Journal of Advanced Research (2015) **6**, 609–619

Cairo University

Journal of Advanced Research

ORIGINAL ARTICLE

A hybrid approach for efficient anomaly detection using metaheuristic methods

Tamer F. Ghanem [a,*], Wail S. Elkilani [b], Hatem M. Abdul-kader [c]

[a] *Department of Information Technology, Faculty of Computers and Information, Menofiya University, Shebin El Kom, Menofiya, Egypt*
[b] *Department of Computer Systems, Faculty of Computers and Information, Ain Shams University, Cairo, Egypt*
[c] *Department of Information Systems, Faculty of Computers and Information, Menofiya University, Shebin El Kom, Menofiya, Egypt*

ARTICLE INFO

Article history:
Received 20 October 2013
Received in revised form 26 February 2014
Accepted 27 February 2014
Available online 5 March 2014

Keywords:
Intrusion detection
Anomaly detection
Negative selection algorithm
Multi-start methods
Genetic algorithms

ABSTRACT

Network intrusion detection based on anomaly detection techniques has a significant role in protecting networks and systems against harmful activities. Different metaheuristic techniques have been used for anomaly detector generation. Yet, reported literature has not studied the use of the multi-start metaheuristic method for detector generation. This paper proposes a hybrid approach for anomaly detection in large scale datasets using detectors generated based on multi-start metaheuristic method and genetic algorithms. The proposed approach has taken some inspiration of negative selection-based detector generation. The evaluation of this approach is performed using NSL-KDD dataset which is a modified version of the widely used KDD CUP 99 dataset. The results show its effectiveness in generating a suitable number of detectors with an accuracy of 96.1% compared to other competitors of machine learning algorithms.

Introduction

Over the past decades, Internet and computer systems have raised numerous security issues due to the explosive use of networks. Any malicious intrusion or attack on the network may give rise to serious disasters. So, intrusion detection systems (IDSs) are must to decrease the serious influence of these attacks [1].

IDSs are classified as either signature-based or anomaly-based. Signature-based (misuse-based) schemes search for defined patterns, or signatures. So, its use is preferable in known attacks but it is incapable of detecting new ones even if they are built as minimum variants of already known attacks. On the other hand, anomaly-based detectors try to learn system's normal behavior and generate an alarm whenever a deviation from it occurs using a predefined threshold. Anomaly detection can be represented as two-class classifier which classifies each sample to normal or abnormal [2]. It is capable of detecting previously unseen intrusion events but with higher false

[*] Corresponding author. Tel.: +20 1004867003.
E-mail address: tamer.ghanem@ci.menofia.edu.eg (T.F. Ghanem).

Peer review under responsibility of Cairo University.

Production and hosting by Elsevier

2090-1232 © 2014 Production and hosting by Elsevier B.V. on behalf of Cairo University.
http://dx.doi.org/10.1016/j.jare.2014.02.009

positive rates (FPR, events incorrectly classified as attacks) compared to signature-based systems [3].

Metaheuristics are nature inspired algorithms based on some principles from physics, biology or ethology. Metaheuristics are categorized into two main categories, single-solution-based and population-based metaheuristics [4]. Population-based metaheuristics are more appropriate in generating anomaly detectors than single-solution-based metaheuristics because of the need to provide a set of solutions rather than a single solution. Evolutionary Computation (EC) and Swarm Intelligence (SI) are known groups of population-based algorithms. EC algorithms are inspired by Darwin's evolutionary theory, where a population of individuals is modified through recombination and mutation operators. Genetic algorithms, evolutionary programming, genetic programming, scatter search and path relinking, coevolutionary algorithms and multi-start framework [5] are examples of EC algorithms. On the other hand, SI produces computational intelligence inspired from social interaction between swarm individuals rather than purely individual abilities. Particle swarm Optimization and Artificial Immune Systems are known examples of SI algorithms.

Genetic algorithms (GAs) are widely used as searching algorithm to generate anomaly detectors. It is an artificial intelligence technique that was inspired by the biological evolution, natural selection, and genetic recombination for generating useful solutions for problem optimization [6]. GAs use data as chromosomes that evolve through the followings: selection (usually random selection), cross-over (recombination to produce new chromosomes), and mutation operators. Finally, a fitness function is applied to select the best (highly-fitted) individuals. The process is repeated for a number of generations until reaching the individual (or group of individuals) that closely meet the desired condition. GAs are still being used up untill the current time to generate anomaly detectors using a fitness function which is based on the number of elements in the training set that is covered by the detector and also the detector volume [7,8].

Negative selection algorithm (NSA) is one of the artificial immune system (AIS) algorithms which inspired by T-cell evolution and self-tolerance in human immune system [9]. The principle is achieved by building a model of non-normal (non-self) data by generating patterns (non-self-detectors) that do not match an existing normal (self) patterns, then using this model to match non-normal patterns to detect anomalies. Despite this, self-models (self-detectors) could be built from self-data to detect the deviation from normal behavior [10]. Different variations of NSA have been used to for anomaly detection [11]. Although these newly developed NSA variants, the essential characteristics of the original negative selection algorithm [9] still remain, including negative representation of information, distributed generation of the detector set which is used by matching rules to perform anomaly detection based on distance threshold or similarity measure [12].

Generating anomaly detectors requires a high-level solution methods (metaheuristic methods) that provide strategies to escape from local optima and perform a robust search of a solution space. Multi-start procedures, as one of these methods, were originally considered as a way to exploit a local or neighborhood search procedure (local solver), by simply applying it from multiple random initial solutions. Some type of diversification is needed for searching methods which are based on local optimization to explore all solution space, otherwise, searching for global optima will be limited to a small area, making it impossible to find a global optimum. Multi-start methods are designed to include a powerful form of diversification [13].

Different data representation forms and detector shapes are used in anomaly detector generation. Input data are represented by either binary or real-valued [14]. Binary representation [15] is easy to use in finite problem space but it is hardly applicable to problems of real valued space [11]. As an alternative, real-valued representation [16] provides more expressiveness and scalability [17]. NSA detectors are formed with different geometric shapes such as hyper-rectangles, hyper-spheres, hyper-ellipsoids or multiple hyper-shapes [14]. The size and the shape of detectors are selected according to the space to be covered.

In this paper, a hybrid approach for anomaly detection is proposed. Anomaly detectors are generated using self- and non-self-training data to obtain self-detectors. The main idea is to enhance the detector generation process in an attempt to get a suitable number of detectors with high anomaly detection accuracy for large scale datasets (e.g., intrusion detection datasets). Clustering is used for effectively reducing large training datasets as well as a way for selecting good initial start points for detector generation based on multi-start metaheuristic methods and genetic algorithms. Finally, detector reduction stage is invoked so as to minimize the number of generated detectors.

The main contribution of this work is to prove the effectiveness of using multi-start metaheuristics methods in anomaly detector generation benefiting from its powerful diversification. Also, addressing issues arises in the context of detector generation for large scale datasets. These issues are related to the size of the reduced training dataset, its number of clusters, the number of initial start points and the detector radius limit. Moreover, their effect on different performance metrics is evaluated. Observations prove that performance improvement occurs compared to other machine learning algorithms.

The rest of this paper is organized as follows: Section 2 presents some literature review on anomaly detection using negative selection algorithm. Section 3 briefly describes the principal theory of the used techniques. Section 4 discusses the proposed approach. Experimental results along with a comparison with six machine learning algorithms are presented in Section 5 followed by some conclusions in Section 6.

Related work

Anomaly detection approaches can be classified into several categories. Statistics-based approaches are one of these categories that identify intrusions by means of predefined threshold, mean and standard deviation, and probabilities [18,19]. Rule-based approaches are another category which use If–Then or If–Then–Else rules to construct the detection model of known intrusions [20,21]. In addition to these categories, state-based approaches exploit finite state machine derived from network behaviors to identify attacks [22,23]. The last category is heuristic-based approaches [24–26], which are inspired by biological concepts as mentioned in the previous section [1].