DOI: https://doi.org/10.30837/csitic52021232208

Development of Computer State Identification Method Based on Boosting Ensemble

Viktor Chelak[1]

[1] National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv UA-61002, Ukraine , e-mail: victor.chelak@gmail.com

Svitlana Gavrylenko [2]

[2] National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv UA-61002, Ukraine , e-mail: gavrilenko08@gmail.com

Abstract. *This work is about developing a modification of boosting method by using a special preprocessing procedure to improve the accuracy of computer system state identification. The aim of the research is to develop method for detection computer threats, malware, etc. Experimental research have confirmed the effectiveness of the proposed method, which makes it possible to recommend it for practical use in order to improve the accuracy of identifying the state of the computer system. Prospects for further research may be to develop an ensemble of fuzzy decision trees based on the proposed method, optimizing their software implementation.*

Keywords: *computer system state identification, data processing, decision tree boosting ensembles.*

I. Introduction and Problem Statement

The task of information security is becoming more difficult every year and is demanding on the accuracy and speed of performance of methods for detecting threats and their prevention. [1] A computer system can be defined by a large number of features: the workload of various components, the amount of processed memory per second, etc. If we add to this a huge number of processes that are performed in processor races, we get a gigantic set of data, factors and conflicting information that needs to be processed. In order to find our potential features, it is necessary to analyze our data and make optimization according to criteria, perform preliminary data processing. [2]

In [3], a comparative analysis was carried out on various applied problems using ensemble methods: Gradient Boosting, Extreme Gradient Boosting, etc. The results of this considered work were good enough for boosting algorithms. The work [4] presents a method aimed at solving applied problems in the field of medicine and diagnostics. However, while this model works perfectly in tasks related to signals and images, for non-uniform data (such as in the task of intrusion detection) it showed a high variance error. There are also interesting solutions to the problems of classifying web pages using boosting algorithms [5].

The analysis made it possible to formulate the problem: build a model which is capable of making multiple classification of the state of a computer system. Basic requirements for such a model:

- The bias error must be 0%.
- The variance error must not exceed 10%.
- For cases of binary classification, the system output "-1" means the normal state of the computer system and "1" - an abnormal state.
- A special data preprocessing procedure is required to remove anomalies and noise in the training set.

II. Problem Solution and Results

A special data preprocessing procedure can be represented in the form of 4 parts:

- Processing conflicting information
- Handling missing values
- Handling anomalous values (strong outliers)
- Noise handling (weak outliers)

Conflicting information means two or more samples that are completely coinciding in all criteria and at the same time have different output classes. No such samples were found in the training set used in this work. However, it is necessary to take into account the fact that the data is obtained in real time and the likelihood of collisions is quite possible.

The way to deal with such data is quite simple - to estimate the probabilities. If conflicting information tends in most cases (80% or more) to one of the classes, delete the data that does not include this sample in this class. If the data does not have such a single class or, in the worst case, is distributed across classes, such data must be deleted.

In the process of monitoring the indicators of a computer system, a failure may occur, due to which some criteria in the sample are lost. An example would be hard disk load values. At the time of the request for information, the system did not have access to it, which is why an empty string or NaN was returned. In the case of training, such data is initially not allowed in the training set. In any case, the system must correctly process the input data in the classification mode. It is necessary to store and calculate the average value of each criterion, and in case of its absence, supply the average value to the model instead of an empty datum. This approach will help to avoid a strong distortion of statistical characteristics (the mathematical expectation will remain the same).

Strong and weak outliers can be found and removed from the training set using machine learning methods such as One-class support vector machines, DBSCAN, etc. The proposed system uses several methods at once to detect anomalies. In addition to those already listed, a standard deviation (coefficient is 3.0) is used to detect strong outliers or anomalies. The standard deviation is calculated for each criterion and only if many criteria are outside the specified range - the sample is considered abnormal and removed. As a result of such cleaning, about 15-25% of the data will be eliminated.

In addition to preprocessing the samples, the proposed solution analyzes the criteria themselves for optimization. The criteria are not considered in case a large number of high absolute values of the correlation coefficients (0.8 or higher).

Criteria for which the variance value is close to zero will also be excluded, as such criteria are insignificant. The data is also being scaled to the normalized range (0, 1). If such normalization results in a low variance, the data are scaled in the range from 0 to 10^n, where n starts at 1 and is increased until the variance has an acceptable value or the criteria is detected as insignificant and deleted.

A software prototype of the proposed model was developed. In order to check the effectiveness of the approach, it is necessary to compare not only the proposed method with the classical ones. But also make a comparison with data preprocessing (optimization of criteria and removal of inconsistent data, filling in the gaps in information, etc.) and without. Figure 1 shows a comparative histogram for bias error.

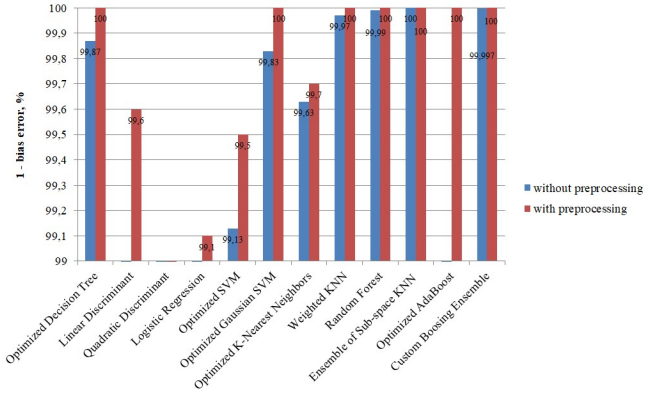

Figure 1. Accuracy comparison on training data

Figure 2 shows a comparative histogram for variance error.

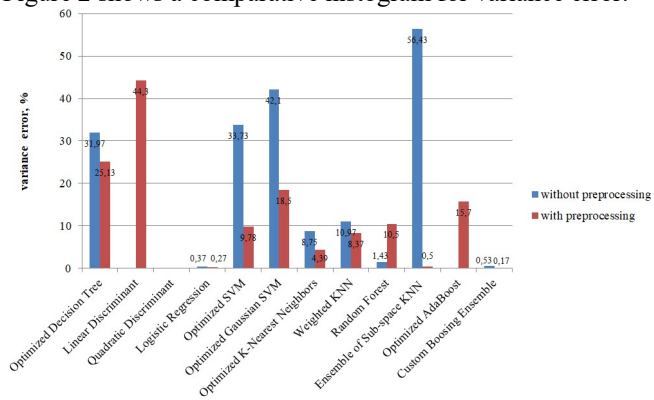

Figure 2. Variance error comparison

III. CONCLUSIONS

The ensemble boosting method was modified with data preprocessing procedure, which made it possible to increase the accuracy and reduce bias and variance errors.

It was found that the use of the proposed classifier makes it possible to reduce the variance to 10%.

The practical significance lies in the fact that the developed method is implemented in software and researched during the solution of the real problem of identifying the state of a computer system.

Prospects for further research may be to develop an ensemble of fuzzy decision trees based on the proposed method, optimize its software implementation and improve the quality of classification

REFERENCES

[1] Inter American Development Bank and Organization of American States, "Cybersecurity Risks, Progress, and the Way Forward in Latin America and the Caribbean", July 2020, pp. 19-38, doi: http://dx.doi.org/10.18235/0002513.

[2] J. Dutta, Y. W. Kim and D. Dominic, "Comparison of Gradient Boosting and Extreme Boosting Ensemble Methods for Webpage Classification," 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Bangalore, India, 2020, pp. 77-82, doi: 10.1109/ICRCICN50933.2020.9296176.

[3] A. H. Mirza, "Online boosting algorithm for regression with additive and multiplicative updates," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404455.

[4] Z. Li and D. Wang, "Classification on Point-cloud of Shoe-last Curvature using Weight-updated Boosting based Ensemble Learning," 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2020, pp. 2073-2077, doi: 10.1109/ITAIC49862.2020.9338947.

[5] P. Kumkar, I. Madan, A. Kale, O. Khanvilkar and A. Khan, "Comparison of Ensemble Methods for Real Estate Appraisal," 2018 3rd International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2018, pp. 297-300, doi: 10.1109/ICICT43934.2018.9034449.