A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters

Slobodan Petrović

NISlab, Department of Computer Science and Media Technology,
Gjøvik University College, P.O. box 191, 2802 Gjøvik, Norway
E-mail: slobodan.petrovic@hig.no

Abstract. One of the most difficult problems in the design of an anomaly based intrusion detection system (IDS) that uses clustering is that of labelling the obtained clusters, i.e. determining which of them correspond to ”good” behaviour on the network/host and which to ”bad” behaviour. In this paper, a new clusters' labelling strategy, which makes use of a clustering quality index is proposed for application in such an IDS. The aim of the new labelling algorithm is to detect compact clusters containing very similar vectors and these are highly likely to be attack vectors. Two clustering quality indexes have been tested and compared: the Silhouette index and the Davies-Bouldin index. Experimental results comparing the effectiveness of a multiple classifier IDS with the two indexes implemented show that the system using the Silhouette index produces slightly more accurate results than the system that uses the Davies-Bouldin index. However, the computation of the Davies-Bouldin index is much less complex than the computation of the Silhouette index, which is a very important advantage regarding eventual real-time operation of an IDS that employs clustering.

Key words: Intrusion detection system, Anomaly detection, IDS benchmarking, Clustering, Silhouette index, Davies-Bouldin index.

1 Introduction

Intrusion detection systems (IDS) are security tools designed to detect and classify attacks against computer networks and hosts. They can operate in two ways: either by searching for specific patterns in data (misuse based IDS) or by recognising certain deviations from expected behaviour (anomaly based IDS).

In anomaly based IDS, clustering algorithms are often used for recognition of ”abnormal” behaviour. The number of clusters into which the input data may be classified is arbitrary, but as the essential goal of these systems is to distinguish between ”normal” and ”abnormal” behaviour, it is very common to partition the incoming resource access requests into two classes that correspond to these two types of behaviour.

We consider a Denial-of-Service (DoS) attack scenario in which attack resource access requests arrive to the monitored network/host in bursts. An anomaly based IDS analyzes N resource access requests at a time and if it detects

that many of these requests correspond to attacks then it should generate a special alert. We call such a scenario a *massive attack*. Sometimes, other network monitoring tools (firewalls etc.) can detect such attacks, but the advantage of an anomaly based IDS regarding all kinds of attacks (including massive attacks as defined in this paper) is in the capability of detecting a completely new attack.

If clustering is used for classification of resource access requests in an IDS, the main problem is the interpretation of clustering results, so called ”labelling” of clusters. Namely, without additional information it is difficult to decide whether the data classified in one cluster correspond to ”normal” behaviour in the monitored network or to ”abnormal” behaviour. Cardinalities of clusters are often used as a decision parameter for this purpose (see, for example, [8]) because the mathematical expectation of ”normal” behaviour is considered greater than that of ”abnormal” behaviour. However, this approach fails to detect massive attacks. Solving this problem requires a more complex clusters' labelling algorithm.

In this paper, we analyze a clusters' labelling strategy based on application of clustering evaluation techniques. The first option is to use the Silhouette index and clusters' silhouettes [9]. The second option is to combine the Davies-Bouldin index [2] and the comparison of centroid diameters of the clusters. The goal of such combinations is to respond adequately to the properties of attack vectors. We consider the compactness of the corresponding clusters and the separation between them the principal parameters that distinguish ”normal” from ”abnormal” behaviour in the analyzed network. The Silhouette index and the Davies-Bouldin index take into account these parameters and because of that we apply them in our IDS. In the experiments, we test the response of a multiple classifier IDS (see, for example, [4]) with the new labelling strategy to artificial data. We express the IDS quality through Receiver Operating Characteristics (ROC) curves. The effectiveness of the IDS that uses the Silhouette index is compared with that of a system that uses the Davies-Bouldin index.

In the experiments, we tested our labelling algorithms on the well known KDD CUP 1999 artificial data set [3], which was used as the traffic source. Although this source was criticized in the literature (see, for example, [7]), we found it convenient as a source of massive attacks, against which we have tested our labelling strategies. The experimental results show that the labelling strategy that uses the Silhouette index gives slightly more accurate results than the strategy that employs the Davies-Bouldin index. However, the computation of the Davies-Bouldin index is much less complex than the computation of the Silhouette index, which is a very important advantage regarding eventual real-time operation of such an IDS.

2 General description of the system

We concentrate on the basic sensor-assessor structure of the multiple classifier IDS shown in Fig. 1. The sensors actually perform the clustering of the incoming resource access requests, whereas the assessors perform the clustering quality evaluation.

We selected the well known K-means algorithm (see for example [6]) for implementation in the sensors of the IDS, because we consider this algorithm the best trade-off between accuracy and efficiency. The input resource access requests are encoded in such a way that vectors of the same length are produced. The Euclidean metric is used in our system as a distance measure between vectors.

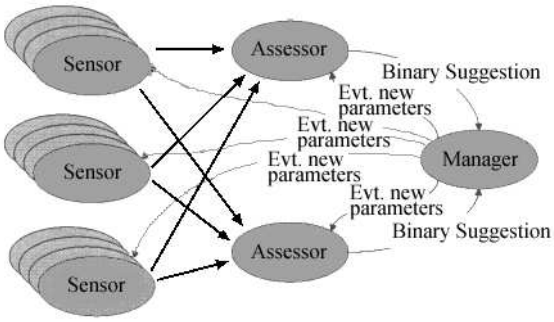


Fig. 1. A multiple classifier IDS

3 The clusters' labelling algorithm

Having obtained clusters from the sensors, the task of the assessors is to label them, i.e. to determine which clusters correspond to "normal" behaviour, and which to "abnormal" behaviour. Since there is no learning on labelled data in the system, the assessors must use other criteria to decide on this. There are at least two problems related to the cardinality-based labelling strategy that considers the cluster of the greatest cardinality the normal one [8]: first, normal data transmitted by means of a less frequently used protocol (such as *ftp* or *telnet*) might produce clusters of very different cardinalities, which could mislead such an assessor. Second, there are some Denial-of-Service attacks, such as *syn-Flood*, that can mislead this labelling strategy by making the mathematical expectation of the attack much greater than that of "normal" behaviour. To overcome the problems related to the labelling strategy described above, we propose clustering evaluation techniques to be used in the assessors of the IDS.

We use the Silhouette index and the Davies-Bouldin clustering evaluation index [1, 5] and compare them in an implementation of an anomaly detection

IDS based on clustering. In one of our IDS assessing algorithms, the global Silhouette index of the clustering is combined with the comparison of the silhouettes of the clusters. In another algorithm, the Davies-Bouldin index of the clustering is combined with the centroid diameters comparison between clusters. In the computation of the Davies-Bouldin index, the centroid linkage is used as the inter-cluster distance. The centroid inter-cluster and intra-cluster measures are selected for compatibility with the K-means clustering algorithm used in the sensors (which essentially computes centroids of clusters at each iteration).

We now present formal definitions of these two clusters' quality indexes.

Let $\mathbf{X}_\tau = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ be the data set and let $\mathcal{C} = (C_1, \ldots, C_K)$ be its clustering into K clusters. Let $d(\mathbf{X}_k, \mathbf{X}_l)$ be the distance between $\mathbf{X}_k$ and $\mathbf{X}_l$. Let $\mathcal{C}_j = \{\mathbf{X}_1^j, \ldots, \mathbf{X}_{m_j}^j\}$ be the j-th cluster, $j = 1, \ldots, K$, where $m_j =\mid \mathcal{C}_j \mid$. The average distance a_i^j between the i-th vector in the cluster $\mathcal{C}_j$ and the other vectors in the same cluster is given by the following expression [1, 5, 9]:

$$a_i^j = \frac{1}{m_j - 1} \sum_{\substack{k=1 \\ k \neq i}}^{m_j} d(\mathbf{X}_i^j, \mathbf{X}_k^j), \quad i = 1, \ldots, m_j. \tag{1}$$

The minimum average distance between the i-th vector in the cluster $\mathcal{C}_j$ and all the vectors clustered in the clusters $\mathcal{C}_k$, $k = 1, \ldots, K$, $k \neq j$ is given by the following expression:

$$b_i^j = \min_{\substack{n=1,\ldots,K \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{k=1}^{m_n} d(\mathbf{X}_i^j, \mathbf{X}_k^n) \right\}, \quad i = 1, \ldots, m_j. \tag{2}$$

Then the silhouette width of the i-th vector in the cluster $\mathcal{C}_j$ is defined in the following way:

$$s_i^j = \frac{b_i^j - a_i^j}{\max\left\{a_i^j, b_i^j\right\}} \tag{3}$$

From the expression (3), it follows that $-1 \leq s_i^j \leq 1$. We can now define the silhouette of the cluster $\mathcal{C}_j$:

$$S_j = \frac{1}{m_j} \sum_{i=1}^{m_j} s_i^j \tag{4}$$

Finally, the global Silhouette index of the clustering is given by:

$$S = \frac{1}{K} \sum_{j=1}^{K} S_j \tag{5}$$

It is easy to see that both a cluster's silhouette and the global silhouette take values between -1 and 1 (both inclusive).

Let $\mathbf{X}_\tau = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ be the data set and let $\mathcal{C} = (C_1, \ldots, C_K)$ be its clustering into K clusters. Let $d(\mathbf{X}_k, \mathbf{X}_l)$ be the distance between $\mathbf{X}_k$ and $\mathbf{X}_l$. Then the Davies-Bouldin index is defined in the following way [1, 2, 5]:

$$DB(\mathcal{C}) = \frac{1}{K} \sum_{i=1}^{K} \max_{i \neq j} \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\}. \tag{6}$$

where $\Delta(C_i)$ is the intra-cluster distance and $\delta(C_i, C_j)$ is the inter-cluster distance. In the observed IDS, the centroid diameter is used for $\Delta(C_i)$. It is defined in the following way [1]:

$$\Delta(C_i) = 2 \left(\frac{\sum_{\mathbf{X}_k \in C_i} d(\mathbf{X}_k, s_{C_i})}{\mid C_i \mid} \right), \quad i = 1, \ldots K, \tag{7}$$

where $s_{C_i} = \dfrac{1}{\mid C_i \mid} \displaystyle\sum_{\mathbf{X}_k \in C_i} \mathbf{X}_k$.

The centroid linkage inter-cluster distance is used for $\delta(C_i, C_j)$. It is defined in the following way [1]:

$$\delta(C_i, C_j) = d(s_{C_i}, s_{C_j}), \tag{8}$$

where $s_{C_i} = \dfrac{1}{\mid C_i \mid} \displaystyle\sum_{\mathbf{X}_k \in C_i} \mathbf{X}_k$ and $s_{C_j} = \dfrac{1}{\mid C_j \mid} \displaystyle\sum_{\mathbf{X}_k \in C_j} \mathbf{X}_k$.

For the remainder of this paper, we shall limit ourselves to studying the case with 2 clusters, of which one corresponds to "normal" and the other to "abnormal" behaviour in the analyzed network. The reason for this is that, whatever the number of clusters we use in the sensors, we must finally decide which of them will be considered "normal", leading us to a case with 2 "superclusters".

The main idea of our clusters' labelling algorithm, which uses a clustering quality index is the following:

The attack vectors are often mutually very similar, if not identical. Because of that, we expect that the attack cluster in the case of a massive attack will be extremely compact. The value of the Silhouette index of such a clustering is either 1 or very close to 1. The value of the Davies-Bouldin index of such a clustering is either 0 or very close to 0. Having in mind the expected mutual

similarity among attack vectors, the silhouette of the attack cluster is expected to be greater than the silhouette of the non-attack cluster. Likewise, the centroid diameter of the attack cluster is expected to be smaller than that of the non-attack cluster.

The case in which one of the clusters is empty must be treated in a special way. For a clustering containing an empty cluster, the global Silhouette index is undefined - we assign the value -1 to such a clustering for convenience. The Davies-Bouldin index in the case of a clustering containing an empty cluster is 0. If the non-empty cluster is extremely compact, then a natural conclusion is that it is the attack cluster, i.e. there were only attacks in the analyzed data set. Thus if the Silhouette index is used, relabelling of the clustering should be performed if the value of the global Silhouette index is -1 and the cluster labelled with ”2” (the label reserved for the attack cluster) is empty; if the Davies-Bouldin index is used, relabelling of the clustering should be performed if the Davies-Bouldin index of the clustering is equal to 0 and the cluster labelled with ”2” is empty.

For clusterings without empty clusters, higher values of the global Silhouette index indicate the presence of a massive attack, whereas higher values of clusters' silhouettes indicate attack clusters. Lower values of the Davies-Bouldin index indicate the presence of a massive attack, whereas small values of the centroid diameter in these cases indicate the attack cluster.

When the global Silhouette index takes lower values, i.e. when there is no massive attack, the silhouette of the non attack cluster (labelled with ”1”) is expected to be higher than the silhouette of the attack cluster (labelled with ”2”). Likewise, when the Davies-Bouldin index takes higher values, i.e. when massive attack is not present, the centroid diameter of the cluster labelled with ”1” is expected to be smaller than that of the cluster labelled with ”2”. This is because isolated attacks (non-massive) are expected to be less similar among themselves.

The study above gives rise to the following labelling algorithms:

Algorithm 1a (using the Silhouette index)
Input:

- Clustering $\mathcal{C}$ of N vectors into clusters C_1 and C_2, with arbitrary labelling.
- The global Silhouette index threshold, Δ_S.
- The clusters' silhouette thresholds, Δ_{S_1} and Δ_{S_2}.

Output:

- The eventually relabelled input clustering, if relabelling conditions are met.

begin
$S \longleftarrow GlobalSilhouetteIndex(\mathcal{C})$;
$s_1 \longleftarrow Silhouette(C_1)$;
$s_2 \longleftarrow Silhouette(C_2)$;
if $(S = -1)$ **and** $(IsEmpty(C_2))$ **then**
$Relabel(\mathcal{C})$
else if $(S < \Delta_S)$ **and** $(s_1 < s_2 + \Delta_{S_1})$ **then**
$Relabel(\mathcal{C})$
else if $(S > \Delta_S)$ **and** $(s_1 + \Delta_{S_2} > s_2)$ **then**
$Relabel(\mathcal{C})$;
end.

Algorithm 1b (using the Davies-Bouldin index)
Input:

- Clustering $\mathcal{C}$ of N vectors into clusters C_1 and C_2, with arbitrary labelling.
- The Davies-Bouldin index threshold, Δ_{DB}.
- The centroid diameters difference thresholds, Δ_{CD_1} and Δ_{CD_2}.

Output:

- The eventually relabelled input clustering, if relabelling conditions are met.

begin
$db \longleftarrow DaviesBouldinIndex(\mathcal{C})$;
$cd_1 \longleftarrow CentroidDiameter(C_1)$;
$cd_2 \longleftarrow CentroidDiameter(C_2)$;
if $(db = 0)$ **and** $(IsEmpty(C_2))$ **then**
$Relabel(\mathcal{C})$
else if $(db > \Delta_{DB})$ **and** $(cd_1 > cd_2 + \Delta_{CD_1})$ **then**
$Relabel(\mathcal{C})$
else if $(db < \Delta_{DB})$ **and** $(cd_1 + \Delta_{CD_2} < cd_2)$ **then**
$Relabel(\mathcal{C})$;
end.

The relabelling procedure simply exchanges labels between the two clusters.

□

The behaviour of the algorithms 1a and 1b depends on the choice of the parameters. These should be determined in advance. One of the ways to do that is to use a network/dataset with known characteristics. In a real network, one could start with the parameters of the algorithms obtained in a controlled network scenario (e.g. with those obtained with the KDD CUP 1999 database) and then fine tune the parameters over time.

Example 1: In the KDD CUP 1999 data set, many attack vectors correspond to the so called ”smurf” attack, which is a sort of DoS attack. Table 1 shows the differences between the coordinates of two attack vectors that correspond to the ”smurf” attack. Table 2 shows the differences between two ”normal” vectors. In this particular example it is easy to see that the difference between two attack vectors is much smaller than the difference between two ”normal” vectors.

Table 1. The differences between two attack vectors in the KDD CUP 1999 data base (records 7635 and 7636 of the reduced (10%) data set). The rest of 41 coordinates are equal to 0.

Coord. id.	Rec. 7635	Rec. 7636
protocol_type	2	2
service	50001	50001
flag	10	10
src_bytes	1032	1032
count	511	511
srv_count	511	511
same_srv_rate	100	100
dst_host_count	228	238
dst_host_srv_count	83	93

Table 2. The differences between two ”normal” vectors in the KDD CUP data base (records 6 and 7 of the reduced (10%) data set). The rest of 41 coordinates are equal to 0.

Coord. id.	Rec. 6	Rec. 7
service	80	80
flag	10	10
src_bytes	212	159
dst_bytes	1940	4087
logged_in	1	1
count	1	5
srv_count	2	5
same_srv_rate	100	100
srv_diff_host_rate	100	0
dst_host_count	1	11
dst_host_srv_count	69	79
dst_host_same_srv_rate	100	100
dst_host_same_src_port_rate	100	0

□

4 Experimental work

Extensive simulation of the basic sensor-assessor structure of a multiple classifier IDS was carried out in order to study its response to the attack data. To this end, the following instance of this structure was built:

1. In the sensor, the 2-means clustering algorithm was implemented.
2. Two types of assessors were tested:
   - 2.1 The assessor implementing the Silhouette index of the clustering and the silhouettes of the clusters, according to the Algorithm 1a. The global Silhouette index threshold, Δ_S, and the clusters' silhouette thresholds, Δ_{S_1} and Δ_{S_2}, were used as parameters of the assessing algorithm.
   - 2.2 The assessor implementing the Davies-Bouldin index of the clustering and the clusters' diameters, according to the Algorithm 1b. The Davies-Bouldin index threshold, Δ_{DB}, and the centroid diameters difference thresholds, Δ_{CD_1} and Δ_{CD_2}, were used as parameters of this assessing algorithm.

We selected the KDD CUP 1999 database as the traffic source for our experiments for two reasons, in spite of the criticism (see [7]): First, it is an artificially generated test data set, which guarantees that no unknown attacks can appear in it, consequently ensuring the possibility of accurate determination of the number of false alarms. Second, it is the best source of massive attacks known so far.

Our aim was to compare the results obtained by applying the two variants of the proposed labelling strategy, with and without the presence of massive attacks. Because of that, the attacks from the KDD CUP database were filtered out in the same way as in [8]. The filtering percentage of 0%, 98% and 99% was used over all the resource access request records of the database. Without filtering out the attacks (0%), the database simulates many massive attacks, whereas if the filtering of 98% and 99% of attacks is applied it simulates a situation in which attacks are rare events. The effectiveness of the system was measured by means of the ROC (Receiver Operating Characteristic) curves for the filtered data set mentioned above. A ROC curve depicts the relationship between false positive rate FPR and true positive rate TPR, where:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \qquad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

In the equation (9), FP is the number of false positive outcomes of the intrusion detection on a fixed data set, i.e. the number of decisions in which a non-existing attack is signalled, TP is the number of true positive outcomes, i.e.

successful detections, TN is the number of true negative outcomes, i.e. the number of decisions, in which a non-existing attack is not signalled, and FN is the number of false negative outcomes, i.e. the number of decisions, in which an existing attack is not signalled.

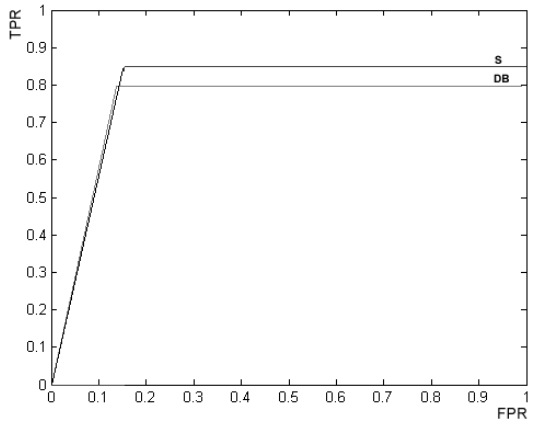
The results concerning the effectiveness of the IDS using the Algorithm 1a are compared with those obtained using the Algorithm 1b (Fig. 2). The best results with the Algorithm 1a over the KDD CUP '99 database were obtained with $\Delta_{S_1} = \Delta_{S_2} = 0.0001$. The best results with the Algorithm 1b over the same database were obtained with $\Delta_{CD_1} = 500$ and $\Delta_{CD_2} = 0$. These parameters for the algorithms 1a and 1b were chosen in order to maximize the system performance on the given data set. Although it may result in overestimation of the algorithms' performance, the fact that the test data set contains many DoS attacks makes us expect a similar performance of the algorithms in a real network containing many similar DoS attacks.

From the Fig. 2, it can be seen that without attack filtering (Fig. 2a), the Algorithm 1a gives better results than the Algorithm 1b. With 98% of the attacks from the KDD CUP 1999 database filtered out (Fig. 2b), the results obtained with the Algorithm 1a are still somewhat better. If even more attacks (99%, Fig. 2c) are filtered out from the KDD CUP 1999 database, the behaviour of the Algorithm 1a and 1b is approximately the same. It is also worth mentioning that the cardinality-based labelling strategy fails completely with 0% attack filtering (TPR achieved is below 10%). It behaves better with 98% and 99% attack filtering but that was expected since in those cases the most of the massive attacks are filtered out.

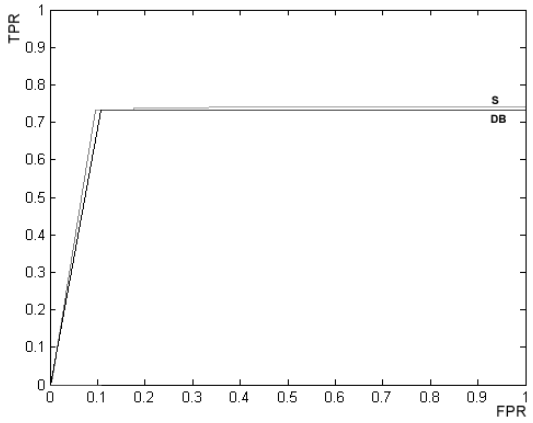
The time complexity of the Silhouette index computation is quadratic in the number of vectors involved in the clustering, whereas the time complexity of the Davies-Bouldin index computation is linear in the number of clustered vectors. In the case of the labelling algorithm that uses the Silhouette index, a relatively small improvement in correctness of the results over the labelling algorithm that uses the Davies-Bouldin index is penalized with a significant increase in computational complexity. This may make the labelling algorithm that uses the Silhouette index unsuitable for real-time IDS operation.

5 Conclusion

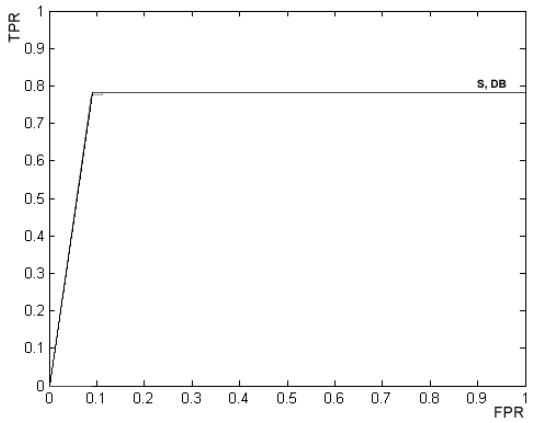
In this paper, a new clusters' labelling strategy was proposed for application in a multiple classifier intrusion detection system (IDS). That strategy combines the computation of a quality index of the clustering and the comparison of certain parameters of the clusters. Two variants of the labelling algorithm were tested. The first one uses the Silhouette index of the clustering and the silhouettes of the



a)



b)



c)

Fig. 2. ROC curves of the IDS. S - labelling using the Silhouette index; DB - labelling using the Davies-Bouldin index. Attack filtration: a) 0%, b) 98%, c) 99%

clusters. The second one uses the Davies-Bouldin index of the clustering and the centroid diameters of the clusters. The aim of the labelling algorithm is to detect compact clusters containing very similar vectors that are highly likely to be attack vectors. The response of an IDS using such a labelling strategy to a massive attack (for example, a Denial-of-Service attack) was tested. In the experiments, the KDD CUP 1999 database was used as the traffic source because it is the best source of massive attacks available. Besides, being an artificial test data source, it guarantees the correct determination of the number of false alarms during the testing. It was shown experimentally, via ROC curves obtained by applying the IDS over the KDD CUP 1999 database, that the labelling algorithm that uses the Silhouette index produces more accurate results than the one that uses the Davies-Bouldin index. However, the time complexity of the Silhouette index computation is much greater than the time complexity of the Davies-Bouldin index computation. Thus in an anomaly detection IDS that uses clustering as a classification method, the Davies-Bouldin index used in a clusters' labelling algorithm has a great advantage over the Silhouette index, regarding the overall performance.

References

1. Bolshakova N. and Azuaje F., "Cluster Validation Techniques for Genome Expression Data", *Signal Processing*, 83, 2003, pp. 825-833.
2. Davies D. and Bouldin D., "A Cluster Separation Measure", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 1, No 2, 1979, pp. 224-227.
3. Elkan C., "Results of the KDD'99 Classifier Learning", *ACM SIGKDD Explorations*, Vol. 1, No. 2, 2000, pp. 63-64.
4. Giacinto G. and Roli F., "Pattern Recognition for Intrusion Detection in Computer Networks", D Chen and X. Cheng (Eds.) *Pattern Recognition and String Matching*, Kluwer Academic Publishers, Dordrecht, 2002, pp. 187-209.
5. Günter S. and Bunke H., "Validation Indices for Graph Clustering", J. Jolion, W. Kropatsch, M. Vento (Eds.) *Proceedings of the 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition*, CUEN Ed., Italy, 2001, pp. 229-238.
6. Jain A., Murty M. and Flynn P., "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, 1999, pp. 264-323.
7. McHugh J., "Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory", *ACM Trans. on Information and System Security*, Vol. 3, No. 4, November 2000, pp. 262-294.
8. Portnoy L., Eskin E. and Stolfo S., "Intrusion Detection with Unlabeled Data Using Clustering", *Proceedings of the ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001)*, Philadelphia, PA, November 5-8, 2001.
9. Rousseeuw P., "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis", *J. Comput. Appl. Math.*, 20, 1987, pp. 53-65.