EVALUATION OF THE EFFECTIVENESS OF ANOMALY IDS BASED ON THE CLUSTERING ALGORITHM AND DATA MINING TECHNIQUES

V. Jecheva[*]*, Burgas Free University, vessi@bfu.bg*
E. Nikolova[*]*, Burgas Free University, enikolova@bfu.bg*

Abstract: *The purpose of this paper is to examine the feasibility of clustering-based approach to anomaly-based intrusion detection systems (IDS). The examined methodology includes a 2-means clustering algorithm with and without data mining techniques, i.e. classification trees. With purpose to evaluate the effectiveness of the methodology, Jaccard index was applied. Davies-Bouldin index, Dunn index and C-index were applied in order to compare the performance results of the two models.*

Keywords: *Anomaly based IDS, 2-means clustering, classification tree, Wagner-Fischer distance, Jaccard index, Davies-Bouldin index, Dunn index, C-index*

INTRODUCTION

Intrusion detection systems (IDS) play an important role in the network security systems. Their purpose is to protect the system by raising an alarm when an intrusive activity is found, as well as to start some proactive mechanisms. As new exploits and attacks appear every day and amount of audit data, which has to be processed, constantly increases, IDSs should develop in order to counteract them.

According to the intrusion detection method, IDS could be broadly divided into two categories: anomaly-based and misuse-based. Misuse-based IDS uses specifically known patterns, referred to as signatures, of unauthorized behavior to detect intrusions. However, it has low degree of accuracy in detecting unknown intrusions since it relies on signatures extracted by human experts [13]. The anomaly based detection relies on preliminarily made description of acceptable and inacceptable user behavior. The network behavior is in accordance with the predefined behavior, then it is accepted or else it triggers the event in the anomaly detection. The accepted network behavior is prepared or learned by the specifications of the network administrators [10]. The major advantage of this approach is the ability to detect novel attacks, or deviations of existing attacks without prior knowledge of the attack nature. The major challenges that anomaly IDS have to solve are the improvement of the detection process and the reduction of the number of the false alarms ([1]).

MOTIVATION

The task of IDS could be modelled with various methods and in various levels of description of normal behaviour and current activity monitoring. A typical supervised anomaly recognition model will analyze data, compare to a known profile, run statistical analysis to determine if any deviation is significant, and flag the event(s) as a normal activity or an attack. On contrary, unsupervised approaches do not need any description of normal user activity, since they try to create a real-time model of legal activities in current data traces. All data, which does not conform to the described model, is marked as anomalous.

The present paper describes an adaptive approach for anomaly intrusion detection using 2-means clustering with purpose to examine the effectiveness of both approaches – supervised and unsupervised. The paper compares the results of evaluations of the performance of the

[*] This work was supported under project 6/2013 of Burgas Free University.

following two models – the first one, denoted with A, applies a 2-means clustering anomaly detection technique; and the second, denoted with B, applies 2-means clustering algorithm with combination with some data mining techniques, i.e. classification trees. As cluster distance in both cases was used Wagner-Fischer distance.

THE DESCRIPTION OF THE METHODOLOGY

Distance metric

The intrusion detection itself is performed using different distances, which measure the degree of proximity between normal and real-time data sequences. We stop our attention to the Wagner-Fischer distance.

Wagner-Fischer distance (*WFD*) [15] is a string metric between two strings, which stands for the minimum number of operations (insertion, deletion, substitution of a single character, transposition of two characters) needed to transform one string 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 $w(\varepsilon,b)$ is the cost of inserting *b*. The *WFD* are computed using the following recurrence relation:

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

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*.

Proposed clustering algorithm

K-means clustering [12] is the algorithm of cluster analysis, which groups the objects in K disjoint clusters, based on the distance function. In our case, the goal is to divide them into two classes, one of the normal data, and the other – for the anomalies. The algorithm in this case consists of the following steps:

- Two arbitrary different objects for centres – one of the normal observations, and the other – from the anomalies, are selected.
- When all observations are classified in their closest clusters, the centres of clusters are recalculated. The *j* new centre is determined 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*.
- The Step 3 is repeated until to find the exact centre of each cluster.

Classification tree

Classification tree is a frequently applied data mining technique in the field of intrusion detection [4], [11]. In our approach the implementation of the classification trees is performed through the process of description of the normal system activity. The normal activity patterns compose a set *Q* with *N* states: $q_1, q_2, \ldots, q_N$ which the system passes through its work in the discrete moments of time *t*=1, ... ,*T*. We assume that the probability of occupying a state is determined solely by the preceding state. Each state transition probability represents the probability of transitioning from a given state to another possible state. Based on the state transition probabilities, we construct classification trees of level *L*, whose roots are all possible states q_k, *k=1,…N*. The inheritors for each vertex are the states for which the corresponding transition probabilities from their predecessor are non-zero.

By traversing the tree from the root to the leaves we can receive all possible state sequences with length *L* along with the corresponding transition probabilities. The obtained lists of system calls consist of all possible sequences with given state in k^{th} position and contain states for which the transition probabilities for each couple of neighbors is non-zero.

Within the created classification trees the *WFD* was applied between the received sequence and normal sequences for the number of errors calculation. The obtained value indicates the degree of similarity between the normal and observed sequences, which is considered as basis for the current activity classification.

EXPERIMENTAL RESULTS

Simulation data

Extensive empirical testing of the proposed methodology was performed on the data, generated and published by the researches in Immune Systems Project from the Computer Science Department, University of New Mexico [14]. The data are obtained from Unix system examination during a period of time and consist of normal user activity patterns of some privileged processes executed on behalf of the root account as well as some anomalous data. The methods for pattern generation are described in [7] and [8]. They substantiate that the short sequences of system calls are reliable discriminator between normal and anomalous activities in the system. Each pattern is a sequence of system calls, which are the results of the examined process. The input data files are sequences of ordered pairs of numbers, where each line consists of one pair. The first number in each pair is the process ID (PID) of the process executed, and the second one is the system call number.

As a first stage of the experiments, a 2-means clustering algorithm, described above, was applied and the current activity sequences were divided into two different clusters. As a second stage of testing model B, based on the normal user activity patterns, the state transition probabilities for the sequences of the normal system activity were evaluated and the normal database, which consists of the classification trees of level *L*, was created. These trees compose the normal program behavior profiles. During the last stage, which is the intrusion detection itself, the anomalous data were divided into portions of length *L* and compared to the lists, extracted by the trees in normal database. The testing data contain both normal and anomalous patterns for the following processes: inetd, login, named and synthetic sendmail.

Performance measures

One of the measures of quality of a cluster algorithm using external criterion is the Jacard index. The *Jaccard index* [9] is used to quantify the similarity between two datasets. An index of 1 means that the two dataset are identical and an index of 0 indicates that the datasets have no common elements. The Jaccard index is defined by the following formula:

$$J = \frac{TP}{TP + FP + FN},$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives and *FN* is the number of false negatives.

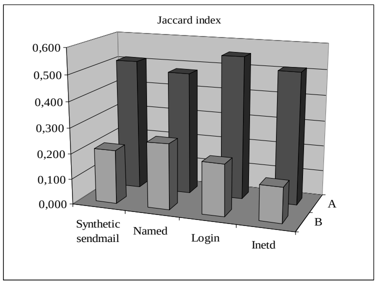

Figure 1. Jaccard index

Figure 1 shows the values of Jaccard index. This index computes the probability that two nodes belonging to a same cluster in a partition also belong to a same cluster in the other partition. The obtained values indicate the good quality of data separation into clusters, i.e. classification of normal or anomalous traffic.

The receiver operating characteristic (*ROC*) [6] curve is a method of graphically demonstrating the relationship between sensitivity and specificity, where sensitivity evaluates intrusion correctly detected and specificity evaluates how well a binary classification test correctly identifies the negative cases. Mathematically, its are expressed as follows

$$Sensitivity = \frac{TPR}{TPR + FNR} \qquad Specificity = \frac{TNR}{TNR + FPR},$$

where the false negative rate (*FNR*) represents undetected attacks on a system, the true positive rate (*TPR*) represents intrusion correctly detected, the false positive rate (*FPR*) represents the frequency with which the IDS reports malicious activity in error and the true negative rate (*TNR*) represents an IDS that is correctly reporting that there are no intrusions.

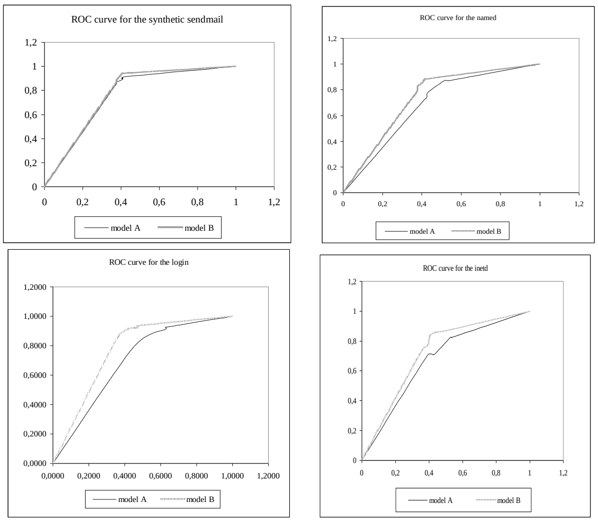

Figure 2. ROC curve

An ROC space is defined by 1-sensitivity and specificity as *x* and *y* respectively, which depicts relative trade-offs between true positive and false positive. Maximal sensitivity is realized when all tests are reported as abnormal. Specificity moves in concert from 0 (no true negatives) to one (no false positives). Maximal specificity is achieved by reporting all tests as normal. The best possible prediction method would yield a point in upper left corner (0,1) of the ROC space, representing 100% sensitivity (all true positives are found) and 100% specificity (no false positives are found). This point is called a perfect classification. The diagonal line (from the left bottom to the right corner) divides the ROC space in areas of good and bad classification. Points above this line indicate good classification results, while points below the line indicate wrong results.

Each sensitive value can be plotted against its corresponding specificity value to create the diagrams for the processes synthetic sendmail, named, login and inetd in the case of models *A* and *B* in Figure 2. Since the graphs are over the diagonal line this methodology gives good classification results such as the better results are given by the model *B* for the all processes. When we compare the ROC curves for all examined processes we can make the conclusion that it is recommended to use the model *B*.

Cluster validity assessment

When analyzing the cluster it is natural to assume that the cluster with a greater number of vectors is a cluster comprising a vector of the normal operation, and the other contains anomalies. Vectors in the same cluster are similar, which usually means that they are “close” to each other. Although it seems illogical, in the case of large-scale attacks could be seen that more vectors are generated by anomalies of the normal vectors. For this reason for a better classification should be analyzed the structures of the clusters. For this purpose, the size and the distance between the clusters are calculated. The compactness is used to describe similarities between objects in the same class. As measure for cluster compactness intra-cluster distance is applied

$$\Delta(K_i) = \max_{x,y \in K_i} \{d(x,y)\},$$

where $\Delta(K_i)$ - intra-cluster distance, $d(,)$ - the measure of the distance between two vectors in cluster K_i.

It is small when the objects are close to their cluster-centroids. It increases if the number of clusters decreases.

The separability measure provides an evaluation of distances between the classes. Inter-cluster distance are used for measure the separability. It is small when there are a few large clusters. One way to calculate it is to find shortest distance between two observations belonging to two different clusters. Higher the inter cluster distance indicating much better distance between the center of the clusters.

In order to evaluate validation through which assesses the compactness of the clusters and the distances between them, we use *Davies-Bouldin index* and C-index.

- *Dunn index*. The Dunn index [3] is a metric for evaluating clustering algorithms. It is defined as the ratio between the minimal inter-cluster distance to maximal intra-cluster distance. The Dunn index is limited to the interval $[0,\infty)$ and its higher value indicates better clustering.
- *Davies-Bouldin index* [2] takes into accout 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 clusters that are compact and whose centers are far away from each other.

- *C-index.* The C-index [5] is a cluster similarity measure expressed as:

$$C = \frac{S - S_{\min}}{S_{\max} - S_{\min}},$$

where S is the sum of all distances between pairs of observations in the same cluster over all clusters. Let n be the number of these pairs. S_{min} and S_{max} are the sums of n lowest/highest distances across all pairs of observations. The C-index is limited to the interval $[0,1]$ and should be minimized.

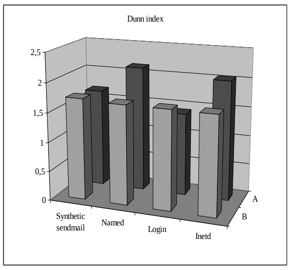

Figure 3. Dunn index

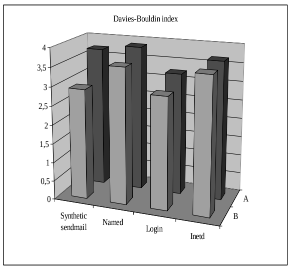

Figure 4. Davies-Bouldin index

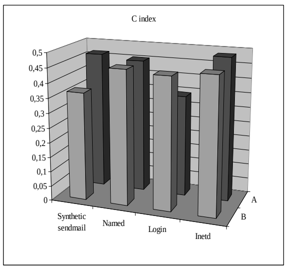

Figure 5. C-index

Figure 3 shows Dunn validity index values. The main goal of the measure is to determine whether the proposed method achieved maximize inter-cluster distance and minimize the intra-cluster distance. From Figure 3 it is clear that we are getting good performance results in terms of the Dunn index.
Figure 4 shows Davies-Bouldin validity index. This index attempts to minimize the average distance between each cluster and the one most similar to it. The presented values reveal that the applied clustering algorithm yields good performance in data separation into clusters.
Figure 5 contains the obtained values of C validity index. C-index values should be minimized. From Figure 5 could be observed, that all obtained values belong to the interval (0.3, 0.5), which means reliable clusterization of the examined data.

CONCLUSIONS

The results, obtained for models *A* and *B* and presented in Figures 1-5, indicate reliable and stable classification results for both models. Some of the values indicate that model B yields better performance results than model A, but it has to be mentioned it takes more time and resources than model A. The purpose of the future work could be comparison with different anomaly detection techniques and obtained results.

REFERENCES

1. Dagorn N., WebIDS: A Cooperative Bayesian Anomaly-Based Intrusion Detection System for Web Applications, Recent Advances in Intrusion Detection, LNCS, Vol. 5230/2008, Springer Berlin / Heidelberg, pp. 392-393.
2. Davies, D.L., Bouldin, D.W., (2000) A cluster separation measure, *IEEE Trans. Pattern Anal. Machine Intell.*, 1(4), 1979, 224-227.
3. Dunn, J. (1974) Well separated clusters and optimal fuzzy partitions, *Journal of Cybernetics* , 4, 1974, 95-104.
4. Gorodetsky, V., Karsaeyv, O., Samoilov, V., Multi-agent technology for distributed data mining and classification, In Proceedings of IEEE/WIC International Conference on Intelligent Agent Technology, 2003. IAT 2003, pp. 438- 441.
5. Hubert L, Schultz J. Quadratic assignment as a general data-analysis strategy . *British Journal of Mathematical and Statistical Psychologie*, 1976; 190-241.

6. Ferri C., N. Lachinche, S. A. Macskassy, A. Rakotomamonjy, *Second Workshop on ROC Analysis in ML*, 2005.
7. Forrest S., S.A. Hofmeyr, A. Somayaji, T.A. Longtaff, A Sense of Self for Unix Processes, In Proceedings of the 1996 IEEE Symposium on Security and Privacy, IEEE Computer Society Press, Los Alamitors, CA, pp.120-128.
8. Forrest S., S.A. Hofmeyr, A. Somayaji, Intrusion detection using sequences of system calls, Journal of Computer Security, Vol. 6, 1998, pp. 151-180.
9. Jaccard P., Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regions voisines, *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37, 1901, 241-272.
10. Jyothsna V., V. V. Rama Prasad, K. Munivara Prasad, A Review of Anomaly based Intrusion Detection Systems, International Journal of Computer Applications, Volume 28– No.7, August 2011, pp. 26-35.
11. Li X. B., A scalable decision tree system and its application in pattern recognition and intrusion detection, Decision Support Systems, Vol. 41, Issue 1, November 2005, pp. 112-130.
12. 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.
13. Marinova V., A Short Survey of Intrusion Detection Systems. Problems of Engineering Cybernetics and Robotics, 58:23–30, 2007.
14. [UNM] University of New Mexico’s Computer Immune Systems Project, http://www.cs.unm.edu/~immsec/systemcalls.htm.
15. Wagner R. A., M. J. Fischer, The string-to-string correction problem, *Journal of the Association for Computing Machinery* 21, pp. 168-173, 1974.