A HYBRID FRAMEWORK BASED ON NEURAL NETWORK MLP AND K-MEANS CLUSTERING FOR INTRUSION DETECTION SYSTEM

Mazyar Mohammadi Lisehroodi[1], Zaiton Muda[1], and Warusia Yassin[2]

[1]University Putra Malaysia, Malaysia, mohammadi.mazyar@gmail.com
[1]University Putra Malaysia, Malaysia, zaiton@fsktm.upm.edu.my
[2]University Technical Malaysia, Malaysia,aisuraw@yahoo.com

ABSTRACT. Due to the widespread use of Internet and communication networks, in case a reliable and secure network plays a crucial role for information technology (IT) service providers and users. The hardness of network attacks, as well as their complexity, has also increased lately. High false alarm rate is a big issue for majority of researches in this area. To overwhelm this challenge a hybrid learning approach is proposed, employing the combination of K-means clustering and Neural Network Multi-Layer Perceptron (MLP) classification. Concerning the robustness of K-means method and MLP algorithms benefits, this research is the part of an effort to develop a hybrid information detection system (IDS) which is able to detect high percentage of novel attacks while keep the false alarm at low rate. This paper provides the conceptual view and a general framework of the proposed system.

Keywords: intrusion detection system, K-means clustering, neural network classifier, Multi-Layer Perceptron

INTRODUCTION

Intrusion detection has emerged to gather and analyze a number of key points in computer systems and networks, to find if there are abnormal behaviors against the policy of system or violent sign in the network. The combination of software and hardware for detection intrusion is called IDS.

Intrusion detection technologies aim to identify two major groups of attack: misuse detection and anomaly detection. Anomaly detection attempts to ‘learn’ the features of event patterns that found normal behavior, and, by observing patterns that deviate from established norms, detect when an intrusion has occurred. Misuse detection comprises the comparison of user’s activities and the known behaviors of assailants to infiltrate a network. Misuse detection method using a rule-based approach to detect known attacks by matching attack’s pattern with the list of signatures, greatly similar to antivirus applications. Since the signatures should be updated regularly, if the signature is not included in its library this type of IDS is unable to detect the zero-day attacks. Contrasting misuse detection, anomaly based detection is involved monitoring user’s activities to catch any deviation from normal behavior profile. Despite being able to detect unknown attacks, the probability of high false alarm is considerable. Due to the vast number of vulnerabilities of mobile and ad hoc networks in term of security, utilizing neural network as anomaly based IDS can be effective to touch near zero false positive and false alarm rate (Jabbehdari et al., 2012).

No one can deny the signature based IDS advantages, but since most signature-based IDSs are suffering to detect the zero-day attacks, researchers are more interested in anomaly detection techniques.

However many of these approaches resulted in high detection rate and accuracy, most of them encounter high false alarm rates. As the result of this falsely classification, authentic users cannot access to the network. Therefore, IDS research area is in desperate need of focusing on false alarm to properly identify such intrusions.

In the following study, this research is organized as follows in section II related works have been discussed to explore the current techniques in detection rate, accuracy and false alarm. In section III proposed framework is a combination of Neural Network classification and K-means to enhance the detection rate and accuracy, and decrease the false alarm rate in anomaly detection technique. And finally in the last section conclusion and discusses about the result.

RELATED WORK

IDS has become a significant area of research regarding safe and secured IT communication. Many of previous researches proposed unsupervised anomaly detection approaches with various classification algorithm(Aneetha & Bose, 2012) . Most researches employed Knowledge Discovery and Data Mining (KDD) cup'99 for evaluation of their proposed approaches.

Hybrid learning approaches have been broadly discovered such as in (Kavitha, Karthikeyan, & Sheeba Maybell, 2012; Li et al., 2012).False alarm (FA), true positive (TP), false positive (FP), false negative (FN), detection rate (DR) and accuracy rate have been issued in the most of these approaches. Some approaches showed high number of detection but unable to improve false alarm rate. To boost the accuracy, a Fuzzy Support Vector Machine (SVM) which is called (FSVM) was proposed aiming to build a new training set using centroids (Teng, Du, Wu, Zhang, & Su, 2010). FSVM is used for training the new set to gain a support vector. The experimental result showing a reasonable increase in accuracy but still rooms to improve it.

The combination of various neural network and clustering method has been proposed to improve anomaly detection for instance Self-organizing map (SOM) is modified to eliminate the drawbacks of the traditional SOM. In this method the network is allowed to be developed with the connection strength to recognize neighbor nodes. The results of the modified SOM are grouped by K-means by natural behavior and centroids (Bose.S & Aneetha.A.S, 2012; S. Lee & Kim , 2011).

Modification of SOM engaged with K-means clustering can enhance anomaly detection rate in a considerable range. However, 2% false alarm still is a big issue(Bose. S, Aneetha. A.S, 2012).

Many of previous works employed artificial neural networks (ANN) to resolve difficult real-world problems. For instance, Formal Concept Analysis ANN (FC-ANN) approach used Fuzzy clustering method alongside neural network to enrich anomaly detection rate (Wang, Hao, Ma, & Huang, 2010). Training dataset is generated by various models of ANN and Fuzzy clustering is used to collect the results. Though this model has been successful to detect R2L and U2R attack, it has been unable to show a reasonable rate in Probe detection.

SOM algorithm can be modified to overcome fixed architecture. To achieve this goal, a new neighborhood updating rules is included and the learning phase is done dynamically. As the primary step, algorithm starts with empty network and develops with the original data space. Distance threshold measure is used to create new nodes and their neighborhoods will be found using connection strength. The results claim 98% detection rate and 2% false alarm rate (Bose. S, Aneetha.A.S, 2012; S.Lee et al., 2011). A new hybrid model applied C-means clustering, Fuzzy neural network and radial basis function (RBF). This study defined four steps. First step for analysis, next C-means is used for data clustering. Then neuro-fuzzy classifier related with the clusters to train each of the nodes. Finally, RBF-Support Vector Machine (RBF-SVM) classifies data to detect intrusions. Comparisons showing the efficiency of this techniques in detection rate and F-measure but high percentage of false alarm is a limitation(Chandrashekhar & Raghuveer, 2013). In this study, a hybrid approach is proposed to detect all types of attack (R2L, U2R, Probe, DoS) using the combination of K-means clustering and neural network Multi-Layer Perceptron (MLP) classifier. After training with KDD cup'99 training data set, 17 new types of attacks will come by Testing data set. This system can detect such anomalies by distinguishing them from normal behaviors. The false positive and false alarm rate are near to zero while high rate of detecting new attacks.

PROPOSED FRAMEWORK

Hybrid learning Approach

Accuracy and detection rate are mostly increasing but false alarm still is a big concern in IDS research area. Therefore, in this study a framework has been proposed which is combination of K-means clustering algorithm with artificial neural network to decrease false alarm and increase accuracy and detection rates at the same time. Our proposed method called KM-NEU. The KM-NEU approach involves two modules. First, data is clustered based on their natural behavior using K-Means as clustering component. Then clustered data are classified by Neural Network MLP algorithm. *K-means clustering module*. K-means clustering is using to group data into attack and normal instances (Muda & Yassin, 2011). A cluster is a collection of records that are similar to one another and dissimilar to records in other clusters. The similarity of the records within the cluster is maximized, and the similarity to records outside this cluster is minimized. Standard classification of network intrusion fall into four major categories: DoS, Probe, U2R and R2L. K-means clustering partition raw dataset into k-cluster based on the initial value which called seed-points into each cluster's centers. Centroids is specified by the mean value of numerical data contain in each single cluster. K-means clustering module can be summarized as following Pseudo code:

```
Initialize $\boldsymbol{m}_i, i = 1,\ldots,k$, for example, to $k$ random $x^t$
Repeat
    For all $\boldsymbol{x}^t \epsilon X$
        $b_i^x \leftarrow \begin{cases} 1 & \text{if } \|\boldsymbol{x}^t - \boldsymbol{m}_i\| = \min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\| \\ 0 & \text{otherwise} \end{cases}$
    For all $\boldsymbol{m}_i, I = 1,\ldots,k$, for example, to $k$ random $x^t$
$\boldsymbol{m}_i \quad \sum_t \leftarrow b^t x^t / \sum_t b^t$
Until $m_i$converge
```

Neural network classifier module. As the related works indicated there are several researchers used the various clustering techniques like C-means, Fuzzy clustering and K-means. Most of them were unable to differ attack records and normal records appropriately. In this study regarding the advantages of neural networks which is able to characterize both nonlinear and linear functions and their ability to learn these relationships straight from the data being modeled, a combination of K-means clustering with neural network classifier has been employed.

Neural network divided into two architectures: SVM and MLP. In this study MLP is used as the classifier algorithm. In such neural network every neuron's input of the next layer is joined to neuron's output of the previous layer. MLP architecture consist of at least one hidden layer. Due to signal transition establishment through the network from input to output, this architecture is called feed forward.

MLP is known as feed forward neural network was initially provided for the non-linear XOR, and was then effectively used to different combinatorial problems. MLP is mostly used for information managing and pattern recognition in prediction of seismic activities. It is incredibly used and analyzed with different problems such as in time series prediction and function approximation (Shah, Ghazali, & Nawi, 2011). Thus MLP can be used as a nonlinear model for regression as well as for classification. As the arriving information in the network could be failed randomly, these characteristics are vital in IDS. Moreover, as multiple attacker may target the system with a synchronized assault, being able to process data from different sources in a non-linear manner is crucial. Figure 1 shows the design of MLP with two hidden layers, one output layer, and one input layer.

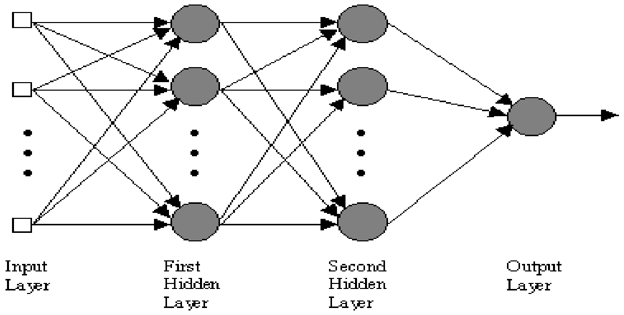

Figure 1. Multi layered Perceptron Neural Network

$$yi = fi\,(\sum_{i=1}^{n} wij\ xi + bi) \qquad (1)$$

Eq. (1) calculates y_i which is the output of the node, x_i stands for the ith input for the node while w_{ij} represents the connection weight between input and output node Θ_i is threshold of the node, and the node transfer function showed by f_i. The network error function summarized Eq. (2):

$$E\big(w(t)\big) = \frac{1}{n}\sum\nolimits_{j=1}^{n} \sum_{k=1}^{k} (dk - Ot) \qquad (2)$$

E(w(t)) is the error function of tth iteration and w(t) is the weight of connection in tth iteration. dk is the wanted output node and Ot is the real value of kth output node's is the number of output node and n is the number of pattern and T isoptimization goal to minimize the objective function by optimizing the weight of network w(t).

Framework design

The previous learning methodologies offered the high range in accuracy and intrusion detection rate for unknown attacks(W. Lee & Stolfo, 2000).On the other hand, these approaches were unable to decrease the false alarm rate. In this study it has been believed that the suggested hybrid framework can satisfy both objectives in an acceptable range. Figure 2 illustrates the structural design of this study for implementation of KM-NEU approach. Phase I is aimed for preparing data set while phase II and III are designed for clustering and classification respectively.

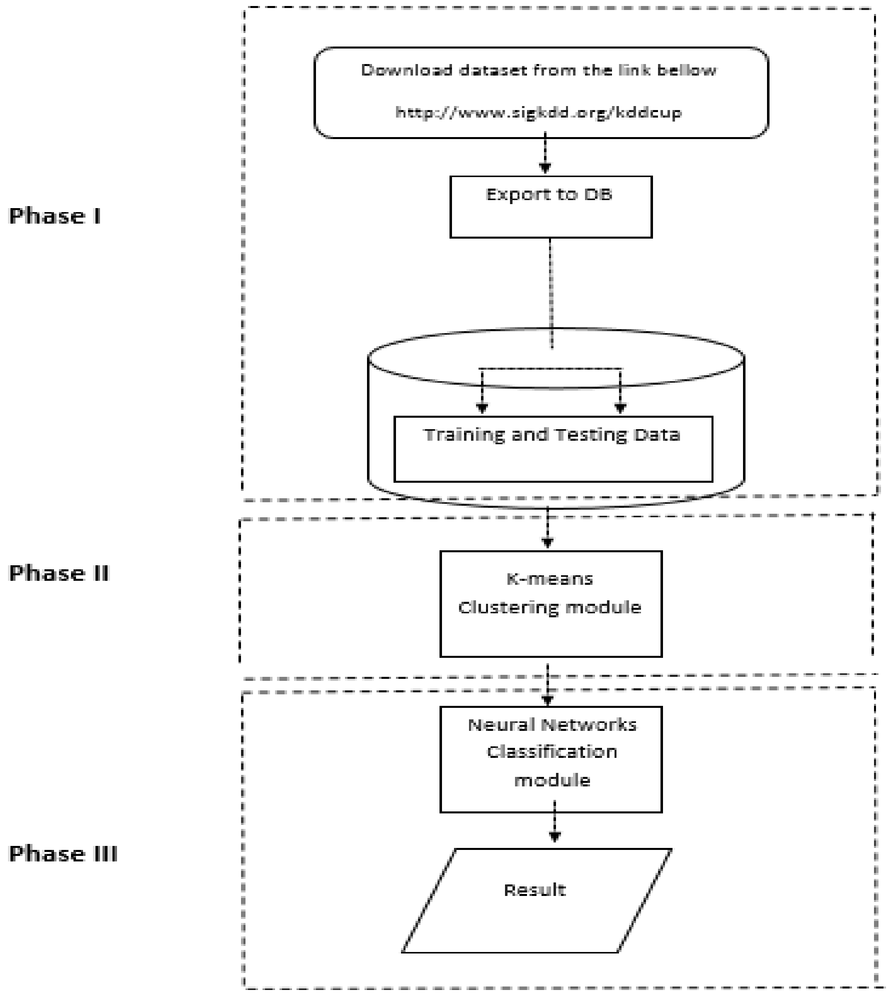

Figure 2. Structural Design

Phase I: Data Preparation In this phase, KDDcup 99'("KDD Cup 1999 Data," n.d.) Data set that mentioned in Figure 2.Then the data is converted from the text format to comma separated value and will be uploaded to simulator.

Phase II: K-means clustering: The data which provided in phase I is clustered by K-means clustering algorithm. As the result of grouping data into K number of clusters, the normal data is accurately separated from the attack one.

Phase III: Neural Network Classifier: Most of the previous works in this area used a single classifiers like neural network, SVM, NB-Tree, GA and Zero-R. Due to the fact that, they have been unable to increase accuracy and detection rate while keep false alarm in low range, in this research the neural network classifier is selected to apply to the dataset just after the clustering phase. Neural network classifies the data into five specific categories which are Normal, Dos, U2R, Probe and R2L.This approach aim to reduce the false alarm while keeping accuracy and detection at high range.

Evaluation Measurement: Regarding the previous researches in IDS area, the performance of IDS is measured and evaluated by value of accuracy, detection rate and false alarm which defined in Eq. (3), (4) and (5):

$$Accuracy = (TP + TN)/(TP + TN + FP + FN \tag{3}$$

$$Detection\ Rate = TP/(TP + FP) \tag{4}$$

$$False\ Alarm = FP/(FP + TN) \tag{5}$$

RESULTS AND DISCUSSION

Two different experiments have been done in this research to compare the single classifier and hybrid approach using the training and testing datasets. In classification matter neural network MLP is selected as single classifier and K-means clustering pre-processing is combined with MLP to produce this approach as hybrid K-means+Neural Network(KM-NEU). Evaluation of this approach is based on the comparison of single ANN and Hybrid approach in terms of accuracy, detection rate and false alarm. Moreover, KM-NEU approach is compared with the other related approaches which used same KDD Cup'99 dataset. Table 3 represents how different types of data classified by Neural network and KM-NEU using the testing data set.

Table 3. Classification Result for Each Type of Data Using Testing Dataset

Data	Normal	Probe	Dos	U2R	R2L
Neural Network	97.92%	97.17%	100%	0%	70%
KM-Neu	99.99%	99.97%	99.99%	99.99%	99.98%

Table 4. Detection Results for the Normal and Attack Classes Using Testing Dataset (neural networks)

Actual	Predicted Normal	Predicted Attack
Normal	97.9%	2.1%
Intrusion	0.77%	99.03

Table 5. Detection Results for the Normal and Attack Classes Using Testing Dataset (KM-Neu)

Actual	Predicted Normal	Predicted Attack
Normal	99.9%	0.01%
Intrusion	0.001%	99.9%

As it observed in Tables 4 and 5 Neural Network resulted less false positives and negatives.

CONCLUSION

Due to the related works about IDS, the challenges and gaps in this area are how to increase accuracy and detection rate, in the contrary decreasing the false alarm rate. It is believed that, due to the robustness of K-means clustering in grouping data, it can be effective to employ it as the pre-classification algorithm. Thus, MLP neural network as the latter algorithm have a better classification over the preprocessed dataset. This approach can be helpful to overcome the single neural networks' difficulties in terms of false alarm. Considering the weakness of single MLP which is unable to detect novel attacks in an acceptable range, implementation of the proposed framework using the generally accepted dataset such as KDD Cup '99 and DARPA is considered as future work.

REFERENCE

Aneetha, A., & Bose, S. (2012). The combined approach for anomaly detection using neural networks and clustering techniques. *Computer Science & Engineering*. Retrieved from http://airccse.org/journal/cseij/papers/2412cseij04.pdf

Bose. S, Aneetha. A.S, R. S. (2012a). Dynamic network anomaly intrusion detection system using modified SOM. *Proceedings of Second International Conference of Computer science and Engineering - 2012, New Delhi* (pp. 27–34).

Chandrashekhar, A., & Raghuveer, K. (2013). Fortification of Hybrid Intrusion Detection System Using Variants of Neural Networks and Support Vector Machines. *International Journal of Network Security & Its Applications*, *5*(1), 71–90. doi:10.5121/ijnsa.2013.5106

Jabbehdari, S., Talari, S. H., & Modiri, N. (2012). A Neural Network Scheme for Anomaly Based Intrusion Detection System in Mobile Ad hoc Networks, *4*(2), 61–66.

Kavitha, B., Karthikeyan, D. S., & Sheeba Maybell, P. (2012). An ensemble design of intrusion detection system for handling uncertainty using Neutrosophic Logic Classifier. *Knowledge-Based Systems*, *28*, 88–96. doi:10.1016/j.knosys.2011.12.004

KDD Cup 1999 Data. (n.d.). Retrieved May 13, 2013, from http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Lee, S., Kim, G., & Kim, S. (2011). Self-adaptive and dynamic clustering for online anomaly detection. *Expert Systems with Applications*, *38*(12), 14891–14898. doi:10.1016/j.eswa.2011.05.058

Lee, W., & Stolfo, S. (2000). A framework for constructing features and models for intrusion detection systems. *ACM Transactions on Information and System Security*, *3*(4), 227–261.

Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X., & Dai, K. (2012). An efficient intrusion detection system based on support vector machines and gradually feature removal method. *Expert Systems with Applications*, *39*(1), 424–430. doi:10.1016/j.eswa.2011.07.032

Muda, Z., & Yassin, W. (2011). Intrusion detection based on k-means clustering and OneR classification. *Information Assurance and Security (IAS)*, 192–197. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6122818

Shah, H., Ghazali, R., & Nawi, N. (2011). Using artificial bee colony algorithm for mlp training on earthquake time series data prediction. *arXiv preprint arXiv:1112.4628*, 1–8. Retrieved from http://arxiv.org/abs/1112.4628

Teng, S., Du, H., Wu, N., Zhang, W., & Su, J. (2010). A Cooperative Network Intrusion detection Based on Fuzzy SVMs. *Journal of Networks*, *5*(4), 475–483. doi:10.4304/jnw.5.4.475-483

Wang, G., Hao, J., Ma, J., & Huang, L. (2010). A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. *Expert Systems with Applications*, *37*(9), 6225–6232. doi:10.1016/j.eswa.2010.02.102