IJICTR
International Journal of Information & Communication Technology Research

Volume 11- Number 3 – Summer 2019 (42 -48)

A Method for Anomaly Detection in Big Data based on Support Vector Machine

Masoud Harimi
Department of Computer Engineering
University of Science and Culture
Tehran, Iran
msdharimi@gmail.com

Mohammad Javad Shayegan*
Department of Computer Engineering
University of Science and Culture
Tehran, Iran
shayegan@usc.ac.ir

Received: 5 March 2019 - Accepted: 13 June 2019

Abstract—**In recent years, data mining has played an essential role in computer system performance, helping to improve system functionality. One of the most critical and influential data mining algorithms is anomaly detection. Anomaly detection is a process in detecting system abnormality that helps with finding system problems and troubleshooting. Intrusion and fraud detection services used by credit card companies are some examples of anomaly detection in the real world. According to the increasing volumes of the datasets that creates big data, traditional data mining approaches do not have efficient enough results. Various platforms, frameworks, and algorithms for big data mining have been presented to account for this deficiency. For instance, Hadoop and Spark are some of the most used frameworks in this field. Support Vector Machine (SVM) is one of the most popular approaches in anomaly detection, which—according to its distributed and parallel extensions—is widely used in big data mining. In this research, Mutual Information is used for feature selection. Besides, the kernel function of the one-class support vector machine has been improved; thus, the performance of the anomaly detection improved. This approach is implemented using Spark. The NSL-KDD dataset is used, and an accuracy of more than 80 percent is achieved. Compared to the other similar approaches in anomaly detection, the results are improved.**

Keywords-Anomaly detection; support vector machine; big data; improvement of anomaly detection; one-class support vector machine; Mutual Information

I. INTRODUCTION

There is a high volume of information in the 21st century, and various data mining methods have been proposed to analyze this data. A data mining technique called Anomaly detection has attracted much attention in recent years since it plays a crucial role in many domains. Anomalies are patterns in data that do not conform to the expected behaviors. Anomaly detection is the process of finding patterns of data that meet this condition [1]. The history of anomaly detection can be traced back to studies conducted by the Statistics community at the beginning of the nineteenth century. Because of the importance of anomaly detection at the time, many researchers from various domains have noted this problem, and a broad range of techniques methods have been proposed [2].

The amount of information is developing in various fields, for instance, sensors (IoT), interpersonal organizations, frameworks with vast system streams, such as energy systems, banking, and so forth. Hence, the analysis of big data becomes a challenge [3]. High

* Corresponding Author

volume, assortment, and rapid information produced in the system have made the information examination cycle to identify assaults by customary procedures entangled [4], so enormous information approaches and stages are utilized to manage big data. Peculiarity identification of big data has basic applications in numerous fields, for example, interruption location [5] and misrepresentation recognition in the banking and protection businesses. [3, 4].

Various approaches have been proposed for anomaly detection, among which the one-class support vector machine (OCSVM) is one of the most popular methods [1].

In this paper, we propose an anomaly detection based on OCSVM. First, a preprocessing method is used to standardize the labels and normalize the data. Second, Mutual Information (MI) is used to reduce dimensionality on the dataset to further improve the training by using the most useful features. Third, improved OCSVM is used for modeling. More specifically, the kernel function of the OCSVM, which is a combination of two basic kernel functions, has been changed. Also, parameters of the kernel function have been improved. The proposed method works with less features compared with various strategies. because of the improved bit work, it has higher accuracy than the related works.

The paper is organized as follows. Section 2 introduces related work. In section 3, we present the proposed method. Also, each step in this method is described. Results and experiment settings are mentioned in section 4. Finally, section 5 concludes this paper and describes future work.

II. Related Works

Besides the support vector machine, several techniques and frameworks have been proposed for anomaly detection in big data [5, 6]. As previously mentioned, this research uses the one-class support vector machine. The rest of this section reviews the recent anomaly detection methods based on the support vector machine.

Tran KP et al. [7] investigated the application of OCSVM to detect anomalies in WSNs with data-driven hyperparameter optimization. The IBRL dataset was used to test the proposed method, to which they achieved a high-level of detection accuracy and a low false alarm rate.

Othman et al. [4] presented the Flash Chi-SVM model for interruption recognition. The creators manufactured an interruption recognition model by the SVM classifier on the Apache Flash huge information stage. In this model, numTopFeatures is applied to dataset highlights. The analysts utilized the KDD dataset, and the proposed technique is contrasted and the strategic relapse classifier. The consequences of the Sparkle Chi-SVM indicated that the model has elite and speed.

Amraee et al. [8] proposed a new method for detecting abnormal events in public surveillance systems. First, they extracted candidate regions and eliminated the redundant information. Then, they calculated HOG-LBP (histogram of the oriented gradients-local binary pattern) and HOF (histogram of oriented optical flow) for each region. Finally, abnormal events are detected using two distinct OCSVM models. Experimental results showed that the proposed method outperforms existing methods based on the UCSD anomaly detection video datasets.

Amer et al. [9] applied two modifications to make OCSVM more suitable for unsupervised anomaly detection. The main idea of both modifications is that anomalies should contribute less to the decision boundary as regular instances. Robust OCSVM and eta OCSVM are the proposed methods in this research. The main modification of robust OCSVM is concerning the slack variables, while eta OCSVM uses an explicit outlier suppression mechanism. Compared with other standard unsupervised anomaly detection algorithms, the enhanced OCSVMs are superior on two out of four UCI machine learning repository [10] datasets. In particular, the proposed eta OCSVM has shown the most promising results.

Erfani et al. [11] proposed an unaided inconsistency location method for high-dimensional enormous scope datasets. The strategy is a mix of a profound conviction arrangement (DBN) and OCSVM. In particular, an unaided DBN is prepared to extricate nonexclusive hidden highlights, and an OCSVM is set up from the highlights learned by the DBN. Since a direct piece can be filled in for nonlinear ones in this crossover model without loss of precision, this model is versatile and computationally effective. They exhibited that contrasted and an autoencoder, the proposed half and half strategy was executed quicker in preparing and testing time.

Chalapathy et al. [12] proposed a one-class neural network (OC-NN) model to detect anomalies in complex data sets. OC-NN combines the ability of deep networks to extract a progressively rich representation of data with the one-class objective of creating a tight envelope around normal data. In this approach, an autoencoder is used for learning deep features and then feeding it to a different anomaly detection method like OCSVM. Experimental results demonstrated that the proposed method outperforms conventional shallow methods in a variety of scenarios.

Garg et al. [16] proposed a hybrid data processing model for network anomaly detection. This strategy utilizes Dark Wolf Streamlining (GWO) and Convolutional Neural System (CNN) and contains two stages. In the main stage, improved GWO is utilized for highlight choice, and the following stage, improved CNN, is utilized for oddity recognition. These methods are improved by recharging their separate standard techniques. CNN is extemporized regarding dropout layer usefulness, while GWO is changed concerning upgrading the underlying populace. The aftereffects of their crossover model assessed on benchmark and

manufactured datasets showed the prevalence of the proposed model looked at over the current models.
From these examinations, we can see that SVM has been generally utilized for peculiarity discovery. All the more explicitly, specialists improved SVM by consolidating it with other cutting edge techniques and applying alterations on independent SVM. In this paper, we present a half breed technique that utilizes Shared Data for include choice and manufactures an inconsistency identification model by utilizing improved OCSVM on the apache flash enormous information stage.

III. PROPOSED METHODOLOGY

In this section, we describe the proposed method and the tools and techniques used in this method. Our proposed method is based on OCSVM, and it uses MI, which is one of the best feature selection methods [13]. Figure 1 shows the proposed method. The steps of the proposed method can be summarized as follows:

1. Platform configuration and loading data
2. Preprocessing
3. Feature selection
4. Train modified OCSVM with the training dataset
5. Test and evaluate the model

Apache spark

In this work, we use Apache Spark [14], an open-source distributed cluster computing platform developed for big data processing. Spark uses a multi-staged in-memory processing scheme, which results in 100 times faster processing than map-reduce processing [15]. In particular, we use Spark's stand-alone cluster mode.

Preprocessing

In the preprocessing phase of this work—due to different types of labels in different data sets— we first standardized the labels of the records for the application of this research. Then, labels of the records were saved separately for evaluation. Finally, for better performance of modeling, data was normalized, ranging from zero to one by using a logarithm function.

Feature selection

Feature selection is an important stage of processing, although it is largely overlooked in the field of anomaly detection [16]. In the forward feature selection method, features are selected based on maximizing MI between each data point and its label. Then, the selected features will be used for modeling.

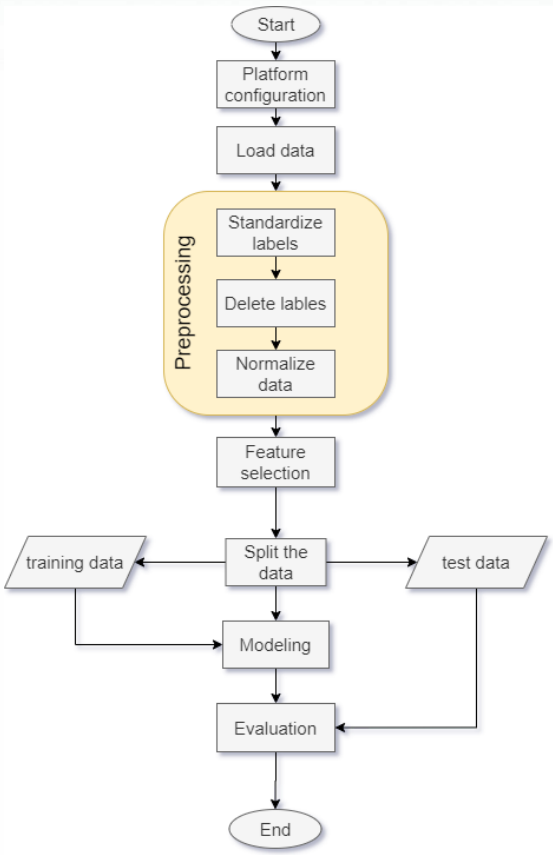

Figure 1. The sequence of the proposed method in this research.

The shared data can be utilized to ascertain any self-assertive reliance between irregular factors in the data hypothesis. Think about two irregular factors X and Y—a proportion of the measure of information on Y provided by X or the other way around—is the MI among X and Y. The MI between these factors will be zero if X and Y are autonomous, i.e., every factor contains no data about the other. By applying highlight determination, just pertinent highlights will be utilized; consequently, abnormality recognition preparing time will be diminished, and its presentation will be improved [17].

Furthermore, we used the analysis of variations (ANOVA) feature selection for evaluation and comparison of the results. ANOVA is a technique for analyzing experimental data, measuring one or more response variables under different conditions. One or more classification variables identified these conditions [17].

Modeling

In this paper, a new kernel function is proposed for the OCSVM model. RBF and Polynomial are two classic kernel functions that are widely used for OCSVM classification. Both of these kernel functions have their characteristics. Our proposed kernel function is a combination of both the RBF and Polynomial kernels. Therefore, their abilities are applied to the new kernel function. Apache Spark distributed platform is used to implement this method. Apache Spark shortens the task executions and is appropriate for big data processing. Additionally, parameters of the proposed kernel function are

optimized by the experiment to enhance the performance of anomaly detection.

The support vector machine (SVM) is a classic machine learning method and is widely used for classification. However, the standard SVM may not be useful for anomaly detection as a special classification problem, because we only have one-class positive samples (and no negative samples) from which to learn. OCSVM [18] has solved this problem, as it maps the data into the kernel space and separates them from the origin with the maximum margin. OCSVM has been formulated as the following optimization problem:

$$min\frac{1}{2}.\|w\|^2 + C\sum_{n=1}^{N}\varepsilon_n - b$$
$$s.t.\ w^T.\varphi(x_n) \geq b - \varepsilon_n. \quad n = 1.2.\ldots.N$$
$$\varepsilon_n \geq 0. \quad n = 1.2.\ldots.N \quad (1)$$

Where C is the penalty parameter, and ξn are slack variables. Slack variables permit some errors during the training phase [1]. Figure 2 shows the geometry interpretation of the OCSVM in which filled squares stand for support vectors, filled circles and squares stand for normal data, and empty circles stand for abnormal data.

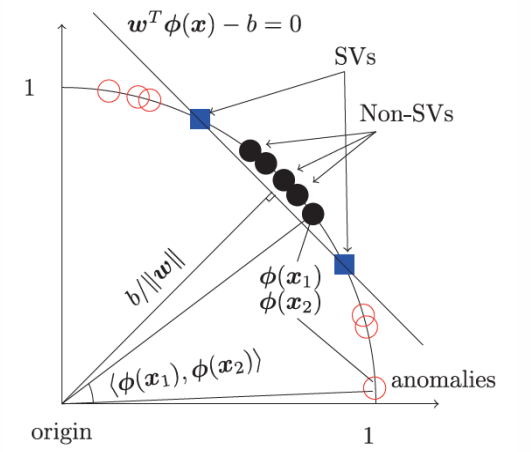

Figure 2. geometry interpretation of the OCSVM [1].

The main issue with the SVM is choosing a kernel function, and yet there is no perfect theory to manufacture the most appropriate kernel function. Nonetheless, the structure of the feature space is determined by the selection of kernel functions directly. RBF kernel is a nonlinear kernel, which is why it is often used. The only parameter of the RBF kernel is s, which is inversely correlated with learning from the data set, so the learning ability of RBF kernel increases with the decrease of s and vice versa. Polynomial is another SVM kernel function that has stronger generalization ability. This ability is increased with the decline of the order number [23]. Thinking about the learning and speculation abilities, a novel part work, which is a curved mix of two conventional portion capacities dependent on the standard of piece work development, is characterized as follows:

$$k(x.x^*) = k_{poly}(x.x^*) + k_{rbf}(x.x^*) \quad (2)$$

Where kpoly is the polynomial portion capacity, and krbf is the RBF piece work. Boundaries of the proposed portion work appear in Table 2: d is the boundary of the polynomial bit work, s is the boundary of RBF part work, C is the punishment Coefficient.

TABLE I. PROPOSED KERNEL FUNCTION PARAMETERS.

parameter	value
d	1
s	0.156
c	16

As it appears in the proposed piece work, the impact of each base part work is equivalent to one another.

As it is evident in the proposed strategy portion work, the computational unpredictability of this technique is equivalent to polynomial and RBF piece capacities' computational intricacy. Also, the computational multifaceted nature of these portion capacities has an immediate connection with measurements. In this way, if the components of the dataset decline, the computational unpredictability of the piece work diminishes. Since in the proposed technique, the component decrease has been applied to the dataset before displaying, in this way the computational multifaceted nature of the proposed strategy is diminished in contrast and the strategies without include decrease.

Contribution

The main contributions of this paper are listed as follows:

a) MI has been used for selecting related features in this method. Therefore, this method can be applied to high dimensional big data.

b) The kernel function of proposed OCSVM is a combination of two classical functions: RBF and Polynomial. Parameters of the kernel function have been optimized by experiment on the NSL-KDD data set. Therefore, the performance of the anomaly detection has been improved, which is shown by experimental results on different data sets.

c) This method has been made efficient for big data by using Spark, which is a distributed platform to execute many tasks in a short amount of time.

IV. RESULTS AND DISCUSSION

This section shows the results of the proposed anomaly detection method. The proposed method was implemented in python programming in apache spark using Databricks [19]. Besides NSL-KDD, three other data sets have been used for evaluation and comparison. The details of all data sets used in this research are provided in Table 1. The proposed method is compared with three other methods on these data sets in terms of Accuracy, Precision, Recall, and F-score.

Dataset

The KDD99 informational collection was made for the KDD Cup Challenge in 1999, and it was one of the most broadly utilized informational collections for cybersecurity research utilizing information mining strategies [25]. Tavallaee et al. [26] investigated the whole KDD dataset and inferred that the assessments of oddity recognition approaches have been inferior quality. Consequently, they proposed the NLS-KDD informational collection dependent on the past KDD informational collection. The NSL-KDD information has the accompanying enhancements:

- Redundant records are excluded from the preparation set. This guarantees classifiers won't be one-sided towards more successive records.
- Duplicate records are avoided in the testing set; subsequently, the presentation of the learning period of the AI calculations is improved since each test record is assessed just a single time. Additionally, the arrangement inclination brought about by successive records is dispensed with.
- Selected records from every trouble level gathering are conversely corresponding to the level of records in the first KDD informational index. Thusly, the presentation of the distinctive AI techniques shifts generally, which makes it more effective in exact assessment of the diverse learning strategies.
- There is a sensible number of records in the preparation and testing sets, which makes it moderate to run the trials on the total set without regularly choosing a little example of the informational collection haphazardly. Therefore, the assessment consequences of the work done by various scientists will be predictable and similar.

In spite of the fact that the NSL-KDD informational index has a few issues, it is an extremely helpful informational index that can be utilized for research purposes [25]. In this paper, notwithstanding NSL-KDD, three more informational collections are utilized for assessment and examination. The depiction of the informational collections utilized in this exploration is appeared in Table 3. These datasets are considered as large information as per their previous use [27-30]. It ought to be referenced that 70 percent of the records in every informational collection is utilized for the preparation stage, and the stay 30 percent is utilized for the testing stage.

TABLE II. DESCRIPTION OF DATA SETS WHICH ARE USED IN THIS RESEARCH.

data set	no. of records	no. of features	anomalies (%)
NSL-KDD	4898431	36	32
Shuttle	58000	9	7
Mnist	70000	100	9.2
Cover	286048	10	0.9

Results

Dong et al. [20] have provided results of implementing three OCSVM anomaly detection methods that differ in their kernel function on the NSL-KDD data set. Figure 3 compares the results of the proposed method with the other three methods based on accuracy, precision, recall, and f-score. It is observed from Figure 3 that the improvement in accuracy and precision is greater than other evaluation measures. The results in f-score and recall rates are slightly better, except for recall in anomaly detection with RBF kernel function, which is the highest.

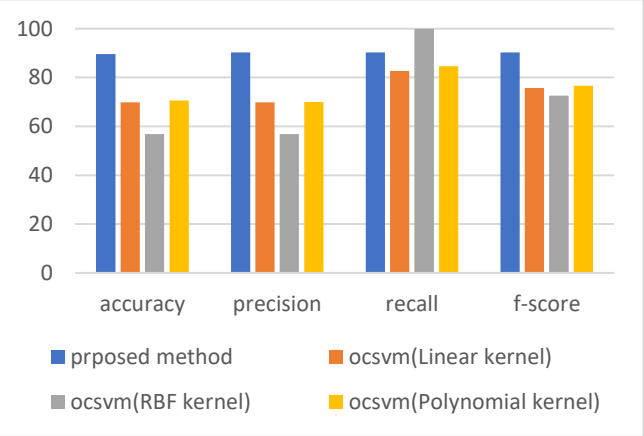

Figure 3. Compare the proposed method with other methods [20] on NSL-KDD.

In spite The results of the experiment method are illustrated in Table 3, along with other plans that are implemented for comparison. Table 3 compares the result of implementing OCSVM anomaly detection with Polynomial kernel and no feature selection, proposed kernel and no feature selection, the proposed kernel with ANOVA feature selection, and the proposed hybrid method which selects relative features using MI based on Accuracy, Precision, Recall and F-score.

According to Table 3, anomaly detection in the Cover data set has the highest performance among all data sets. This can be explained by the abnormally low rate of its training data set, supposedly. Generally, the order of these methods based on the best results obtained on these four data sets is the proposed hybrid method, proposed kernel function with ANOVA feature selection, proposed kernel function stand-alone, and polynomial kernel function without feature selection.

TABLE III. RESULTS OF VARIOUS METHODS.

	Proposed method without feature selection				Proposed method				The proposed method with ANOVA				Polynomial kernel without feature selection			
	Acc	Pre	Rec	F	Acc	Pre	Rec	F	Acc	Pre	Rec	F	Acc	Pre	Rec	F
NSL-KDD	80.37	71	90.37	79.86	89.65	90.27	90.33	90.3	85.08	85.71	86.37	86.04	70.57	69.97	84.62	76.6
Shuttle	85.43	92.41	91.87	92.14	86.06	92.48	92.54	92.51	85.3	92.18	91.95	92.06	-	-	-	-
Mnist	91.71	95.94	94.97	95.45	95.06	96.07	98.63	97.33	94.54	95.29	98.83	97.03	89.15	96	93.96	93.96
Cover	99.12	99.56	99.55	99.55	99.11	99.54	99.56	99.55	99.02	99.48	99.52	99.5	99.1	99.52	99.56	99.54

Figure 4 illustrates the average of all data sets used in this research in terms of Accuracy, Precision, Recall, and F-score. The results of the experiment showed that the proposed method has high performance in anomaly detection.

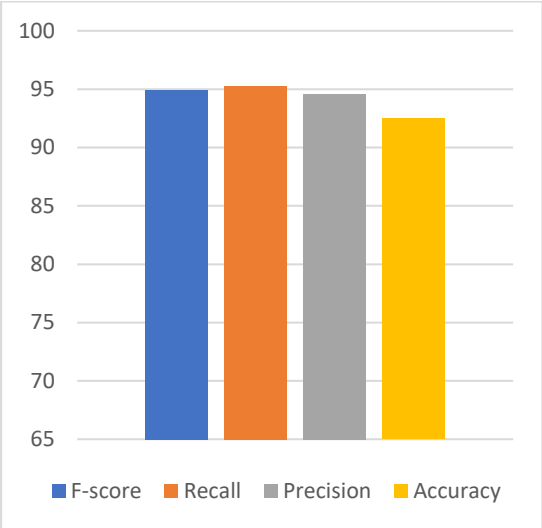

Figure 4. Average of evaluation measures in the proposed method on all data sets.

V. CONCLUSION

This paper proposes a new hybrid method for anomaly detection based on OCSVM that can deal with Big Data by using the Spark platform. High dimensionality in Big Data makes the anomaly detection process more complex. Therefore, in the proposed method, MI is used to select relative features and used them for modeling. Additionally, a new kernel function is proposed in this method for improving anomaly detection performance. The new kernel function is a convex combination of two traditional kernel functions based on the principle of kernel function construction; the new kernel function has characteristics of its consistent functions. Experimental results on the NSL-KDD data set have optimized parameters of the proposed kernel function. The results show that the proposed hybrid method has higher performance in anomaly detection compared to similar methods.

In future work, we will replace the feature selection method with other methods and compare the results. Besides, we will extend the model to a multi-class model that has several normal classes and one abnormal class WSNs.

REFERENCES

[1] Miao, X., Y. Liu, H. Zhao, and C. Li, "Distributed Online One-Class Support Vector Machine for Anomaly Detection Over Networks". IEEE Transactions on Cybernetics, 2018(99): p. 1-14.

[2] Tian, Y., M. Mirzabagheri, S.M.H. Bamakan, H. Wang, and Q. Qu, "*Ramp loss one-class support vector machine; a robust and effective approach to anomaly detection problems*". Neurocomputing, 2018. 310: p. 223-235.

[3] Wang, Y. and V. Ng. "*Anomaly Detection with Attribute Conflict Identification in Bank Customer Data*". in *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*. 2017. IEEE.

[4] Tran, P.H., K.P. Tran, T.T. Huong, C. Heuchenne, P. HienTran, and T.M.H. Le. "*Real Time Data-Driven Approaches for Credit Card Fraud Detection*". in *Proceedings of the 2018 International Conference on E-Business and Applications*. 2018. ACM.

[5] Alghussein, I., W.M. Aly, and M.A. El-Nasr, "*Anomaly detection using Hadoop and MapReduce technique in cloud with sensor data*". Int. J. Comput. Appl, 2015. 125: p. 22-26.

[6] Fontugne, R., J. Mazel, and K. Fukuda. "*Hashdoop: A MapReduce framework for network anomaly detection*". in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. 2014. IEEE.

[7] Akyildiz, Tran, K.P. and T.T. Huong. "*Data driven hyperparameter optimization of one-class support vector machines for anomaly detection in wireless sensor networks*". in *2017 International Conference on Advanced Technologies for Communications (ATC)*. 2017. IEEE.

[8] Amraee, S., A. Vafaei, K. Jamshidi, and P. Adibi, "*Abnormal event detection in crowded scenes using one-class SVM*". Signal, Image and Video Processing, 2018. **12**(6): p. 1115-1123.

[9] Amer, M., M. Goldstein, and S. Abdennadher. "*Enhancing one-class support vector machines for unsupervised anomaly detection*". in *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*. 2013. ACM.

[10] Lichman, M., *UCI machine learning repository*. 2013, Irvine, CA.

[11] Erfani, S.M., S. Rajasegarar, S. Karunasekera, and C. Leckie, "*High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning*". Pattern Recognition, 2016. **58**: p. 121-134.

[12] Chalapathy, R., A.K. Menon, and S. Chawla, "*Anomaly detection using one-class neural networks*". arXiv preprint arXiv:1802.06360, 2018.

[13] Amiri, F., M.R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "*Mutual information-based feature selection for*

intrusion detection systems". Journal of Network and Computer Applications, 2011. 34(4): p. 1184-1199.
[14] *Apache Spark*. 2019 [16 December 2019]; Available from: https://spark.apache.org/.
[15] Kulariya, M., P. Saraf, R. Ranjan, and G.P. Gupta. "*Performance analysis of network intrusion detection schemes using Apache Spark"*. in *2016 International Conference on Communication and Signal Processing (ICCSP)*. 2016. IEEE.
[16] Pascoal, C., M.R. De Oliveira, R. Valadas, P. Filzmoser, P. Salvador, and A. Pacheco. "*Robust feature selection and robust PCA for Internet traffic anomaly detection"*. in *2012 Proceedings IEEE INFOCOM*. 2012. IEEE.
[17] Sheikhan, M., M. Bejani, and D. Gharavian, "*Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method"*. Neural Computing and Applications, 2013. 23(1): p. 215-227.
[18] Schölkopf, B., J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson, "*Estimating the support of a high-dimensional distribution"*. Neural computation, 2001. 13(7): p. 1443-1471.
[19] *Databricks*. 2019 [10 January 2020]; Available from: https://databricks.com/.
[20] Dong, G. and S.K. Pentukar, "*OCLEP+: One-class Anomaly and Intrusion Detection Using Minimal Length of Emerging Patterns"*. arXiv preprint arXiv:1811.09842, 2018.

Masood Harimi received his M.Sc. degree from of Computer Engineering Department at the University of Science and Culture, Tehran, Iran. His research interests include Data Science and Distributed Systems.

Mohammad Javad Shayegan is an Associate Professor in Computer Engineering Department at the University of Science and Culture, Tehran, Iran. He is the founder of the Web Research Center, International Conference on Web Research, and International Journal of Web Research in Iran. His research interests include Web Research, Data Science and Distributed Systems.