International Journal of Information Security and Privacy
Volume 15 • Issue 1 • January-March 2021

A Smart System of Malware Detection Based on Artificial Immune Network and Deep Belief Network

Dung Hoang Le, University of Information Technology, Vietnam National University, Vietnam
Nguyen Thanh Vu, Ho Chi Minh City University of Food Industry, Vietnam
Tuan Dinh Le, Long An University of Economics and Industry, Vietnam

ABSTRACT

This paper proposes a smart system of virus detection that can classify a file as benign or malware with high accuracy detection rate. The approach is based on the aspects of the artificial immune system, in which an artificial immune network is used as a pool to create and develop virus detectors that can detect unknown data. Besides, a deep learning model is also used as the main classifier because of its advantages in binary classification problems. This method can achieve a detection rate of 99.08% on average, with a very low false positive rate.

KEYWORDS

Artificial Immune Network, Artificial Immune System, Clonal Selection Algorithm, Deep Belief Network, Negative Selection Algorithm, Portable Executable

INTRODUCTION

For many years, virus recognition and elimination have become critical and important problems because of the complicated change in the development of malware systems, which causes antivirus programs to inefficiently detect and remove viruses with traditional algorithms. In the past, the most exact method for malware detection was signature analysis, which made antivirus systems with outdated virus signature databases become powerless in the face of new security threats.

In recent years, the artificial immune system (AIS) (Read, Andrews, & Timmis, 2012) has been developed as a prospective model because of its ability to adapt naturally to the environment where it is applied. The AIS is grounded on basic principles of the biological immune system and has powerful information processing capabilities, such as future extraction, pattern recognition, learning, adaptability, memory and distributive nature. All aforesaid features made AIS attractive for many computer security researchers. Computer security systems which are based on AIS principles allow to detect unknown malicious code, as certain research projects indicated (Al-Sheshtawi, & Hatem, 2010; Khang et al., 2015).

In addition, deep learning (DL) has become a promising and potential technology in the problem solving of large volumes of data, especially identification problems. Some DL architectures such as convolutional deep neural networks, deep belief networks (DBNs), recurrent neural networks, have been used successfully in the field of computer virus detection. In the past, although these models had significant performances, they still had deficiencies in unknown malware detection and almost

DOI: 10.4018/IJISP.2021010101

required very large amount of data, in order to perform better than other techniques. Thus, they have had to be retrained, due to the quickly increased number of viruses, in recent years.

This paper presents a hybrid approach to build a virus detection system that can detect unknown malware based on the idea of the human immune system, in order to recognize unknown viruses and increase accurate detection rate. By using algorithms of the AIS, instead of training the model directly with a dataset, the authors generate a detector set and use it for training the system. In this method, the authors extract the features from clean and malware files, then use the artificial immune network (AiNet) to generate malicious detectors to detect unknown elements. After that, they run a DL model to train the detectors and compare its efficiency with other classifier models.

In the following, the paper is organized in four sections: The first section describes related work; the second section details the authors' model for virus detection; the third section presents the authors' experimental results; the fourth and last section concludes the paper.

RELATED WORK

Wang, Wu, & Hsieh (2009) proposed a support-vector machine (SVM) model for detecting unseen malware. Using static analysis, these authors extracted portable executable (PE) header entries and trained the SVM classifier using selected features. Wang et al.'s classification model detected viruses and worms with considerable accuracy, but the detection accuracy was lower for trojans and backdoors.

Nguyen et al. (2014) integrated an artificial neural network (ANN) with a clonal selection algorithm (CSA) to create a new virus detection approach, which aimed to handle virus detection. In this approach, these authors used some ANNs as the detectors; also, they used the CSA to find the best ANN's structure and weights. The CSA is used to train a pool of immature detectors for an adaptation with the problem-space. However, the authors had not examined the coverage of the detector, so they obtained many irrelevant detectors and, thereby, a low detection rate.

Shah, Jani, Shetty and Bhowmick (2013) used Fisher score to select best features. By this way, they extracted the PE features and proceeded to use an ANN for classifying. Although their approach could identify unknown virus patterns, they used only one deployed ANN as learning model, which was not efficient in training cost nor in performance for large data.

Liao (2012) extracted attributes in a PE file including PE header, optional header and section header. Then, this researcher proceeded to select the top five properties with the most obvious difference between clean files and virus infections. This research achieved 99% accuracy and 0.2% false positive rate (FPR) for unknown malware. The limitation of this approach is that it could not detect all malware from the dataset. Furthermore, Liao used a decision table of five properties to detect the malware with some rules he observed in his samples without using any classifier. Liao did not consider other features of the PE file which could also be used to for detection system. It became ineffective with the current virus problem.

Tobiyama, Yamaguchi, Shimada, Ikuse, and Yagi (2016) used convolutional deep neural networks to make the classification of behavioral characteristics extracted from the recurrent neural networks model. The combination of these two models brought about quite high accuracy, about 96%. However, extracting feature images slowed the performance during the training, and the authors could not utilize the network on a large scale because of the small amount of data.

Saxe and Berlin (2015) introduced a deep neural network. They took the four 256-dimensional feature vectors and concatenated them into a single 1024-dimensional feature vector. The model achieved a usable detection rate at an extremely low false positive (95% detection rate at 0.1% FPR), but the performance decayed significantly in the time split validation because they relied on syntactic, rather than semantic, features for achieving a very low FPR.

Nguyen, Nguyen, Khang, and Le (2014) extracted files into binary strings of size L (L = 16, 32 or 64 bit), and two consecutive strings overlapped ½ L. They used an AiNet to improve detectors'