A Similarity Normal Clustering Labelling Algorithm for Clustering Network Intrusion Detection

Zulaiha Ali Othman, Azuraliza Abu Bakar, Afaif Mustah 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 assumption is 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 clusters that 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 that the 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 Ngamwirthayanon 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 ART approach (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 Ngamwirthayanon 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 (Ngamwithayanon 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 typos in KDD cup 99 data sets fall 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 rules (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 aims 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 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 from the initial cluster and reassigning each instance 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 iteration 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

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

Fig. 3: Size of cluster notation

Fig. 4: Area of selection of normal clusters

has the largest size and other cluster centroids, using the Euclidean distance (Afaf Mufiah, 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 (Thamaraivelvi 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

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% \( d_{\text{max}} \), 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

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>2.97</td>
</tr>
<tr>
<td>C3</td>
<td>0.82</td>
</tr>
<tr>
<td>C4</td>
<td>4.42</td>
</tr>
<tr>
<td>C5</td>
<td>3.74</td>
</tr>
</tbody>
</table>

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}{N} \sum_{i=1}^{N} \text{ or } \end{equation}

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 evaluated by 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 Ngamwittayanon 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.
of the complete KDD data set. Details about the NSL-KDD data set can be obtained from Tavallae et al. (2009).

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Attack type</th>
<th>Original record</th>
<th>Data set 1</th>
<th>Data set 2</th>
<th>Data set 3</th>
<th>Data set 4</th>
<th>Data set 5</th>
<th>Data set 6</th>
<th>Data set 7</th>
<th>Data set 8</th>
<th>Data set 9</th>
<th>Data set 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>--</td>
<td>67343</td>
<td>1970</td>
<td>3940</td>
<td>5911</td>
<td>7861</td>
<td>9852</td>
<td>11822</td>
<td>15792</td>
<td>17563</td>
<td>17733</td>
<td>19703</td>
</tr>
<tr>
<td>DoS</td>
<td>Back</td>
<td>956</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>15</td>
<td>5</td>
<td>21</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Land</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Neptune</td>
<td>41214</td>
<td>20</td>
<td>46</td>
<td>61</td>
<td>83</td>
<td>103</td>
<td>14</td>
<td>30</td>
<td>14</td>
<td>45</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Smurf</td>
<td>2646</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>25</td>
<td>15</td>
<td>11</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Pod</td>
<td>201</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Teardrop</td>
<td>892</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>14</td>
<td>17</td>
<td>15</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Probe</td>
<td>Ipsweep</td>
<td>3599</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>11</td>
<td>6</td>
<td>9</td>
<td>19</td>
<td>15</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Nmap</td>
<td>1493</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>11</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Portswamp</td>
<td>2931</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>8</td>
<td>15</td>
<td>18</td>
<td>28</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Satan</td>
<td>3633</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>25</td>
<td>13</td>
<td>12</td>
<td>26</td>
</tr>
<tr>
<td>U2R</td>
<td>Buffer_overflow</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>6</td>
<td>25</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Loadmodule</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Rootkit</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>R2L</td>
<td>Ftp-write</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Guess-password</td>
<td>53</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Inap</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>MultiHop</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Fih</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Spy</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Warclient</td>
<td>890</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>34</td>
<td>34</td>
<td>18</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Warcruiser</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>125970</td>
<td>2000</td>
<td>4000</td>
<td>6000</td>
<td>8000</td>
<td>10000</td>
<td>12000</td>
<td>14000</td>
<td>16000</td>
<td>18000</td>
<td>20000</td>
</tr>
</tbody>
</table>

Table 3: Example of results collected in each experiment

<table>
<thead>
<tr>
<th>Vigilance value</th>
<th>Detection rate (DR %)</th>
<th>False alarm rate (FAR %)</th>
<th>No. of clusters</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>26.6667</td>
<td>0</td>
<td>4</td>
<td>0.482778</td>
</tr>
<tr>
<td>0.55</td>
<td>80</td>
<td>1.9797</td>
<td>5</td>
<td>0.476198</td>
</tr>
<tr>
<td>0.60</td>
<td>93.3333</td>
<td>3.0457</td>
<td>9</td>
<td>0.510910</td>
</tr>
<tr>
<td>0.65</td>
<td>3.0457</td>
<td>2.4873</td>
<td>16</td>
<td>0.630552</td>
</tr>
<tr>
<td>0.70</td>
<td>14.685</td>
<td>4.5808</td>
<td>22</td>
<td>0.629027</td>
</tr>
<tr>
<td>0.75</td>
<td>16.4975</td>
<td>4.6798</td>
<td>36</td>
<td>0.792853</td>
</tr>
<tr>
<td>0.80</td>
<td>20.9645</td>
<td>2.4976</td>
<td>47</td>
<td>0.903794</td>
</tr>
<tr>
<td>0.85</td>
<td>29.7462</td>
<td>2.4976</td>
<td>67</td>
<td>1.114748</td>
</tr>
</tbody>
</table>

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, number of 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 algorithms and 8 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 that are 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 "*" is marked as win, while superscript "+" marked as loose.

![ROC graph for the experiment using fuzzy art and SNC labelling for data set](image)

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 labelling algorithm yielded better intrusion detection results based on the DR 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. Applying SNC labelling improves the FAR by 12.1% and the DR by 0.5%, the FAR by 11.87% and the DR by 1.5% 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 FAR by 6.51%, the FAR by 14.9% and the DR by 6.14% and the FAR by 4.5% 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>
<thead>
<tr>
<th>Vigilance parameter</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DR</td>
<td>FAR</td>
<td>DR</td>
<td>FAR</td>
<td>DR</td>
<td>FAR</td>
<td>DR</td>
<td>FAR</td>
<td>DR</td>
<td>FAR</td>
</tr>
<tr>
<td>0.50</td>
<td>26.67</td>
<td>0.000</td>
<td>30.23</td>
<td>0.23</td>
<td>17.98</td>
<td>2.06</td>
<td>10.08</td>
<td>0.27</td>
<td>44.60</td>
<td>0.71</td>
</tr>
<tr>
<td>0.55</td>
<td>30.00</td>
<td>1.980</td>
<td>16.28</td>
<td>0.66</td>
<td>1.12</td>
<td>1.17</td>
<td>21.01</td>
<td>2.09</td>
<td>24.32</td>
<td>1.46</td>
</tr>
<tr>
<td>0.60</td>
<td>93.33</td>
<td>3.005</td>
<td>97.67</td>
<td>2.03</td>
<td>22.47</td>
<td>3.38</td>
<td>14.29</td>
<td>3.43</td>
<td>36.49</td>
<td>1.79</td>
</tr>
<tr>
<td>0.65</td>
<td>93.33</td>
<td>2.490</td>
<td>74.42</td>
<td>4.34</td>
<td>13.48</td>
<td>4.30</td>
<td>29.41</td>
<td>6.83</td>
<td>41.89</td>
<td>2.50</td>
</tr>
<tr>
<td>0.70</td>
<td>93.33</td>
<td>14.570</td>
<td>97.67</td>
<td>4.92</td>
<td>96.63</td>
<td>8.09</td>
<td>26.89</td>
<td>8.25</td>
<td>24.32</td>
<td>4.55</td>
</tr>
<tr>
<td>0.75</td>
<td>100.00</td>
<td>16.500</td>
<td>100.00</td>
<td>14.54</td>
<td>38.20</td>
<td>12.23</td>
<td>40.34</td>
<td>11.27</td>
<td>35.81</td>
<td>10.27</td>
</tr>
<tr>
<td>0.80</td>
<td>100.00</td>
<td>20.970</td>
<td>100.00</td>
<td>18.25</td>
<td>38.20</td>
<td>16.17</td>
<td>100.00</td>
<td>16.38</td>
<td>97.30</td>
<td>11.39</td>
</tr>
<tr>
<td>0.85</td>
<td>100.00</td>
<td>29.750</td>
<td>100.00</td>
<td>24.24</td>
<td>40.45</td>
<td>25.68</td>
<td>100.00</td>
<td>21.23</td>
<td>100.00</td>
<td>19.99</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Vigilance parameter</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DR</td>
<td>FAR</td>
<td>DR</td>
<td>FAR</td>
<td>DR</td>
<td>FAR</td>
<td>DR</td>
<td>FAR</td>
<td>DR</td>
<td>FAR</td>
</tr>
<tr>
<td>0.50</td>
<td>100.00</td>
<td>40.97</td>
<td>100.00</td>
<td>41.57</td>
<td>100.00</td>
<td>37.30</td>
<td>100.00</td>
<td>36.96</td>
<td>100.00</td>
<td>34.02</td>
</tr>
<tr>
<td>0.55</td>
<td>100.00</td>
<td>38.72</td>
<td>100.00</td>
<td>34.07</td>
<td>100.00</td>
<td>29.42</td>
<td>100.00</td>
<td>30.81</td>
<td>100.00</td>
<td>24.71</td>
</tr>
<tr>
<td>0.60</td>
<td>100.00</td>
<td>35.94</td>
<td>100.00</td>
<td>23.86</td>
<td>42.70</td>
<td>19.10</td>
<td>43.70</td>
<td>15.65</td>
<td>39.87</td>
<td>14.75</td>
</tr>
<tr>
<td>0.65</td>
<td>100.00</td>
<td>33.10</td>
<td>100.00</td>
<td>19.82</td>
<td>40.45</td>
<td>12.99</td>
<td>36.14</td>
<td>8.51</td>
<td>42.57</td>
<td>8.46</td>
</tr>
<tr>
<td>0.70</td>
<td>93.33</td>
<td>19.85</td>
<td>93.02</td>
<td>11.07</td>
<td>37.08</td>
<td>8.53</td>
<td>26.95</td>
<td>5.98</td>
<td>39.87</td>
<td>14.75</td>
</tr>
<tr>
<td>0.75</td>
<td>93.33</td>
<td>19.85</td>
<td>93.02</td>
<td>6.65</td>
<td>23.60</td>
<td>4.48</td>
<td>28.57</td>
<td>4.12</td>
<td>25.00</td>
<td>3.55</td>
</tr>
<tr>
<td>0.80</td>
<td>93.33</td>
<td>19.85</td>
<td>93.02</td>
<td>6.65</td>
<td>23.60</td>
<td>4.48</td>
<td>28.57</td>
<td>4.12</td>
<td>25.00</td>
<td>3.55</td>
</tr>
<tr>
<td>0.85</td>
<td>93.33</td>
<td>19.85</td>
<td>93.02</td>
<td>6.65</td>
<td>23.60</td>
<td>4.48</td>
<td>28.57</td>
<td>4.12</td>
<td>25.00</td>
<td>3.55</td>
</tr>
<tr>
<td>0.90</td>
<td>93.33</td>
<td>19.85</td>
<td>93.02</td>
<td>6.65</td>
<td>23.60</td>
<td>4.48</td>
<td>28.57</td>
<td>4.12</td>
<td>25.00</td>
<td>3.55</td>
</tr>
<tr>
<td>0.95</td>
<td>93.33</td>
<td>19.85</td>
<td>93.02</td>
<td>6.65</td>
<td>23.60</td>
<td>4.48</td>
<td>28.57</td>
<td>4.12</td>
<td>25.00</td>
<td>3.55</td>
</tr>
<tr>
<td>1.00</td>
<td>93.33</td>
<td>19.85</td>
<td>93.02</td>
<td>6.65</td>
<td>23.60</td>
<td>4.48</td>
<td>28.57</td>
<td>4.12</td>
<td>25.00</td>
<td>3.55</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Data set</th>
<th>Fuzzy art</th>
<th>Fuzzy art K-means</th>
<th>Euclidean-art</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNC</td>
<td>NMR</td>
<td>SNC</td>
</tr>
<tr>
<td>1</td>
<td>2.49</td>
<td>93.33</td>
<td>1.98</td>
</tr>
<tr>
<td>2</td>
<td>2.03</td>
<td>97.67</td>
<td>23.00</td>
</tr>
<tr>
<td>3</td>
<td>8.09</td>
<td>96.63</td>
<td>8.54</td>
</tr>
<tr>
<td>4</td>
<td>16.40</td>
<td>100.00</td>
<td>13.30</td>
</tr>
<tr>
<td>5</td>
<td>11.40</td>
<td>97.30</td>
<td>19.40</td>
</tr>
<tr>
<td>6</td>
<td>5.2</td>
<td>80.34</td>
<td>23.80</td>
</tr>
<tr>
<td>7</td>
<td>9.19</td>
<td>83.17</td>
<td>25.30</td>
</tr>
<tr>
<td>8</td>
<td>6.66</td>
<td>82.70</td>
<td>27.30</td>
</tr>
<tr>
<td>9</td>
<td>6.59</td>
<td>91.01</td>
<td>24.80</td>
</tr>
<tr>
<td>10</td>
<td>5.15</td>
<td>86.82</td>
<td>27.20</td>
</tr>
<tr>
<td>Average</td>
<td>7.32£</td>
<td>90.90</td>
<td>19.50</td>
</tr>
<tr>
<td>Best ROC</td>
<td>2.03£</td>
<td>97.76</td>
<td>8.54</td>
</tr>
</tbody>
</table>

Fig. 8: 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 in better 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 a parameter setting, the alpha level, or the significance level, is set to 0.05 (α = 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.

The t-test is also conducted to determine the performance of each clustering algorithm when applied with each labelling algorithm. Table 8 shows the t-test results of FAR and DR between the clustering methods using SNC. The t-test result shows 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 algorithms in seconds.
Table 9: t-test result of FAR and DR among clustering techniques using NMF

<table>
<thead>
<tr>
<th>Variable</th>
<th>t-value</th>
<th>p-value</th>
<th>t-value</th>
<th>p-value</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAR</td>
<td>1.953</td>
<td>0.054</td>
<td>0.033</td>
<td>0.900</td>
<td>-9.366</td>
<td>0.000</td>
</tr>
<tr>
<td>DR</td>
<td>-1.252</td>
<td>0.214</td>
<td>2.305</td>
<td>0.024</td>
<td>3.288</td>
<td>0.002</td>
</tr>
</tbody>
</table>

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 a graph 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 of Fuzzy ART based on K-means with the SNC labelling algorithm yields 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


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.
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.
networks. Proceedings of the World Congress on
Engineering and Computer Science, October 22-24,
2008, San Francisco, USA.
clustering method for intrusion detection. Pattern
Recogn., 37: 927-942.
MacQueen, J., 1967. Some methods for classification and
multivariate of observations. Proc. 5th Berkeley Symp.
Ngamwittayanon, N., N. Wattanapongsakorn and
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.
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.
detection with unlabeled data using clustering.
Proceedings of ACM CSS Workshop on Data
Mining Applied to Security, November 5-8, 2001,
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
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: 579-784.
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.
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. Sivathasan, J. Imayavendhan,
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,
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.
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.
Clustering-based network intrusion detection.