Jurnal Teknologi Maklumat & Multimedia 10(2011): 1 - 11

Improvement Anomaly Intrusion Detection using Fuzzy-ART Based on K-means based on SNC Labeling

Zulaiha Ali Othman
Afaf Muftah Adabashi
Suhaila Zainudin
Saadat M. Al Hashmi

ABSTRACT

Intrusion detection has received a lot of attention from many researchers, and various techniques have been used to identify intrusions or attacks against computers and networks. Data mining is a well-known artificial intelligence technique to build network intrusion detection systems. However, numerous data mining techniques have been successfully applied in this area to find intrusions hidden in large amounts of audit data through classification, clustering or association rule. Clustering is one of the promising techniques used in Anomaly Intrusion Detection (AID), especially when dealing with unknown patterns. This paper presents our work to improve the performance of anomaly intrusion detection using Fuzzy-ART based on the K-means algorithm. The K-means is a modified version of the standard K-means by initializing the value K from the value obtained after data mining using Fuzzy-ART and SNC labeling technique. The result has shown that this algorithm has increased the detection rate and reduced the false alarm rate compared with Fuzzy-ART.

Keywords: intrusion detection, anomaly detection, data mining, NSL-KDD dataset, Fuzz-ART, K-means, labeling

ABSTRAK

Bidang pengesanan pencerobohan telah menjadi perhatian ramai penyelidik dan pada masa yang sama berbagai teknik telah dibangunkan untuk mengesan pencerobohan atau serangan ke atas komputer dan rangkaian. Perlombongan data adalah teknik kecerdasan buatan yang telah dikenal pasti untuk pembangunan sistem pengesanan pencerobohan rangkaian. Berbilang teknik perlombongan data telah berjaya diaplikasikan di dalam bidang ini untuk melombong pencerobohan yang tersembunyi di dalam data audit menerusi pengkelasan, pengelompokan atau petua sekutuan. Pengelompokan adalah satu daripada teknik yang berpotensi untuk Pengesanan Anomali Pencerobohan, khasnya bila berhadapan dengan corak yang tidak diketahui. Kertas kerja ini memperihalkan kajian untuk meningkatkan prestasi pengesanan anomali pencerobohan menggunakan Fuzzy-ART berasaskan alkhwarizmi K-means. Kaedah K-means ini adalah versi K-means piawai yang telah ditambahbaik dengan menggunakan nilai K awal berdasarkan kepada nilai yang dilombong menggunakan Fuzzy-ART yang menggunakan teknik perlabelan SNC. Keputusan menunjukkan bahawa alkharizmi ini telah meningkatkan kadar pengesanan dan mengurangkan kadar penggera palsu berbanding penggunaan Fuzzy-ART semata-mata.

Kata Kunci: Pengesanan pencerobohan, pengesanan anomali, perlombongan data, set data NSL-KDD, Fuzzy-ART, K-means, perlabelan

INTRODUCTION

Intrusion detection is defined as a process of observing any event occurring in computer systems or networks, and analyzing the events for any signs of intrusion (Bauer & Koblentz 1988). An Intrusion Detection System (IDS) is a tool to provide security in computers or the network environment by stopping unlawful access to the system resources and the information stored within the system. An IDS can be a software or hardware system or a combination of both that aims to monitor the different events occurring in the network.

IDS can be classified into two categories, namely misuse detection and anomaly detection (Wang & Megalooikonomu 2005). In the misuse detection approach, each instance in training data is labeled as "normal" or "intrusion." It analyzes the gathered information and compares it to large databases of attack signatures. In contrast, anomaly detection (behavior-based) methods build models of normal data and then make an effort to detect deviation from the normal profile; any deviation is considered as an anomaly. According to Chandola et al. (2007), anomaly can be defined as a pattern in a data that is not consistent with the specific concept of normal behavior. Anomaly detection can make use of supervised or unsupervised methods to identify the anomalies. Unsupervised Anomaly Detection (UND) methods make some assumptions about the data that motivate the general approach (Portnoy et al. 2001). The first assumption is that the majority of the network connections are normal traffic,

with only a small number of traffic being anomalies. The second assumption is that the attack traffic is statistically different from normal traffic. Since the anomalies are different from the normal traffic, they will appear as outliers in the data which can be detected.

Data mining techniques can deal with a large volume of data required by both IDS categories to find intrusions hidden in network data. Clustering is one of the data mining techniques that have been used in anomaly intrusion detection, especially when dealing with unlabeled data. Several clustering algorithms has been proposed by Guan et al. (2003), Abdul Samad et al. (2008) and Ngamwitthayanon et al. (2009), that aim to have a high detection rate and low false alarm rate for intrusion detection. Among the clustering algorithms, Fuzzy-ART has shown the best clustering technique for intrusion detection. However, there is still much room for improvement. Some research have shown that a high detection rate with low false alarm rate limit the research into specific types of attack only.

This paper aims to improve the anomaly intrusion detection system using Fuzzy-ART based on the k-means clustering algorithm. K-means uses Fuzzy-ART for getting the initial stated value for K-means, while Fuzzy-ART uses K-means to reassign the data instances in the obtained clusters in order to achieve a better result as we named it as K-means.

CLUSTERING

Clustering is a common method for data analysis, which is used in many fields including pattern recognition, machine learning, and data mining. Clustering is the separation of data into subsets of similar objects, each subset being called a cluster. The cluster consists of objects that are related between themselves and dissimilar to the objects in other clusters (Lappas & Pellechrinis 2007).

Clustering is a well known data mining technique for unsupervised data. There are a lot of clustering methods and algorithms, which can be categorized into four categories: partitioning methods, hierarchical methods, density-based methods and grid-based methods (Leung & Leckie 2005).

Clustering techniques are beneficial in intrusion detection, which have the ability to cluster malicious activity together, and separate them from non malicious activity. Clustering techniques provide some considerable advantages compared with classification techniques, in that they do not require the use of a labeled data for training.

Applying clustering techniques in network intrusion detection has received much attention in the network academic community. Smith et al. (2002) has applied the algorithms K-means, Self-Organizing Maps (SOM), and Expected Maximization (EM). The results show that SOM has the same complexity regardless of the data volume or clusters used. However, SOM can incorrectly classify the input data that will correspond to points not found in training data. As Smith et al. (2002) claimed the K-means algorithm has a predictable performance. However, it poorly manipulates highly dimensional data sets. The EM algorithm can tolerate missing and unlabeled data and can offer information about how close a data point is to each cluster since the data point has a varying membership to all clusters.

Guan et al. (2003) introduced a K-means based clustering algorithm, called Y-means. It overcomes the deficiencies of the traditional K-means algorithm. The Y-means algorithm has some advantages. It does not need to determine the initial number of clusters in advance. Additionally, it can directly use the raw log data of information systems as training data without being manually labeled. This advantage provides the Y-means algorithm with the ability to detect known intrusions as well as unknown intrusions.

Abdul Samad et al. (2008) applied the Fuzzy Adaptive Resonance Theory (Fuzzy-ART) method in KDD Cup'99 benchmark data set to build an anomaly intrusion detection classifier. They also used Principal Component Analysis (PCA) for data reduction. Moreover, they proposed a new cluster labeling algorithm to label clusters as normal or abnormal (attack). The result shows that this research has a good performance for classifying the network connections.

Ngamwitthayanon et al. (2009) investigated a one shot fast learning option of Fuzzy-ART. They used a subset of KDD Cup'99 data set for experiments which contains a normal record connection and two types of attack categories, which are Dos and Probe. The results found that this can be applied to provide a real-time detection system with high detection rate and low false alarm rate.

FUZZY ADAPTIVE RESONANCE THEORY (FUZZY-ART)

The ART neural networks were developed by Carpenter et al. (1987). The ART neural networks can learn without forgetting past learning. The first version is ART-1 which can only deal with a binary data.

Carpenter et al. (1987) developed Fuzzy-ART systems which is an extension of ART-1. Fuzzy-ART incorporates computations from Fuzzy set theory into ART systems. This is by replacing the crisp inter-section operator ($\cap$) by the Fuzzy AND operator (^) of the Fuzzy set theory. Fuzzy-ART has several advantages over the ART-1 such as less implementation cost and processing time, and the capability of handling both binary and analog data. It has an unsupervised learning feature and does not require a predefined number of clusters. Figure 1 shows the Fuzzy-ART algorithm flowchart.

The Fuzzy-ART algorithm has been summarized by Carpenter et al. (1991) as follows:

Input vector: Every input vector I is an M-dimensional vector, where each component of I is in the interval [0, 1].

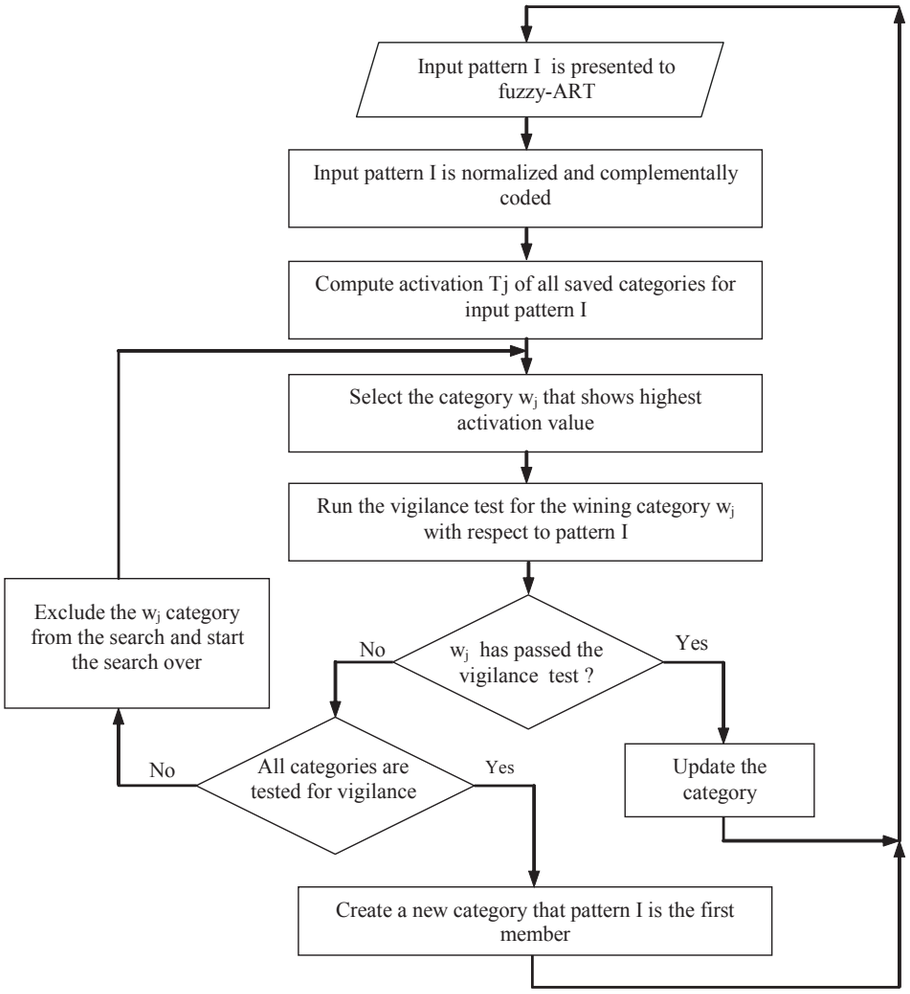

FIGURE 1. Fuzzy-ART algorithm flowchart

Weight vector: Each cluster or category J corresponds to a weight vector $w_j = (w_{ji}, ..., w_{ji})$ of adaptive weights. Initially, all weights are set to one and each category is said to be uncommitted.

$$w_{ji} = ... = w_{jm} = 1 \quad (1)$$

After a category is selected it becomes committed.

Parameters: Fuzzy-ART networks depend on these three main parameters:

1. Vigilance parameter ρ [0,1].
2. Choice parameter $\alpha > 0$.
3. Learning Rate parameter β [0,1].

Category Choice: For each input *I* and category *J*, the choice function of T_j is defined by:

$$T_j(I) = \frac{|I \wedge w_j|}{\alpha + |w_j|} . \quad (2)$$

Where the operator is fuzzy AND which is defined by:

$$(x \wedge y)_i = \min (x_i, y_i). \quad (3)$$

Activation function: The $T_j(I)$ in Equation (2) calculates the similarity between the input pattern I and all of saved categories. After evaluating the T_j for every category, the one with the largest value of choice function is selected:

$$T_j = \max\{T_j : j = 1,2,\ldots\ldots\ldots\ldots,N\}. \quad (4)$$

Where N is the number of categories. If there is more than one category with the same maximum value then the category with the smallest j is chosen.

Resonance or rest: category w_j that is chosen from the activation function in Equation (4) is compared with the input pattern I.

$$\frac{|I \wedge w_j|}{|I|} > \rho. \quad (5)$$

If the category w_j meets the vigilance parameter ρ then the w_j^{th} category is selected. Otherwise the j^{th} category is rested. If no category can be selected, a new category is committed with weights equal to 1, and the input vector is assigned to it.

Learning: After selecting a category j, the weight vector w_j is updated according to the next equation.

$$w_j^{(new)} = \beta\,(I \wedge w_j^{(old)}) + (1-\beta)\,w_j^{(old)}. \quad (6)$$

Input normalization option: To avoid a category proliferation problem in a Fuzzy-ART network, the inputs must be in normalized form. An alternative normalization rule is named as complement coding. The complements of the data are denoted by a^c, where:

$$a^c = 1\text{-}a. \quad (7)$$

Then, the complement coded input vector I is 2M-dimentional vector:

$$I = (a, a^c) = (a_1, ...,a_m, a_1^c, \ldots, a_m^c). \quad (8)$$

In the Fuzzy-ART algorithm, the fast learning occurs when the learning rate set to 1. Learning in the Fuzzy-ART can be interpreted as the extension of the category region towards the input sample. Furthermore, the vigilance parameter is the most important that can affect the Fuzzy-ART clustering quality, which controls the categorization by employing it as the criterion for a maximum tolerable value between the input data pattern and prototype of a category.

K-MEANS

The K-means clustering algorithm is a method used to partition a dataset into groups (MacQueen 1967). This algorithm classifies objects to a predefined number of clusters, which is the *k* from its name. Each cluster is represented by the mean value of the objects in the cluster, which is known as the center of the cluster.

The K-means clustering algorithm is one of the simplest unsupervised learning algorithms that have been adapted to solve many problem domains. The procedure of this algorithm follows a simple and easy way to classify a given data set to a certain number of clusters. The steps of the K-means algorithm are given as follows:

1. Select **k** points randomly to be the initial centroids of *k* clusters.
2. Assign each object to the closest centroid.
3. Recalculate the centroid of each cluster, which is an average of all attribute values of the examples belonging to the same cluster.
4. Repeat steps 2 and 3 until the centroids do not change.

FUZZY-ART BASED ON K-MEANS

Recent researches have supported the use of neural networks such as Fuzzy-ART, due to their flexibility and stability. However, Fuzzy-ART networks have some limitations such as the unrestricted growth of clusters, which is known as category proliferation. Fuzzy-ART networks often produce numerous clusters, each with only a small number of members.

In this paper, the Fuzzy-ART algorithm is used in order to find the groups of normal and attack instances, whereby each group shares common interests and behaviors. However, the groups or clusters that are obtained as a result of this algorithm still have many of the normal and attack instances in the same cluster. For this reason, this paper has applied K-means clustering algorithm to reassign the data instances in these clusters in order to achieve a better clustering quality.

Fuzzy-ART based on the K-means clustering algorithm consists of two phases (Figure 2). After generating clusters in the first phase using Fuzzy-ART depending on the vigilance parameter, K-means in the second phase takes these clusters as the initial clusters and reassigns each instance or object in these clusters to the cluster with the nearest centroid. This means that the cluster that has a minimum distance to this object using Euclidean distance, takes this object as a member. After reassigning all instances to the nearest clusters, the updates of the cluster centroids are calculated by the mean value of the objects in each cluster. This iteration was repeated until the centroids do not change any more.

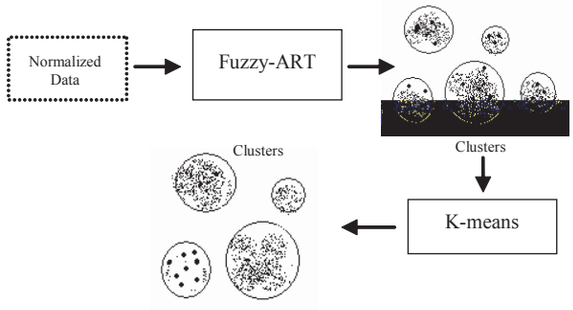

FIGURE 2. Architecture of Fuzzy-ART based on K-means clustering algorithm

One of the major problems of the K-means algorithm is that it may end with some empty clusters. The empty clusters are meaningless and they prolong the time of calculations. For these reasons, in this research the empty clusters have been removed within each iteration. Figure 3 shows the Fuzzy-ART based on K-means algorithm flowchart.

LABELING CLUSTERS

Labeling clusters is a technique to label the clusters as normal or attack after applying a clustering process. Portnoy et al. (2001) and Abdul Samad et al. (2008) proposed a labeling algorithm known as Normal Membership Factor (NMF) and this paper introduces a labeling algorithm known as Similarity Normal Clusters (SNC) (Provost et al. 2001).

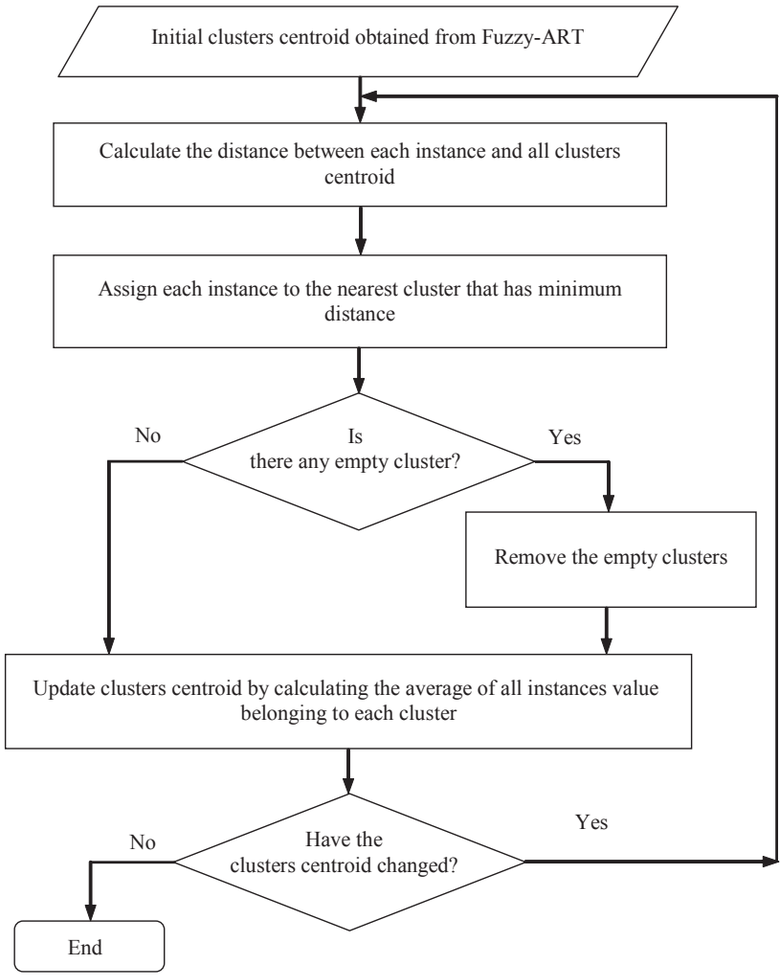

FIGURE 3. Flowchart of Fuzzy-ART based on K-means Algorithm

NORMAL MEMBERSHIP FACTOR (NMF) LABELING ALGORITHM

The aims of the NMF labeling algorithm is to identify the other clusters that may be having normal patterns and they are gathered into a normal group in order to reduce the false alarm rate. This algorithm uses the concept of calculating the weighting degree of probability of the clusters belonging to a normal group. The calculation of the weight of clusters and the NMF of the cluster is made as follows:

1. Weight of clusters (c_i): the weight (size) of the cluster is considered to greatly reduce the effect of anomalies. The calculation of the weight of clusters is given as Equation 9.

$$WC(c_i) = \frac{1}{d(Normal, c_i)} \times \frac{\text{Number of instance in } c_i}{\text{Number of instance}}. \quad (9)$$

Where d(Normal, c_i) is the distance between the normal cluster and the other clusters, and i is the number of clusters.

2. Normal membership factor (c_i): the normal membership factor is calculated as Equation 10.

$$NMF(c_i) = \frac{WC(c_i)}{\sum_{i=1}^{0} c_i}. \quad (10)$$

Where $\sum_{i=1}^{0} c_i$ summarizes all the weight of the cluster.

If *NMF* values are greater than 40 percent, then these are gathered into those clusters which are in the normal group. The percentage was determined from previous experiment performed for this approach.

SIMILARITY NORMAL CLUSTER LABELING TECHNIQUE

The SNC is a cluster labeling algorithm which is built based on the assumptions introduced by Portnoy et al. (2001). SNC calculates the similarity percentage between the normal cluster centroid, which has the largest size and other cluster centroids using the Euclidean distance. The aim of this algorithm is to identify the other normal

clusters that may have a small number of instances. Figure 4 shows example of five clusters. The labeling algorithm are illustrated as follows:

Step 1: Identify the largest cluster, the one with the large number of instances and label it as normal as shown in Figure 4.

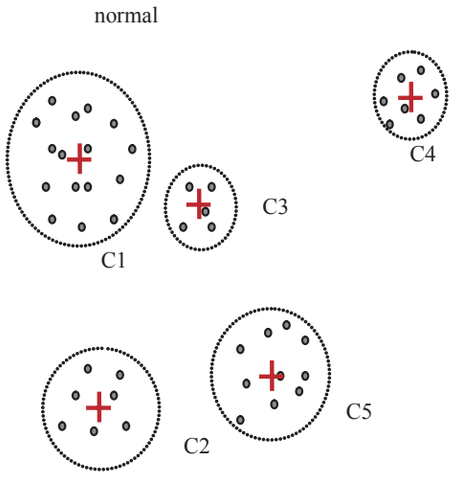

FIGURE 4. Size of cluster notation

Step 2: Calculate the distances between the normal cluster centroid C1 and other cluster centroids using the Euclidean distances as shown in Figure 5. The Euclidean distances equation as shown in Equation 11.

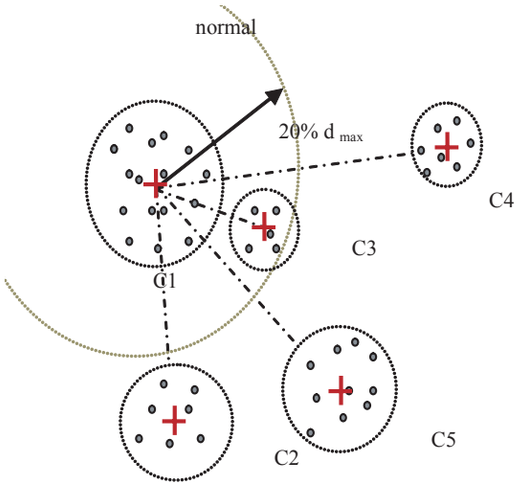

FIGURE 5. Area of selection of normal clusters

Step 3: If the distances are less than 20%, then gather these clusters into the normal group. The 20% value was concluded as the best percentage after several rounds of experiments for this proposed approach.

Step 4: Find the largest cluster in the remainder clusters.

Step 5: If the size of the largest cluster is larger than $\frac{1}{i}N$, then label this cluster as normal and go back to step 2. Otherwise label this cluster and all the rest of the clusters as attack.

Where i is the number of attack categories and N is the number of attack instances in the data set.

Where $\boldsymbol{i}$ is the number of attack categories and $\boldsymbol{N}$ is the number of attack instances in the dataset. The chosen size of the largest cluster is based on the analysis result.

The remaining clusters from the last step are C2, C4 and C5. The largest one is C5. Suppose that the size of this cluster is larger than the threshold, then label C5 as normal and go back to Step 2.

SIMILARITY MEASURES

Similarity is a fundamental concept in clustering. It measures the similarity between two patterns drawn from the same feature space. One of the challenges in clustering is to choose suitable similarity measurement techniques based upon the feature type. The concept of similarity in most clustering techniques is based on distances. Some well-known distance functions include Euclidean distance, Manhattan distance, and cosine distance. The most used distance measure to calculate the similarity between two objects is Euclidean distance. Euclidean distance is sufficient to successfully group similar data objects.

Simply, Euclidean distance is the geometric distance in multidimensional space. For instance, for two vectors x and y in an n-dimensional space, it is defined as the square root of the sum of the difference of the corresponding dimensions of the vector. It is computed:

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \quad (11)$$

Where n is the number of dimensions in the data vector.

EXPERIMENTAL SETUP

The experiments conducted follow the standard data mining process (Han & Kamber 2001), which consists of collecting the dataset, preprocessing, normalization and mining.

The experiments used NSL-KDD data collected from the website via the link (http://nsl.cs.unb.ca/NSL-KDD). NSL-KDD data is considered as a modified dataset for KDD Cup '99 intrusion detection benchmark dataset (KDD Cup '99). This dataset was suggested by Tavallaee et al. (2009) to solve some of the inherent problems of the KDD Cup '99 dataset.

This research deals with 20% NSD-KDD training dataset, which consists of 25,191 records which were reduced to 13,650 records with 41 attributes. The number of attacks was randomly selected (Abdul Samad et al.

2008), and reduced to about 1.48% of the complete dataset to match the assumption about the distribution of normal and attack instances in the dataset (Han & Kamber 2001). The experimental dataset contains all types of attack which belong to one of the four attack categories. For each of Dos, Probe, U2R, and R2L attacks, there were 83, 70, 11, and 38 records, respectively. Three data sets were generated randomised from the original data set as Data Set1, Data Set 2 and Data Set 3. Table 1 shows the description of datasets used in this research. The first row shows the description of original data set, the other three are the description of other data sets, where the distribution number of records are same in all experiments.

Each connection record in the experimental dataset is described by 41 attributes; some of these attributes are continuous and some are symbolic or discrete. Since most clustering algorithms require a normalized data, these attribute values have been converted into a normalized form using Min-Max normalization (Zong et al. 2001) in order to feed these algorithms.

Before the normalization all attribute values are discretized. For symbolic features like protocol-type, service, and flag they are mapped manually to integer values ranging from 1 to N where N is the number of symbols (Chimphlee et al. 2006). These attributes were numbered according to their distribution in the data set after sorting the attribute values in a descending order (Shirazi 2009). Attributes like duration, source bytes, etc. are discretized using the ChiMerge discretization technique (Kerber 1992).

All experiments were carried out using MATLAB r2007b on a laptop with Microsoft Windows Vista operating system. Finally, the evaluation of anomaly intrusion detection methods is done using two main measures, Detection Rate (DR) and False Alarm Rate (FAR). DR is defined as the percentage of attacks correctly identified by the system while the FAR is the percentage of normal instances wrongly identified as attack by the system.

The experiments are trained with the number of epoch value of 100, learning rate equal to 1, choice parameter of 0.000001, and with different number of vigilance value. The result of the experiments recommended the range from 0.5 to 0.85 as the values of vigilance parameter. These experiments were conducted for Fuzzy-ART based on K-means versus Fuzzy-ART using the SNC labeling algorithm and Fuzzy-ART based on K-Means versus Fuzzy-ART using NMF labeling algorithm.

TABLE 1. Data set descriptions

Data	Attach types	Number of Original Records	Number of Record Data Set 1	Number of Record Data Set 2	Number of Record Data Set 3
Normal	--	67,343	13,448	13,448	13,448
Dos	Back	956	10	10	10
	Land	18	1	1	1
	Neptune	41,214	40	40	40
	Smurf	2646	15	15	15
	Pod	201	5	5	5
	Teardrop	892	12	12	12
Probe	Ipsweep	3599	25	25	25
	Nmap	1493	10	10	10
	Portsweep	2931	15	15	15
	Satan	3633	20	20	20
U2R	buffer_overflow	30	6	6	6
	Perl	3	0	0	0
	Loadmodule	9	1	1	1
	Rootkit	10	4	4	4
R2L	ftp-write	8	1	1	1
	guess-password	53	10	10	10
	Imap	11	5	5	5
	Multihop	7	2	2	2
	Phf	4	2	2	2
	Spy	2	1	1	1
	Warezclient	890	10	10	10
	Warezmaster	20	7	7	7
Total		125,973	13,650	13,650	13,650

EXPERIMENT STEPS

The experimental steps follow the standard data mining process using clustering for anomaly detection proposed by Abdul Samad et al. (2008) which consists of data preprocessing, mining, labeling and evaluation. Figure 6 shows the experiment steps conducted in this research.

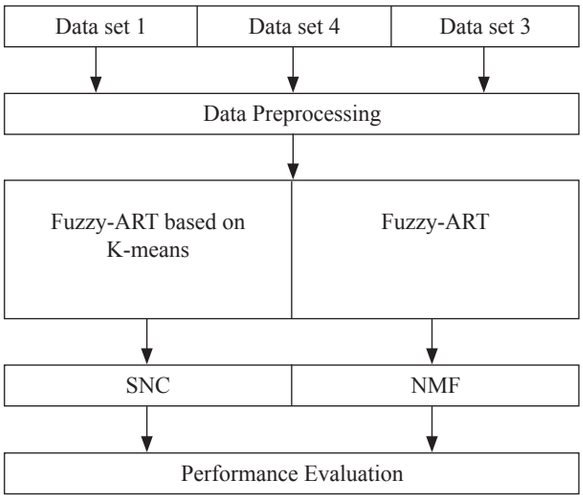

FIGURE 6. Experiment Steps

EXPERIMENTAL RESULT

The vigilance parameter decides the number of categories (clusters) and the degree of the similarity of each cluster. The network becomes more sensitive to the small changes in input patterns with increases of the vigilance value. Thus, it will increase the number of resulting clusters. Table 2 shows the results of Fuzzy-ART based on K-means algorithm with different vigilance values for Data Set 1.

The results show that the false alarm rate increased with the increase of the vigilance value because more numbers of clusters are classified as attack. At the same time, the number of clusters increases. The results also show that the best performance of Fuzzy-ART based on K-means with SNC algorithm is obtained when the value of the vigilance parameter is equal to 0.8.

TABLE 2. Results of Fuzzy-ART based on K-means algorithm with different vigilance values for data set 1

Vigilance ρ	Detection Rate %	False Alarm Rate %	No. of Clusters
0.50	34.15	0.27	24
0.55	45.54	1.10	37
0.60	41.08	1.79	48
0.65	36.63	2.80	53
0.70	46.03	3.29	71
0.75	72.77	5.49	84
0.80	85.64	8.96	71
0.85	89.11	14.19	97

In an effort to show more explanation on the performance of Fuzzy-ART based on K-means algorithm, Table 3 presents the detection rate of each category. The results show that the algorithm with vigilance values of 0.8 and 0.85 has the best level of detection rate for Dos, Probe, U2R, and R2L. However, the detection rate for normal category with these two vigilance values is lower than the detection rate with the other values of vigilance parameter, which led to an increase in the false alarm rate.

TABLE 3. Detection rate of Fuzzy-ART based on K-means algorithm for each category

Vigilance ρ	Normal	Dos	Probe	U2R
0.50	99.72	35.01	31.42	0.00
0.55	98.89	48.19	51.42	45.45
0.60	98.20	49.39	44.28	45.45
0.65	97.17	53.01	31.42	18.18
0.70	96.70	50.60	51.42	54.54
0.75	94.50	84.33	60.00	63.68
0.80	91.03	85.54	91.42	63.63
0.85	85.80	89.15	95.71	90.90

Table 4 shows comparison result of Fuzzy-ART based on K-means versus Fuzzy-ART algorithm. The results show that the Fuzzy-ART based on K-means has better performance in terms of both the detection rate and the false alarm rate than the Fuzzy-ART. That means the K-means algorithm can enhance the performance of the Fuzzy-ART.

TABLE 4. Comparison results of Fuzzy-ART based on K-means and Fuzzy-ART using SNC labeling Technique

Method	Detection Rate %	False Alarm Rate %
Fuzzy-ART	81.68	9.59
Fuzzy-ART based on K-means	85.64	8.96

The performance of the algorithm was also measured using ROC curves. Figure 7 shows the comparison results of Fuzzy-ART based on K-means versus Fuzzy-ART algorithm using ROC curves. The figure shows the percentage of detection rate versus percentage false alarm rate. According to Provost and Fawcett (2001), the upper left point (0, 1) represents the ideal IDS, which has a 100% detection rate and 0% false alarm rate.

The figure shows that, the curve of Fuzzy-ART based on K-means is higher than the Fuzzy-ART. This means that the Fuzzy-ART based on K-means algorithm is more accurate and more effective than the Fuzzy-ART alone. The figure also shows that, the Fuzzy-ART based on K-means has a detection rate about 70% with 5% false alarm rate, while the Fuzzy-ART has detection rate of about 30% with the same false alarm rate.

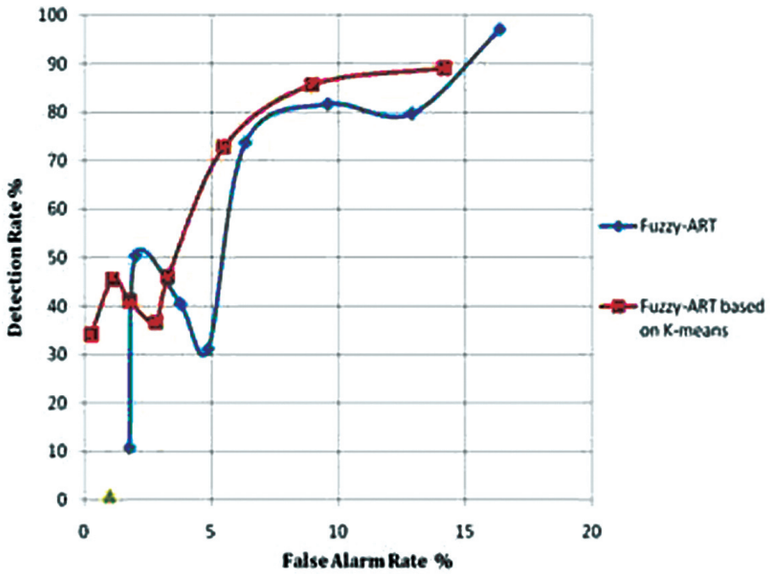

FIGURE 7. ROC Curves for Fuzzy-ART versus Fuzzy-ART based on K-means algorithm

The performance of the SNC labeling algorithm is then further evaluated by comparing it with the NMF labeling algorithm using the Fuzzy-ART based on K-means clustering algorithm. Table 5 and Table 6 show the mining result for the SNC and NMF labeling algorithms based on vigilance numbers, respectively. The highlight cells shows the best result among various vigilance parameters.

Table 5 and Table 6 shows that the vigilance parameter controls the degree of similarity between the instances in each cluster. The results show that the false alarm rate increases with the increase of the vigilance parameter because there are a greater number of normal instances which are classified as attack. Figure 5, Figure 6 and Figure 7 show the Receiver Operating Characteristic (ROC) curve for SNC and NMF using the Fuzzy-ART algorithm for Data Set 1, Data Set 2 and Data Set 3, respectively. ROC curves are a way of visualizing the trade-offs between percentage of detection rate and the false alarm rate.

TABLE 5. Performance Fuzzy-ART based on K-means using SNC for the three data sets

Vigilance Parameter	Data Set 1		Data Set 2		Data Set 3	
	Detection Rate %	False Alarm Rate %	Detection Rate %	False Alarm Rate %	Detection Rate %	False Alarm Rate %
0.50	34.1584	0.2751	19.1919	0.5724	14.8515	0.6916
0.55	45.5446	1.1005	22.7273	0.6542	36.6337	0.2305
0.60	41.0891	1.7995	34.3434	0.5947	43.0693	1.4128
0.65	36.6337	2.8257	37.8788	1.2340	49.5050	2.4465
0.70	46.0396	3.2942	73.7374	3.5459	74.2574	2.1862
0.75	72.7723	5.4952	88.3838	3.1594	86.6337	5.1234
0.80	85.6436	8.9679	93.9394	6.7797	95.5446	8.8192
0.85	89.1089	14.1954	99.4949	15.6705	97.5248	14.68682

TABLE 6. Performance Fuzzy-ART based on K-means using NMF for the three data sets

Vigilance Parameter	Data Set 1		Data Set 2		Data Set 3	
	Detection Rate %	False Alarm Rate %	Detection Rate %	False Alarm Rate %	Detection Rate %	False Alarm Rate %
0.50	93.5644	50.4090	94.9495	49.0485	97.0297	49.0928
0.55	95.5446	55.5250	95.4545	52.7357	97.5248	50.2156
0.60	95.5446	57.5476	95.4545	49.7770	97.5248	56.3430
0.65	98.5149	55.5770	97.4747	60.8980	92.5743	47.5907
0.70	97.0297	58.4176	97.9798	58.8017	98.5149	59.2579
0.75	100	58.1499	100	59.7532	100	59.5999
0.80	100	57.9194	100	58.3779	100	49.6208
0.85	100	62.3215	100	60.0059	100	57.9789

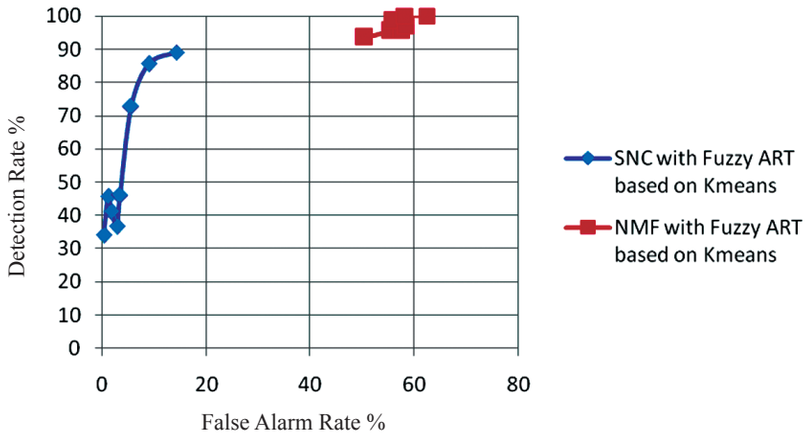

FIGURE 8. ROC Fuzzy-ART K-Means using SNC labeling versus NMF labeling for Data Set 1

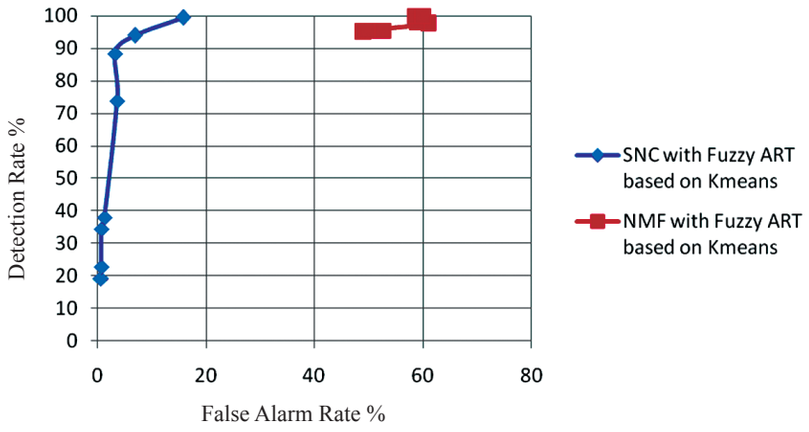

FIGURE 9. ROC Fuzzy-ART K-Means using SNC labeling versus NMF labeling for Data Set 2

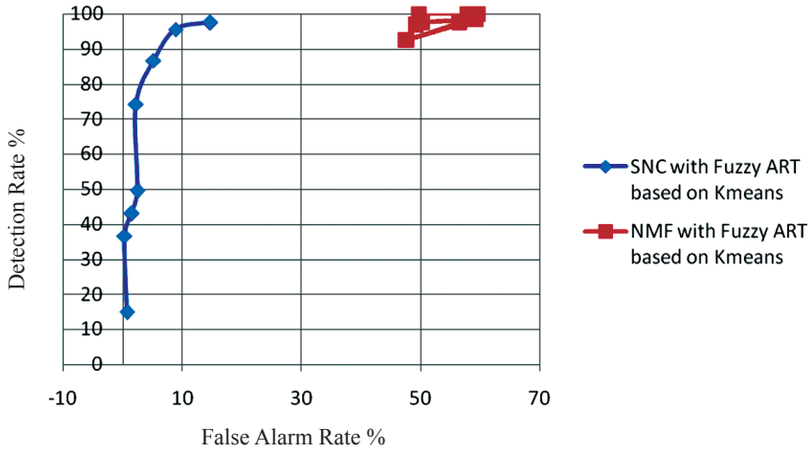

FIGURE 10. ROC Fuzzy-ART K-Means using SNC labeling versus NMF labeling for Data Set 3

CONCLUSION

This paper has improved an anomaly intrusion detection by applying two clustering techniques: Fuzzy-ART and K-means algorithms. The Fuzzy-ART algorithm is first applied to a subset of NSL-KDD network data set in order to find groups of normal and attack instances, which each group sharing common interests and behaviours. However, the obtained clusters still have many normal and attack instances in the same cluster. Therefore, K-means algorithm has been applied to reassign the data instances in these clusters to achieve a better clustering quality. The experimental results showed that the Fuzzy-ART based on K-means algorithm has been able to increase the detection

rate and reduce the false alarm rate compared with Fuzzy-ART algorithm. Future research direction will investigate whether the use of feature selection techniques to extract the optimum feature subset from the whole feature in the dataset will increase the performance of anomaly intrusion detectors.

REFERENCES

Abdul Samad, H.I., Abdul Hanan, A., Kamalrulnizam, A.B., Md Asri, N., Dahliyusmanto, D. & Chimphlee, W.A. 2008. *Novel Method for Unsupervised Anomaly Detection using Unlabelled Data*. Proceedings of the International Conference on Computational Sciences and Its Applications (ICCSA).

Afaf Muftah Adabashi. 2011. Similarity Normal Cluster Labeling for Network Traffic, Master Thesis, FTSM, UKM.

Anderson, J.P. 1980. *Computer Security Threat Monitoring and Surveillance*. Technical Report. James P Anderson Co. Fort Washington. Pennsylvania.

Bauer, D.S. & Koblentz, M.E. 1988. NIDX - an Expert System for Real-time Network Intrusion Detection. *Proceedings of the Computer Networking Symposium*: 98-106.

Carpenter, Gail, Grossberg & Stephen. 1987. A Massively Parallel Architecture for a Self Organizing Neural Pattern Recognition Machine. *Computer Vision, Graphics and Image Processing* 37(1):

Carpenter, G.A., Grossberg, S. & Rosen, D.B. 1991. Fuzzy-ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance Theory. *Neural Networks* 4: 759-771.

Chandola, V., Banerjee, A. & Kumar, V. 2007. *Anomaly Detection: A Survey*. Technical Report, Department of Computer Science and Engineering. University of Minnesota.

Chimphlee, W., Abdullah, A.H., Md Sap, M. Noor, Srinoy, S. & Chimphlee, S. 2006. Anomaly-Based Intrusion Detection Using Fuzzy Rough Clustering. *Proceedings of the International Conference on Hybrid Information Technology (ICHIT)*: 329-334.

Guan, Y., Ghorbani, A. & Belacel, N. 2003. Y-means: A Clustering Method for Intrusion Detection. *Proceedings of Canadian Conference on Electrical and Computer Engineering* 2: 1083-1086.

Han, J. & Kamber, M. 2001. *Data mining concepts and techniques*. United States of America: Academic Press.

KDD Cup '99. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Kerber, R. 1992. ChiMerge: discretization for numeric attributes. *In Proceedings of the Tenth National Conference on Artificial Intelligence*. AAAI-92. San Jose. CA. AAAI Press.

Lappas, T. & Pelechrinis, K. 2007. *Data Mining Techniques for (Network) Intrusion Detection Systems*. Department of Computer Science and Engineering. California: University of California.

Leung, K. & Leckie, C. 2005. *Unsupervised Anomaly Detection in Network Intrusion Detection Using Clusters*. Proceedings of the Twenty-eighth Australasian Conference on Computer Science: 38.

MacQueen, J.B. 1967. *Some Methods for Classification and Analysis of Multivariate Observations*. Proceedings of 5th Berkeley Symposium on Mathematical, Statistics and Probability: 281-297.

Ngamwitthayanon, N., Wattanapongsakorn, N. & Coit, D. W. 2009. *Investigation of Fuzzy Adaptive Resonance Theory in Network Anomaly Intrusion Detection*. Proceedings of the 6th International Symposium on Neural Networks: Advances in Neural Networks. Wuhan. China.

Portnoy, L., Eskin, E. & Stolfo, S. 2001. *Intrusion Detection with Unlabeled Data Using Clustering*. In Proceedings of ACM CSS Workshop on Data Mining Applied to Security.

Provost, F. & Fawcett, T. 2001. Robust classification for imprecise environments. *Machine Learning* 42: 203-231.

Shirazi, H.M. 2009. Anomaly intrusion detection system using information theory, K-NN and KMC algorithms. *Australian Journal of Basic and Applied Sciences* 3(3): 2581-2597.

Smith, R., Bivens, A., Embrechts, M., Palagiri, C. & Szymanski, B. 2002. Clustering approaches for anomaly based Intrusion detection. *Walter Lincoln Hawkins Graduate Research Conference*. New York. USA.

Tavallaee, M., Bagheri, E., Lu, W. & Ghorbani, A. 2009. A Detailed Analysis of the KDD CUP 99 Data Set. *Proceeding of IEEE symposium on Computational Intelligence in Security and Defense Applications (CISDA)*.

Wang, Q. & Megalooikonomou, V. 2005. A Clustering Algorithm for Intrusion Detection, *Proceedings of the SPIE Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security*, 5812: 31-38.

Zhong, S., Khoshgoftaar, T. & Seliya, N. 2007. Clustering-Based Network Intrusion Detection. *International Journal of Reliability, Quality and Safety Engineering* 14: 169-187.

Zulaiha Ali Othman
Afaf Muftah Adabashi
Suhaila Zainudin
Centre of Artificial Intelligence Technology,
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, 43650 Bangi, Selangor
zao@ftsm.ukm.my, amd_shsh@yahoo.com

Saadat M.Alhashmi
School of IT
Monash University, Sunway Campus
Petaling Jaya, Selangor
alhashmi@monash.edu