Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 14 (10): 969-980, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.969.980

A Similarity Normal Clustering Labelling Algorithm for Clustering Network Intrusion Detection

Zulaiha Ali Othman, Azuraliza Abu Bakar, Afaf Muftah Adabashi and Zurina Muda
Centre of Artificial Intelligence Technology, Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, Malaysia

Abstract: Clustering is one of the promising techniques used in Anomaly Intrusion Detection (AID) especially to detect unknown patterns. Two factor influence accuracy IDS using this technique: clustering and labelling algorithm. Fuzzy Adaptive Resonance Theory (Fuzzy ART) is well known algorithm for high accuracy but has high false-alarm rate, while Normal Membership Factor (NMF) is a good labelling algorithm for IDS, but preliminary experiments found that many clusters are labelled incorrectly. Therefore this paper proposed a new labelling algorithm known as Similarity Normal Cluster (SNC) and improved the Fuzzy ART clustering technique using K means. The SNC uses fundamental assumption of NMF, but similarity is measured based on the percentage of similarity among the regular clusters, using the Euclidean distance repeatedly in the cluster and ensuring that all clusters are measured. The performance of the proposed labelling algorithm is evaluated by comparing it with the NMF and with Fuzzy ART and Fuzzy ART with Euclidean ART respectively. The experiment is conducted using 10 data sets collected from the NSL-KDD dataset. The result shows that SNC always deliver better results than NMF whilst Fuzzy Art with SNC obtained the best combination result compare to others.

Key words: Intrusion detection, anomaly detection, NSL-KDD data set, data mining, clustering, labelling technique, NMF labelling algorithm

INTRODUCTION

The demand of having quality Intrusion Detection System (IDS) keep increasing due the increasing use of computer networks and the ubiquitous presence of the internet, networks have become a primary target for hackers, even for those with limited experience in networks, who are able to access networked computers without a trace. An (IDS) detects actions that attempt to compromise the confidentiality, integrity or availability of the system (Folorunso *et al*., 2010). In implementation, IDS can be host-based (HIDS) or network-based (NIDS). HIDS operates on information collected from an individual host machine, whereas NIDS monitors the data packets that pass through the network links. In addition, NIDS attempts to find intrusions hidden in large amounts of network data and can be achieved by employing data mining techniques. On the other hand, IDS can be classified into two categories: misuse detection and anomaly detection (Zhao *et al*., 2009). In the misuse detection approach, each instance in the training data set is labelled as “normal” or “intrusion”. The system analyses the information gathered and compares it with a large database of attack signatures. Anomaly detection methods build models of normal data and attempt to detect deviations from the normal profile. Applying clustering techniques to anomaly intrusion detection has salient advantages. The labelled data are expensive, but the un-labelled data can be obtained easily from a real-world system. The main advantage of using unsupervised learning, or clustering, to detect network attacks is the ability to find new attacks that have not been previously recorded (Zhong *et al*., 2007).

Three factors influence the accuracy of anomaly detection in clustering: The clustering algorithm, the labelling algorithm and the complexity of network traffic data. The focus of this research is on the clustering and labelling techniques. The clustering algorithm is able to gather the data into several groups. The challenge of the clustering algorithm is to produce distinct groupings with the fewest clusters. In IDS, each cluster is later defined as normal or attack in the labelling process.

Corresponding Author: Zulaiha Ali Othman, Centre of Artificial Intelligence Technology,
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

Labelling is an important element in clustering. It influences the interpretation of the clustering results. Without prior knowledge of labelling, it is difficult to determine whether the data are classified in the cluster that corresponds to normal or abnormal (Petrovic *et al*., 2006). Existing labelling techniques are based on various assumptions. The most common assumptionsare that the number of normal instances is much higher than the number of attacks in the data set and that the attack instances are qualitatively different from the normal data set (Thamaraiselvi *et al*., 2009). Some researchers set a percentage assumption, based on the data sets, whereas several others calculate the distance between the normal and abnormal cases, which is the approach proposed by Abdul Samad *et al*. (2008). However, applying the current labelling algorithm shows that there are some abnormal objects fall into the normal cluster and vice versa. Thamaraiselvi *et al*. (2009) labelled the clusters based on the following assumptions: They labelled the clusters that contained the largest number of instances as normal and labelled the remainder of the clusters as attack. Fang and Le-Ping (2005) assumed that if the number of data in some clusters is larger than 10% of the entire data set, these clusters are normal clusters and all the other clusters are attack clusters. Zhong *et al*. (2007) proposed a simple self-labelling heuristic to detect attack by sorting the clusters and the data points in ascending order of the distance to the centroid of the normal cluster that contains the most instances. Abdul Samad *et al*. (2008) proposed a new algorithm for labelling clusters, namely, the Normal Membership Factor (NMF). Prior to applying the NMF, there are many clustersthat have not been classified into either normal or attack clusters. Therefore, it is important to include labelling that considers all of the clusters and assigns them to previously defined categories. These techniques are able to reduce the false-alarm rate.

The goal of the NMF labelling algorithm is to identify the other clusters that may exhibit normal patterns and gather them into a normal group to reduce the false-alarm rate. This algorithm calculates the weighted degree of probability thatthe clusters belong to a normal group. The calculation of the weight of the clusters and the NMF of the cluster is shown by Abdul Samad *et al*. (2008).

However this study proposes a new labelling algorithm that uses the basic assumption stated by Thamaraiselvi *et al*. (2009) and calculates the distance of cluster groups as proposed by Zhong *et al*. (2007) and the process to identify the cluster as normal or abnormal is performed repeatedly in the remaining clusters based units similarity. The similarity measured based on the percentage of similarity among the normal clusters, using the Euclidean distance.

Clustering is a method that groups data objects in clusters based on information found in the data set that describes the objects and their relationships. The objects within a cluster are found to be similar to one another and dissimilar to the objects in other clusters. Clustering algorithms have been widely used in the anomaly intrusion detection field, especially when dealing with unknown patterns.

Clustering techniques are beneficial in intrusion detection because they have the ability to cluster abnormal activity together and separate it from normal activity. Many clustering techniques have been used in this field. Smith *et al*. (2002) used self-organising maps, K-means and the expected maximisation algorithm to develop processing tools for other detection processes in a DARPA data set (Smith *et al*., 2002). Guan *et al*. (2003) proposed Y-means as an improvement of K-means and showed the ability to detect intrusion in the KDD Cup 99 dataset. Liu *et al*. (2004) proposed a genetic algorithm and nearest neighbor as an effective network intrusion detection algorithm and Ngamwitthayanon *et al*. (2009) proposed the Fuzzy-connectedness Clustering (FCC) algorithm, which was able to achieve a detection intrusion rate above 94% and a false-alarm rate below 4% using the KDD Cup 99 data set.

Shirazi (2009) proposed the Fuzzy Rough C-means algorithm, which performs better than the K-means algorithm when applied to the KDD Cup 99 data set. Zhong *et al*. (2007) applied the K-means, mixture-of-spherical Gaussians, self-organising map and neural-gas algorithms to the DARPA 1998 data set. The results show the advantage of clustering-based methods over supervised classification techniques in identifying new or unseen attack types. Abdul Samad *et al*. (2008) proposed a Fuzzy adaptive resonance theory (Fuzzy ART) algorithm for clustering using the KDD Cup 99 data set. The Fuzzy ART approach achieved good results by applying the principal component analysis (PCA) for feature selection. Carpenter *et al*. (1991) and Ngamwitthayanon *et al*. (2009) also applied the Fuzzy ART algorithm to the KDD Cup 99 data set. The results showed that Fuzzy ART has potential for network anomaly intrusion detection applications, with the ability to perform adaptive real-time clustering with a high detection rate and a low false-alarm rate. Later, the ART neural networks is integrated with Fuzzy, which the neural networks can learn without forgetting past learning developed by Carpenter *et al*. (1991) and the algorithm is shows by Kenaya and Cheok (2008).

Fuzzy ART suffers from a few disadvantages, such as sensitivity to noise and representations of fuzzy categories (He *et al*., 2002). Euclidean adaptive resonance theory (Euclidean-ART) is the solution suggested by Kenaya and Cheok (2008) to solve these problems. The approach is an unsupervised learning algorithm that is analogous to Fuzzy Art. The Euclidean adaptive resonance theory (Euclidean-ART) is a clustering method that evaluates the Euclidean distance between input vectors and cluster weights to decide the pattern's clustering (Ngamwitthayanon *et al*., 2009). In this method, the fuzzy operations found in Fuzzy ART are replaced with Euclidean distances. The learning rule is an averaging procedure used to calculate the new cluster centre's position after a new pattern is added to the cluster.

K-means clustering algorithm is used to partition a data set into groups (Sharma *et al*., 2012). The algorithm classifies objects in a predefined number of clusters from which the "K" takes its name. Each cluster is represented by the mean value of the objects in the cluster, also known as the centre of the cluster.

The K-means clustering algorithm is one of the simplest unsupervised learning algorithms and has been adapted to many problem domains. The algorithm follows a simple and easy-to-use procedure to classify a given data set to a certain number of clusters (MacQueen, 1967).

The main objective of an intrusion detection system is to identify attacks on network traffic. Generally, attack typesin KDD cup 99 data setsfall into one of four categories according to their behaviour (Guo *et al*., 2012). These categories are denial of service (DoS), remote to local (R2L), user to root (U2R) and probing. Some of these attacks, such as DoS and probing, may use hundreds of connections, whereas other attacks use only a few connections.

Data mining has the advantage of extracting intrusions from a large network data system and was first applied in the intrusion detection field by Stein *et al*. (2005). Various data mining techniques, such as classification (Stein *et al*., 2005), clustering (Xie *et al*., 2010) and association rule (Zhang, 2005), are used to detect intrusions.

Despite of several clustering algorithms, FuzzyART has shown the best clustering algorithm in IDS (Carpenter *et al*., 1991). This study aim to enhance the FuzzyART based on K-means and apply Euclidean-ART clustering in IDS using a proposed labelling algorithm name Similarity Normal Cluster.

MATERIALS AND METHODS

In this section, propose an improved Fuzzy ART clustering based on K-means clustering algorithm and a new labelling algorithm called the Similarity Normal Cluster (SNC).

Fuzzy ART based on K-means clustering algorithm: Fuzzy ART based on K-means clustering consists of two phases. First, the Fuzzy ART algorithm is used to generate the initial clusters. The clusters generated in this phase may still contain instances of both of normal and attack. Therefore, the second phase is used to re-examine each of these clusters.

In the second phase, K-means is employed, using the clusters result from phase 1 as the initial clusters and reassigning each instance or object in the clusters to the cluster with the nearest centroid. This means that the cluster that has a minimum distance to this object using Euclidean distance and 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 is repeated until the centroids stop changing. Common K-means step apply randomised to get initial cluster. The K-means also very sensitive with initial value, therefore applying the Fuzzy ART aims to get better initial cluster before applying clustering again using K-means. Figure 1 illustrates the proposed integration of the Fuzzy Art and K-means algorithms.

A major problem with the K-means algorithm is that it may result in some empty clusters. The empty clusters are meaningless and prolong the calculation time. Therefore, the empty clusters are removed at each iterations in this study. Figure 2 shows the Fuzzy ART based on the K-means algorithm flowchart.

Similarity normal cluster labelling algorithm: Similarity is a fundamental concept in clustering (Borgatti, 2012). It measures the similarity between two patterns in the same feature space. One of the challenges in clustering is to choose suitable similarity measurement techniques based on feature type (Zhang *et al*., 2006). Similarity in most clustering techniques is based on distances. Similarity Normal Cluster (SNC) is a labelling algorithm used to label a cluster as normal or abnormal. SNC determines which other clusters may contain a small number of normal instances and then labels these clusters as normal to reduce the false-alarm rate.

The SNC labelling algorithm calculates the similarity percentage between the normal cluster centroid, which

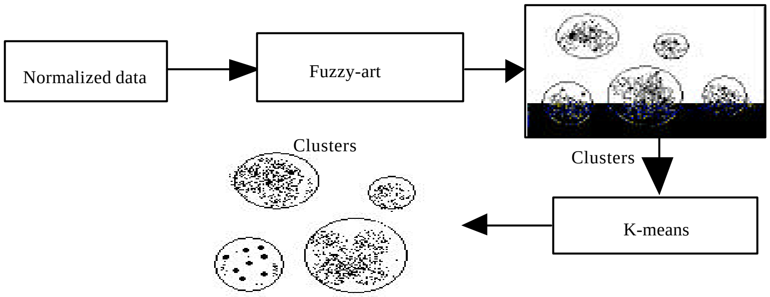

Fig. 1: Architecture of fuzzy art based on K-means clustering algorithm

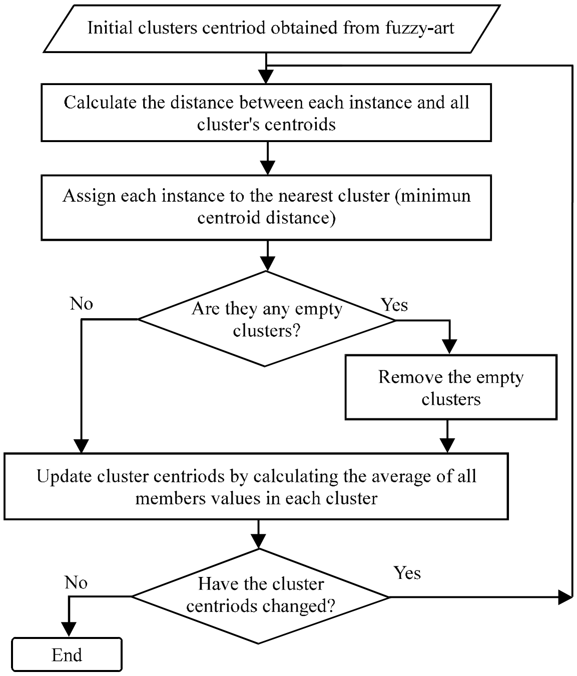
Initial clusters centriod obtained from fuzzy-art
Calculate the distance between each instance and all cluster's centroids
Assign each instance to the nearest cluster (minimun centroid distance)
No
Are they any empty clusters?
Yes
Remove the empty clusters
Update cluster centriods by calculating the average of all members values in each cluster
No
Have the cluster centriods changed?
Yes
End

Fig. 2: Flowchart of fuzzy ART based on K-means algorithm

has the largest size and other cluster centroids, using the Euclidean distance (Afaf Muftah, 2011). The steps of this algorithm are as follows:

- **Step 1:** Identify the largest cluster, that is, the one with the largest number of members and label it as normal, as shown in Fig. 3. Step 1 follows the previous studies (Thamaraiselvi *et al.*, 2009)
- **Step 2:** Calculate the distances between the normal cluster centroid, C1 and other cluster centroids using the Euclidean distances, as illustrated in Fig. 4

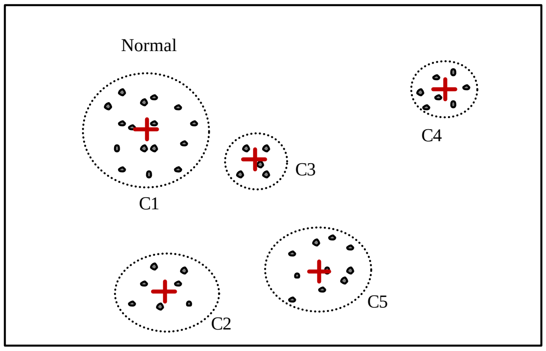

Fig. 3: Size of cluster notation

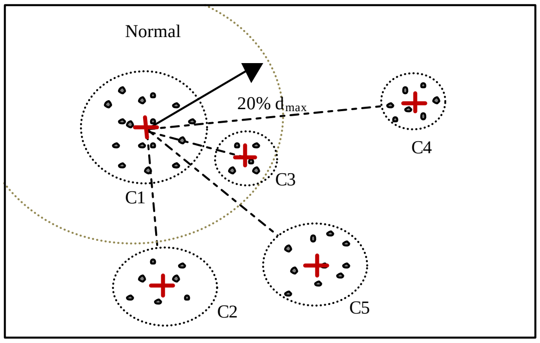

Fig. 4: Area of selection of normal clusters

For example, the Euclidean distance between the normal cluster C1 and C2 is 2.97. Table 1 shows the distance from other cluster centroids:

- **Step 3:** If the distances are less than 20% $d_{max,}$ then gather these clusters into the normal group

For instance, the distance between C1 and C4 is the maximum distance. The 20% from the maximum distance

Table 1: Distance of C1 from the other cluster centroids

Cluster	Distance
C2	2.97
C3	0.82
C4	4.42
C5	3.74

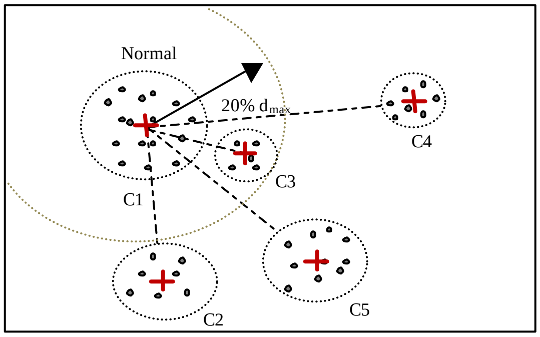

Fig. 5: Area of selection of normal clusters

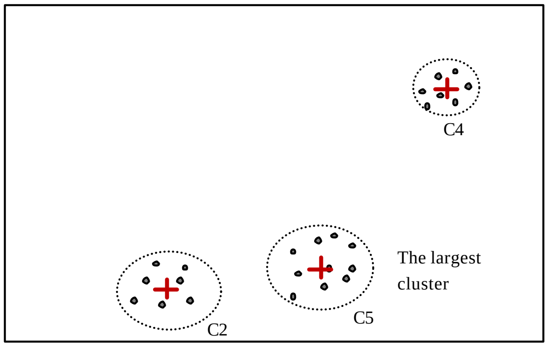

Fig. 6: Remaining clusters

will be 0.88. Thus, only cluster C3 has a distance of less than 20%. Accordingly, the cluster C3 will be normal as shown in Fig. 5.

Several experiments on different percentages of the distance between the normal cluster and other clusters were conducted. The result found that the 20% of the distance yielded better results:

- **Step 4:** Find the largest cluster in the remaining clusters. For instance, from Fig. 3, the number of clusters after Step 2 and 3 becomes as shown in Fig. 6
- **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 the other cluster as attacks

where, i is the number of attack categories and N is the number of attack instances in the data set. The chosen size of the largest cluster is based on the several analysis result on several experiment conducted in (Carpenter *et al*., 1991).

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.

RESULTS AND DISCUSSION

The experiments use standard data mining techniques in IDS, which consist of data collection, pre-processing, mining, labelling and evaluation. The performance of the SNC labelling algorithm is evaluatedby comparing it with the normal membership factor (NMF) labelling algorithm using Fuzzy Art, Fuzzy ART based on K-means and Euclidean-ART. The algorithms are implemented in MATLAB.The experimental steps follow the standard data mining process using clustering for anomaly detection proposed by Ngamwitthayanon *et al*. (2009) which consists of data collection, data pre-processing, mining using three mining techniques, labelling using SNC and NMF and evaluation.

The experiments are conducted in two main phases: applying first the clustering algorithms and then the labelling algorithms. The data sets are pre-processed following standard data pre-processing procedures (Abdul Samad *et al*., 2008). Next, the data are clustered with the proposed Fuzzy ART based on the K-means algorithm. Two benchmark clustering algorithms are used for comparison: (i) The original Fuzzy ART (Carpenter *et al*., 1991) and (ii) The Euclidean-ART (Kenaya and Cheok, 2008). In the second phase, the clusters are labelled using the proposed SNC algorithm. For the purpose of performance evaluation, SNC is compared with the NMF labelling algorithm (Abdul Samad *et al*., 2008). The detection rate, false-alarm rate and execution time, based on accuracy and ROC-graph, are measured for both algorithms.

Data collection: KDD Cup 99 is the mostly widely used data set for intrusion detection. Recently, researchers have conducted a statistical analysis on these data and found some important issues that significantly affect the performance of the systems evaluated. One major deficiency in the KDD Cup 99 data set is the large number of redundant records. Nearly 78% of the training records and 75% of the testing records are duplicated. To resolve these issues, Tavallaee *et al*. (2009) proposed a new data set called NSL-KDD, which consists of selected records

Table 2: No. and type of attacks in the experimental data set

Variable	Attack type	Original record	Data set1 2000	Data set 2 4000	Data set 3 6000	Data set 4 8000	Data set 5 10000	Data set 6 12000	Data set 7 14000	Data set 8 16000	Data set 9 18000	Data set 10 20000
Normal	--	67343	1970	3940	5911	7881	9852	11822	13792	15763	17733	19703
DoS	Back	956	0	0	1	3	2	15	5	21	6	6
	Land	18	0	0	0	0	0	0	0	10	10	5
	Neptune	41214	20	46	61	83	103	14	30	14	45	53
	Smurf	2646	0	2	2	4	6	25	15	11	10	10
	Pod	201	0	0	1	0	0	1	3	4	1	11
	Teardrop	892	2	2	1	2	4	14	17	15	9	9
Probe	Ipsweep	3599	2	2	9	11	6	9	19	15	27	30
	Nmap	1493	1	1	2	0	3	3	11	12	4	10
	Portsweep	2931	2	3	7	10	13	8	15	18	28	16
	Satan	3633	1	2	3	3	7	4	25	13	12	26
U2R	Buffer_overflow	30	0	0	0	0	0	28	6	25	30	30
	Loadmodule	9	0	0	0	0	0	8	1	4	9	9
	Rootkit	10	0	0	0	0	0	9	4	4	10	10
R2L	Ftp-write	8	0	0	0	0	0	0	1	4	4	1
	Guess-password	53	0	0	0	0	1	9	10	12	20	16
	Imap	11	0	0	0	0	0	0	5	4	6	5
	Multihop	7	0	0	0	0	0	0	2	5	5	4
	Phf	4	0	0	0	0	0	0	2	2	3	2
	Spy	2	0	0	0	0	0	1	1	2	1	1
	Warezclient	890	2	2	2	3	3	10	34	34	18	32
	Warezmaster	20	0	0	0	0	0	20	2	8	9	11
Total		125970	2000	4000	6000	8000	10000	12000	14000	16000	18000	20000

Table 3: Example of results collected in each experiment

Vigilance value	Detection rate (DR %)	False alarm rate (FAR %)	No. of clusters	Time (sec)
0.50	26.6667	0	4	0.482778
0.55	80	1.9797	5	0.476198
0.60	93.3333	3.0457	9	0.510910
0.65	93.3333	2.4873	16	0.630552
0.70	93.3333	14.5685	22	0.629027
0.75	100	16.4975	36	0.792053
0.80	100	20.9645	47	0.903794
0.85	100	29.7462	67	1.114748

of the complete KDD data set. Details about the NSL-KDD data set can be obtained from Tavallaee *et al.* (2009).

Ten data sets are randomly generated from the original data set and stored as Data Set1, Data Set2 ... Data Set10 (Table 2). It provides the descriptions of the data sets used in this study. The first row describes the original data set as Normal, DoS, U2R or R2L. The second row describes the type of attack and the third row is the number of attacks for the entiredata set. The remainder of the rows describe the number of attacks or normal instances for the different sizes of data sets. The total for DataSet1 is 2,000 records, Data Set2 is 4,000 records and the number of records is increased by 2,000 until Data Set10, which consists of 20,000 records. The objective to see performance of the algorithm when the data increasing. This study addresses the 20% training data set from each data set. The number of attacks was randomly selected (Abdul Samad *et al.*, 2008) and reduced to approximately 1.48% of the complete data set to match the assumption regardingthe distribution of normal and attack instances in the data set (Portnoy *et al.*, 2001).

Mining: Three clustering techniques, Fuzzy Art, Fuzzy ART based on K-means and Euclidean-ART, are implemented and tested on the data set. The resulting clusters are labelled using the SNC and the NFM labelling algorithms. For each label, eight experiments are conducted, based on a vigilance parameter from 0.5 to 0.9 for Fuzzy ART and Fuzzy ART based on K-means algorithms and a vigilance parameter ranging from 0.5 to 2.25 for the Euclidean-ART algorithm. A total of 480 experiments are conducted following the experimental steps previously described. Table 3 shows a sample of data collected for the experiment using clustering Fuzzy ART for clustering and SNC for labelling on Data Set1, which contains 2,000 records. The experiment collected the detection date, false-alarm rate, numberof clusters generated and time to complete the experiment.

The performance of the labelling algorithm is measured based on percentage of Detection Rate (DR) and False-alarm Rate (FAR). DR is defined as the percentage of attacks correctly identified by the system and FAR is the percentage of normal instances wrongly identified as attacks by the system. In each experiment, the best model is selected. Table 3 shows the best selected at the vigilance value 0.65, where the DAR is 93.3% and FAR is 2.5%. The best result is found in the number of cluster 16, where the time represents on how long the experiment get best result is about 0.63 sec. The best result selected using ROC graph as discusses in next section.

Evaluation: This study evaluates the performance of Fuzzy ART based on the K-means clustering algorithm and a labelling algorithm based on the DR and FAR measure using the ROC curve (Abdul Samad *et al*., 2008) and t-test on DR, FAR and execution time. The ROC curve is a method to visualise the trade-offs between detection rate and the false-alarm rate. ROC is a method to select the best result among different vigilance parameters.

Figure 7 shows an example of an ROC graph for the first experiment, in Table 3. 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. Based on the location of that point and the curve in Figure 7, the best result for the experiment is at a vigilance of 0.65, where the DR is 93.3% and the FAR is 2.5%, highlighted in grey in Table 2.

The t-test is a statistical method used to measure whether there is a significant difference in the performance between two algorithms. The t-test is performed on the results of DR, FAR and execution time for SNC versus the NMF labelling algorithm for the three clustering algorithms for all data sets. It is also performed between the clustering algorithms using the same labelling algorithm. Finally, the new algorithm is evaluated based on execution time.

Analysis result: Table 4 shows an example of summary experiment results for the case using Fuzzy ART and SNC labelling for all data sets and 10 values of the vigilance parameter and Table 5 shows an example of summary experiment results for the case using Euclidean-ART and SNC labelling algorithmsand8 values of the vigilance parameter.The results show that the false-alarm rate increases with the increase of the vigilance parameter because there are more normal instances thatare classified as attack.

Table 6 shows a summary of the results of the experiments. The best result (highlighted in grey) from each experiment is selected using the ROC curve method for all data sets to compare the SNC and NMF labelling algorithms using Fuzzy Art, Fuzzy ART based on K-means and Euclidean-ART.

The row average shows the average of FAR and DR for all data sets and the last row shows the best FAR and DR using ROC curve. The value with add superscript ^{w} is marked as win, while superscript ^{L} marked as loose.

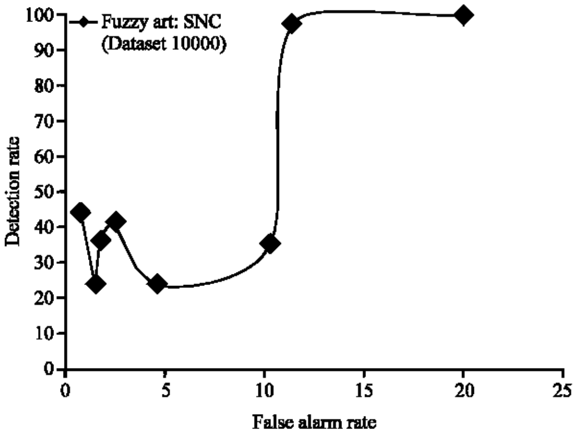

Fig. 7: ROC graph for the experiment using fuzzy art and SNC labelling for data set1

The average row shows that the Fuzzy ART based on K-Means is able to reduce the FAR 1.28% compare with Fuzzy ART and reduce 8.59% compare with Euclidean ART, when applying using SNC labelling, whereas it loose with 0.23% compare with Fuzzy ART and 4.54% compare with Euclidean ART in term of DR.

In concludes that improvement of Fuzzy ART with K-Mean using SNC labelling is able to reduce the FAR but also reduce the DR.

The average row also shows that the Fuzzy ART based on K-Means is yielded better DR compare with Fuzzy ART when using NMF labelling, but less 3.04% compare with Euclidean ART. Fuzzy ART based on K-Means is loose in term of FAR, neither with Fuzzy ART (10.56%) or Euclidean ART (2.54%).

Table 6 shows that the average row shows that the SNC labellingalgorithm yieldedbetter intrusion detection results based on theDR and FAR using all the clustering algorithms: Fuzzy Art, Fuzzy ART based on K-Means and Euclidean ART compare to the NMF labelling algorithm accept the DR applying NMF using Fuzzy Art based on K-Means is higher 1.58% compare applying SNC labelling. ApplyingSNC labelling improvesthe FAR by 12.1% and the DR by 0.5%, the FAR by 11.87% and the DR by 1.9% and the FAR by 23.9% compared with NMF, using Fuzzy Art, Euclidean-ART and Fuzzy ART based on K-means, respectively.

Similarly in the best ROC row shows that SNC labelling algorithm yielded better DR and FAR using all the clustering algorithms accept the DR applying NMF using Fuzzy Art is higher 0.08% compare applying SNC labelling. The SNC also has improved the FARby 6.51%, the FAR by 14.9% and the DR by 6.14% andthe FAR by 6.45% and the DR by 3%, when compared to the NMF applied with Fuzzy ART and the Euclidean-ART and Fuzzy Art based on K-means, respectively.

Table 4: Performance fuzzy art using SNC for ten data sets

	Data sets																			
	1		2		3		4		5		6		7		8		9		10	
Vigilance parameter	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR
0.50	26.67	0.000	30.23	0.23	17.98	2.06	10.08	0.27	44.60	0.71	41.57	1.17	2.88	0.66	5.49	1.21	16.85	1.05	5.41	1.08
0.55	80.00	1.980	16.28	0.66	1.12	1.47	21.01	2.09	24.32	1.46	14.61	3.45	10.58	2.05	38.40	2.35	31.84	2.13	20.95	1.19
0.60	93.33	3.005	97.67	2.03	22.47	3.38	14.29	3.43	36.49	1.79	37.64	2.93	27.40	2.28	36.71	2.73	48.69	3.01	11.82	1.92
0.65	93.33	2.490	74.42	4.34	13.48	4.30	29.41	6.83	41.89	2.50	80.34	5.20	37.50	4.40	64.98	4.74	34.08	4.88	54.05	3.38
0.70	93.33	14.570	97.67	4.92	96.63	8.09	26.89	8.25	24.32	4.55	60.67	8.11	52.40	7.75	82.70	6.65	86.52	6.21	86.82	5.15
0.75	100.00	16.500	100.00	14.54	38.20	12.23	40.34	11.27	35.81	10.27	89.89	10.99	83.17	9.19	90.30	9.15	91.01	6.59	85.47	6.82
0.80	100.00	20.970	100.00	18.25	38.20	16.17	100.00	16.38	97.30	11.39	94.94	12.88	92.79	12.63	90.30	11.61	97.75	11.28	90.20	10.61
0.85	100.00	29.750	100.00	24.24	40.45	25.68	100.00	21.23	100.00	19.99	96.07	17.13	99.52	18.60	97.47	21.82	95.88	16.60	93.92	16.85

Table 5:Performance of euclidean-ART using SNC for 10 data sets

	Data sets																			
	1		2		3		4		5		6		7		8		9		10	
Vigilance parameter	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR
0.50	100.00	40.97	100.00	41.57	100.00	37.30	100.00	36.96	100.00	34.02	100.00	29.01	100.00	29.58	100.00	27.24	98.88	27.62	95.27	24.64
0.75	100.00	37.82	100.00	34.07	100.00	29.42	100.00	30.81	100.00	24.71	98.88	22.47	99.04	20.08	94.09	18.88	95.88	17.88	94.26	17.47
1.00	100.00	38.38	100.00	27.92	100.00	26.80	100.00	22.22	57.43	19.85	98.88	17.86	86.06	14.06	93.67	14.21	96.26	13.62	93.58	13.25
1.25	100.00	35.94	100.00	23.86	42.70	19.10	43.70	15.65	39.87	14.75	97.19	11.87	81.73	11.10	92.41	11.36	91.39	10.03	90.88	9.27
1.50	100.00	33.10	100.00	19.82	40.45	12.99	36.14	8.51	42.57	8.46	91.01	9.42	76.92	7.45	73.42	6.53	82.02	7.40	82.10	6.65
1.75	93.33	19.85	93.02	11.07	37.08	8.53	26.05	5.98	39.87	14.75	78.65	5.52	74.04	4.89	69.20	3.57	78.65	4.20	67.23	3.42
2.00	93.33	10.56	93.02	6.65	23.60	4.48	28.57	4.12	25.00	3.53	62.36	3.41	65.39	2.41	59.49	2.59	66.67	1.75	69.26	1.95
2.25	76.67	7.06	90.70	2.77	21.35	2.32	26.05	2.96	22.97	2.54	56.74	1.83	53.85	1.70	64.14	1.85	59.55	0.90	56.76	0.81

Table 6: DR and FAR for fuzzy art, fuzzy art based on K means and Euclidean art between SNC and NMF for 10 data set

	Fuzzy art				Fuzzy art K-means				Euclidean-art			
	SNC		NMR		SNC		NMR		SNC		NMR	
Data set	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR
1	2.49	93.33	1.98	80.00	8.63	80.00	4.42	76.67	10.56	93.33	13.10	90.00
2	2.03	97.67	23.00	97.67	2.87	86.05	20.41	93.02	6.65	93.02	21.78	95.35
3	8.09	96.63	8.54	97.75	1.20	82.02	29.94	96.63	26.80	100.00	28.93	96.63
4	16.40	100.00	13.30	92.44	5.51	99.16	19.81	94.12	22.22	100.00	22.60	96.64
5	11.40	97.30	19.40	97.97	6.76	96.62	31.85	97.30	24.71	100.00	14.37	90.54
6	5.2	80.34	23.80	73.60	9.15	94.38	42.43	96.07	11.87	97.19	27.25	94.38
7	9.19	83.17	25.30	83.17	7.70	88.94	38.68	90.87	20.08	99.04	25.41	92.79
8	6.66	82.70	27.30	81.01	7.65	89.03	18.72	74.68	11.36	92.41	26.15	90.72
9	6.59	91.01	24.80	98.50	7.10	93.63	50.48	98.13	13.62	96.26	48.90	98.13
10	5.15	86.82	27.20	99.32	5.06	86.82	43.83	94.93	9.27	90.88	47.32	97.64
Average	7.32^w	90.90^w	19.50	90.14	6.16^w	89.67^L	30.06	91.24	15.71^w	96.21^w	27.58	94.28
Best ROC	2.03^w	97.70^L	8.54	97.75	5.51^w	99.16^w	20.41	93.02	6.65^w	93.02^w	13.10	90.00

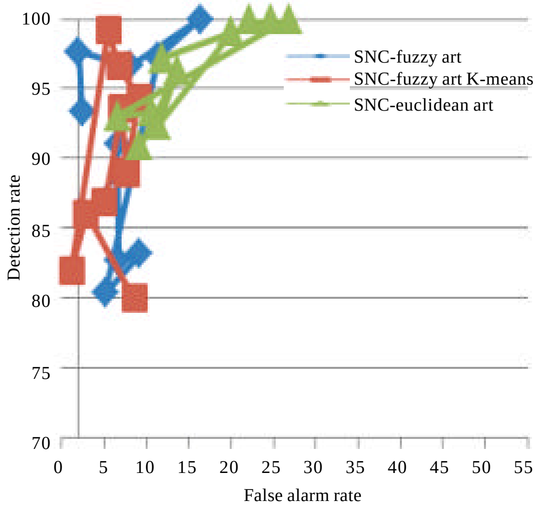

Fig. 8: Performance of SNC using three clustering technique

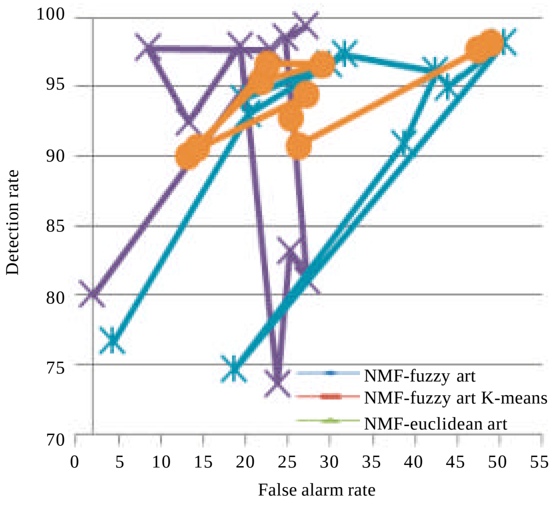

Fig. 9: Performance of SNC using three clustering technique

The performance of algorithms is evaluated using ROC curve. Figure 8 and 9 shows the pattern of ROC curve for SNC labelling and NMF labelling versus the three classification algorithms, respectively. The pattern ROC curve in Fig. 10 shows that most of the points are more towards to the coordinate (0,1) compare with ROC curve in Fig. 11 is more scattered and far from coordinate (0,1). This means that SNC labelling has perform better for IDS compare to NMF labelling. The graph also shows that the combination of SNC labelling with Fuzzy ART based on K-means results inbetter performance for quality IDS as most of the ROC curve are more close to coordinate (0,1) compare to SNC-Fuzzy ART and SNC- Euclidean ART.

Performance of SNC labelling and Fuzzy ART K-Means also evaluated using ROC point based on the average of DR and FAR as stated in Table 6. Figure 10 shows that applying SNC- Fuzzy ART presented the best result followed by SNC-Fuzzy ART based on K-Means, NMF-Euclidean ART, SNC-Euclidean ART, NMF-Fuzzy ART and NMF-Fuzzy ART K-Means.

However, Fig. 11 shows the performance of SNC labelling and Fuzzy ART K-Means also evaluated using ROC point based on the best DR and FAR using ROC curve as stated in Table 6. Figure 13 shows that applying SNC-Fuzzy ART with Fuzzy ART based on K-Means presented the best result followed by SNC-Fuzzy ART, NMF-Euclidean ART, SNC-Euclidean ART, NMF-Fuzzy ART and NMF-Fuzzy ART K-Means.

To investigate the robustness of the SNC and Fuzzy ART based on K-Means, a statistical method t-test is applied. For each algorithm, the t-value and p-value are reported as aparameter setting, the alpha level, or the significance level, is set to 0.05 ($\alpha = 0.05$).

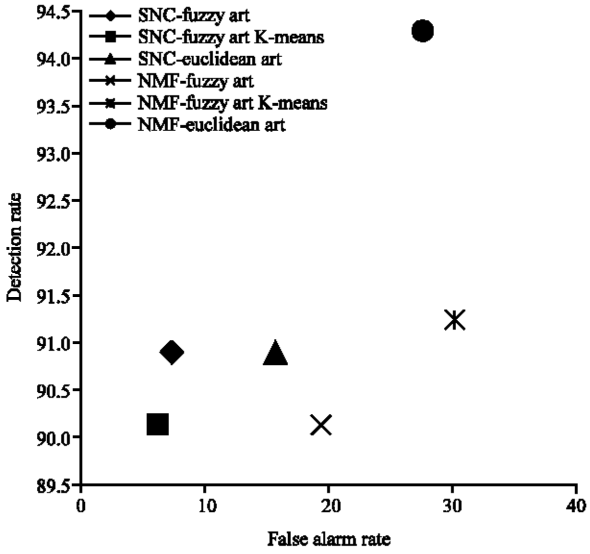

Fig. 10: ROC point labelling algorithm and clustering techniques average DR and FAR

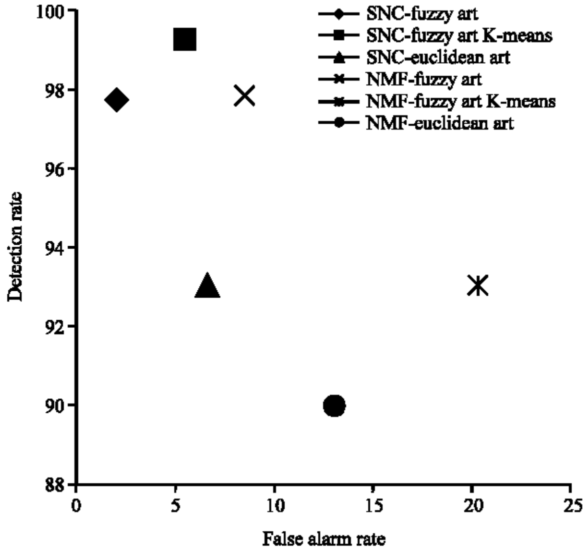

Fig. 11: ROC point labelling algorithm and clustering techniques based on the best result using the ROC curve

The performance of the SNC labelling is also evaluated using the t-test between the NMF labelling algorithm for the detection rate and the false-alarm rate. Table 7 shows the significant-value comparison of DR and FAR between SNC and NMF for each clustering algorithm.

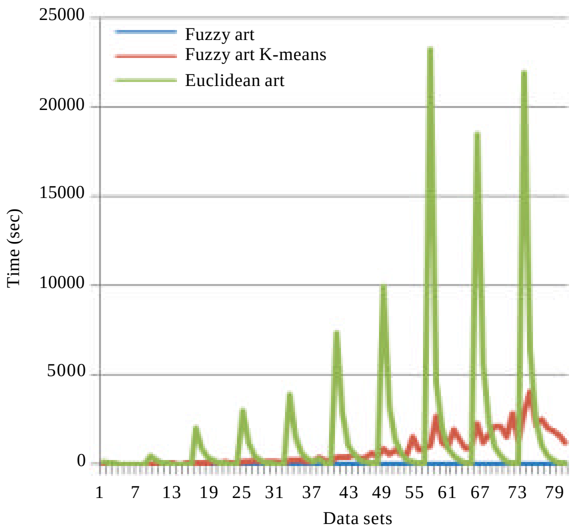

Fig. 12: Execution time for the clustering algorithms using NMF

Table 7: t-test result of DR and FAR between SNC and NMF for each clustering algorithm

	Fuzzy art-K-means		Fuzzy art		Euclidean-art	
Variable	t-value	p-value	t-value	p-value	t-value	p-value
DR	11.608	0.000	9.487	0.000	7.797	0.000
FAR	25.532	0.000	20.07	0.000	15.848	0.000

Table 8: t-test result of FAR and DR among clustering techniques using SNC

	Fuzzy art–fuzzy art K-means		Euclidean-art–fuzzy art K-means		Euclidean-art–fuzzy art k-means	
Variable	t-value	p-value	t-value	p-value	t-value	p-value
FAR	-9.366	0.000	3.834	0.000	5.578	0.000
DR	3.496	0.001	2.926	0.004	4.742	0.000

The t-test is also conducted to determine the performance of each clustering algorithm when applied witheach labelling algorithm. Table 8 shows the t-test results of FAR and DR between the clustering methods using SNC. The t-test result sshow all significant values are less than the alpha value.

Table 9 shows the t-test results of FAR and DR between the clustering methods using NMF. The t-test results show that only the following combinations show significant values less than the alpha value: FAR of Euclidean-ART versus Fuzzy ART based on K-means and Euclidean-ART versus Fuzzy ART and DR of Euclidean-ART versus based on K-means. The result indicates that there is not a significant difference in FAR (p = 0.54) and DR (p = 0.214 between Fuzzy ART and Fuzzy ART based on K-means) for the NMF trials.

The labelling algorithm is also evaluated based on execution time. Figure 12 shows the result of the execution time analysis for the three clustering algorithmsin seconds

Table 9: t-test result of FAR and DR among clustering techniques using NMF

	Fuzzy art–fuzzy art K-means		Euclidean-art–fuzzy art K-means		Euclidean-art–fuzzy art k-means	
Variable	t-value	p-value	t-value	p-value	t-value	p-value
FAR	1.953	0.054	6.033	0.000	-9.366	0.000
DR	-1.252	0.214	2.305	0.024	3.288	0.002

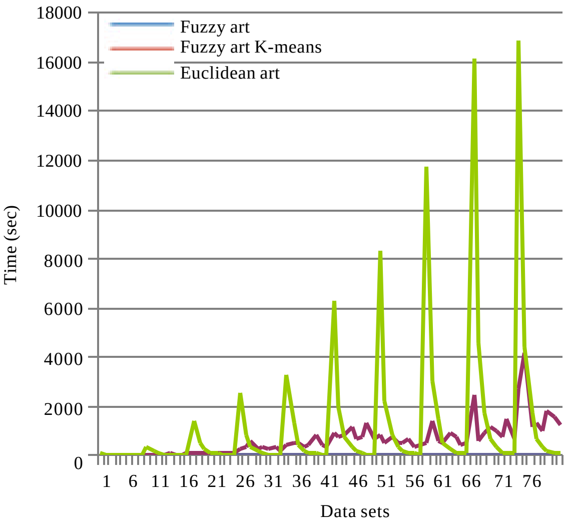

Fig. 13: Execution time for the clustering algorithm using SNC Labelling

per experiment for 10 data sets using NMF. The graph shows that Fuzzy ART performs the fastest, Euclidean-ART performs the slowest and Fuzzy ART based on K-means is faster than Euclidean-ART.

Figure 13 shows agraph of the execution time for each clustering algorithm, in seconds per experiment, for 10 data sets, using the SNC algorithm. The graph shows that the performance of the clustering algorithms exhibits a similar pattern whether using NMF or SNC for labelling. However, a comparison of the two graphs reveals that the three clustering algorithms perform faster using the SNC labelling algorithm (maximum time 18,000 sec) than when using the NMF labelling algorithm (maximum time 25,000 sec).

CONCLUSION

This study proposes a new labelling algorithm known as SNC and a Fuzzy ART based on the K-means clustering algorithm for IDS. The SNC labels clusters as normal or abnormal after performing an initial clustering process, assuming that small clusters are abnormal. The SNC determines the similarity with other clusters by calculating the similarity percentage between the normal cluster centroid (the normal cluster defined as the largest one) and other cluster centroids using Euclidean distance. The experimental results show that the SNC labelling algorithm increases the detection rate and reduces the false-alarm rate compared with the NMF labelling algorithm, using the three mining techniques Fuzzy ART, Fuzzy ART based on K-means and Euclidean-ART. The results also show that the three mining techniques perform faster using SNC than using NMF. Applying the combination ofFuzzy ARTbased on K-means with the SNC labelling algorithmyields the best result; however, in terms of execution time, Fuzzy ART is slightly faster. The percentage of DR and FAR is relatively high compared with past research applied to the KDDCUP 99 data set. This is because of the irrelevant and redundant features in the network packet in the NST-KDD data set, which cause the performance of the anomaly intrusion detectors to be inefficient. This study presents an alternative method for labelling and also presents the best clustering algorithm for IDS.

REFERENCES

Abdul Samad, H.I., A. Abdul Hanan, A.B. Kamalrulnizam, N. MdAsri, D. Dahliyusmanto and W. Chimphlee, 2008. A novel method for unsupervised anomaly detection using unlabelled data. Proceedings of the International Conference on Computational Sciences and Its Applications, June 30-July 3, 2008, Perugia, Italy, pp: 252-260.

Afaf Muftah, S.A., 2011. Labeling algoritm to unsuperivised anomaly intrusion detection. Master's Theses, University of Kebangsaan, Malaysia.

Borgatti, S., 2012. Distance and correlations. Boston College, September 11, 2012.

Carpenter, G.A., S. Grossberg and D.B. Rosen, 1991. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks, 4: 759-771.

Fang, L. and L. Le-Ping, 2005. Unsupervised anomaly detection based n an evolutionary artificial immune network. Proceedings of the Evo Workshop on Applications of Evolutionary Computing, March 30-April 1, 2005, Lausanne, Switzerland, pp: 166-174.

Folorunso, O., O.O. Akande, A.O. Ogunde and O.R. Vincent, 2010. ID-SOMGA: A self organising migrating genetic algorithm-based solution for intrusion detection. Comput. Inform. Sci., 3: 80-92.

Guan, Y., A.A. Ghorbani and N. Belacel, 2003. Y-MEANS: A clustering method for intrusion detection. Proc. Can. Conf. Electr. Comput. Eng., 2: 1083-1086.

Guo, X.C., D.M. Ma, Y.J. Sun and H.Y. Ma, 2012. Research of Network Intrusion Detection System Based on Data Mining Approaches. In: Green Communications and Networks, Yang, Y.H. and M. Ma (Eds.). Springer, Netherlands, pp: 1101-1107.

He, J., A.H. Tan and C.L. Tan, 2002. ART-C: A neural architecture for self-organization under constraints. Proceedings of the International Joint Conference on Neural Networks, Volume 3, May 12-17, 2002, Honolulu, Hawaii, pp: 2550-2555.

Kenaya, R. and K.C. Cheok, 2008. Euclidean ART neural networks. Proceedings of the World Congress on Engineering and Computer Science, October 22-24, 2008, San Francisco, USA.

Liu, Y., K. Chen, X. Liao and W. Zhang, 2004. A genetic clustering method for intrusion detection. Pattern Recogn., 37: 927-942.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Statist. Prob., 1: 281-297.

Ngamwitthayanon, N., N. Wattanapongsakorn and D.W. Coit, 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, May 26-29, 2009, Wuhan, China, pp: 208-217.

Petrovic, S., G. Alvarez, A. Orfila and J. Carbo, 2006. Labelling clusters in an intrusion detection system using a combination of clustering evaluation techniques. Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Volume 6, January 4-7, 2006, Hawaii, USA., pp: 129b-129b.

Portnoy, L., E. Eskin and S. Stolfo, 2001. Intrusion detection with unlabeled data using clustering. Proceedings of ACM CSS Workshop on Data Mining Applied to Security, November 5-8, 2001, Philadelphia, PA., pp: 5-8.

Provost, F. and T. Fawcett, 2001. Robust classification for imprecise environments. Machine Learn., 42: 203-231.

Sharma, S.K., P. Pandey, S.K. Tiwari and M.S. Sisodia, 2012. An improved network intrusion detection technique based on k-means clustering via Naive bayes classification. Proceedings of the International Conference on Advances in Engineering, Science and Management, March 30-31, 2012, Tamil Nadu, India, pp: 417-422.

Shirazi, H.M., 2009. Anomaly intrusion detection system using information theory, K-NN and KMC algorithms. Aust. J. Basic Applied Sci., 3: 2581-2597.

Smith, R., A. Bivens, M. Embrechts, C. Palagiri and B. Szymanski, 2002. Clustering approaches for anomaly based intrusion detection. Proceedings of the Walter Lincoln Hawkins Graduate Research Conference, October 2002, New York, USA., pp: 579-584.

Stein, G., B. Chen, A.S. Wu and K.A. Hua, 2005. Decision tree classifier for network intrusion detection with GA-based feature selection. Proceedings of the 43rd Annual Southeast Regional Conference, Volume 2, March 18-20, 2005, Kennesaw, GA., USA., pp: 136-141.

Tavallaee, M., E. Bagheri, W. Lu and A.A. Ghorbani, 2009. A detailed analysis of the KDD CUP 99 data set. Proceedings of the 2nd IEEE Symposium on Computational Intelligence for Security and Defence Applications, July 8-10 2009, Ottawa, Canada.

Thamaraiselvi, S., R. Srivathsan, J. Imayavendhan, R. Muthuregunathan and S. Siddharth, 2009. Combining naive-bayesian classifier and genetic clustering for effective anomaly based intrusion detection. Proceedings of the 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, December 15-18, 2009, Delhi, India, pp: 455-462.

Xie, L., Y. Wang, L. Chen and G. Yue, 2010. An anomaly detection method based on fuzzy C-means clustering algorithm. Proceedings of the 2nd International Symposium on Networking and Network Security, April 2-4, 2010, Jinggangshan, China, pp: 89-92.

Zhang, G., 2005. Applying Mining Fuzzy Association Rules to Intrusion Detection Based on Sequences of System Calls. In: Networking and Mobile Computing, Lu, X. and W. Zhao (Eds.). Springer, Berlin, Germany, pp: 826-835.

Zhang, Z., K. Huang and T. Tan, 2006. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. Proceedings of the 18th International Conference on Pattern Recognition, Volume 3, August 20-24, 2006, Hong Kong, pp: 1135-1138.

Zhao, Z., S. Guo, Q. Xu and T. Ban, 2009. G-means: A Clustering Algorithm for Intrusion Detection. In: Advances in Neuro-Information Processing, Koppen, M., N. Kasabov and G. Coghill (Eds.). Springer, Berlin, Germany, pp: 563-570.

Zhong, S., T.M. Khoshgoftaar and N. Seliya, 2007. Clustering-based network intrusion detection. Int. J. Reliability Qual. Safety Eng., 14: 169-187.