# Application of Local Outlier Factor Algorithm to Detect Anomalies in Computer Network

Juozas Auskalnis[1], Nerijus Paulauskas[1,2], Algirdas Baskys[1,2]
*[1]Department of Computer Science and Communications Technologies, Vilnius Gediminas Technical University,*
*Naugarduko St. 41, LT-03227 Vilnius, Lithuania*
*[2]Center for Physical Sciences and Technology,*
*Sauletekio al. 3, LT-10257 Vilnius, Lithuania*
*juozas.auskalnis@vgtu.lt*

***Abstract*—Gap between the new attack appearance and signature creation for this attack may be critical. During this time, many computer systems may be affected and valuable resources may be lost. Even after signature creation, many computer systems still stay vulnerable because of bad security practice, i.e. patches and updates are not installed as needed. Therefore, anomaly intrusion detection system (IDS) that is capable to detect new unknown attacks is valuable security tool. This paper analyses the use of Local Outlier Factor (LOF) to detect anomalies in the computer network. The application of the LOF algorithm for the detection of anomalies when only normal network data are used for the model training has been demonstrated. Experimental results of different threshold values influence on the anomaly detection accuracy using NSL-KDD dataset is presented.**

***Index Terms*—Intrusion detection; Anomaly detection; Local outlier factor.**

## I. INTRODUCTION

The growth of the cyber-attacks over computer networks makes difficulties for companies to secure information system resources and to ensure business continuity. This arise the need of means allowing companies to mitigate the security threats and lower possible risks. An intrusion detection system (IDS) is powerful tool that examines system or network activity to identify and alert about malicious activity. There are two main approaches, which are used for the intrusion detection: signature-based and anomaly-based. Signature-based IDS can only detect previously known attacks that have a corresponding signature in the database. For the each new type of attack the signature database has to be updated. The main drawback of signature-based approach is that it fail to detect new, earlier unseen attacks or even variants of known but modified attacks. Anomaly-based IDS search for the attacks, which deviates from created profile of normal activities. Events that exceed specified threshold values are reported as anomalies or attacks. The main drawback of anomaly-based IDS is high false alarm rate.

This paper analyses the use of the Local Outlier Factor (LOF) to detect anomalies in computer network. Experimental results of different threshold values influence on anomaly detection accuracy using NSL-KDD data set is presented. NSL-KDD data set is a refined version of its predecessor KDD CUP 99 data set and is used for evaluation of network-based intrusion detection systems [1].

## II. RELATED WORKS

D. Pokrajac *et al.* in their paper [2] propose an incremental Local Outlier Factor algorithm, which is appropriate for the detecting of outliers in the data streams. The goal of the incremental LOF algorithm is to get results equivalent to the static LOF algorithm. Every time a new point is inserted into a data set in such a way that the asymptotic time complexity of incremental LOF algorithm would be comparable to the static LOF algorithm.

Experiments were performed on several synthetic data (DARPA98 data set [3]) and real life data sets. The obtained performance was the same as iterated static LOF algorithm, however authors admit that performance crucially depends on efficient indexing structures to support *k*-nearest neighbour (kNN) and reverse *k*-nearest neighbour queries. Authors mention that proposed incremental LOF is not applicable when the data have large number of dimensions and that approximate kNN and reverse kNN algorithms might improve the applicability of proposed algorithm with multidimensional data as well.

Adaptive anomaly detection scheme for cloud computing based on LOF is presented by T. Huang *et al.* [3]. The experiment is oriented to computing statistics such as CPU utilization, disk I/O activity, network activity etc. Authors construct knowledge base, which is kept up to date and LOF is computed for each incoming new point according to the constructed knowledge base. Experiments show that the adaptability was enabled without compromising the performance of LOF algorithm and the overhead of adaptability is small enough to enable online surveillance. However, because the adaptability captures the smooth changes of the normal behaviour, authors admit that it also treats some anomalies based on the gradual change as normal behaviours as well (e.g. computer worm that gradually increase of the workload eventually will be ignored).

Manuscript received 14 November, 2017; accepted 11 March, 2018.

J. Zhang *et al*. investigate unsupervised techniques for the anomaly-based network intrusion detection [4], [5]. Authors used the real-time network traffic data for their experiment. The performance of the unsupervised techniques is between the performance of the LOF and Isolation Forest algorithms. Authors state that Isolation Forest performs better then LOF in identifying anomalies. However, they agree that there are limitations in their investigation, because proposed model only looked at four numeric attributes (byte, rate, packet, duration) of the data set and Isolation Forest classifies the most extreme values as outliers. In addition, only http (port 80) traffic for two static IPs was investigated.

This paper presents a novel approach to detect anomalies in computer network using Local Outlier Factor algorithm. We suggest to use for model training only normal dataset that is pre-processed with LOF algorithm to remove outliers, which influence anomaly detection performance. Then prepared normal dataset is used to detect anomalies comparing it with the new data and calculating LOF value, which is direct indicator of data abnormality. Section III describes the processes of data pre-processing and anomaly detection. Section IV outlines the importance of parameters selection for LOF algorithm and presents the results of anomaly detection accuracy. Section V concludes the main points of the paper.

## III. METHODOLOGY

Usually all data consisting of normal and attack records for the model training are used in the classification tasks, where the goal is to detect anomalies. The developed model works more accurate and generates lower amount of the false positives if the variety of data used for training is wider and contains more different attack types. The LOF model based on the normal type of data has been used in the work. In computer networks the amount of generated normal packets is significantly bigger then anomalous, hence in practice it is easier to gather data set, which defines normal network operation, rather data set with complete set of possible attack types.

LOF was originally proposed in [6]. This algorithm evaluates each event's uniqueness based on distance from the *k*-nearest neighbours. LOF algorithm is able to detect outliers regardless the data distribution, since it does not make any assumptions about the distributions of data. The main idea is that the density around an outlier object significantly differs from the density around its neighbours. LOF is an unsupervised outlier detection method. It is an advantage when the data that are analysed are not labelled or cannot be labelled due to big amount of data. This is common in computer networks, were the number of generated network packets is very high.

For our experiment, we use well-known NSL-KDD data set [1]. This data set has following advantages: it does not include redundant and duplicated records, the number of available records in the train and test data sets are reasonable, which enables to execute experiments on the complete set. NSL_KDD training data set consists of 125973 records, from which 67343 are labelled as normal and the rest of the records are labelled as attacks: denial-of-service (DOS), surveillance and other probing (PROB), unauthorized access from a remote to local host (R2L) or unauthorized access to local super user (U2R). Each record of NSL-KDD data set contains 41 main attributes (e.g., protocol type, service, flag, duration) and two additional attributes describing the type and the difficulty level of each record.

Figure 1 shows data processing and anomaly detection scheme. First step consist of splitting NSL-KDD training data into normal and attack data sets. For training, only records corresponding to normal data are used. Next, it is necessary to remove attributes 8 and 20 from the normal data set, since all values of these attributes are equal to 0, i.e. they have no predictive power. Then, the z-standardization is applied to all numerical values and the nominal attributes (2-protocol type, 3-service, and 4-flag) with binary values using dummy coding are replaced. After processing, we get 75 attributes. Before applying the LOF algorithm, it is necessary to specify two parameters: the number of closest neighbours *k* and the threshold value, above which it is considered that the record is outlier. When choosing the number of closest neighbours, it is recommended that *k* value would be equal to square root of all data, used for the model training. According to the LOF algorithm, records whose LOF value is greater than 1 are considered as outliers. For normal data processing, six threshold values are selected: $Th_{\mathrm{c}}$ (cleaning) = {1.5, 1.75, 2, 3, 5, and 10}.

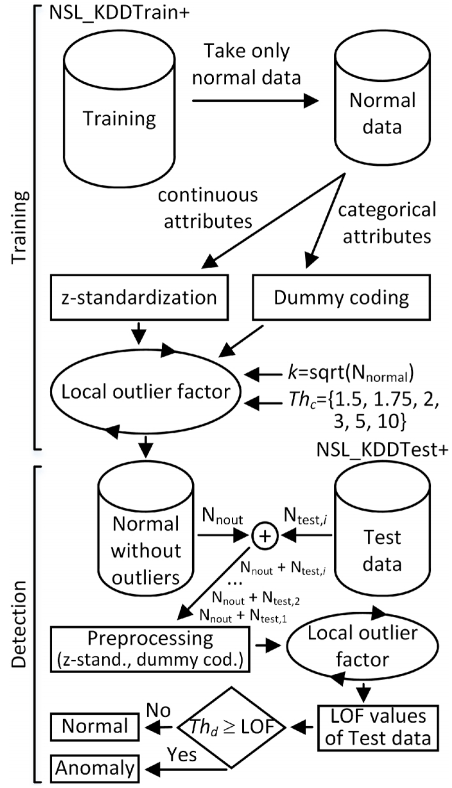

Fig. 1. Data processing and anomaly detection scheme.

The LOF algorithm with selected *k* and threshold values is applied. Based on the calculated values, the records with the LOF value higher or equal to $Th_{\mathrm{c}}$ are removed. Then the *k* value for the normal data type, from which the outliers were removed, are recalculated and again the LOF algorithm is applied. The cycle is repeated until there are no records exceeding the specified threshold value. After the training, the resulting data set is further used to detect anomalies. This dataset is accompanied by a new record from the test data set and verifies whether it is the anomalous or not. The tested record is assigned to anomaly if its LOF value exceeds the detection threshold value $Th_d$.

## IV. EXPERIMENTAL RESULTS

After training phase, six datasets were prepared that consist of normal type data with outliers removed. The number of outliers found and removed using the selected threshold values is shown in Fig. 2. The number in parentheses represents the percentage of records that have been removed from the training dataset. From the figure, it is seen that the first three data sets with threshold values of 1.5, 1.75 and 2.0 should make the highest impact on the detection results, as these data sets decrease by 25, 15 and 11 percent, respectively.

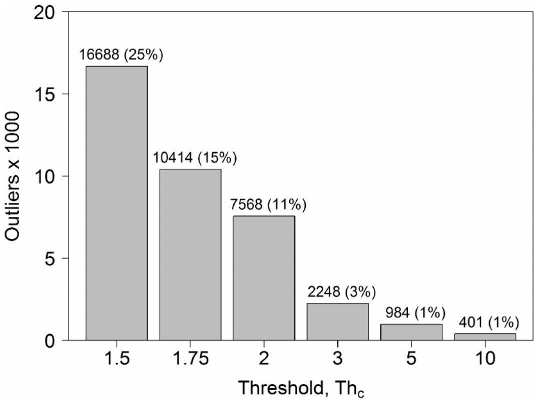

Fig. 2. The number of outliers removed from normal data set using corresponding $Th_c$ values.

The NSL-KDD test data set totals 22544 records, 9711 of which are labelled as normal, and 12833 records are assigned to one of four attack types: DOS (7458), PROB (2421), R2L (2887) or U2R (67). The calculated LOF values of the attack records are shown in Fig. 3. Based on these distributions, it can be seen how many attacks will be detected by selecting a certain threshold value. For example, it is seen that at the selected detection threshold $Th_d = 5$ just one thousand attacks will be detected, when a training normal data set with no outliers removed or removed with $Th_c \geq 3$ threshold is used. By using for the training the normal set of data, from which the outliers are removed with a threshold value lower or equal to 2, the number of detected attacks increases 4 times. It is obvious that in order to increase the number of recognizable attacks, it is necessary to reduce the threshold value for detection. However, in such a case the false positive number, which is related to the distribution of the LOF values of the test normal data in Fig. 4, increases. In this case, on the contrary, in order to reduce the number of false positives, it is necessary to increase the threshold value, above which the record will be recognized as an attack. $Th_d$ threshold value has to be such, which detects the highest number of attacks, with the least false positives. For this purpose, a graph representing the accuracy dependency on detection threshold can be used. The accuracy is a proportion that represents the number of true positives (TP) and true negatives (TN), divided by the total number of predictions: *Accuracy* = (TP+TN)/(TP + TN + FP + FN), where FP – false positive, FN – false negative [7].

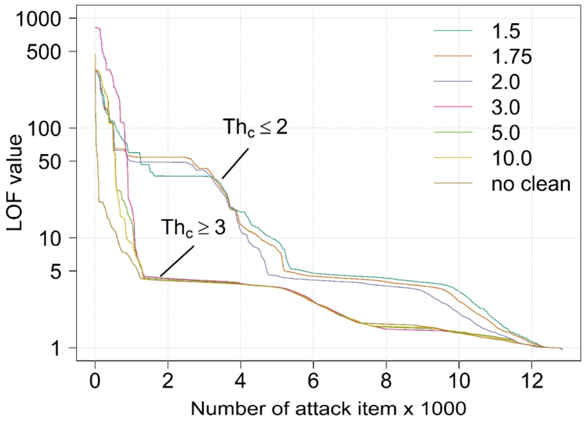

Fig. 3. LOF value distribution of test attack records using various training normal data sets.

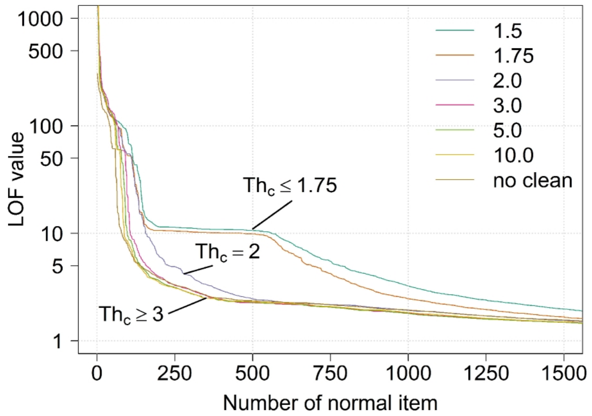

Fig. 4. LOF value distribution of test normal records using various training normal data sets.

Figure 5 shows that as the $Th_d$ increases, accuracy increases in all cases and when a detection threshold value reaches 1.4, accuracy begins to decrease when for detection normal data sets obtained with $Th_c \geq 3$ are used. This is because with higher $Th_d$ value, decreases the number of attacks that are detected, and the false positive number is still high. For detection using data sets obtained with $Th_c \leq 2$, the accuracy remains constant in the range from 1.2 to 3.7. This happens because there is a significant amount of attack detection in this range.

The number of detected attacks remains high (Fig. 3) with small false positives (Fig. 4) at low detection threshold ($Th_d > 2$). Therefore, it is recommended to use the normal data set obtained with $Th_c = 2$ in order to detect attacks. Distribution of true positive and false positive values using normal data set with $Th_c = 2$ is shown in Fig. 6. When $Th_d = 2.3$, the highest value of accuracy (0.84) is reached. Comparison of the accuracy results obtained using LOF algorithm with other Decision Trees, Naïve Bayes and Rule-Based classifiers shows that the achieved accuracy values are the same [8]. The advantage of using LOF algorithm in contrary with mentioned classifiers is that the only normal data set is used for the model training.

As stated in [9] a large set of machine learning algorithms have difficulties detecting U2R type of attacks. In our situation with chosen $Th_d$ = 2.3, all seven U2R types of attack from test data set are detected having different recall values (Table I). The recall is the proportion of attacks detected over the total amount of attacks: *Recall* = TP/(TP + FN) [7].

TABLE I. U2R ATTACK DETECTION RESULTS.

| U2R attack | Detected (TP) | Missed (FN) | Sum | % (Recall) |
|---|---|---|---|---|
| buffer_overflow | 13 | 7 | 20 | 65 |
| loadmodule | 1 | 1 | 2 | 50 |
| perl | 2 | 0 | 2 | 100 |
| ps | 12 | 3 | 15 | 80 |
| rootkit | 12 | 1 | 13 | 92.3 |
| sqlattack | 2 | 0 | 2 | 100 |
| xterm | 8 | 5 | 13 | 61.5 |

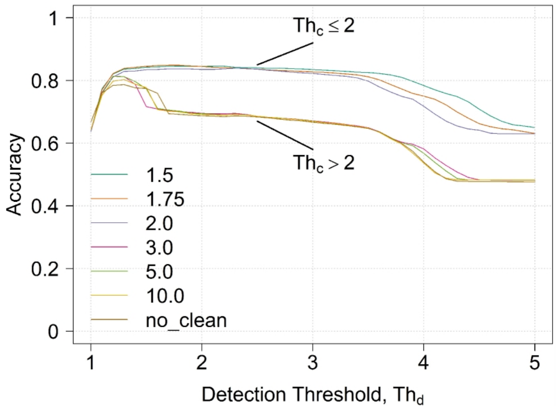

Fig. 5. The dependence of accuracy on the detection threshold $Th_d$.

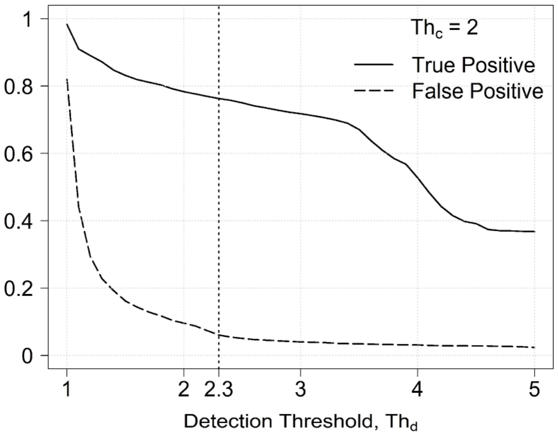

Fig. 6. The distribution of true positive and false positive values based on detection thresholds $Th_d$ when for detection training normal data with $Th_c = 2$ is used.

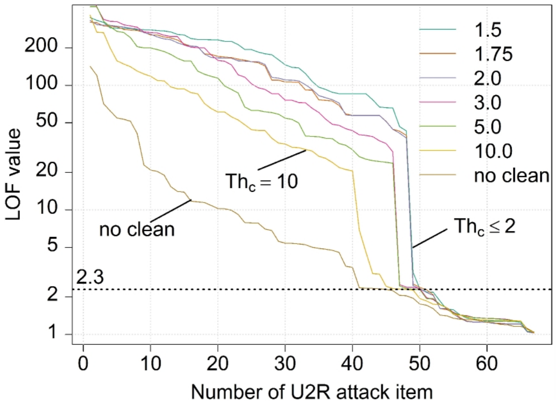

Fig. 7. LOF value distribution of U2R test attack records using various training normal data sets.

Figure 7 shows LOF value distribution of U2R test data set attacks. Depending on training normal data set used, more than 50 (75 %) of U2R attacks have high (> 20) LOF value, which suggest that records corresponding to these attacks are in low density region compared with normal records and is well distinguished by LOF algorithm.

## V. CONCLUSIONS

In this paper, the application results of Local Outlier Factor algorithm to detect anomalies in computer network are presented. The accuracy of anomaly detection highly depends on the training data set preparation and detection thresholds. The highest accuracy (0.84) is achieved when anomaly detection threshold $Th_d = 2.3$ and normal data set with outlier cleaning threshold $Th_c = 2$ is used.

The outlier cleaning process during data preparation step is important because it allows to exclude normal records, which may intersect with density location of anomaly records and affect LOF value. If the selected cleaning threshold is too small ($Th_c \leq 1.75$ in this case), more records are removed and the density location of normal records decreases. This implies, that more normal records fall into low density region (have higher LOF value) and are detected as attacks. If the cleaning is not used or the cleaning threshold is too high, more attacks of which LOF value is close to detection threshold will fall in to density region of normal records and will not be detected as anomalies.

## VI. REFERENCES

[1] M. Tavallaee, E. Bagheri, W. Lu, A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", *Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA 2009)*, 2009. DOI: 10.1109/CISDA.2009.5356528.

[2] D. Pokrajac, A. Lazarevic, L. J. Latecki, "Incremental local outlier detection for data streams", in *Proc. 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)*, Honolulu, 2007, pp. 504–515. DOI: 10.1109/CIDM.2007.368917.

[3] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. P. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, M. A. Zissman, "Evaluating intrusion detection systems: the 1998 DARPA offline intrusion detection evaluation", in *Proc. DARPA Information Survivability Conf. and Exposition (DISCEX 2000)*, Los Alamitos, 2000, pp. 12–26. DOI: 10.1109/DISCEX.2000.821506.

[4] J. Zhang, K. Jones, T. Song, H. Kang, D. E. Brown, "Comparing unsupervised learning approaches to detect network intrusion using NetFlow data", *Systems and Information Engineering Design Symposium (SIEDS 2017)*, Charlottesville, 2017, pp. 122–127. DOI: 10.1109/SIEDS.2017.7937701.

[5] T. Huang, Ya. Zhu, Q. Zhang, Yo. Zhu, D. Wang, "An LOF-based adaptive anomaly detection scheme for cloud computing", *IEEE 37th Annual Computer Software and Applications Conf. Workshops*, Japan, 2013, pp. 206–211. DOI: 10.1109/COMPSACW.2013.28.

[6] M. M. Breuning, H. P. Kriegel, R. T. Ng, J. Sander, "LOF: identyfying density-based local outliers", in *Proc. Int. Conf. Management of Data (ACM SIGMOD 2000)*, Dallas, TX, 2000, pp. 93–104. DOI: 10.1145/342009.335388.

[7] M. H. Bhuyan, D. K. Bhattacharyya, J. K. Kalita, "Network anomaly detection: Methods, systems and tools", *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303–336, 2014. DOI: 10.1109/SURV.2013.052213.00046.

[8] N. Paulauskas, J. Auskalnis, "Analysis of data pre-processing influence on intrusion detection using NSL-KDD dataset", in *Proc. Open Conf. Electrical, Electronic and Information Sciences (eStream 2017)*, 2017, p. 1–5. DOI: 10.1109/eStream.2017.7950325.

[9] M. Sabhnani, G. Serpen, "Why machine learning algorithms fail in misuse detection on KDD intrusion detection data set", *Intelligent Data Analysis*, vol. 8, no. 4, pp. 403–415, 2004.