Engineering Applications of Artificial Intelligence 51 (2016) 171–181

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A population-based incremental learning approach with artificial immune system for network intrusion detection

Meng-Hui Chen [a,b], Pei-Chann Chang [a,b,*], Jheng-Long Wu [b]

[a] School of Software, Nanchang University, Nanchang 330031, China
[b] Innovation Center for Big Data & Digital Convergence and Department of Information Management, Yuan Ze University, Taoyuan 32026, Taiwan

ARTICLE INFO

Available online 20 January 2016

Keywords:
Artificial immune system
Population-based incremental learning
Evolutionary computation
Classification problems
Electronic commerce
Collaborative filtering

ABSTRACT

The focus of this research is to develop a classifier using an artificial immune system (AIS) combined with population-based incremental learning (PBIL) and collaborative filtering (CF) for network intrusion detection. AIS is a powerful tool in terms of extirpating antigens inspired by the principles and processes of the natural immune system. PBIL uses past experiences to evolve into new species through learning and adopting the idea of CF for classification. The novelty of this research is in its combining of the three above mentioned approaches to develop a new classifier which can be applied to detect network intrusion, with incremental learning capability, by adapting the weight of key features. In addition, four mechanisms: creating a new antibody using PBIL, dynamic adjustment of feature weighting using clonal expansion, antibody hierarchy adjustment using mean affinity, as well as usage rates, are proposed to intensify AIS performance. As shown by the comparison carried out with other artificial intelligence and evolutionary computation approaches in network anomaly detection problems, our PBIL-AIS^{CF} classifier can achieve high accuracy for the benchmark problem.

1. Introduction

With the advent and explosive growth of the global mobile computing and electronic commerce environments, network intrusion and anomaly detection in wide area data networks and electronic commerce infrastructures are gaining great attention from academic researchers and industrial practitioners. Recently, the number and impact of attacks has been increasing, as evidenced by recent network attacks against several prominent Web portals as well as many well-known companies. In addition, there are many types of network attacks which can damage an organization. Intrusions, such as advanced persistent threats (APTs), are designed to penetrate networks and surreptitiously steal intellectual property; distributed denial-of-service (DDoS) and flooding attacks can blind servers and shut down web sites. In addition, in the network economy, the problems caused by sustained network attacks are very difficult to handle. People worry about whether a major network disruption could cause confusion at local, national and even global levels. To detect network attacks effectively and automatically, software should be developed such that statistical signatures of network anomalies can be recognized algorithmically. This research focuses on building a classifier for network intrusion and anomaly detection in electronic commerce environments.

Classification models based on statistical pattern recognition approaches have attracted a wide range of interest in response to the growing demand for reliable and intelligent data mining systems that can be used to detect network intrusion attacks. Today, sophisticated classification is required in various domains. The goal of classification is to accurately predict the target class for each case; the determination of correct classes would lead to efficient service provision. Traditional classification systems use statistical approaches based on frequency and attack probability determination. Detection approaches such as fuzzy control, artificial neural network, decision tree, SVM, and so forth, also have good performance in numerous fault detection problems based on classification (Razavi-Far et al., 2009; Baccarini et al., 2011; Przystałka and Moczulski, 2015).

In recent years, another computationally intelligent approach, Evolutionary Computing (EC) has gained wide interest in the fields of computer science and artificial intelligence. EC approaches, such as artificial immune system (AIS), genetic algorithm (GA), evolutionary strategy (ES), genetic programming (GP), ant swarm

* Correspondence to: Department of Information Management, Yuan Ze University, 135 Yuan Tung Road, Chungli 32003, Taiwan. Tel.: +886 3 4638800x2305; fax: +886 3 4352077.
E-mail address: iepchang@saturn.yzu.edu.tw (P.-C. Chang).

http://dx.doi.org/10.1016/j.engappai.2016.01.020
0952-1976/

optimization (ASO) and particle swarm optimization (PSO) mimic the natural evolution mechanism to solve complex problems (Zhao, 2007; Castellani and Rowlands, 2008; Nemati et al., 2009; Agarwal et al., 2012). Hunt et al. (1998) proposed that AIS as a special approach, among these EC approaches, has a powerful evolution mechanism in accordance with the natural immune response. AIS exploits immune systems' characteristics such as feature extraction, pattern recognition and learning and memory capabilities to intelligently modify itself with experience and continuously evolve to achieve higher accuracy. Many researchers have used AIS for parameter combination optimization (Gao, 2010; Aydin et al., 2011). Dasgupta et al. (2011) investigated the AIS approach, which effectively identifies antibodies, creates new antibodies and then incorporates Clonal Selection and Negative Selection Mechanisms. Based on their usage and mean affinity, a qualitative classification of antibodies is also maintained during the training. AIS has also been employed to solve classification problems, such as the study by Watkins et al. (2004); they developed immune inspired supervised learning algorithms and artificial immune recognition systems (AIRS). De Castro and Von Zuben (2002) proposed the computational implementation of the clonal selection principle that explicitly takes into account the affinity maturation of the immune response. Carter (2000) proposed a supervised learning system (Immunos-81) using software abstraction of T cells, B cells, antibodies and their interactions in which artificial T cells control the creation of B-cell populations (clones). However, the performances of the above AIS-based approaches are not competitive with other evolutionary or SVM-based approaches, as shown in Baccarini et al. (2011). As mentioned by Dasgupta et al. (2011), the AIS-based approach can be further improved by embedding more sophisticated mechanisms in controlling the affinity between the antibody and the antigen to improve their classification accuracy.

In this research, we incorporate population-based incremental learning (PBIL) into AIS classification to enhance the performance. Since classification is one of the most important mining tasks, we focus on developing various AIS algorithms, such as a clonal selection algorithm (CLONALG) and an antibody hierarchy adjusting mechanism. Since the existing antibodies may not be sufficient to efficiently extirpate all the antigens, the creation of new antibodies in response to the dynamic requirements becomes necessary. This can be achieved through incremental learning procedures; therefore we use PBIL to evolve new antibodies with higher affinities than the older ones, which by themselves were not capable of correctly identifying the class. The primary objective of this research is to develop a classifier by combining PBIL with AIS. In addition, the antibodies related to the target instance are clustered together using collaborative filtering for classifying the class of the target intrusions.

The rest of the paper is organized as follows. Section 2 presents a review of the literature and defines the classification problem in general. Based on the two specific problems we aim to solve in this paper, we outline a few traditional classification approaches and introduce evolutionary computation methods. Section 3 describes the methodology detailing the core algorithms used in our method. Section 4 presents the experimental results and compares our method with other classification approaches. We then close with conclusions and directions for future research.

2. Literature review

2.1. Classification problems

Classification problems require assigning items in a collection to target categories or classes. The goal is to accurately predict the target class of each case in the data set. The assignment is based on quantitative information derived from certain characteristics, such as antibody or antigen features inherent in the items as antibodies and antigens. During the classification process, a group of marked classes is provided and a training set is used to learn the definitions of these classes. Classification rules are determined, and then these rules are treated as benchmarks to identify the most likely label and the class of a new pattern (Woolley and Milanovic, 2011). Classification problems stem from various real-world situations such as those found in biological fields (Alves et al., 2010), financial services (Twala, 2010), remote sensing (Zhong et al., 2007) and inventory classification (Thomas, 2000).

Whether used to detect legitimate uses against attacks in today's ubiquitous use of e-commerce, or the need to accurately distinguish cancer cells with the rising demands for sophisticated health, or for industrial activities such as expensive petroleum drilling operations, identification of impeccable classifiers is needed everywhere. An evaluation of classifiers thus becomes very important. Criteria such as classification precision or accuracy, the scalability of learning and classification on large data sets, the robustness to noise and the ability for incremental learning are often used to evaluate classification techniques. Scalability and incremental learning ability have become increasingly important with the collection of large amounts of data resulting from modern computing and information technologies. Since patterns embedded in large data sets generated through industrial processes may be dynamic, a classification technique should have incremental learning ability to update existing patterns with the collection of new data, and also be scalable to process data in large volumes.

An intrusion detection system (IDS) is a monitoring or a protection mechanism against various malicious activities or policy violations. With the proliferation of computers, networks and the Internet, security has become a primary concern. In general, there are two main types of IDS: network intrusion detection system (NIDS) and host-based intrusion detection system (HIDS). The intrusion detection approach usually uses statistical analysis and pattern recognition, and is capable of detecting anomaly intrusions without any prior knowledge; therefore, the model is able to generalize and extract intrusion rules during training. IDSs can reliably identify intrusion attacks in correspondence to known signatures of discovered vulnerabilities. Abadeh et al. (2011) referred to the use of genetic fuzzy systems (GFSs) in hybrid models to solve intrusion attacks problems. They presented three kinds of genetic fuzzy systems and an iterative rule learning approach to deal with intrusion detection. Altwaijry and Algarny (2012) presented a Bayesian intrusion detection system based on the Bayesian probability theory. Horng et al. (2011) presented a hierarchical clustering and support vector machines hybrid model to build an IDS. Afzali and Azmi (2014) presented a multi-agent AIS-based distributed intrusion detection system; the characteristics of MAIS-IDS are cloning, mutation, migration, collaboration and randomness. All of these researches solved the KDD Cup 1999 dataset, which is very popular in numerous studies. In this paper, we also use the KDD Cup 1999 dataset to evaluate our classification model and hence, to compare results.

Other popular classification problems also include credit card approval. With the recent growth of the credit industry, a need for an actively managed credit scoring model has emerged. A credit scoring technique is a set of decision models that assist lenders in granting consumer credit (Thomas, 2000). The models help in making decisions on whether to grant credit to new applicants based on customer characteristics such as age, income and marital status. In recent years, it has been extensively used for credit admission. The basic principle of credit scoring is to assess those who apply for fresh credit by predicting through the analyses of repayment performance on the part of previous consumers.