A Novel Cluster-based Intrusion Detection Approach Integrating Multiple Learning Techniques

Hossein Shapoorifard
Department of Computer & Electrical Engineering
Shiraz branch, Islamic Azad University, Shiraz, Iran

Pirooz Shamsinejad
Department of Computer Engineering and information technology,
Shiraz University of Technology, Shiraz, Iran

ABSTRACT
In order to make computer systems completely secure, in addition to firewalls and other intrusion protection devices, other systems called intrusion detection systems (IDS) are needed to detect intrusion and provide solutions to counter the intruder if he penetrated through firewall, antivirus and other security devices. Many IDS have been developed based on machine learning techniques. Specifically, advanced detection approaches created by combining or integrating multiple learning techniques have shown better detection performance than general single learning techniques. This paper proposes an improvement for a feature representation approach, namely the cluster center and nearest neighbor (CANN) approach.

Keywords
Intrusion Detection System; Data Mining; Hybrid Intrusion Detection System; anomaly detection; cluster center, nearest neighbor.

1. INTRODUCTION
More powerful hardware, more advance mobile devices, improvements in computing and network technology made the Internet an important part of our daily life and the extensive growth in using the Internet in social networking (e.g., social media apps, video conferences, etc.), healthcare, e-commerce, bank transactions, and many other services is undeniable. These Internet applications need a satisfactory level of security and privacy[1]. On the other hand, the amount of people who have access to the Internet is increasing rapidly. therefore, the high popularity of world-wide connections has led to more complex security issues, our computers are under attacks and vulnerable to many threats. There is an increasing availability of tools and tricks for attacking and intruding networks. An intrusion can be defined as any set of actions that threaten the security requirements (e.g., integrity, confidentiality, availability) of a computer/network resource (e.g., user accounts, file systems, and system kernels)[2].

Traditionally, some techniques, such as user authentication, data encryption, firewalls, and antiviruses, are used to protect computer security. Intrusion detection systems (IDS), which use specific analytical technique(s) to detect attacks, identify their sources, and alert network administrators, have recently been developed to monitor attempts to break security[3][4]. In general, there are two types of Intrusion IDS; signature detection systems and anomaly detection systems. for anomaly detection, IDS should identify normal behaviors and define a specific pattern for them. Behaviors which fallow these patterns will be considered as normal and behaviors which deviate from these patterns more than a threshold will be considered as attack. On the other hand, in signature detection, intrusion patterns are predefined as rules. Each pattern represents different varieties of a certain type of intrusion. In this method, detecting system normally insists on a database that contains pattern and signature of intrusions, and any match with signatures is reported as a possible attack.

Both of the mentioned approaches have their own drawbacks. *"TABLE I"* shows a comparison between the two types of intrusion detection:

TABLE II. Comparison between signature detection and anomaly detection

	IDS Type	
	signature detection	***anomaly detection***
drawbacks	- False negatives - Unable to detect new attacks - Need signatures update frequently	- False positives. - Has to study sequential interrelation between transactions - Overwhelming security analysts

We can conclude from *"TABLE III"* that traditional IDS suffer from many limitations. This has led to an increased interest in improving IDS. There have been many recent studies, which focus on combining or integrating different techniques in order to improve detection performance, such as accuracy, detection, and/or false alarm rates. KDDCUP'99 is the mostly widely used dataset for the evaluation of these systems[5]. statistical analysis on this data set, showed important issues which highly affects the performance of evaluated systems, and results in a very poor evaluation of intrusion detection approaches. To solve these issues, we have used a newer data set, NSL-KDD, according to L.Dhanabal (2015) NSL-KDD data set is a refined version of KDD. It contains essential records of the complete KDD data set[6].

This paper proposes an improvement to a feature representation approach named CANN which is based on cluster center and nearest neighbor[4], therefore we decided to call it ICANN. In this approach, two distances are measured and summed:

- distance between each data sample and its cluster center.
- distance between data and its nearest neighbor in the same cluster.

As a result, it will induce a one-dimensional distance based feature which will be used to represent each data sample for intrusion detection by a *k*-nearest neighbor (*k*-NN) classifier. As our experimental results based on the NSL-KDD dataset shows, in terms of classification accuracy, detection rates, and false alarms the ICANN classifier performs better than CANN, and in the worst state it performs similar to CANN.

2. PREVIOUS RESEARCHS

Data mining approaches have several applications and inspirations in various domains fields such as in text mining, image processing, bioinformatics, search engines and on [7]–[13]. One important application of data mining is for fraud detection which especially in the two recent decades many researchers have investigated the deployment of data mining algorithms and techniques and combining or integrating them for intrusion detection systems (e.g. [4], [14]–[16]). For instance, Aslahi Shahri et al.[16] proposed a hybrid method of support vector machine and genetic algorithm (GA) and implementing that in intrusion detection. this algorithm reduces the number of features from 45 to 10 and categorizes them into three priorities using GA algorithm. the most important is the first priority and the least important is placed in the third priority.

Muda et al.[17] combined K-MEANS clustering and ONE-R classification and used it for IDS and they named it KM+IR.it uses K-MEANS for clustering the data and then ONE-R to classify the data in each cluster. Wang et al.[18] presented A new approach using Artificial Neural Networks (ANN) and fuzzy clustering, called FC-ANN, firstly fuzzy clustering technique is used to generate different training subsets. Subsequently, based on different training subsets, different ANN models are trained to formulate different base models. Finally, a meta-learner, fuzzy aggregation module, is employed to aggregate these results. Gisung Kim et al. [19]presents a new hybrid intrusion detection method hierarchically integrates a misuse detection and anomaly detection in a decomposed structure. The misuse detection model is built based on C4.5 decision tree algorithm and is used to decompose the normal training data into smaller subsets. The one-class SVM is used to create anomaly detection for the decomposed region.C4.5 decision tree does not form a cluster, which can degrade the profiling ability. Vahid Golmah. [20]proposed an efficient hybrid intrusion detection method based on C5.0 and SVM .This method achieves a better performance compared to the individual SVM. Amuthan Prabakar Muniyandi et al. [21] proposed an anomaly detection method using K-Means+C4.5 , a method to cascade k-means clustering and the C4.5 decision tree methods. This method achieves better performance in comparison to the K-Means, ID3, Naïve Bayes, K-NN, and SVM. Adel Sabry Eesa et al. [22] represented A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Cuttlefish. The model uses the cuttlefish algorithm (CFA) is used to find optimal subset of features and ID3 classifier as a judgment on the selected features that are produced by the CFA.

3. PROPOSED APPROACH

The most important purpose of IDS is increasing the rate of detecting suspicious behaviors and decreasing failure alarm rate. We discussed some previous approaches in this field. Obviously, methods which are insisting on integrating and combining different techniques are showing better results. in this paper we use two data mining machine learning algorithms

- *k*-MEANS classification
- *k* nearest neighbor clustering

3.1 The ICANN process

As we said before the proposed approach is based on two distances which are used to determine the new features, between a specific data point and its cluster center and nearest neighbor respectively. The steps are shown in *"Fig. 1,"*

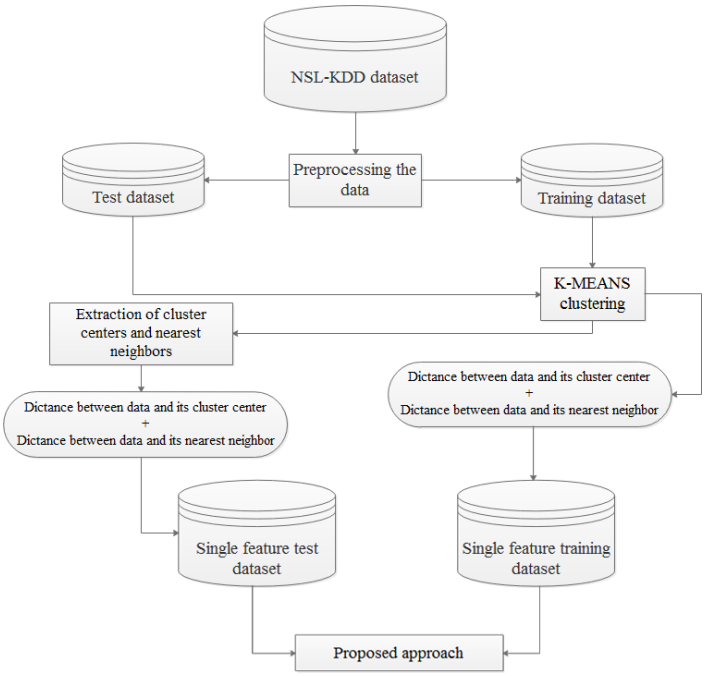

Figure I. The ICANN process

At first our data from NSL-KDD dataset goes for preprocessing or normalization.in this step, process starts with linear transformation of the data. Assume that minA and maxA are maximum or minimum amounts of a certain feature. With minimum-maximum normalization, the v amount from A will be maped to v' amount in [$new_$minA , $new\_$ maxA] with this *"equation. 1"*

$$v' = \frac{v - minA}{maxA - minA}(new_{minA} - new_{maxA}) + new_{minA} \qquad \text{(I)}$$

Then data goes for clustering. Here we used *k*-MEANS algorithm for clustering our test and training data, assigning the most similar data to a same cluster is our purpose. The number five is considered for the number of clusters *(k)* because In the NSL-KDD data set beside the normal traffic there are four types of attack

- Probe
- Dos
- U2R
- R2L

Therefore we are dealing with total of five types of behavior. One type of normal behavior and four types of attack. Then we need to use *k* nearest neighbor algorithm to find the nearest neighbor of our data in its cluster. To do this we use Euclid distance to find distances between our data and its nearest neighbor, after that we choose the minimum of them. Then we need to calculate the distance between our data and five cluster centers, again we use Euclid distance. All of this distances are shown in *"fig. II"*

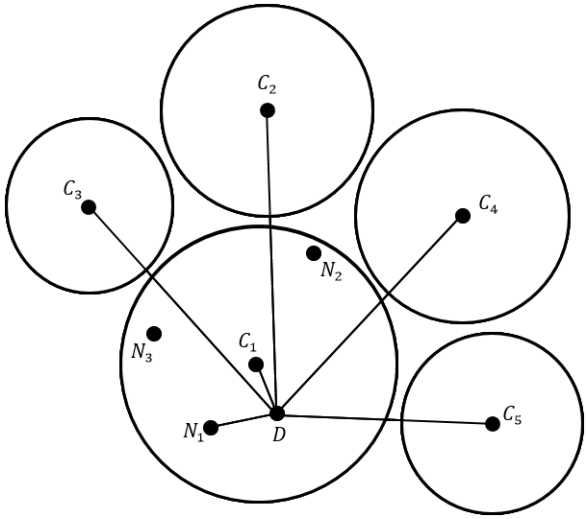

Figure II. An example for distances between data and cluster centers, data and its nearest neighbor

Assume that *D* in *"fig. II"* is our data, C_1 to C_5 are cluster centers and N_1 is the nearest neighbor for *D*. now we can calculate the summation of distances between *D* and cluster centers (C_1 to C_5) and *D* and N_1, we can call it D_i

$$D_i = \overline{DC_1} + \overline{DC_2} + \overline{DC_3} + \overline{DC_4} + \overline{DC_5} + \overline{DN_1} \quad \text{(II)}$$

Now D_i is a feature which can represent all of the features of *D* because all of them affected *D* during the mentioned process, so D_i is a representative feature for *D*. we can do this for all of our data in the dataset, having just one representative feature for each data. Now with this process, our training and test datasets are transformed to single representative feature datasets.

In a distance based IDS such as our proposed approach we assume that the distance between normal and abnormal data is big enough to make them distinguishable. We can use similarities and differences between our test data and our training data to find out whether it is an attack or not. In this approach not only the most similar data to our test data affects the result but the most different data is effective too. Now we can classify our data to attack or normal behavior easily by *k*-NN under our mentioned basic assumption, (e,g we want to classify our test data and the most similar data to our data is normal and the most different one is an attack, so we can classify our data as a normal behavior)

4. EXPERIMENTAL RESULTS

In this study, we consider the rates of accuracy, detection and false alarms, which are widely used in literatures, to evaluate the performance of intrusion detection, They can be calculated by a confusion matrix as shown in *"TABLE IVV"*.

TABLE VIVII. Confusion matrix

actual \ predicted	Normal	Attacks
Normal	TN	FP
Attacks	FN	TP

- True Positives (TP): the number of malicious executables correctly classified as malicious;
- True Negatives (TN): the number of benign programs correctly classified as benign;
- False Positives (FP): the number of benign programs falsely classified as malicious;
- False Negative (FN): the number of malicious executables falsely classified as benign.

Then, the rates of accuracy, detection and false alarm can be obtained by:

$$Detection\ Rate = \frac{TP}{TP+FP} \quad \text{(III)}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(IV)}$$

$$False\ Alarm\ Rate = \frac{FP}{FP+TN} \quad \text{(V)}$$

To determine our new approaches performance and compare it with the previous approach we examined both of them with the same random input from NSL-KDD containing about 30000 records, we repeated or experiment ten times, And we noticed significant improvement in accuracy, detection, and false alarm rates. As you can see in *"fig. III" "fig. IV" "fig. V"* after ten times of repeating the experiment, most of the time ICANN performs better than CANN and in the worst cases it works similar to CANN.

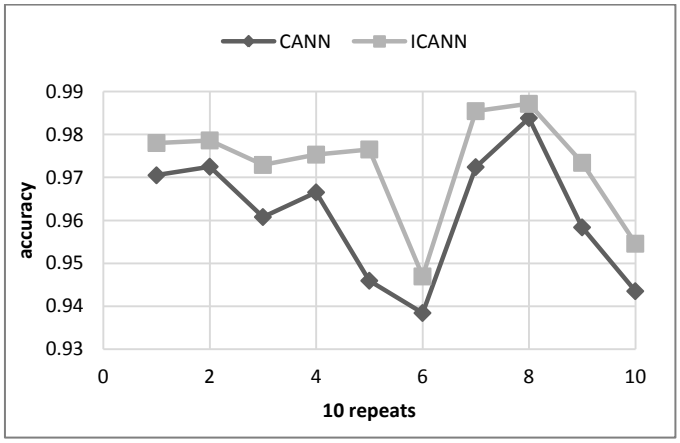

Figure III. Accuracy, ICANN vs CANN(higher is better)

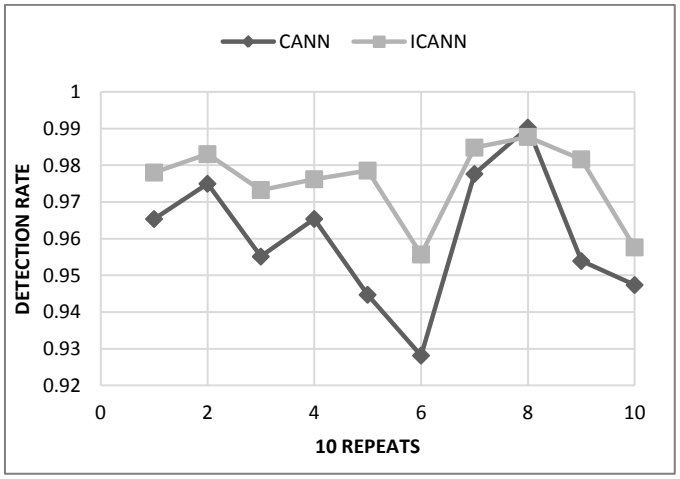

Figure IV. Detection rate, ICANN vs CANN(higher is better)

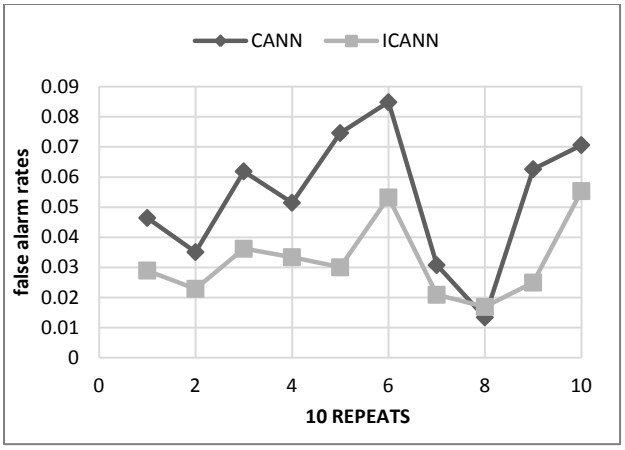

Figure V. False alarm rates, ICANN vs CANN(lower is better)

5. CONCLUSION

This paper presents an improved feature representation approach that combines cluster centers and nearest neighbors for effective and efficient intrusion detection. The ICANN approach first transforms the original feature representation of a given dataset into a one-dimensional distance based feature. Then, this new dataset is used to train and test a *k*-NN classifier for classification. The experimental results show that ICANN performs better than CANN, providing higher accuracy and detection rates and a lower false alarm rate. But CANN requires a little bit less computational effort than the ICANN. Another thing that should be mentioned here is that CANN cannot effectively detect U2L and R2L attacks, but the ICANN can detect this types of attacks effectively and that's because of considering the most different data along with the most similar data. Finally, as ICANN is applicable to the 5-class intrusion detection problem, other domain datasets including different numbers of dimensions and classes can be used to examine the effectiveness of ICANN.

6. REFERENCES

[1] A. Youssef and A. Emam, "Network Intrusion Detection Using Data Mining and Network Behaviour Analysis," *Int. J. Comput. …*, vol. 3, no. 6, pp. 87–98, 2011.

[2] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[3] Y. Chen, A. Abraham, and B. Yang, "Hybrid flexible neural-tree-based intrusion detection systems," *Int. J. Intell. Syst.*, vol. 22, no. 4, pp. 337–352, 2007.

[4] W. C. Lin, S. W. Ke, and C. F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-Based Syst.*, vol. 78, no. 1, pp. 13–21, 2015.

[5] M. K. Siddiqui and S. Naahid, "Analysis of KDD CUP 99 Dataset using Clustering based Data Mining," *Int. J. Database Theory Appl.*, vol. 6, no. 5, pp. 23–34, 2013.

[6] L. Dhanabal and S. P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 6, pp. 446–452, 2015.

[7] A. Taheri and M. Shamsfard, "Mapping farsnet to suggested upper merged ontology," in *Asia Information Retrieval Symposium*, 2011, pp. 604–613.

[8] M. H. Dashtban and P. Moradi, "A novel and robust approach for iris segmentation," *Int. J. Comput.*, 2011.

[9] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, 2017.

[10] S. Olyaee, Z. Dashtban, M. H. Dashtban, and A. Najibi, "Hybrid analytical-neural network approach for nonlinearity modeling in modified super-heterodyne nano-metrology system," in *Telecommunications (ConTEL), Proceedings of the 2011 11th International Conference on*, 2011, pp. 525–530.

[11] A. Taheri and M. Shamsfard, "SBUEI: results for OAEI 2012," in *Proceedings of the 7th International Conference on Ontology Matching-Volume 946*, 2012, pp. 189–196.

[12] M. H. Dashtban, Z. Dashtban, and H. Bevrani, "A novel approach for vehicle license plate localization and recognition," *Int. J. Comput. Appl.*, vol. 26, no. 11, 2011.

[13] S. Olyaee, Z. Dashtban, and M. H. Dashtban, "Design and implementation of super-heterodyne nano-metrology circuits," *Front. Optoelectron.*, vol. 6, no. 3, pp. 318–326, 2013.

[14] M. A. Aydin, A. H. Zaim, and K. G. Ceylan, "A hybrid intrusion detection system design for computer network security," *Comput. Electr. Eng.*, vol. 35, no. 3, pp. 517–526, 2009.

[15] R. M. Elbasiony, E. A. Sallam, T. E. Eltobely, and M. M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted k-means," *Ain Shams Eng. J.*, vol. 4, no. 4, pp. 753–762, 2013.

[16] B. M. Aslahi-Shahri *et al.*, "A hybrid method consisting of GA and SVM for intrusion detection system," *Neural Comput. Appl.*, vol. 27, no. 6, pp. 1669–1676, 2016.

[17] Z. Muda, W. Yassin, M. N. Sulaiman, and N. I. Udzir, "Intrusion detection based on K-means clustering and OneR classification," in *Proceedings of the 2011 7th International Conference on Information Assurance and Security, IAS 2011*, 2011, pp. 192–197.

[18] G. Wang, J. Hao, J. Ma, and L. Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6225–6232, 2010.

[19] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Syst. Appl.*, vol. 41, no. 4 PART 2, pp. 1690–1700, 2014.

[20] V. Golmah, "An Efficient Hybrid Intrusion Detection System based on C5. 0 and SVM.," *Int. J. Database Theory Appl.*, vol. 7, no. 2, pp. 59–70, 2014.

[21] A. P. Muniyandi, R. Rajeswari, and R. Rajaram, "Network anomaly detection by cascading k-Means clustering and C4.5 decision tree algorithm," in *Procedia Engineering*, 2012, vol. 30, pp. 174–182.

[22] A. S. Eesa, Z. Orman, and A. M. A. Brifcani, "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2670–2679, 2015.