A Clustering-Based Unsupervised Approach to Anomaly Intrusion Detection

Evgeniya Nikolova
Faculty for Computer Science and Engineering
Burgas Free University
Burgas, Bulgaria
enikolova@bfu.bg

Veselina Jecheva
Faculty for Computer Science and Engineering
Burgas Free University
Burgas, Bulgaria
vessi@bfu.bg

Abstract— **In the present paper a 2-means clustering-based anomaly detection technique is proposed. The presented method parses the set of training data, consisting of normal and anomaly data, and separates the data into two clusters. Each cluster is represented by its centroid - one of the normal observations, and the other - for the anomalies. The paper also provides appropriate methods for clustering, training and detection of attacks. The performance of the presented methodology is evaluated by the following methods: Recall, Precision and F1-measure. Measurements of performance are executed with Dunn index and Davies-Bouldin index.**

Keywords- anomaly based IDS, 2-means clustering, Recall, Precision, F_1 measure, Dunn index, Davies-Bouldin index

I. INTRODUCTION

Intrusion detection is a very important and critical area of research and practice, since it is relevant to various business and non-profit organizations. Intrusion detection relies on the assumption that intrusive activities are notably different from normal system activities and therefore detectable [3]. There are two major detection methods, according to the intrusion detection technique: misuse detection, based on the search of preliminarily collected signature database in the current system activity; and anomaly detection, which is based on behavioral analysis. It relies on created profiles for the legal users on the system, which describe the acceptable user behavior. During the current system activity the intrusion detection system (IDS) scans the network and host audit data and looks for significant deviations from the baseline, which describes normal user activity.

The most essential advantage of the anomaly based IDS is their potential to reveal previously unknown exploits and attacks, or deflections of acceptable activities of system services. The sources of the attacks, that the anomaly based IDS could detect, can be both legal internal users or unauthorized external ones, unlike misuse IDS, which can detect only preliminarily known intrusions. Another priority of anomaly detection approaches over misused ones is the ability to automate the description of the normal activity process. Moreover, this methodology has the potential to discover unauthorized activity from insiders or account theft. Since the detection method relies on preliminarily defined and tuned-up profiles of acceptable user activity, the attacker could not be aware previously what exact action would trigger the an alarm, since the performed actions would not conform to the normal user activity profiles. As a drawback of the approach could be considered the high rate of false alarms, compared to those of misused based approach, since the normal activity description and the determination of deviations from the profiles is not a trivial task.

Analyzing the sequences of system calls is a wide-spread approach in host-based anomaly intrusion detection ([6], [14], [2]). The system calls, performed by a critical process, during some period of time, are a dependable classifier, which can be utilized to separate the audit data into two sets, respectively for normal and abnormal behavior. The purpose is to monitor system activity and find out sequences that do not comply with the preliminarily defined profiles of normal behavior.

Clustering is an important technique, which could be applied in many research topics, in particular in the field of the intrusion detection. During the past years various supervised learning methodologies have been applied to anomaly detection ([9], [13]). It is a methodology of separation of the examined elements into non-intersecting groups, according to a preliminarily chosen feature. The separation is executed so that the elements from the same set are similar to each other, and the elements from different sets are quite different from each other, according to the selected feature. Consequently, the clustering methodology can be helpful for classifying system current user data and detecting deviations from normal user activity, which is the basis of the anomaly detection approach.

II. RELATED WORK

There are many previous publications in the examined area. Anomaly detection technique with fast stepwise clustering method (ADWIC) is presented in [1]. This algorithm use advanced BIRCH [16] algorithm, which performs clustering with fast, scalable and adaptive anomaly detection scheme. In this approach, the clustering is applied in order to train the normal user activity model.

In [15] Yang et al. propose a method for anomaly detection, based on clustering and classification of the examined audit data. Clustering is performed for grouping the examined data patterns into clusters. Then some clusters are selected as normal or anomalous according to previously selected criterion. Those examined data, extracted from the profile, are used to define a specific classifier. The authors apply an influence-based classification method in order to classify the current network data.

In [7] a clustering based supervised anomaly detection technique is proposed. The set of current data patterns, containing only normal data, is divided into clusters. Each cluster is represented by its profile and the normal model is created as a result. Any significant variation from the described normal activity model is considered as an intrusive activity.

The present paper proposes an unsupervised method of anomaly detection, based on 2-means clustering algorithm. The main purpose is to create and test a methodology for the anomalous activity discovery during the regular system activities. The major goal is to distinguish the regular user activity sequences from intrusive ones, so the intrusion detection process is considered as a binary classification problem and a 2-means clustering approach is applied. The proposed method works on unlabeled data and does not need any previous knowledge about the current activity data in order to detect new intrusions.

III. Description of the Methodology

A. Detection method

The distances between the observed sequence and each cluster centroid cluster are calculated using the Wagner-Fischer distance (WFD). An object is considered as normal, if the distance to the centroid of the normal cluster data is less than to the anomalous one, and vice versa. In our case with two clusters: if the observation *P* is closer to the normal cluster, therefore *P* is normal. This distance-based classification algorithm can detect arbitrary kinds of anomalies without any previous knowledge about them.

B. Wagner-Fischer distance

Wagner-Fischer distance (*WFD*) [12] is a string metric between two sequences, which calculates the minimum number of operations (an insertion, deletion, substitution of a single character, a transposition of two characters) needed to transform one sequence into the other. Let the weighting for the cost of transforming symbol *a* into symbol *b* be denoted by $w(a,b)$. Then $w(a,b)$ is the cost of a symbol substitution $a \to b, w(a,\varepsilon)$ is the cost of deleting *a* and $w(a,\varepsilon)$ is the cost of inserting *b*. The WFD are computed using the following recurrence relation:

$$d_{WF}(i,j) = \min\begin{Bmatrix} d(i-1,j)+w(x_i,\varepsilon), d(i,j-1)+w(\varepsilon,y_j), \\ d(i-1,j-1)+w(x_i,y_j) \end{Bmatrix}.$$

It calculates the cost of the optimal string alignment, which does not equal the edit distance. The cost of the optimal string alignment is the number of edit operations needed to make the strings equal under the condition that no substring is edited more than once. This value is referred to as *restricted edit distance*.

C. Proposed clustering algorithm

K-means clustering [8] is an unsupervised algorithm of cluster analysis, which executes objects grouping into *K* disjoint clusters based on the distance function. In our case, the purpose is to perform a binary classification, so we want to divide them into two classes, one of the normal data, and the other – from the anomalies. The algorithm in this case consists of the following steps:

- Two arbitrary different objects for centers – one of the normal observations, and the other – from the anomalies, are selected.
- When all observations are classified in their closest clusters, the centers of clusters are recalculated. We determine the j new center by

$$\xi_j = \arg\min_{\xi} \sum_{i:\pi_i = j} d(x_i, \xi),$$

where $\pi_i = \arg\min_j d(x_i, \xi_j)$, $d(,)$ - the measure of the distance between two vectors, in this case – *WFD*.

- Step 3 is repeated until the exact center of each cluster is found.

Clustering is carried out for the observation, which was extracted from the current audit data and has to be determined as normal or anomalous. The sequences, which are not inserted in the first cluster, containing normal activity patterns, are considered as anomalies, and those, which are inserted, are treated as normal.

IV. Simulation Experiments and Results

A. Dataset description

With purpose to test the proposed methodology, considerable amount of simulation experiments were accomplished. The experimental data were collected from the Computer Immune Systems Project, which was performed by the researchers from the Computer Science Department, University of New Mexico [11].

The data were collected from different systems examination during some period of time. They contain system calls executed by active processes, which include different kinds of programs, running with administrative rights, as well as different types of intrusions. The privileged processes are of special interest of the attacker as they perform some services which require access to system resources that are inaccessible to ordinary users. The examined data consist of normal user activity patterns of some privileged processes, as well some anomalous data. The methods for pattern generation are described in [6].

Each pattern is a sequence of system calls, which are the results of the examined process. The text files with input data represent sequences of ordered pairs of numbers. Each line contains one pair, where the first number is the process ID (PID) of the executed process, and the second one in each pair is the system call number.

The experimental data include traces of user activity during some period of time. The described methodology was applied with purpose to separate the acceptable user activity

from the inacceptable one for the following privileged processes: synthetic sendmail, login, inetd, named. The examined data contain respectively 297, 705, 308 and 610 sequences of system calls for the enumerated processes.

B. Performance Measures

Evaluation of the performance of the proposed model is based on the number of data sequences correctly and incorrectly predicted by the model – four points of concern. These four points are true positives *TP* (the IDS correctly identifies an intrusion attempt), false positives *FP* (the IDS consideres normal activity as intrusive), true negatives *TN* (the IDS correctly identifies normal activity) and false negatives *FN* (the IDS fails to detect an intrusive activity).

Recall and Precision are two commonly used measures in evaluating quality of classifiers. They are defined as follows:

$$\text{Pr } ecision = \frac{TP}{TP + FP},$$
$$\text{Re } call = \frac{TP}{TP + FN}.$$

The F_1 is a measure of a classification accuracy, which summarizes the measures Precision and Recall into single indicator [10]. F_1 measure is defined as follows:

$$F_1 = 2.\frac{\text{Pr}ecision\text{Re}call}{\text{Pr}ecision + \text{Re}call}$$

If F_1-measure achieves high value it provides that both precision and recall are reasonably high. It is one of the measures of quality of a cluster algorithm using external criterion.

C. Cluster validity assessment

When analyzing the cluster, it is natural to assume that the cluster with a greater number of cluster vectors comprising a sequence of normal activities, while the other consists of anomalies. It can also assume that vectors in one cluster are close to each other. But in the case of large-scale attacks could be seen that more vectors are generated by an abnormality of the normal vectors. Therefore, the structures of the clusters must be analyzed in order to reach better classification. For this purpose, the size and the distance between the groups should be calculated, as well as the cluster compactness. The compactness is used to describe similarities between objects in the same class. As a measure for cluster compactness is used the intra-cluster distance. It is small, when the objects are close to their cluster-centroids and it increases, if the number of clusters decreases.

The divisibility measure submit an estimate of distances between the clusters. As a measure for cluster divisiveness is used inter-cluster distance. One method to calculate it is to find the shortest distance between two observations belonging to two different clusters. Higher the inter cluster distance indicates much better remoteness of the centers of the clusters. If it is small, there are a few large clusters.

Some of the methods of the compactness validation of the clusters and the distances between them are:

- *Dunn index*. It is defined by dividing the minimal inter-cluster distance and maximal intra-cluster distance. The Dunn index takes values from the interval $[0,\infty)$. Its higher value shows better clustering.
- *Davies-Bouldin index* [4] takes into account both the error caused by representing the data vectors with cluster centroids and the distance between clusters. It is defined as:

$$DB(K) = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left\{\frac{\Delta(K_i)+\Delta(K_j)}{\delta(K_i,K_j)}\right\}$$

where n is the number of clusters, $\Delta(K_i)$ - intra-cluster distance, $\delta(K_i,K_j)$ - inter-cluster distance. Small values of Davies-Bouldin index correspond to compact clusters whose centers are distant from each other. Its main difference from Dunn index is that it considers the average case by using the average error of each cluster.

D. Experimental Results

Table 1 contains the obtained values of F_1 measure for the examined data.

TABLE I. F_1 MEASURE FOR THE EXAMINED DATA

Processes	F_1-measure
Synthetic sendmail	0,610
Named	0,875
Login	0,861
Inetd	0,677

Table 2 summarizes the effects of the *WFD* on the calculation of the Dunn and Davies-Bouldin cluster validation indices for the examined processes.

TABLE II. DUNN AND DAVIES-BOULDIN INDICES FOR THE EXAMINED DATA

Processes	Dunn index	Davies-Bouldin index
Synthetic sendmail	1,733	3,921
Named	1,696	3,584
Login	1,690	3,590
Inetd	1,685	3,587

Analyzing the data, represented in Tables 1 and 2, we can conclude that the presented methodology gives good classification results and stable parameters of the result clusters. The results for all examined data sets are similar and represent reliable and firm quality of the clustering.

V. CONCLUSIONS

IDS play a major role in the field of information security as they can reveal intrusions and other malicious activities that are not detected by the prevention systems. This article presented a 2-means clustering-based unsupervised methodology for anomaly-based intrusion detection. In the

proposed system no manually or otherwise classified data is necessary for initial training. According to the obtained results, the proposed algorithm achieves qualitative level of clustering quality and demonstrates the reliability of the approach. Future work will concentrate on improving the results using different distance metrics and comparing algorithms implementation. It also could contain experiments with different experiment data with different structure, which could examine the proposed method at different level of system monitoring.

ACKNOWLEDGEMENT

The paper was supported under project 6/2013 of Burgas Free University.

REFERENCES

[1] Burbeck K., S. Nadjm-Tehrani, ADWICE: Anomaly Detection with Real time Incremental Clustering, Proceedings of 7th International Conference on Information Security and Cryptology (ICISC 04), Springer Verlag, December 2004.

[2] Call A., Review of Database Intrusion Detection Methodologies using Attribute Dependence, Technical Report #TR-20130606-1, March 2013.

[3] Chebrolu, S.; Abraham A.; Thomas, J. P., Feature deduction and ensemble design of intrusion detection systems, Computers & Security, Volume 24, Issue 4, 1 June 2005, pp. 295-307.

[4] Davies, D.L., Bouldin, D.W., (2000) A cluster separation measure, IEEE Trans. Pattern Anal. Machine Intell., 1(4), 1979, pp. 224-227.

[5] Dunn, J. (1974) Well separated clusters and optimal fuzzy partitions, Journal of Cybernetics , 4, 1974, 95-104.

[6] Forrest S., S.A. Hofmeyr, A. Somayaji, Intrusion detection using sequences of system calls, Journal of Computer Security Vol. 6, 1998, pp. 151-180.

[7] Gogoi Prasanta, B. Borah, D. K. Bhattacharyya, Supervised Anomaly Detection using Clusteringbased Normal Behaviour Modeling, International Journal of Advances in Engineering Sciences, Vol.1, Issue 1, Jan, 2011, pp. 12-17

[8] MacQueen J., Some methods for classification and analysis of multivariate observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 1967, pp. 281-297.

[9] Portnoy L., E. Eskin , S. Stolfo, Intrusion Detection with Unlabeled Data Using Clustering, In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001), pp.5-8.

[10] Tan Pang-Ning, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining", Pearson Education, Inc., 2009.

[11] University of New Mexico's Computer Immune Systems Project, http://www.cs.unm.edu/~immsec/systemcalls.htm.

[12] Wagner R. A., M. J. Fischer, The string-to-string correction problem, Journal of the Association for Computing Machinery 21, 1974, pp. 168-173.

[13] Wei M., L. Xia, J. Jin, C. Chen, Research of Intrusion Detection Based on Clustering Analysis, Proceedings of the 2012 International Conference on Cybernetics and Informatics, Lecture Notes in Electrical Engineering, Volume 163, 2013, pp. 1973-1979.

[14] Xu Z. et.al., A multi-module anomaly detection scheme based on system call prediction, 8th IEEE Conference on Industrial Electronics and Applications (ICIEA), Melbourne, VIC, June 2013, pp. 1376 – 1381.

[15] Yang H., F. Xie, Y. Lu, Clustering and Classification Based Anomaly Detection, L. Wang et al.(Eds.): FSKD 2006, LNAI 4223, pp. 1082-1091.

[16] Zhang T., R. Ramakrishnan and M. Livny, "BIRCH: an effective data clustering method for very large databases", SIGMOID Record 1996 ACM SIGMOID International Conference on Management of Data, 25(1996), pp. 103-14.