Future Generation Computer Systems 37 (2014) 127–140

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Mining network data for intrusion detection through combining SVMs with ant colony networks

Wenying Feng [a,b,*], Qinglei Zhang [c], Gongzhu Hu [d], Jimmy Xiangji Huang [e]

[a] Department of Computing & Information Systems, Trent University, Peterborough, ON, Canada, K9J 7B8
[b] Department of Mathematics, Trent University, Peterborough, ON, Canada, K9J 7B8
[c] Department of Computing and Software, McMaster University, Hamilton, ON, Canada, L8S 4L8
[d] Department of Computer Science, Central Michigan University, 115 Pearce Hall, Mount Pleasant, MI 48859, USA
[e] School of Information Technology, York University, Toronto, ON, Canada, M3J 1P3

HIGHLIGHTS

- A new machine-learning-based data-classification algorithm is introduced.
- The new algorithm combines a Support Vector Machine with an Ant Colony Network.
- The method is applied to network intrusion detection.
- Experiments show improvements of the system in both classification accuracy and run-time efficiency.

ARTICLE INFO

Article history:
Received 18 May 2012
Received in revised form
2 April 2013
Accepted 28 June 2013
Available online 11 July 2013

Keywords:
Data mining
Data classification
Intrusion detection system (IDS)
Machine learning
Support vector machine
Ant colony optimization

ABSTRACT

In this paper, we introduce a new machine-learning-based data classification algorithm that is applied to network intrusion detection. The basic task is to classify network activities (in the network log as connection records) as normal or abnormal while minimizing misclassification. Although different classification models have been developed for network intrusion detection, each of them has its strengths and weaknesses, including the most commonly applied Support Vector Machine (SVM) method and the Clustering based on Self-Organized Ant Colony Network (CSOACN). Our new approach combines the SVM method with CSOACNs to take the advantages of both while avoiding their weaknesses. Our algorithm is implemented and evaluated using a standard benchmark KDD99 data set. Experiments show that CSVAC (Combining Support Vectors with Ant Colony) outperforms SVM alone or CSOACN alone in terms of both classification rate and run-time efficiency.

1. Introduction and motivation

In today's information system management, large-scale data clustering and classification have become increasingly important and a challenging area. Although various tools and methods have been proposed, few are sufficient and efficient enough for real applications due to the exponential growing-in-size and high-dimensional data inputs.

As a particular application area, Intrusion Detection Systems (IDSs) are designed to defend computer systems from various cyber attacks and computer viruses. IDSs build effective classification models or patterns to distinguish normal behaviors from abnormal behaviors that are represented by network data. There are two primary assumptions in the research of intrusion detection: (1) user and program activities are observable by computer systems (e.g. via system auditing mechanisms), and (2) normal and intrusion activities must have distinct behaviors [1].

1.1. Data-mining-based approaches for IDSs

Researchers have proposed and implemented various models that define different measures of system behavior. IDSs have been developed based on these models. Many of the existing IDSs, however, cannot adequately deal with new types of attack or changing computing environments, and hence the installed IDSs always need to be updated. As it is an energy and time consuming job for security experts to update current IDSs frequently by manual encoding, using data mining approaches to network intrusion

* Corresponding author at: Department of Computing & Information Systems, Trent University, Peterborough, ON, Canada K9J 7B8. Tel.: +1 705 7481011x7249; fax: +1 705 7481066.
E-mail addresses: wfeng@trentu.ca (W. Feng), zhangq33@mcmaster.ca (Q. Zhang), hu1g@cmich.edu (G. Hu), jhuang@yorku.ca (J.X. Huang).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.06.027

detection provides an opportunity for IDSs to learn the behaviors of networks automatically by analyzing the data trails of their activities. Data mining has been widely used in many application areas. Two key advantages of using a data mining approach to IDSs are the following. (1) It can be used to automatically generate the detection models for IDSs, so that new attacks can be detected automatically as well. (2) It is general, so it can be used to build IDSs for a wide variety of computing environments.

The central theme of data mining approaches is to take a data-centric point of view and consider intrusion detection as a data analysis process [2]. This includes four essential steps.

(1) Capturing packets transferred on the network.
(2) Extracting an extensive set of features that can describe a network connection or a host session.
(3) Learning a model that can accurately describe the behavior of abnormal and normal activities by applying data mining techniques.
(4) Detecting the intrusions by using the learned models.

In our research, we assume that the steps (1) and (2) have been developed and are already available for the further training and testing phases. Approaches relevant to step (3) in data mining, in general, are by classification [3], link analysis, and sequence analysis [1]. In the rest of the paper, we will use SVM to denote either the concept or the algorithm when there is no confusion.

1.2. Motivation and contribution

Support Vector Machines (SVMs) have been widely accepted as a powerful data classification method (Section 4). On the other hand, the Self-Organized Ant Colony Network (CSOACN) has been shown to be efficient in data clustering (Section 5). Our work aims to develop an algorithm that combines the logic of both methods to produce a high-performance IDS.

One challenge of developing IDSs is to realize real-time detection in high-speed networks. There are two important issues for this problem. First, in order to reduce the cost of deploying a model, we must be able to minimize the amount of clean data that is used by the data mining process. The machine-learning-based SVM method [4,5] is a good choice for learning with little volume of data [6–8]. Second, when new information is added into a system, updating of the old model is required immediately to ensure that the system is properly protected. As retraining may take weeks, or even months, it is impractical to retrain the new model on all available data. Thus, a mechanism is needed to generate an adaptive model that can be updated by cooperation of the old model with the new information. We take advantage of the clustering-based Ant Colony Networks [9] in updating the models. Clustering in intrusion detection is used to resolve the multiple classification problem. However, the general method always involves expensive computation, especially if the set of training data is large. In order to save extensive computations, we modify the traditional CSOACN to control the clustering processes by clustering around certain objects (Section 6.2). This significantly reduced the retraining process and therefore the training time. Considering both of the issues for real-time detection problems, an active learning SVM and the modified CSOACN are chosen as the two basic components for our new classification algorithm.

The main contributions of this paper include the following.

(1) Modifications to the supervised learning SVM and the unsupervised learning CSOACN so they can be used together interactively and efficiently.
(2) A new algorithm, CSVAC, that combines the modified SVM and CSOACN to minimize the training data set while allowing new data points to be added to the training set dynamically.

The idea of combining supervised learning and unsupervised learning was applied previously [10]. However, the combined two algorithms of [10] are closely related by the neural dissimilarity. In our model, the supervised learning and the unsupervised learning algorithms have no relation, and they follow totally different logic. Their combination has the advantages of applying the logic of both sides and providing a more reliable solution to today's data-intensive computing processes.

2. Related work

Issues related to intrusion detection can be categorized into two broad areas: (1) network security and intrusion detection models, and (2) intrusion detection methods and algorithms based on artificial intelligence (mostly machine learning) techniques. In this section we shall briefly review some related work in the second area, and leave area (1) to the next section, when we discuss the background of IDSs.

Intrusion detection as a classification problem has been studied for decades using machine learning techniques, including traditional classification methods (*single classifier*) such as K-Nearest Neighbor (K-NN), Support Vector Machines (SVMs), Decision Trees (DTs), Bayesian, Self-Organized Maps (SOMs), Artificial Neural Networks (ANNs), Generic Algorithms (GAs), and Fuzzy Logic, as well as *hybrid classifiers* that combine multiple machine learning techniques to improve the performance of the classifier. A review of using these approaches was given in [11], which also included statistics of the use of these techniques reported in 55 research articles during the period 2000–2007. The review indicates that SVM and K-NN were the most commonly used techniques while the use of a hybrid increased significantly after 2004 and became mainstream. Another more recent review [12] provided a thorough survey of intrusion detection using computational intelligence. It presented the details of the classification algorithms and swarm intelligence methods to solve problems using the decentralized agents. Most recently, an IDS was introduced by integrating On Line Analytical Processing (OLAP) tools and data mining techniques [13]. It is shown that the association of the two fields produces a good solution to deal with defects of IDSs such as low detection accuracy and high false alarm rate.

As stated in [12], as one of the swarm intelligence approaches, Ant Colony Optimization (ACO), has been applied in many fields to solve optimization problems, but its application to the intrusion detection domain is limited. Several methods were reported using ACO for intrusion detection. For example, an ant classifier was proposed in [14] that used more than one colony of ants to find solutions in multiclass classification problem. Another ant-based clustering algorithm applied to detect intrusions in a network presented in [15] showed that the performance was comparable to some traditional classification methods like SVM, DT, and GA. In [16,17], the authors evaluated the basic ant-based clustering algorithms and proposed several improvement strategies to overcome the limitations of these clustering algorithms that would not perform well on clustering large and high-dimensional network data. The work presented in [18] also used ACO for intrusion detection in a distributed network. The basic ingredient of their ACO algorithm was a heuristic for probabilistically constructing solutions. All these ACO-based intrusion detection approaches are single classifiers as categorized by [11].

Hybrid intrusion detection approaches involving SVM have been studied in the past, such as the one reported in [4] that uses the Dynamically Growing Self-Organizing Tree (DGSOT) algorithm for clustering to help in finding the most qualified points to train the SVM classifier. It starts with an initial training set and expands the set gradually so that the training time for the SVM classifier is significantly reduced. Another hybrid intrusion detection approach was recently reported in [19] that combines hierarchical clustering and SVM. The purpose of using the hierarchical clustering algorithm is to provide the SVM classifier with fewer but higher quality training data that may reduce the training time and improve the performance of the classifier.