ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 4, April 2018

# Design Of An Intrusion Detection System Based On Distance Feature Using Ensemble Classifier

[1] R.Radhika, [2] B. Sundarraj
[1] PG Student ,Dept. of CSE, BIHER, Chennai, Tamil Nadu.
[2] Assistant Professor, Dept.of CSE, BIHER, Chennai, Tamil Nadu.

*Abstract*: ***The focus of Intrusion Detection System (IDS) is used to determine the computer usage and detect any malicious network traffic. These activities cannot be detected by conventional firewall. Various IDS have been developed using advanced detection approaches which is created by integrating different techniques which shown better detection performance than existing techniques. The major disadvantage of IDS is it often provides false report of malicious activities. Sometimes it also misses out major malicious threat or intrusion which is nothing but it (IDS) needs to extract more features for normal connections and needs more reasonable ways to detect the attack. This paper proposes a better representation namely, the cluster center and nearest neighbor (CANN) approaches. In this approach, we will measure and sum up the distance between each data and its cluster center. Initially the distance between data and cluster is measured and then the data and its nearest neighbor in same cluster is identified. This CANN classifier performs like k-Nearest Neighbor (k-NN) in identifying the defects and reduces false alarm also CANN provides efficient training and testing to find the defect.***

***Keywords: Intrusion Detection, Feature Representation, Cluster center, Nearest Neighbor.***

## I. INTRODUCTION

Accessing the internet has become the important part of our daily life due to advancement in network technology and computing techniques. There has been a rapid improvement in number of people connected to internet which led to security problems. Traditionally, we have few methods and techniques to protect computer security such as firewall, authentication, and encryption and so on.

Intrusion Detection System (IDS) are used to detect attacks and identify the main sources by means of specific analytical techniques and alert the network administrator and monitor the attempts to break security. IDS are developed for (i) anomaly detection which uses behavior patterns to indicate malicious activities, it will analyze the past activities to check the observed activity is normal, (ii) signature detection will have a sequence of commands to determine the indicative of attack, it will check the data packets or audit logs with previously observed sequence of command. As signature detection is used by early IDS to detect the attacks by validating from signature database, it has high rate of false alarm.

Many studies and research have been carried out to combine or integrate various techniques to improve detection such as false alarm rate, performance and so on. But, there are few limitations in these studies. First, amongst various sophisticated and advanced studies and approaches, very few have focused on feature representation for attacks. However, there is no clear view about which dataset is more representative among the different dataset used such as KDD-Cup 99 and DARPA 1999. Second, the efficiency of on-line detection is degraded as training the system for detection takes much time for validation is not considered in evaluation methods.

## II. EXISTING SYSTEM

In this study, we propose combining cluster centers and nearest neighbors (CANN) for novel feature representation for efficient and effective intrusion detection. It uses k-means cluster algorithm to extract the cluster centers and then the sum of distance between cluster center and specific data sample. The new dataset will be one dimensional to indicate the distance using k-Nearest neighbor classification which will provide effective intrusion detection.

## III. MACHINE LEARNING

It requires a system capable of k knowledge integration and autonomous acquisition. It is mainly used to improve analytical observation for the system as it can continuously self-improve and increase the systems effectiveness.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 4, April 2018

Algorithms are developed for the system to learn and there are two types of techniques to learn, supervised and unsupervised.

**Supervised Learning:**

Supervised learning uses a given set of training data for machine learning which has a pair of input and output objects where output can be a continuous value and predict the input object's class label.

K-Nearest Neighbor algorithm is used for machine learning which is conventional non-parameter classifier. It will assign unlabeled data sample to the k nearest neighbor class where k is taken as integer. If there are 5 nearest neighbors for unlabeled data and it is split into 3 patterns for class C2 and 2 for class C1, then on majority basis it will enable class C2. It is represented in below diagram.

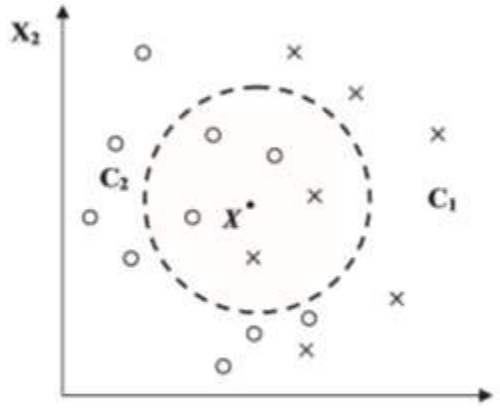



Fig. 1. A k-Nearest Neighbor rule for k = 5.

**Unsupervised Learning:**

This type of learning is different from supervised learning by means of prior output. In this type of learning it will have a set of input objects and the objects will be treated as a set of variables. The machine receives inputs $x_1, x_2 \ldots x_n$ which is not supervised target output. The machine will not get any feedback from the environment and hence it will be mysterious how machine will learn.

The most commonly used machine learning technique for unsupervised learning is k-means clustering algorithm since it is easy to classify the given dataset through number of clusters.

$$J = \sum_{j=1}^{k} \sum_{n \in S_j} |x_n - \mu_j|^2$$

**Comparison of related work:**

As there are various datasets available for intrusion detection, DARPA1998 and KDD-Cup99 are most commonly used and SVM and k-NN are popularly used in novel techniques. In current phase of technological improvement, there are studies to combine or integrate two different techniques to improve intrusion detection.

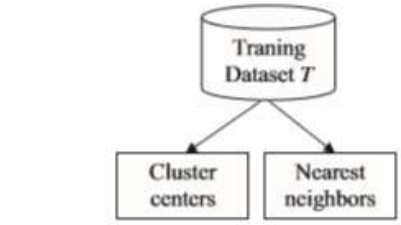
Every study will use different techniques with same dataset which led to unclear about which feature are more representative in two datasets.

We often examine evaluation measurements by means of false positive, false alarm, false negative, detection rate and true positive. The run time for computational process should be as short as possible for intrusion detection which is also very critical issue.

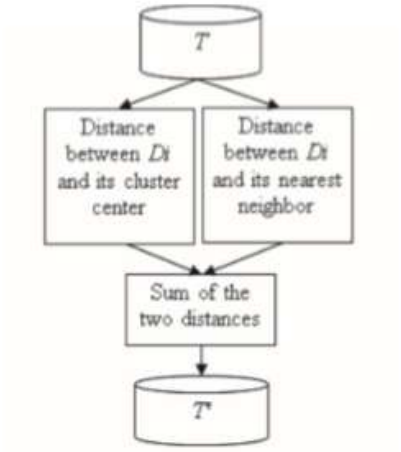
**CANN - Proposed system:**

CANN is based on two distances used between cluster center and specific data point and its nearest neighbor. The pictorial representation of CANN is shown in fig.2.

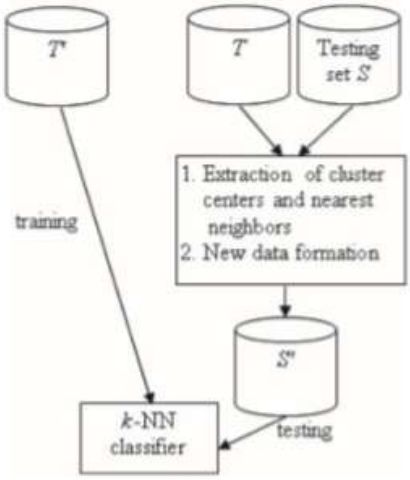
ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 4, April 2018




(a) Extraction of cluster centers and nearest neighbors




(b) New data formation




(c) Classifier training and testing

**Fig. 2.** The CANN process.

From the given training dataset, extract the cluster center which is nothing but the number of classes that needs to be classified. This dataset will already have a defined number of classes since intrusion detection is one classification problem. By calculating the distance between one specific data point and other data that are available in same cluster, cluster centers can be extracted using dataset and nearest neighbor in same cluster.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 3, Issue 4, April 2018

Next the distance is summed by using the measurements, distance of $D_i$ and its cluster center along with its nearest neighbor which will lead to the new distance-based feature value which is represented by T in the picture. These new distance-based feature is used for the cluster data formation that will be used for data integrity.

To test the new unknown data for intrusion detection, the test set S is joined with the original training set T. while this process is carried out, only test data S is considered, and hence new test result based on S is obtained. Hence these two-dataset results T and S are used to train and test the k-NN classifier for intrusion detection.

**Extraction of cluster centers and nearest neighbors:**

Clustering technique is used to extract the cluster centers and for this k-means clustering algorithm is used.

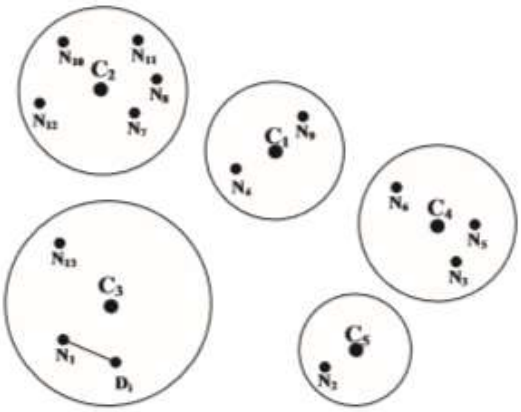



Fig. 3. An example of extracting cluster centers and nearest neighbors.

In above diagram we can see 12 data samples have been chosen ($N_1$ to $N_{12}$) is a five-class classification problem. The number of cluster is five (k=5). For the k-means clusters algorithm we have 5 set of clusters and each contain a cluster centers that is $C_1$, $C_2$, $C_3$, $C_4$, $C_5$.

Nearest neighbor can be identified by using $D_i$ in the above diagram as the nearest cluster for the k-NN algorithm is identified to be $N_1$ as it is nearest to the $D_i$

**New data formation**

once the nearest neighbor and cluster center for each data point is extracted, distance are calculated and summed. Distance from each data point to the cluster center is identified. If there are three cluster centers, then there will be three distance data point to the three-cluster center. Next type is distance from each data point to its nearest neighbor. The below diagram represents the data formation.

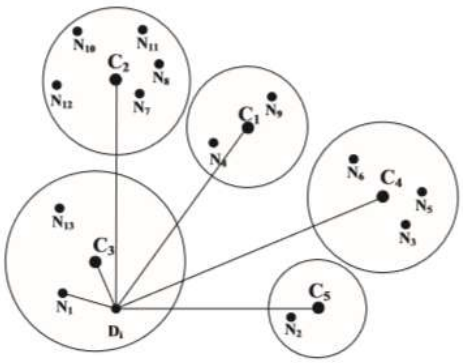



Fig. 4. An example of new data formation.

The two types of distance for the data point is obtained by below calculation,

$$D_i = \overline{D_iC_1} + \overline{D_iC_2} + \overline{D_iC_3} + \overline{D_iC_4} + \overline{D_iC_5}$$

The distance between two data points is calculated based on the Euclidean distance. On the given data A and B contain n-dimensional features, their Euclidean distance is based on

$$dis\ AB = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

$$Dis(D_i) = \sum_{j=1}^{5} dis(D_i, C_j) + \sum_{k=1} dis(D_i, N_k)$$

To construct classifier which is the final step of CANN, the training and testing dataset is divided with the new data set to train and test a specific classifier. In this study, we consider the k-NN classifier since it is easy to implement and widely used as a baseline classifier in many applications.

## IV. CONCLUSION

This paper represents a novel feature approach which combines nearest neighbor and cluster center for efficient intrusion detection CANN. This CANN will transform the original feature of representation of the dataset given into one-dimensional distance-based feature. Results of CANN thus shows it can perform better than existing k-NN and SVM classifier which is originally 6-dimensional dataset which provides high accuracy and low false alarm rate.

The main advantage of CANN is it requires less computational effort compared to that of k-NN or SVM classifiers, meaning CANN requires additional computation to extract the distance-based features, the training and testing time is highly reduced since the new dataset only contains one dimension.

## REFERENCES:

[1] M.S. Abadeh, J. Habibi, Z. Barzegar, M. Sergi, A parallel genetic local search algorithm for intrusion detection in computer networks, Eng. Appl. Artif. Intell. 20 (8) (2007) 1058–1069.
[2] Z.A. Baig, S.M. Sait, A. Shaheen, GMDH-based networks for intelligent intrusion detection, Eng. Appl. Artif. Intell. 26 (7) (2013) 1731–1740.
[3] Y. Chen, A. Abraham, B. Yang, Hybrid flexible neural-tree-based intrusion detection systems, Int. J. Intell. Syst. 22 (2007) 337–352.
[4] E. de la Hoz, E. de la Hoz, A. Ortiz, J. Ortega, A. Martinez-Alvarez, Feature selection by multi-objective optimisation: application to network anomaly detection by hierarchical self-organising maps, Knowl.-Based Syst. 71 (2014) 322–338.
[5] W. Feng, Q. Zhang, G. Hu, J.X. Huang, Mining network data for intrusion detection through combining SVMs with ant colony networks, Future Gener. Comput. Syst. 37 (2014) 127–140.
[6] A.S. Eesa, Z. Orman, A.M.A. Brifcani, A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems, Expert Syst. Appl. 42 (5) (2015) 2670–2679.
[7] P. Garcı ´a-Teodoro, J. Dı ´az-Verdejo, G. Macia ´-Ferna ´ndez, E. Va ´zquez, Anomalybased network intrusion detection: techniques, systems and challenges, Comput. Secur. 28 (2009) 18–28.
[8] G. Giacinto, R. Perdisci, M. Del Rio, F. Roli, Intrusion detection in computer networks by a modular ensemble of one-class classifiers, Inf. Fusion 9 (2008) 69–82.
[9] H. Guan, J. Zhou, M. Guo, A class-feature-centroid classifier for text categorization, in: Proceedings of the International Conference on World Wide Web, 2009, pp. 201–209.
[10] E.-H. Han, G. Karypis, Centroid-based document classification: analysis and experimentalresults,in:ProceedingsoftheEuropeanConferenceonPrinciples of Data Mining and Knowledge Discovery, 2000, pp. 424–431.
[11] J.V. Hansen, P.B. Lowry, R.D. Meservy, D.M. McDonald, Genetic programming for prevention of cyberterrorism through dynamic and evolving intrusion detection, Decis. Support Syst. 43 (2007) 1362–1374.