International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 1 | Issue – 5

An Interruption Discovery Structure Depend on Cluster Centres and Adjacent Neighbours

V.Ravi Kishore
Assistant Professor, CSE Department

Dr.V.Venkata Krishna
Professor, CSE Department

ABSTRACT

The aim of an interruption discovery structure (IDS) is to notice various types of hateful network transfer and computer usage, which cannot be detected by a straight firewall. Many IDS have been urban based on engine learning techniques. Specifically, advanced finding approaches created by combining or integrating multiple learning techniques have shown better finding act than general single learning techniques. The feature image way is an important model classifier that facilitates correct classifications, still, there have been very few correlated studies focusing how to extract more agent features for normal connections and effective detection of attacks. This paper proposes a novel feature representation approach, namely the cluster centre and nearest neighbour (CANN) approach. In this approach, two distances are measured and summed, the first one based on the distance between each data sample and its cluster centre, and the second distance is between the data and its nearest neighbour in the same cluster. Then, this new and one-dimensional distance based mark is used to represent each data sample for interruption detection by a k-Nearest Neighbour (k-NN) classifier. The experimental results based on the KDD-Cup 99 dataset show that the CANN classifier not only performs better than or similar to k-NN and support vector machines trained and tested by the original feature representation in terms of classification correctness, discovery rates, and false alarms. I also provides high computational competence for the time of classifier training and testing (i.e., detection).

Keywords: *Intrusion detection, Anomaly detection, Feature representation, Cluster Center, Nearest neighbour*

1. INTRODUCTION

Advancements in computing and network technology have made the activity of accessing the Internet an important part of our daily life. In addition, the amount of people connected to the Internet is increasing rapidly. However, the high popularity of world-wide connections has led to security problems.

Traditionally, some techniques, such as user authentication, data encryption, and firewalls, are used to protect computer security. Interruption detection systems (IDS), which use specific logical technique(s) to detect attacks, identify their sources, and alert network administrators, have recently been developed to monitor attempts to break security. In general, IDS are developed for sig-nature and/or anomaly detection. For signature detection, packets or audit logs are scanned to look for sequences of commands or events which are previously determined as indicative of an attack. On the other hand, for anomaly detection, IDS use behaviour

Patterns which could indicate malicious activities and analyzes past activities to recognize whether the observed behaviours are normal. As early IDS largely used signature detection to detect all the attacks captured in their signature databases, they suffer from high false alarm rates. Recent innovative approaches including behaviour-based modelling have been proposed to detect anomalies include data mining,

statistical analysis, and artificial intelligence techniques.

Much related work in the literature focuses on the task of anomaly detection based on various data mining and machine learning techniques. 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

However, there are two limitations to existing studies. First, although more advanced and sophisticated detection approaches and/or systems have been developed, very few have focused on feature representation for normal connections and attacks, which is an important issue in enhancing detection performance. There is a huge amount of related studies using either the KDD-Cup 99 or DARPA 1999 dataset for experiments, however there is not an exact answer to the question about which features of these datasets are more representative. Second, the time taken for training the systems and for the detection task to further validate their systems is not considered in many evaluation methods. Recent systems that combine or integrate multiple techniques require much greater computational effort. As a result, this can degrade the efficiency of 'on-line' detection.

Therefore, in this study, we propose a novel feature representation method for effective and efficient intrusion detection that is based on combining cluster centres and nearest neighbours, which we call CANN. Specifically, given a dataset, the k-means clustering algorithm is used to extract cluster centres of each pre-defined category. Then, the nearest neighbour of each data sample in the same cluster is identified. Next, the sum of the distance between a specific data sample and the cluster centres and the distance between this data sample and its nearest neighbour is calculated. This results in a new distance based feature that represents the data in the given dataset. Consequently, a new dataset containing only one dimension (i.e., distance = based feature representation) is used for k-Nearest Neighbour classification, which allows for effective and efficient intrusion detection.

The idea behind CANN is that the cluster centres or cancroids' for a given dataset offer discrimination capabilities for recognition both similar and dissimilar classes [9, 10, 5]. Therefore, the distances between a data sample and these identified cluster centres are likely to provide some further information for recognition. Similarly, the distance between a specific data sample and its nearest data sample in the same class also has some discriminatory power.

The rest of this paper is organized as follows. Section 2 reviews related literature including offering brief descriptions of supervised and unsupervised machine learning techniques. The techniques used in this paper are also described. Moreover, the techniques used, datasets and evaluation strategies considered in related work are compared. The proposed approach for intrusion detection is introduced in Section 3. Section 4 presents the experimental setup and results. Finally, some conclusions are provided in Section 5.

2. LITERATURE REVIEW

2.1. Machine learning

Machine learning requires a system capable of the autonomous acquisition and integration of knowledge. This capacity includes learning from experience, analytical observation, and so on, the result being a system that can continuously self-improve and thereby offers increased efficiency and effectiveness. The main goal of the study of machine learning is to design and develop algorithms and techniques that allow computers to learn. In general, there are two types of machine learning techniques, supervised and unsupervised which are described in greater detail below.

2.2. Supervised Learning

Supervised learning or classification is one common type of machine learning technique for creating a function from a given set of training data. The training data are composed of pairs of input objects and their corresponding outputs. The output of the function can be a continuous value, and can predict a class label of the input object. Particularly, the learning task is to compute a classifier that approximates the mapping between the input– output training examples, which can correctly label the training data with some level of accuracy.

The k-Nearest Neighbour (k-NN) algorithm is a conventional non-parametric classifier used in machine learning. The purpose of this algorithm is to assign an unlabelled data sample to the class of its k nearest neighbours (where k is an integer). Fig. 1

shows an example for a k-NN classifier where k = 5. Consider the 5 nearest neighbours around X for the unlabelled data to be classified. There are three 'similar' patterns from class C_2 and two from class C_1. Taking a majority vote enables the assignment of X to the C_2 class.

According to Jain et al., k-NN can be conveniently used as a benchmark for all the other classifiers since it is likely to provide a reasonable classification performance in most applications. Other well-known supervised learning techniques used in intrusion detection include support vector machines, artificial neural net-works, decision trees, and so on.

2.3. Unsupervised Learning

Unsupervised learning or clustering is a method of machine learning where a model is fit to observations. It differs from supervised learning in the absence of prior output. In unsupervised learning, a data set of input objects is gathered first. The input objects are typically treated as a set of random variables. A joint density model is then built for the data set.

The machine simply receives the inputs x_1, x_2, . . . , x_n, obtaining neither supervised target outputs, nor rewards from its environment. It may seem somewhat mysterious to imagine what the machine could possibly learn given that it does not get any feedback from its environment. However, it is possible to develop a formal framework for unsupervised learning based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc.

The k-means clustering algorithm is the simplest and most commonly used unsupervised machine learning technique being a simple and easy way to classify a given dataset through a certain number of clusters. The goal of the k-means algorithm is to find k points of a dataset, which can best represent this dataset in a certain number of groups. The point, k, is the cluster center or centroid of each cluster.

In the literature, it can be seen that some clustering techniques are combined with specific supervised learning techniques for intrusion detection. For example, Khan et al. combined self-organizing maps (SOM) and support vector machines, Xiang et al. combined Bayesian clustering and decision trees, and C-means clustering and artificial neural networks are combined in Zhang et al.

2.4. Comparison of Related Work

A number of related intrusion detection systems are compared and the results shown in Table 1. In particular, we compare the machine learning techniques used for developing the detection systems, datasets used for experiments, evaluation methods considered, baseline classifiers for comparisons, etc. in relevant studies.

3. CANN: the proposed approach

3.1. The CANN process

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 neighbour respectively. CANN is comprised of three steps as shown in Fig. 2.

Given a training dataset T, the first step is to use a clustering technique to extract cluster centers. The number of clusters is based on the number of classes to be classified. Since intrusion detection is one classification problem, the chosen dataset has already defined the number of classes to be classified. Therefore, for example, if the given dataset is a three-class problem, then the number of clusters is defined as three. Besides extracting

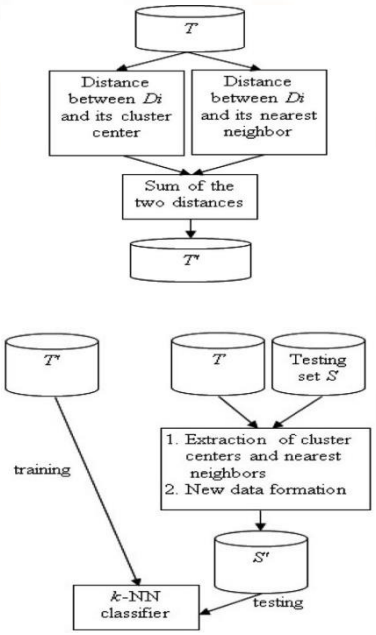

Fig. 1 the CANN Process

Cluster centers, each data point of the given dataset and its nearest neighbour in the same cluster is identified. This can be done by calculating the distances between one specific data point (Di) and all of the other data in the same cluster. Then, the shortest distance between two data examples representing Di and its nearest neighbour can be found.

The second step is to measure and sum the distance (dis1) between all data of the given dataset and the cluster centers and the distance (dis2) between each data point and its nearest neighbour in the same cluster. This leads to a new distance based feature value to represent each data point of the given dataset, which is T^0. That is, the original features (i.e., the number of dimensions is usually larger than one) are replaced by one new dimension feature.

To test the new unknown data for intrusion detection, the testing set S is combined with the original training set T. Then, the pro-cesses of extracting cluster centers and nearest neighbours (Fig. 2(a)) and new data formation are executed. During these processes, only the data samples in S are considered. As a result, the new distance based feature dataset S^0 is obtained.

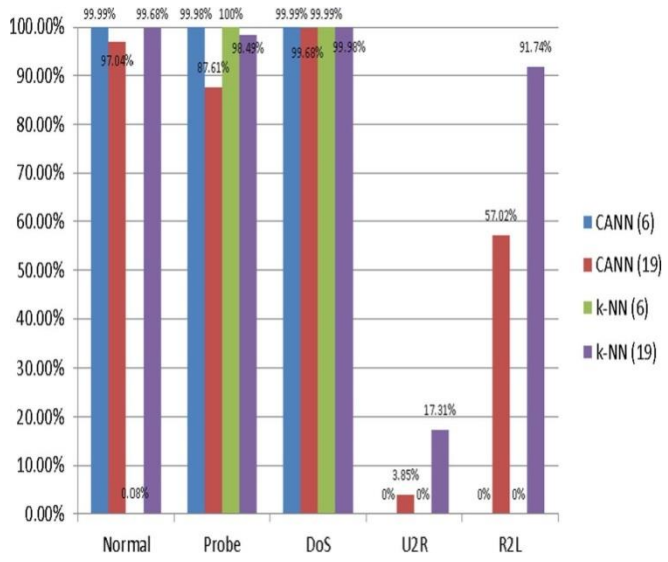

Fig. 2 The performances of CANN and k-NN over 5 classes

CONCLUSION

This paper presents a novel feature representation approach that combines cluster centers and nearest neighbors for effective and efficient intrusion detection, namely CANN. The CANN 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 CANN performs better than the k-NN and SVM classifiers over the original 6-dimension data-set, providing higher accuracy and detection rates and a lower false alarm rate. On the other hand, CANN performs similar to the k-NN and SVM classifiers over the original 19-dimension dataset. How-ever, the advantage of CANN is that it requires less computational effort than the k-NN or SVM classifiers trained and tested by the two original datasets. In other words, although CANN requires additional computation to extract the distance based features, the training and testing (i.e., detection) time is greatly reduced since the new dataset only contains one dimension.

As to the limitations of this research CANN cannot effectively detect U2L and R2L attacks, which means that this one-dimensional distance based feature representation is not able to well represent the pattern of these two types of attacks. This is an issue that future work can look into. One possibility is to consider the weight for the distances between the data to each of the cluster centers and its nearest neighbor. Alternatively, before performing CANN, outlier detection and removal can be employed in order to first filter out noisy or bad data from the given dataset. Finally, as CANN is applicable to the 5-class intrusion detection problem, other domain datasets including different numbers of dimensions and classes can be used to examine its effectiveness.

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, Anomaly-based 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 experimental results, in: Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery, 2000, pp. 424–431.