Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 124 (2017) 672–681

Procedia Computer Science

www.elsevier.com/locate/procedia

4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia

A Study on Intrusion Detection Using Centroid-Based Classification

Bambang Setiawan*, Supeno Djanali, Tohari Ahmad

Department of Informatics, Institut Teknologi Sepuluh Nopember, Sukolilo, Surabaya, 16011, Indonesia

Abstract

The ultimate goal of intrusion detection system (IDS) development is to accomplish the best possible accuracy for detection attacks. Various hybrid machine learning techniques were developed for IDS. The centroid-based classification method is a particular hybrid learning approach that highly efficient in the training and classification stages. This paper studies 60 associated papers in the period between 2010 and 2016 concentrating on developing IDS using hybrid classifiers, which 11 papers used centroid-based classification. Similar studies are compared by the algorithm used in hybrid machine learning, the dataset used, the establishment of the representative feature, the stages of pre-processing data, and evaluation methods considered. The accomplishments and limitations in developing IDSs using hybrid machine learning and centroid-based classification were presented and discussed. Several future research opportunities were provided that may encourage interested researchers to work in this area.

Keywords: Intrusion Detection System; Hybrid Classifiers; Centroid-Based Classification; Representative Feature

1. Introduction

Network security has become an important focus of computer security research. Computer security researchers intend to maintain the confidentiality, integrity, and availability of information by various security systems that shield equipment computers and networks from hackers who will want to intrude on the system. Traditional approaches to network protection used firewalls and authentication mechanisms in network servers. Those methods are used as the first line of protection against network security holes, the second line used IDS. IDSs identify intrusions based on data traffic network so that attacks can be detected and the area of damage can be restricted.

* Corresponding author. Tel.: +62-31-5999944; fax: +62-31-5999944.
E-mail address: setiawan@is.its.ac.id

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 4th Information Systems International Conference 2017
10.1016/j.procs.2017.12.204

The IDS system has many weaknesses, especially if they fail to conclude to recognize new attacks because signatures are unknown [1]. The potential weakness of this approach related to the false alarm rate. It can happen because normal system behavior that not seen (not stored in the database) previously, will be recognized as an anomaly and marked as intrusion [2]. Shon et al. [3] indicate that the IDS will also potentially be a cause of system failure when the IDS are disabled because it will provide an opportunity for the attackers to compromise the system, and then the attackers get to obtain a footing to infiltrate the system. For handling the problems, several anomaly detection systems were developed, among others, using machine learning techniques. This system uses normal behavior network traffic to identify attacks. In particular, anomaly detection is applied using supervised learning techniques, unsupervised learning techniques, and their hybrids. In this paper, we focus on the various anomaly detection using hybrid machine learning techniques and centroid-based classification. The centroid-based classification is a particular hybrid learning approach [4]. This model is highly efficient in the training and classification stages.

Many studies have focused the review on intrusion detection techniques [5, 6]. In [7], Hodge and Austin classified anomaly detection techniques in four categories: statistics, neural networks, machine learning, and hybrid approach. Several studies of data mining-based IDS are conducted by [9-11]. Moreover, in [12-14] conduct study on IDS using machine learning. Shon and Moon in [11] do a study of network anomaly detection using hybrid machine learning approach; focus on the Support Vector Machine (SVM) among various machine learning algorithms. In [12] Tsai et al. review the single classifier, the ensemble classifier, and hybrid classifier using machine learning. Mohamad Tahir et al. in [13] conduct the IDS using a hybrid classifier that combination of k-means clustering and SVM classification. To the best of our knowledge, there is no comprehensive review of intrusion detection using centroid-based classification techniques. Hence, the aim of this paper is to conduct a literature review covering 60 similar studies about IDS published from 2010 to 2016 by exploring what methods have been applied, what IDS dataset, pre-processing and evaluation methods have been used. Moreover, discuss what should be considered for future work if using the hybrid machine learning and centroid-based classification techniques on IDS.

This paper is organized as follows. Section 2 provides an overview of hybrid machine learning techniques and briefly describes some related techniques for intrusion detection. Section 3 compares and discusses similar work based on the algorithm used in hybrid classifier design, IDS datasets and data pre-processing used for experiments, the chosen evaluation methods, the type of centroid-based classification, and the representative feature process. Finally, Section 4 provides the conclusions for future research.

2. Hybrid Machine Learning

The final goal of IDS development is to accomplish the best possible accuracy for detection attack. In the use of machine learning, this goal can be achieved by using a hybrid approach. The hybrid machine learning improves system performance by combining several machine learning techniques.

Hybrid machine learning combines supervised and unsupervised learning techniques to detect attacks. For the objective of increasing the results of the clustering process, the hybrid model uses the supervised learning technique (classifier) as the first component and the second uses unsupervised technique (clustering). The classifier initially added to the training process first, and the result then be used as input for the clustering process. This strategy was used because clustering, which is an unsupervised learning, cannot accurately distinguish data such as a supervised learning method.

Hybrid IDS models can also be used to improve the performance of classifiers by using the unsupervised technique (clustering) as the first component and the second using the supervised technique (classifier) [15-17]. The first component will take the input data and generate intermediate results, called the new representative feature. Moreover, the second component will process that intermediate result to produce the final result [18]. This method is known as the centroid-based classification method, which is a particular hybrid learning approach. This model is highly efficient in the training and classification stages. The calculation process is based only on a much smaller number of centroids excluding the entire training data, so the number of calculations performed is much less. Finally, the processing time becomes shorter.

In this paper, beside discusses hybrid machine learning, we will discuss more in the research of IDS using centroid-based classification (CBC) method.