on the system operation with attack on it and without attack.

Noise is generated based on training set with introduction of entropy of 1%. We changed number of automatically generated sets for the neural network and received graphs of dependence of errors sum at tenfold priority of second kind error and attack efficiency on multiplicity of noising. Experiments were carried out for system without and for attacked system.
Figure 9. Dependence of error to multiplicity of noising, system without attack and tenfold priority of second kind error

Figure 10. Dependence of error to multiplicity of noising, system with attack and tenfold priority of second kind error

Figure 11. Dependence of attack success to multiplicity of noising, percentage of include is fixed

It is seen, increase multiplicity of noising over 5 doesn’t affect error rates which are also near some constant indicator

4. Discussion of results
The obtained dependences were analyzed and an attempt was made to explain them. Summing up the study we inferred optimal number of neurons, significant different from which leads to a decrease of system efficiency. Change in the number of neurons doesn’t affect system efficiency for legal users that in the presence of attack in the absence of. And extremum has blurred position
boundaries, although compared to the previously recorded level error value and rate don’t change. This shows the problematic nature of attack detection for system. It’s impossible to detect using methods of changing system parameters. Also efficiency of attack doesn’t depend number of neurons. Also it’s seen, system efficiency doesn’t change under the influence of introduced sets. We can see increased attack efficiency from percentage inclusion. And minimum level of inclusion (1%) increase attack efficiency to 30 % which is significant for system efficiency. Inclusion of 10-20% of sets make access to the system for attacker.

Study of depend errors to numbers of neurons shows that general index doesn’t change in investigated range, and probability of successful attack with fixed level of poisoning set decreases for a single user. This can be explained by two factors: firstly, increase of number of users is an increase of total number of sets, as a result, total share of included sets is decrease. Secondly, total number of classes is increase, and all of them can output for system. Attacker needs to be identified only as attacked user for a successful attack. Therefore, increase of total number of classes decrease probability of this match. However, corresponding to attack users output is stay more possible.

Similar in general concept attacks to the neural network were carried out by Jiang, Wenbo, et al. [26]. This neural network was created to recognize road signs. Jiang, Wenbo, et al modeled situation in which attacker included own set in training set. Inclusive set is similar set of attacked user. Inclusive set is modificated to visually practically do not differ to legal set. This inclusion doesn’t allow the neural network to correctly identify inclusive set. Varying of inclusion degree was make possible to strike dependence of system accuracy to varations of number of inclusive set to total set. As a result, system accuracy decrease from 95 to 33 for one user set and from 97 to 63 for another user when varying of poison rate from 0% to 10%.

Similar on approach to poisoning, but different in poisoning set generation method study was created by Biggio, Battista, et al. [27]. Varying of poison rate from 0% to 10% was in their publication too. Biggio, Battista, et al used different from the proposed Jiang algorythm to generate poisoning set and received increase of error from 0-4 to 20-28 to different trainig sets. Unfortunately for us, systems in the presented publications don’t have a calculating statistical errors. But we can compare received by us values of statistical second kind errors on a specific user with system efficiency and percentage of test error, which is received in the presented publications.

Increase of percentage of inclusive set rate from 1% to 10% changes value of statistical second kind error from 30% to 87% in our system. We can make a conclusion that attack in this study has a high impact on the system. But the attack can be easier to detect because it involves more aggressive intervention in the system.

5. Conclusion
The study was presented on the effect of changing various parameters of neural network training to final classification accuracy.

The study was carried out for the task of on-line signature authentication using a perceptron to make a decision.

Results of the study confirm the inherent for similar attacks on machine learning methods sensitivity to poisoning inclusion.

Even 1% of include is enough to multiple reduction in quality of system operation in relation to the detection of an attacking user.

Further studies in countering to attack on machine learning methods is relevant for such areas of information security as a authentication and detecting invasion [24].
6. Acknowledgment
This research was funded by the Ministry of Science and Higher Education of Russia, Government Order for 2020–2022, project no. FEWM-2020-0037 (TUSUR).

References
[1] Goodfellow I J, Shlens J and Szegedy C 2015 Explaining and Harnessing Adversarial Examples arXiv:1412.6572 [cs, stat]
[2] Rakhmanenko I, Shelupanov A and Kostyuchenko E 2019 Fusion of biLSTM and GMM-UBM systems for audio spoofing detection International Journal of Advanced Trends in Computer Science and Engineering 8
[3] Chen B, Carvalho W, Baracaldo N, Ludwig H, Edwards B, Lee T, Molloy I and Srivastava B 2018 Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering arXiv:1811.03728 [cs, stat]
[4] Paudice A, Muñoz-González L, Gyorgy A and Lupu E C 2018 Detection of Adversarial Training Examples in Poisoning Attacks through Anomaly Detection arXiv:1802.03041 [cs, stat]
[5] Kipyatkova I S and Karpov A A 2016 Variants of Deep Artificial Neural Networks for Speech Recognition Systems Тp. СПИИРАН 6 80
[6] Iskhakov S Y, Shelupanov A A and Meshcheryakov R V 2017 Simulation modelling as a tool to diagnose the complex networks of security systems J. Phys.: Conf. Ser. 803 012057
[7] Tahir R, Cheng K, Eric L and Khokhar M S 2019 Multi-domain Cross-dataset and Camera Style Transfer for Person Re-Identification International Journal of Advanced Trends in Computer Science and Engineering 8 2035–41
[8] Roli F, Biggio B and Fumera G 2013 Pattern Recognition Systems under Attack Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications Lecture Notes in Computer Science ed J Ruiz-Shulcloper and G Sanniti di Baja (Berlin, Heidelberg: Springer) pp 1–8
[9] Kloft M and Laskov P Online Anomaly Detection under Adversarial Impact 8
[10] Zhao W, Li P, Zhu C, Liu D and Liu X 2019 Defense Against Poisoning Attack via Evaluating Training Samples Using Multiple Spectral Clustering Aggregation Method Computers, Materials & Continua 59 817–32
[11] Steinhardt J, Koh P W and Liang P 2017 Certified Defenses for Data Poisoning Attacks arXiv:1706.03691 [cs]
[12] Kwon H, Yoon H and Park K-W 2019 Selective Poisoning Attack on Deep Neural Networks Symmetry 11 892
[13] Xiao H, Biggio B, Brown G, Fumera G, Eckert C and Rolli F 2018 Is feature selection secure against training data poisoning? arXiv:1804.07933 [cs, stat]
[14] Zhang J, Chen J, Wu D, Chen B and Yu S 2019 Poisoning Attack in Federated Learning using Generative Adversarial Nets
[15] Umer M, Frederiksson C and Polikar R 2019 Vulnerability of Covariate Shift Adaptation Against Malicious Poisoning Attacks 2019 International Joint Conference on Neural Networks (IJCNN) 2019 International Joint Conference on Neural Networks (IJCNN) pp 1–8
[16] Jiang W, Li H, Liu S, Ren Y and He M 2019 A Flexible Poisoning Attack Against Machine Learning ICC 2019 - 2019 IEEE International Conference on Communications (ICC) ICC 2019 - 2019 IEEE International Conference on Communications (ICC) pp 1–6
[17] Li P, Liu Q, Zhao W, Wang D and Wang S 2018 Chronic Poisoning against Machine Learning Based IDSs Using Edge Pattern Detection 2018 IEEE International Conference on Communications (ICC) 2018 IEEE International Conference on Communications (ICC) pp 1–7
[18] Chen S, Xue M, Fan L, Hao S, Xu L, Zhu H and Li B 2018 Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach Computers & Security 73 326–44
[19] Matveev Y 2013 The Problem of Voice Template Aging in Speaker Recognition Systems Speech and Computer Lecture Notes in Computer Science ed M Železný, I Habernal and A Ronzhin (Cham: Springer International Publishing) pp 345–53
[20] Kostyuchenko E Y, Viktorovich I, Renko B and Shelupanov A A 2018 User Identification by the Free-Text Keystroke Dynamics 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC) 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC) pp 1–4
[21] Even S, Goldreich O and Micali S 1990 On-Line/Off-Line Digital Signatures Advances in Cryptology — CRYPTO’89 Proceedings Lecture Notes in Computer Science ed G Brassard (New York, NY: Springer) pp 263–75
[22] Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia, Samsonov S S, Filimonenko I V and Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia 2018 Dynamic-signature-based user authentication using a fuzzy classifier CO 42 657–66
[23] MATLAB — MathWorks — MATLAB & Simulink https://www.mathworks.com/products/matlab.html
[24] Kostyuchenko E, Gurakov M, Krivonosov E, Tomyshev M, Mescheryakov R and Hodashinskiy I 2016 Integration of Bayesian Classifier and Perceptron for Problem Identification on Dynamics Signature Using a Genetic Algorithm for the Identification Threshold Selection Advances in Neural Networks – ISNN 2016 Lecture Notes in Computer Science ed L Cheng, Q Liu and A Ronzhin (Cham: Springer International Publishing) pp 620–7
[25] Reed R and II R J M 1999 Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks (Cambridge, Mass.: Bradford Books)
[26] Jiang W, Li H, Liu S, Luo X and Lu R 2020 Poisoning and Evasion Attacks Against Deep Learning Algorithms in Autonomous Vehicles IEEE Transactions on Vehicular Technology 69 4439–49
[27] Biggio B, Nelson B and Laskov P 2013 Poisoning Attacks against Support Vector Machines arXiv:1206.6389 [cs, stat]
[28] Hayakawa Y, Marumoto A and Sawada Y 1995 Effects of the chaotic noise on the performance of a neural network model for optimization problems Phys. Rev. E 51 R2693–6
[29] Proshkin N, Barannik E, Lapina M, Alla K and Lapin V 2020 Developing models of IoT infrastructures to identify vulnerabilities and analyse threats IOP Conference Series: Materials Science and Engineering 873 012018