International Journal of Security and Its Applications
Vol. 10, No. 1 (2016), pp.215-228
http://dx.doi.org/10.14257/ijsia.2016.10.1.20

Some Clustering-Based Methodology Applications to Anomaly Intrusion Detection Systems

Veselina Jecheva and Evgeniya Nikolova

[1,2]*Burgas Free University*
[1]*vessi @bfu.bg,* [2]*enikolova @bfu.bg*

Abstract

The present paper introduces some clustering-based methodology applications to the anomaly and host-based intrusion detection. The proposed methodologies include fuzzy clustering, fuzzy clustering by local approximation of memberships and 2-means clustering algorithms. The presented anomaly-based frameworks are evaluated by simulation experiments and comparison of the obtained results.

Keywords: *intrusion detection, anomaly-based IDS, host-based IDS, fuzzy clustering, fuzzy clustering by local approximation of memberships (FLAME) algorithm, clustering validity index.*

1. Introduction

The contemporary economy and everyday life highly depend on the data, which is stored and transferred from one location to another through public and unsecure networks. The critical data is vulnerable to various attacks when transferred or stored due to the presence of errors or weaknesses in the software systems and network protocols, as the computer networks are now becoming the central nervous systems of our physical world—even of highly critical infrastructures such as the power grid [54].

Intrusion detection is a promising and highly important issue in the network security systems. Intrusion Detection Systems (IDS) are essential parts of the security systems, which purpose is to strengthen the security of communication and information systems. Their purpose is to detect any unauthorized and unusual behavior in the protected system, as it damages the network by violating confidentiality, integrity, availability, authenticity, non-repudiation or privacy [32].

Based on their location, Intrusion Detection Systems (IDS) are primarily classified into two types *i.e*. host-based IDS (HIDS) and network-based IDS (NIDS) [42]. HIDS looks for particular host activity while NIDS watches network traffic in the protected network or subnetwork. According to the applied techniques, IDS could be broadly divided into the following categories: misuse-based IDS, which rely on matching the current data with preliminarily collected known attack patterns or signatures; and anomaly-based IDS, which attempt to estimate the ‘‘normal’’ behavior of the system to be protected, and generate an anomaly alarm whenever the deviation between a given observation at an instant and the normal behavior exceeds a predefined threshold [18]. The major drawback of the misused-based IDS is their inability to detect novel attacks or variations of known attacks, since they rely on the preliminarily composed database of patterns or signatures of known exploits. On contrary, the anomaly-based IDS has the potential to detect previously unknown attacks or zero-day attacks, even though system is not updated [33]. Another advantage is the ability of customizing the normal activity profile and therefore making it very difficult for an attacker to know with certainty what activities it can carry out without getting detected [2]. Unfortunately, the maintenance of this type of IDS is very difficult because updating the behavior for which system is trained can’t be done

ISSN: 1738-9976 IJSIA

without losing the previous one ([1], [5]). The detection accuracy of this type of IDSs is also relatively low, which means the rate of the false positives is relatively high [26].

Various approaches have been applied to the anomaly-based intrusion detection during the recent years. There are various techniques, applying anomaly-based intrusion detection approach: statistics [39], Markov processes [51], genetic algorithms [29], neural networks [50], machine learning [22] and many others.

Machine learning is a special branch of artificial intelligence that acquires knowledge from training data based on known facts. It is defined as a study that allows computers to learn knowledge without being programmed mentioned by Arthur Samuel in 1959. Machine learning mainly focuses on prediction. It could be broadly divided into two major categories: supervised, which requires previously labeled instances in order to correctly classify the data; and unsupervised, which performs classification of the objects without any previous knowledge about them. Both approaches have been applied into anomaly-based IDS [12, 13, 28, 40] with various details – single or hybrid classifier design type, decision trees, SVM, neural networks, fuzzy logic, K-nearest neighbour (KNN) and many others.

Another issue, that addresses the IDS creation, is the selection of the level, at which the IDS should monitor the protected system, since it should be both stable over the time and sensitive enough to change in the case of unexpected behavior. Forrest [15] and Kosoresow [25] proved that the short sequences of system calls, generated by privileged processes during some period of time are stable and reliable and could be applied as a reliable discriminator between normal and anomalous behavior. The privileged processes are of special interest to the hackers, since they are executed with raised privileges, compared with the ordinary processes, executed by ordinary users.

The proposed framework focuses its attention on the clustering techniques, which are unsupervised machine learning approach. The simulation experiments have been performed on host-based data in order to examine the proposed methodology correctness.

2. Some Fuzzy Clustering Techniques for Intrusion Detection

Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. It is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics [47]. Cluster analysis is sensitive to both the distance metric selected and the criterion for determining the order of clustering. The choice of clustering algorithm depends on the type of data available and on the particular purpose.

W. Ren [45] has proposed intrusion detection system based on fuzzy c-means clustering algorithm applying to detect network intrusion. His experiment for separating normal data and intrusions shows the feasibility and validity of fuzzy c-means clustering algorithm. In [34] E. Narayan, *et al* proposed algorithms on expectation maximization fuzzy c-means clustering (EMFCM), which provide better result to fuzzy c-means clustering by avoiding the looping problems and saves time.

Linquan Xie, *et al* [58] presented hybrid algorithms, which combined the average information entropy, support vector machine and fuzzy genetic algorithm. The authors strive not only to reduce the rate of the false alarm and the rate of the failing alarm, but with optimizing of the fuzzy c-means clustering algorithm to achieve additional detection efficiency. Another hybrid fuzzy clustering algorithm that uses Quantum-behaved Particle Swarm Optimization (QPSO) algorithm and combines with fuzzy c-means (FCM) for abnormally detection is proposed by H. Wang [57]. In his paper [21] F. Guorui developed a semi-supervised learning algorithm for intrusion detection which is combined with the

fuzzy c-Means algorithm. The experimental results proved that the detection of attacks in the network data can be more efficiently by using the semi-supervised FCM clustering algorithm.

J. Visumathi, *et al.,* [55] proposed intrusion detection system based on weighted fuzzy c–means clustering and immune genetic algorithm, which solves the high dimensionality problem in the data set. T. Fries [14] in his paper presents a fuzzy-genetic approach to intrusion detection. In comparison to other GA-based techniques, the method demonstrates improved stability. Chenghua *et al* [6] improved the fuzzy C-means clustering algorithm by using the hierarchical clustering algorithm and genetic algorithm, creating AGFCM algorithm. The feature attribute data sets of network attack connection are sorted through the information gain algorithm and the Youden index is determined in order to reduce the data sets. The improved FCM clustering algorithm works over the reduced data sets. In [19] Goni and Lawal proposed Neuro-fuzzy Genetic Intrusion Detection System.

Rachnakulhare, *et al* [41] proposed fuzzy c-means algorithm using probabilistic neural network. The method not only reduces the training time but also increases the detection accuracy. Another algorithm, which combine the fuzzy c-means algorithm and the neural network are proposed by S. Chittineni, *et al* [7]. The method involves two steps. First, to determine cluster centers the authors used Enhanced K-means Fast Leaning Artificial Neural Network (KFLANN). Secondly, fuzzy c-means uses these cluster centers to generate fuzzy membership functions.

In [59], Yu-Ping Zhou has proposed a modified fuzzy c-means clustering algorithm in which Principal Component Analysis (PCA) neural network is used and a hierarchical neuro-fuzzy classifier is developed. The evaluations of the proposed method indicate the high detection accuracy for intrusion attacks and low false alarm rate.

A new fuzzy clustering method, which excludes outlier points by giving them extremely small membership values in existing clusters, is proposed in [52]. B. Thomas, *et al* incorporated the positive aspects of K-means algorithm in calculating the new cluster centers in the algorithm and they received more efficient approach than the c-means method. Om H [37] proposed a hybrid intrusion detection system that combines k -Means, and two classifiers: K Nearest neighbor and Naïve Bayes for anomaly detection. The classification performance of the modify algorithm is better than simple k-means algorithm. Solanki and Dhamdhere [30] presented a four level intrusion detection method using K-means clustering, neuro-fuzzy models, Support vector machine (SVM) and C4.5 algorithm. The first procedure concerns to generate different training subsets by using K-means clustering, the second procedure is based on different neuro-fuzzy models, the third procedure is classification using SVM and radial SVM, at the last procedure is being constructed the decision tree using C4.5 decision tree algorithm.

In [30] are compared three fuzzy clustering algorithms: FCM algorithm, Gath-Geva algorithm and Gostafson-Kesel algorithm, which are developed for intrusion detection systems. The performance of intrusion detection techniques is evaluated based on two criteria: detection rate and false positive rate.

The articles above are presented network intrusion detection system that is most often tested with the experimental data set KDD CUP99.

3. Description of the Proposed Methodology

3.1. The FLAME Algorithm

The clustering problem is to partition a data set into clusters so that the elements within a cluster are closer to each other than elements in different clusters. In [49] is presented a general classification of clustering (Figure 1).

One large group of clustering algorithms are fuzzy clustering [46]. These algorithms aim to partition the observations into distinct clusters as membership is probabilistic, not absolute and data elements can be member of more than one cluster. There are three types of fuzzy clustering methods based on: fuzzy relation, k-nearest neighbor rule and objective function. First of them can be divided into two categories based on objective functions and based on a relation matrix such as correlation coincident, equivalence relation, similarity relation and fuzzy relations, *etc*. [31]. In [9] is presented a full taxonomy of the fuzzy nearest neighbor classification. As examples from the third type fuzzy clustering method can be mentioned Fuzzy C-Means, Fuzzy Gaussian Mixture Modals, *etc*. Fuzzy C-means algorithm is based on the classic c-means. The membership is represented by the membership matrix, on which each row presents the membership of all *n* objects to a certain fuzzy subset and each column presents the membership of an object to all *c* fuzzy subsets.

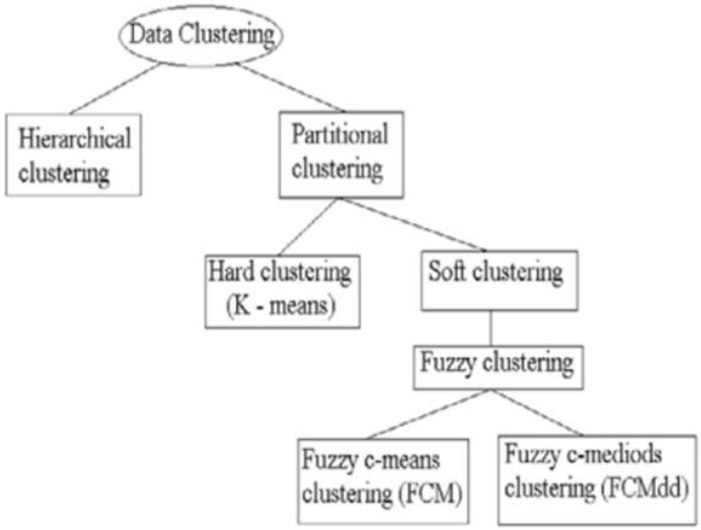

Figure 1. Classification of Clustering

Cluster analysis tries to subdivide a data set *Y* into *c* clusters. Let $Y = \{y_1, y_2, ..., y_N\}$ be a set of *N* elements in *n*-dimensional Euclidean space and *c* is an integer, $2 \le c < n$. A *c*-partition of *Y* is *c*-tuple $(Y_1, Y_2, ..., Y_n)$ of subsets of *Y* that satisfies the following properties

$$\begin{aligned} &Y_i \neq \varnothing, 1 \le i \le c; \\ &Y_i \cap Y_j = \varnothing, i \neq j; \\ &\bigcup_{i=1}^{c} Y_i = Y. \end{aligned} \tag{1}$$

While in non-fuzzy clustering each data point belongs to exactly one cluster, in fuzzy clustering (also referred to as soft clustering), the data points can belong to more than one cluster, and each point is associated with a membership grade, who indicates the degree of belonging to the different clusters. In terms of membership functions, a partition can be represented by the partition matrix $U = (\mu_{ij})_{c \times N}$ whose rows contain values of the membership function μ_i of the i[th] subset Yi of Y. The elements of U must satisfy the following conditions:

$$\mu_{ij} \in [0,1], 1 \le i \le c, 1 \le j \le N,$$
$$\sum_{i=1}^{c} \mu_{ij} = 1, 1 \le j \le N, \quad (2)$$
$$0 < \sum_{j=1}^{N} \mu_{ij} < N, 1 \le i \le c.$$

The fuzzy partitioning space for Y is the set

$$\left\{ U \in \mathbb{R}^{c \times N} \middle| \mu_{ij} \in [0,1], \forall i, j; \sum_{i=1}^{c} \mu_{ij} = 1, \forall j; 0 < \sum_{j=1}^{N} \mu_{ij} < N, \forall i \right\}. \quad (3)$$

One of Fuzzy clustering is Fuzzy clustering by Local Approximation of Memberships (FLAME) algorithm [17] on which is focused our attention. The three basic steps of the FLAME algorithm are:

- Definition of the neighborhood of each object and constructing a neighborhood graph to connect each object. In this step is used the K-Nearest Neighbors (KNN) clustering by which the objects are classified into three sets:
  1. Set of objects with density higher than all its neighbors - Cluster Supporting Objects (CSO);
  2. Set of objects with density lower than all its neighbors and lower than a defined threshold - Cluster Outliers;
  3. Set of the rest objects.
- Iterative converging process for Local/Neighborhood approximation of fuzzy memberships. First, in this step is initialized the fuzzy membership:
  1. Each CSO is assigned with fixed and full membership to itself to represent one cluster;
  2. All outliers are assigned with fixed and full membership to the outlier group;
  3. The rest are assigned with equal memberships to all clusters and the outlier group.

Second, update the fuzzy memberships of all three sets via the Local/Neighborhood Approximation of Fuzzy Memberships, in which the fuzzy membership of each object is updated by a linear combination of the fuzzy memberships of its nearest neighbors.

At the end of this process, objects can be assigned to one of the established clusters around the CSO or to the Cluster Outliers, based on their approximate memberships.

3.2. Fuzzy Clustering Techniques for Intrusion Detection System

The described algorithm has been applied in order to create an anomaly and host-based IDS. The purpose was to divide the current activity patterns into two clusters – one for the normal and one for anomalous data. The first step of FLAME contains the definition of the neighborhood of each object and the classification of the objects into three sets. This step is based on KNN algorithm. In the proposed framework the applied metrics is Wagner-Fischer distance (WFD). *Wagner-Fischer distance* [56] is a string metric between two strings, 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:

$$d_{WF}(i,j) = \min \begin{Bmatrix} d(i-1,j) + w(x_i,\varepsilon), d(i,j-1) + w(\varepsilon, y_j), \\ d(i-1,j-1) + w(x_i, y_j) \end{Bmatrix}. \quad (4)$$

The second step includes the iterative procedure Local/Neighborhood Approximation of Fuzzy Memberships. In our case the fuzzy memberships are determined based on the matrix of transition probabilities, created in the following way. Let's consider the system work in the discrete moments of time *t=1, 2, .. ,T*, when the system occupies some of the following *N* states: $S_1, S_2, \ldots, S_N$. Let $O=(O_1, O_2, \ldots O_T)$ is the observation sequence at the moments *t=1, 2, ..., T*. The partition matrix $U=\{\mu_{ij}, 1 \le i \le N, 1 \le j \le N\}$, $0 \le \mu_{ij} \le 1$ and $\sum_{j=1}^{N} \mu_{ij} = 1$ is a square matrix, whose elements are the state transition probabilities. Each element μ_{ij} represents the probability of transitioning from given state to another possible state.

3.3. Fuzzy Clustering Methodology

Fuzzy logic was introduced as a means to the model of uncertainty of natural language. And due to the uncertainty nature of intrusions fuzzy sets are strongly used in discovering attack events and reducing the rate of false alarms at the same time [48].

Cluster analysis tries to subdivide a data set Y into c clusters. Let $Y = \{y_1, y_2, \ldots, y_N\}$ be a set of N elements in n-dimensional Euclidean space and c is an integer, $2 \le c < n$. A c-partition of Y is c-tuple $(Y_1, Y_2, \ldots, Y_n)$ of subsets of Y that satisfies the following properties

$$Y_i \ne \varnothing, 1 \le i \le c;$$
$$Y_i \cap Y_j = \varnothing, i \ne j;$$
$$\bigcup_{i=1}^{c} Y_i = Y. \quad (5)$$

While in non-fuzzy clustering each data point belongs to exactly one cluster, in fuzzy clustering (also referred to as soft clustering), the data points can belong to more than one cluster, and each point is associated with a membership grade, who indicates the degree of belonging to the different clusters. In terms of membership functions, a partition can be represented by the partition matrix $U = (\mu_{ij})_{c \times N}$ whose rows contain values of the membership function μ_i of the ith subset Yi of Y. The elements of U must satisfy the following conditions:

$$\mu_{ij} \in [0,1], 1 \le i \le c, 1 \le j \le N,$$
$$\sum_{i=1}^{c} \mu_{ij} = 1, 1 \le j \le N, \quad (6)$$
$$0 < \sum_{j=1}^{N} \mu_{ij} < N, 1 \le i \le c.$$

The fuzzy partitioning space for Y is the set

$$\left\{ U \in \mathbb{R}^{c \times N} \middle| \mu_{ij} \in [0,1], \forall i, j; \sum_{i=1}^{c} \mu_{ij} = 1, \forall j; 0 < \sum_{j=1}^{N} \mu_{ij} < N, \forall i \right\}. \quad (7)$$

There are two main types of fuzzy clustering algorithms: fuzzy c-means=fuzzy logic+k-means partition and fuzzy k-medoids=fuzzy logic+k-medoids partition.

The Fuzzy C-Means Clustering (FCM) algorithm was introduced by Jim Bezdek in 1981 [3]. The algorithm is based on minimization of the fuzzy c-means functional:

$$\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}^{m}\left\|y_j - x_i\right\|_A^2, \tag{8}$$

where U is a fuzzy partition matrix of Y, $X=\left[x_1, x_2, ..., x_c\right], x_i \in \mathbb{R}^n$ is a vector of cluster centers, which have to be determined, $\left\|y_j - x_i\right\|_A^2 = \left(y_j - x_i\right)^T A\left(y_j - x_i\right)$ is a squared inner-product distance norm, $m \in [1,\infty)$ is a parameter which determines the fuzziness of the resulting clusters. The norm-inducting matrix A can be

- identity matrix;
- a diagonal matrix that accounts for different variances in the directions of the coordinate axes of *Y;*
- the inverse of the covariance matrix of *Y*,

which gives the standard Euclidean norm, a diagonal norm and the Mahalanobis norm on $\mathbb{R}^n$ respectively. The norm influences the clustering criterion by changing the measure of dissimilarity. In the third case,

$$A = R^{-1}, \quad R = \frac{1}{N}\sum_{j=}^{N}\left(y_j - \bar{y}\right)\left(y_j - \bar{y}\right)^T, \tag{9}$$

where $\bar{y}$ denotes the mean of the data.

The vector norms $\left\|y_j - x_i\right\|$ with respect to the Damerau-Levenshtein distance (DLD) ([8], [27]). It is a string metric between two strings, which means the minimum number of operations (an insertion, deletion, substitution of a single character, a 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 $w(a,b)$. $w(a,b)$ 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 DLD are computed using the following recurrence relation

$$d_{DL}(i,j) = \min\left\{d(i-1,j) + w(x_i,\varepsilon), d(i,j-1) + w(\varepsilon,y_j), d(i-1,j-1) + w(x_i,y_j)\right\},$$ where cost function has the following definition

$$w(a,\varepsilon)=1, \quad w(\varepsilon,b)=1, \quad w(a,b)=1, \text{ if } a=b \text{ or 0, otherwise.} \tag{10}$$

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.

The basic approach of fuzzy k-medoids algorithm, according to [23, 24], is to find k clusters in n data points by first finding the medoids (central points) for each cluster. The method uses representative points as reference and each remaining points are clustered by computing the minimum Euclidean distances.

In the proposed methodology, the short sequences of system calls with length T, describing current system activity, were examined. Since the desired method performs a binary classification, the purpose is to divide the test data into two clusters: one for the normal data and one for intrusive data. Let's consider the different system calls S_1, S_2, …, S_N, as a set S with N states in number, which the system passes through its work in discrete

moments of time t=1, 2, .. ,T. The state transition probability matrix A={a_{ij}, 1≤i≤N, 1≤j≤N}, $0 \le a_{ij} \le 1$ and $\sum_{j=1}^{N} a_{ij} = 1$ is a square matrix with the elements, which represent the probability of transitioning from given state to another possible state.

The partition matrix A is determined in the following way: each element μij are the transition probabilities of the elements, belonging to the set S.

4. Clustering Validity Indices

Since clustering is mostly unsupervised process, it is very important to find an appropriate metric for measuring if the found cluster configuration is acceptable or not. In order to achieve better classification, the size and the distance between the clusters should be calculated, as well as the cluster compactness.

Measures, which appreciate how well different data clustering algorithms perform on a set of data, based on the data that was clustered itself, is called internal evaluation. The most commonly used methods for measuring the quality of clustering algorithms, based on internal criterion are the Dunn index and Davies-Bouldin index.

Dunn index [10, 11] is defined by dividing the minimal inter-cluster distance (a measure of the average density in the region among clusters in relation to the density of the clusters) and maximal intra-cluster distance (a measure of the average of dispersal of clusters). It takes values from the interval [0, ∞). Its higher value shows better clustering.

Davies-Bouldin index ([4], [20]) takes into account both the error, caused by representing the data vectors with cluster centroids, and the distance between clusters. It is defined as follows:

$$DB(C) = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\} \tag{11}$$

where *n* is the number of clusters, $\Delta(C_i)$ - intra-cluster distance, $\delta(C_i, C_j)$ - inter-cluster distance, C_k - *k* cluster. It takes values from the interval [0, ∞) . Small values of Davies-Bouldin index correspond to clusters that are compact and whose centers are far away from each other.

Measures, which evaluate clustering results based on data that was not used for clustering, are called external evaluation. Some of the measures of quality of a cluster algorithm using external criterion are the Rand index and *F*-measure.

The Rand index (*RI*) [44] computes how similar the clusters, returned by the clustering algorithm, are. It can be computed using the following formula:

$$RI = \frac{true \quad positives + true \quad negatives}{true \quad positives + true \quad negatives + false \quad positives \; + \; fasle \quad negatives} \tag{12}$$

The Rand index gives a value between 0 and 1, where 1 means the two clustering outcomes match identically.

Recall and *Precision* are two commonly used measures in evaluating quality of classifiers, which are defined as follows:

$$\text{precision} = \frac{true\quad positives}{true\quad positives\quad +\quad fasle\quad positives}$$
$$\text{recall} = \frac{true\quad positives}{true\quad positives\quad +\quad fasle\quad negatives}, \tag{13}$$

where true positives mean the IDS correctly identifies an intrusion attempt, false positives occur when the IDS considers normal activity as intrusive, true negatives mean the IDS correctly identifies normal activity and false negatives occur when the IDS fails to detect an intrusive activity.

A measure of a classification accuracy, which summarizes the measures precision and recall into single indicator, is *F-measure* [38]. If it achieves high value, both precision and recall are reasonably high.

$$F-measure = \frac{2.precision.recall}{precision + recall} \tag{14}$$

Another clustering validity index is the *SD* validity index. It is defined as

$$SD(c) = a \cdot Scat(c) + Dis(c) \tag{15}$$

where

- a is a weighting factor equal to $Dis(c_{max})$, c_{max} - the maximum number of input clusters;
- the average scattering for clusters

$$Scatt(c) = \frac{1}{c}\sum_{i=1}^{n_c}\frac{\|\sigma(v_i)\|}{\|\sigma(Y)\|} \tag{16}$$

where $\sigma(Y)$ - variance of the dataset, $\sigma(v_i)$ - variance of a cluster;

- the total scattering between clusters

$$Dis(c) = \frac{D_{\max}}{D_{\min}}\sum_{k=1}^{c}\left(\sum_{z=1}^{c}\|v_k - v_z\|\right)^{-1}, \tag{17}$$

where v_k - the representative point of the cluster, $D_{max} = \max(\|v_i - v_j\|)$, $\forall i, j \in \{1, 2, 3,\ldots, c\}$ is the maximum distance between cluster center points, the $D_{min} = \min(\|v_i - v_j\|)$, $\forall i, j \in \{1, 2,\ldots, c\}$ is the minimum distance between cluster center points.

A small value of the first term in *SD* validity index indicates compact clusters. If clusters become less compact, the value of $Scat(n_c)$ increases. The second term is an indication of inter-cluster distance and it increases with the number of clusters. Lower *SD* index means better cluster configuration as in this case the clusters are compacts and separated.

There are validity indices suitable for fuzzy clustering [43] such as the Classification Entropy and Partition Index.

Classification Entropy (*CE*) measures the fuzzyness of the cluster partition only.

$$CE = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log\mu_{ij} \tag{18}$$

The indicator values vary in [0, $log_a c$]. The closer the value of *CE* to 0, clusters are more clearly separated.

Partition Index (*SC*) is the ratio of the sum of compactness and separation of the clusters.

$$SC = \sum_{i=1}^{c}\frac{\sum_{j=1}^{N}\mu_{ij}^{m}\left\|y_j - x_i\right\|^2}{N_i\sum_{s=1}^{c}\left\|x_s - x_i\right\|^2} \tag{19}$$

SC is useful when comparing different partitions having equal number of clusters. A lower value of *SC* indicates a better partition.

5. Performance Analysis

The experimental data were obtained from Unix-based system examination during some period of time from a project, performed by the researches in the Computer Science Department, University of New Mexico [53]. The experimental data include normal user activity traces of some privileged processes that run with administrative rights (synthetic ftp, synthetic lpr, login, named and xlock) and their child processes patterns, as well as intrusion data. The input data files are sequences of ordered pairs of numbers, where the first number is the process ID (PID) of the process executed, and the second one is the system call number. Forks are considered as separate processes and their execution results are examined as normal user activity. The methods for pattern generation are described in [15] and [16]. They have proved that short sequences of system call traces produced by the execution of the privileged processes are a good discriminator between the normal and abnormal operating characteristics of programs.

The experimental data include traces of user activity during some period of time. The described methodology was applied in order to distinguish normal user activity from abnormal one for the following privileged processes: synthetic sendmail, login, inetd, named.

Table 1. Dunn Index of the Algorithms on Four Datasets

Processes	Method 1	Method 2	Method 3
Synthetic sendmail	1,35	1.37	1.30
Named	1,47	1.36	1.35
Login	1,42	1.28	1.24
Inetd	1,21	1.31	1.26

The following validity indices were calculated: *Dunn index*, *F-measure* and *Partition Index*, given in Section 4, on four datasets - synthetic sendmail, login, inetd, named. Table 1 summarizes the effects of the *WFD* on the calculation of the Dunn cluster validation indices for the examined processes. Since there are only two intra-cluster distances and one inter-cluster distance in the case of two clusters, the Dunn index provides stable clustering quality evaluations. The presented values reveal that method 1 yields the best results, compared to the other two methods. It also could be seen that method 2 yields slightly better results, compared to these of method 3. We'll mention that all methods

yield a relatively low values with the dataset inetd, which should be examined as a future work.

Table 2 contains the obtained values of F-measure for the examined data. The table shows that all proposed intrusion detection methods could detect most of intrusions. The presented values reveal that method 1 yields the best results, compared to the other two methods. It also could be seen that method 2 yields slightly better results, compared to these of method 3. We'll mention that all methods yield a relatively high values with the dataset inetd and their lowest values with the dataset synthetic sendmail, which should be examined as a future work.

Table 2. F-Measure of the Algorithms on Four Datasets

Processes	Method 1	Method 2	Method 3
Synthetic sendmail	0.917	0.904	0.902
Named	0.934	0.914	0.927
Login	0.912	0.908	0.906
Inetd	0,971	0.954	0.961

Table 3 presents the values of Partition Index for the examined processes. The presented values reveal that method 1 yields the best results, compared to the other two methods, although all methods produce stable and reliable results.

Table 3. Partition Index of the Algorithms on Four Datasets

Processes	Method 1	Method 2	Method 3
Synthetic sendmail	0,2648	0.2694	0.2526
Named	0,2714	0.2567	0.2702
Login	0,2992	0.2614	0.2478
Inetd	0,2846	0.2471	0.2574

6. Comparisons of Diffferent Categorical Clustering Algorithm

An advantage of the described method is its potential to detect an unknown attack the first time it appears, since it is based on unsupervised clustering algorithm. As a result of the algorithm we obtain the set of sequences, which contain intrusive activity patterns.

Our method is based on the definition of normal behavior in terms of short sequences of system calls, described by Forrest et.al. ([15] and [16]). With the purpose of simplicity, this method ignores the parameters passed to the system calls, and look only at their temporal orderings. We should mention this definition of normal behavior ignores many other important aspects of process behavior, such as timing information instruction sequences between system calls, and interactions with other processes.

We used the FLAME algorithm - method 1, the fuzzy k-medoids algorithm – method 2 and a model that characterizes the expected/acceptable behavior of the system using a clustering algorithm based on a 2-means clustering anomaly detection technique and a classification tree, presented in [36] - method 3, for perform the clustering. The method 3 consists of two stages – the first one consists of the clustering algorithm, which divides the current activity data into two non-interceptive sets, containing the normal and intrusive activity, respectively. The second one includes the comparison of the marked as anomalous activity with preliminarily composed classification trees with normal activity sequences, using the Damerau-Levenshtein distance. A comparison performed between the results of three methods is shown in Figure 2.

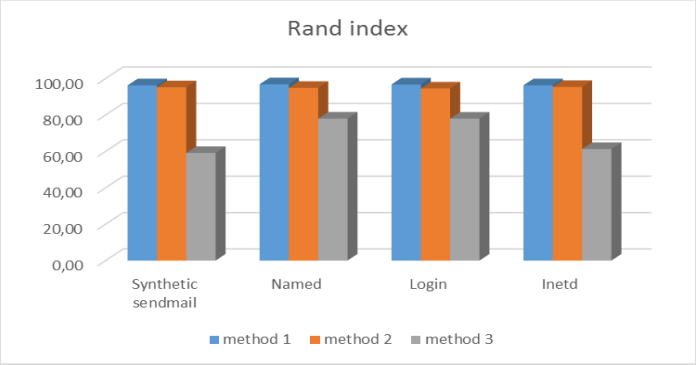

Figure 2. Comparison of our Proposed Classification via Method 1, Method 2 and Method 3

7. Conclusion

The present paper introduces a host-based scenario of fuzzy clustering application into anomaly-based IDS. The application of the fuzzy clustering methodology yields tolerance for imprecision and uncertainty, which are natural problems in IDS creation, relatively low solution cost and robustness. The dataset, which is used to examine the proposed methodology, the simulation results and their evaluations were also described in this work.

References

[1] S. Agrawal, J. Agrawal, "Survey on Anomaly Detection using Data Mining Techniques", 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science 60 **(2015),** pp. 708 – 713.
[2] P. Animesh ; P. Jung-Min, "An overview of anomaly detection techniques: Existing solutions and latest technological trends", Elsevier, Science Direct, Computer Networks, vol. 51, pp. 3448-3470, 2007.
[3] Bezdek, J. C., Pattern Recognition with Fuzzy Objective Function Algorithms, Springer Publishers, New York, **(1981)**.
[4] N. Bolshakova, F. Azuaje, "Cluster Validation Techniques for Genome Expression Data", Signal Processing, 83, **(2003)**,pp. 825-833,
[5] V. Chandola , A. Banerjee, V. Kumar, "Anomaly detection: A survey", ACM Computing Surveys (CSUR); vol. 41, no. 3; **(2009**;),p. 15 .
[6] T. LChenghua, Pengcheng, T. Shensheng, X. Yi, "Anomaly Intrusion Behavior Detection Based on Fuzzy Clustering and Features Selection, Journal of Computer Research and Development", vol. 52 Issue (3): **(2015)**, pp, 718-728,
[7] S. Chittineni S., Dr. R. B. Bhogapathi, "Neural Network Based Fuzzy C MEANS Clustering Algorithm", Available online at www.interscience.in
[8] Damerau F. J., A technique for computer detection and correlation of spelling errors, Communications of the ACM, 1964.
[9] Derrac J., S. García, F. Herrera, Fuzzy nearest neighbor algorithms: Taxonomy, experimental analysis and prospects, Journal Information Sciences: an International Journal, Volume 260, March, 98-119, (2014).
[10] Dunn J., Well separated clusters and optimal fuzzy partitions, Journal of Cybernetics, 4, 95–104, (1974).
[11] Dunn J. C., A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybernetics, 3(3), 32-57, (1973).
[12] Eesa A.S., Z. Orman, Brifcani A.M.A., (2014). A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Systems with Applications, 42(5), pp. 2670-2679.
[13] Feng W., Q. Zhang, G.Hu, J.X. Huang, (2014). Mining network data for intrusion detection through combining SVMs with ant colony networks. Future Generation Computer Systems, 37:127–140.
[14] Fries T.P., "A Fuzzy-Genetic Approach to Network Intrusion Detection", Department of Computer Science Coastal Carolina University Conway, South Carolina.
[15] Forrest S., S.A. Hofmeyr, A. Somayaji, Intrusion detection using sequences of system calls, Journal of Computer Security Vol. 6, 1998, pp. 151-180.
[16] Forrest S., Hofmeyr, S.A., Longstaff, T.A.: A Sense of Self for UNIX Processes. Proc. IEEE Symposium on Security and Privacy, Los Alamitos, CA. (1996) 120–128.

[17] Fu L., E. Medico, FLAME: A Novel Fuzzy Clustering Method for the Analysis of DNA Microarray Data, BMC Bioinformatics. Vol. 8, No. 3, (2007).
[18] Garcıa-Teodoro P., J.Dıaz-Verdejo, G. Macia-Fernandez, E.Vazquez, Anomaly-based network intrusion detection: Techniques, systems and challenges, COMPUTERS & SECURITY, Vol. 28(1-2), 2009, pp.18-28.
[19] Goni I., A. Lawal, A Propose Neuro-Fuzzy-Genetic Intrusion Detection System, International Journal of Computer Applications, 115(8), 5-9, April 2015.
[20] G¨unter S., 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, 229-238, (2001).
[21] Guorui F., "Intrusion detection based on the semi-supervised Fuzzy C-Means clustering algorithm", 2nd IEEE International Conference on Consumer Electronics, Communications and Networks (CECNet), 2012
[22] Haq N. F., M. Rafni, A. R. Onik, F. M. Shah, M. A. K. Hridoy, D. Md. Farid, Application of Machine Learning Approaches in Intrusion Detection System: A Survey, International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.3, 2015, pp. 9-18.
[23] Jain, A.K., M.N. Murty and P.J. Flynn, Data Clustering: A Review, ACM Computing Surveys, Vol. 31, No. 3, Sep. 1999, pp. 264-323.
[24] Kaufman, L. and P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley and Sons, 1990.
[25] Kosoresow A.P., Hofmeyr, S.A.: Intrusion Detection via System Call Traces. IEEE Software. 11 (1997) 35–42.
[26] Kumar U., B. N. Gohil, A Survey on Intrusion Detection Systems for Cloud Computing Environment, International Journal of Computer Applications (0975 – 8887), Volume 109 – No. 1, January 2015, pp. 6-15.
[27] Levenshtein V. I., Binary codes capable of correcting deletions, insertions and reversals, Soviet Physics Doklady, 1966.
[28] Lisehroodi M. M., Z. Muda, W. Yassin. (2013). A Hybrid Framework based On Neural Network Mlp And Kmeans Clustering For Intrusion Detection System. Proceedings of the 4th International Conference on Computing and Informatics, ICOCI 2013 (p. Paper No. 020). Sarawak, Malaysia: Universiti Utara Malaysia, pp. 305-311.
[29] Majeed P. G., S. Kumar, Genetic Algorithms in Intrusion Detection Systems: A Survey, International Journal of Innovation and Applied Studies, ISSN 2028-9324, Vol. 5, No. 3, Mar. 2014, pp. 233-240.
[30] Meghana S., V. Dhamdhere, Intrusion Detection System by using K-Means clustering, C 4.5, FNN, SVM classifier, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 6, November-December 2014, 19-23.
[31] Miin-Shen Yang, Hsing-Mei Shih, Cluster analysis based on fuzzy relations, Fuzzy Sets and Systems, 120, 197–212, (2001).
[32] Mitchell R., I.R. Chen, A survey of intrusion detection in wireless network applications, Computer Communications 42 (2014), pp. 1–23.
[33] Mudzingwa D., R. Agrawal, A study of methodologies used in intrusion detection and prevention systems (IDPS), Proceedings of IEEE Southeast Conference 2012, pp. 1-6.
[34] Narayan E., P. Singh, G. K. Tak, "Intrusion Detection System Using Fuzzy C_ Means Clustering with Unsupervised Learning via EM Algorithms", VSRD-IJCSIT, Vol. 2 (6), 2012, 502-510.
[35] Neda J., J. Bagherzadeh, Comparison of Fuzzy Clustering Algorithms in Intrusion Detection System, Journal of World's Electrical Engineering and Technology, 3(2): 53-58, 2014
[36] Nikolova E., V. Jecheva, An Adaptive Approach of Clustering Application in the Intrusion Detection Systems, Open Journal of Information Security and Applications, Vol. 1, No. 3, December 2014.
[37] Om H, "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system" Recent Advances in Information Technology (RAIT), 2012 1st International IEEE Conference.
[38] Pang-Ning T., M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson Education, Inc., (2009).
[39] Parvathi D.; S. Prasad, "Study of Anomaly Identification Techniques in Large Scale Systems", International Journal of Computer Trends and Technology, Vol.3, Issue 1, 2012.
[40] Patcha A., Park J. M., An overview of anomaly detection techniques: Existing solutions and latest technological trends; Computer Networks; 51(12); 2007; p. 3448-3470.
[41] Rachnakulhare, Divakar Singh, Intrusion Detection System based on Fuzzy C Means Clustering and Probabilistic Neural Network, International Journal of Computer Applications, Volume 74– No.2, July 2013, 30-33
[42] Rajasekhar K., B.Sekhar Babu, P.Lakshmi Prasanna, D.R.Lavanya, T.Vamsi Krishna, An Overview of Intrusion Detection System Strategies and Issues, 127-131.
[43] Ramze Rezaee M., B.P.F. Lelieveldt, J.H.C. Reiber, A new cluster validity index for the fuzzy c-mean, Pattern Recognition Letters, 19, 237-246 (1998).
[44] Rand W. M., Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association, 66 (336), 846–850, (1971).

[45] Ren W., "Application of Network Intrusion Detection Based on Fuzzy C-Means Clustering Algorithm", Intelligent Information Technology Application, 2009.
[46] Sampat Richa, Shilpa Sonawani, A Survey of Fuzzy Clustering Techniques for Intrusion Detection System, International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 1, 2188-2192, (January, 2014).
[47] Santhi M.V.B.T., V.R.N.S.S.V.Sai Leela, P.U.Anitha, D.Nagamalleswari, Enhancing K-Means Clustering Algorithm International Journal of Computer Science & Technology (IJCST), Vol. II, Issue IV, 2011, pp. 73-77.
[48] Sharma S., R. K. Gupta, Intrusion Detection System: A Review, International Journal of Security and Its Applications Vol. 9, No. 5 (2015), pp. 69-76.
[49] Sharma D., Fuzzy Clustering as an Intrusion Detection Technique, International Journal of Computer Science & Communication Networks, Vol 1(1), September-October, 2011
[50] Shuang-can Z., H. Chen-jun, Z. Wei-ming, Multi-Agent Distributed Intrusion Detection System Model Based on BP Neural Network, International Journal of Security and Its Applications, Vol.8, No.2 (2014), pp.183-192.
[51] Sukhwani H., S. Kodesia, S. Sharma, Detection and Classification of Intrusions using Fusion Probability of HMM, International Journal of Computer Applications (0975 – 8887), Volume 103 – No.12, October 2014.
[52] Thomas B., G. Raju, "A Novel Fuzzy Clustering Method for Outlier Detection in Data Mining", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.
[53] University of New Mexico's Computer Immune Systems Project, http://www.cs.unm.edu/~immsec/systemcalls.htm.
[54] Vasilomanolakis E., S. Karuppayah, M. Mühlhäuser, M. Fischer, Taxonomy and Survey of Collaborative Intrusion Detection, ACM Computing Surveys (CSUR), Volume 47 Issue 4, July 2015.
[55] Visumathi J., Dr. K.L.Shanmuganathan and Dr. K.A.Muhamed Junaid, "Misuse and Anomaly-based Network Intrusion Detection System using Fuzzy and Genetic Classification Algorithms", International Conference on Computing and Control Engineering (ICCCE), 2012.
[56] Wagner R. A., M. J. Fischer, The string-to-string correction problem, Journal of the Association for Computing Machinery 21, 1974, pp. 168-173.
[57] Wang H., "Network intrusion detection based on hybrid Fuzzy C-mean clustering", Fuzzy Systems and Knowledge Discovery (FSKD), Seventh International IEEE Conference, 2010.
[58] Xie L., Y. Wang, L. Chen, G. Yue, An Anomaly Detection Method Based on Fuzzy C-means Clustering Algorithm, Proceedings of the Second International Symposium on Networking and Network Security (ISNNS '10) Jinggangshan, P. R. China, 2-4, April. 2010, pp. 089-092
[59] Zhou Yu-Ping, "Research on Neuro-fuzzy Inference System in Hierarchical Intrusion Detection", Information Technology and Computer Science (ITCS), 2009.

Authors

Veselina G. Jecheva was born in 1971 in Burgas, Bulgaria. She obtained her master degree in Computer Science and Economics from Sofia University "St. Kliment Ohridski" in 1995. She received her PhD degree in Computer Science, especially Information Security from National Laboratory of Computer Virology, Bulgarian Academy of Sciences, in 2005. She is an Associate Professor at Burgas Free University, Faculty of Computer Science and Engineering, Bulgaria. Her research interests include information security, e-commerce, programming and Web systems.

Evgeniya P. Nikolova was born in 1968 in Pomorie, Bulgaria. From 1986 to 1991 she studied at Sofia University "St. Kliment Ohridski" and was given her master degree in Probability Theory and Statistics. She received her master degree in Economics from Burgas Free University in 1996. She received her PhD in Computer Science (Proper codes for error detection) from the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, in 2005. Evgeniya is an Associate Professor at Burgas Free University, Faculty of Computer Science and Engineering, Bulgaria. Her research interests include probability theory and coding theory.

