Expert Systems with Applications 42 (2015) 193–202

Contents lists available at ScienceDirect

Expert Systems with Applications

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

On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on Intrusion Detection Systems

Salma Elhag [a], Alberto Fernández [b,*], Abdullah Bawakid [c], Saleh Alshomrani [c], Francisco Herrera [c,d]

[a] *Department of Information Systems, King Abdulaziz University (KAU), Jeddah, Saudi Arabia*
[b] *Department of Computer Science, University of Jaén, Jaén, Spain*
[c] *Faculty of Computing and Information Technology - North Jeddah, King Abdulaziz University (KAU), Jeddah, Saudi Arabia*
[d] *Department of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, Granada, Spain*

ARTICLE INFO

Article history:
Available online 11 August 2014

Keywords:
Intrusion Detection Systems
Genetic Fuzzy Systems
Pairwise learning
One-vs-One
Misuse detection

ABSTRACT

Security policies of information systems and networks are designed for maintaining the integrity of both the confidentiality and availability of the data for their trusted users. However, a number of malicious users analyze the vulnerabilities of these systems in order to gain unauthorized access or to compromise the quality of service. For this reason, Intrusion Detection Systems have been designed in order to monitor the system and trigger alerts whenever they found a suspicious event.

Optimal Intrusion Detection Systems are those that achieve a high attack detection rate together with a small number of false alarms. However, cyber attacks present many different characteristics which make them hard to be properly identified by simple statistical methods. According to this fact, Data Mining techniques, and especially those based in Computational Intelligence, have been used for implementing robust and accuracy Intrusion Detection Systems.

In this paper, we consider the use of Genetic Fuzzy Systems within a pairwise learning framework for the development of such a system. The advantages of using this approach are twofold: first, the use of fuzzy sets, and especially linguistic labels, enables a smoother borderline between the concepts, and allows a higher interpretability of the rule set. Second, the divide-and-conquer learning scheme, in which we contrast all possible pair of classes with aims, improves the precision for the rare attack events, as it obtains a better separability between a "normal activity" and the different attack types.

The goodness of our methodology is supported by means of a complete experimental study, in which we contrast the quality of our results versus the state-of-the-art of Genetic Fuzzy Systems for intrusion detection and the C4.5 decision tree.

1. Introduction

In an era of communications, a lot of effort has been put on filtering out known malware, exploits and vulnerabilities within a network, which could compromise the confidentiality, integrity or availability of the system. Therefore, intrusion detection is an essential part of a complete security policy in information systems. Since a wide number of potential intrusions occur every day, Intrusion Detection Systems (IDS) have been research and developed for addressing these cyber-attack events (Axelsson, 1998). In conjunction with security audit mechanisms, which dynamically monitor logs and network traffic for gathering information of the system use, IDS have the function of analyzing this information and then applying detection algorithms to determine whether these events are symptomatic of an attack or constitute a legitimate use of the system (Denning, 1987).

When referring to IDS, two main categories are clearly emphasized (Debar, Dacier, & Wespi, 1999):

1. **Misuse detection**, which are based on a signature database of already known attacks (Lee & Stolfo, 2000).
2. **Anomaly detection** adopts a complementary procedure: they first define a profile for "normal behavior", and then attacks are detected as deviations from this normal profile (Patcha & Park, 2007).

The former types of IDS are very efficient, but they are limited to the information from which they were trained, i.e. new types of attack might be not identified. On the contrary case, anomaly detection could incur into more false positives, and they strongly depends on the continuity of the user for its "normal activity".

* Corresponding author. Tel.: +34 953 213016; fax: +34 953 212472.
E-mail addresses: salma53ster@gmail.com (S. Elhag), alberto.fernandez@ujaen.es (A. Fernández), abawakid@kau.edu.sa (A. Bawakid), sshomrani@kau.edu.sa (S. Alshomrani), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.eswa.2014.08.002
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

Throughout the years, a wide number of different approaches in the field of Data Mining have been proposed for the area of intrusion detection (Lee, Stolfo, & Mok, 2000). Among them, those based on Computational Intelligence techniques have achieved a high success according to their good properties to detect both known and unseen intrusion attacks and to recognize normal network traffic (Wu & Banzhaf, 2010; Guo et al., 2014).

Our aim in this paper is to develop a misuse detection system to automatically extract optimal classification rules from training data under two main premises. On the one hand, the learnt rule set must be capable of correctly identifying all types of attacks, including rare attack categories, which is a major challenge in the IDS research domain (Khor, Ting, & Phon-Amnuaisuk, 2012). On the other hand, the final model should be linguistically interpretable for human comprehension (Gacto, Alcalá, & Herrera, 2011).

For achieving these goals, we propose the use of linguistic Fuzzy Rule Based Classification Systems (FRBCSs) (Ishibuchi, Nakashima, & Nii, 2004) as baseline classifiers for the development of our proposal. Additionally, in order to enhance in a higher degree the recognition of the minority classes within the IDS, we consider the use of the fuzzy system in synergy with decomposition techniques (Lorena, Carvalho, & Gama, 2008).

This classification scheme is based on a "divide-and-conquer" strategy, in which the original multi-class problem is divided into binary subproblems, which are independently learned by different base classifiers whose outputs are then combined to classify an instance. Proceeding this way, the borderline areas among the classes are simplified and individual concepts can be better identified (Galar, Fernández, Barrenechea, Bustince, & Herrera, 2011).

The choice of FRBCSs is justified by two main reasons: first, the intrusion detection problem involves many numeric attributes, and models which are directly built on numeric data might cause high detection errors. Hence, small deviations in an intrusion might not be detected and small changes in the normal user profile will cause false alarms. Second, security itself includes fuzziness, as the boundary between the normal and abnormal behavior cannot be well defined.

Specifically, as fuzzy learning classifier we have considered the use of a robust FRBCS, i.e. the Fuzzy Association Rule-based Classification for High-Dimensional problems (FARC-HD) (Alcalá-Fdez, Alcalá, & Herrera, 2011). The inner procedure of this algorithm comprises an optimization stage carried out by means of Evolutionary Algorithms (Eiben & Smith, 2003). This type of hybridization is known as a Genetic Fuzzy System (GFS) (Cordón, Gomide, Herrera, Hoffmann, & Magdalena, 2004; Alcalá, Nojima, Ishibuchi, & Francisco, 2012). One of the main reasons for the success of this type of techniques is their ability to exploit the information accumulated about and initially unknown search space in order to bias subsequent searches into useful subspaces, i.e. their robustness (Herrera, 2008). However, to the best of our knowledge only few works on the topic have addressed the problem of IDS with this type of approach (Gomez & Dasgupta, 2001; Özyer, Alhajj, & Barker, 2007; Tsang, Kwong, & Wang, 2007; Abadeh, Mohamadi, & Habibi, 2011).

As pointed out previously, in this paper we will make use of the One-vs-One (OVO) methodology in which the binary subproblems are obtained by confronting all possible pair of classes (Hastie & Tibshirani, 1998). The usage of pairwise learning to deal with real-world applications is frequent, being a simple yet effective way of overcoming multi-class problems. Moreover, empirical results in those papers have shown that the usage of OVO can enhance the results of the direct application of the baseline classifiers with inherent multi-class support (Fürnkranz, 2002; Galar et al., 2011; Sáez, Galar, Luengo, & Herrera, 2014).

The validity of our approach will be tested using the standard KDDCUP'99 dataset (Lee & Stolfo, 2000). This way, the experimental results will be directly comparable with most of the Computational Intelligence approaches for intrusion detection. Specifically, for the evaluation of the goodness of our IDS proposal, we will contrast the experimental results versus the standard FARC-HD algorithm and several GFS approaches that have been developed for misuse detection. In particular, we have selected a multi-objective genetic fuzzy intrusion detection system (MOGFIDS) (Tsang et al., 2007), three different GFS schemes proposed by in Abadeh et al. (2011), and a GFS for boosting fuzzy association rules (Özyer et al., 2007). Finally, we will complement our comparison with the classical C4.5 decision tree (Quinlan, 1993).

In short, the main contributions of this work are enumerated below:

1. We consider the use of a GFS for the intrusion detection problem. This kind of soft computing technique provides two main advantages: (1) obtaining a better separability of the different types of alarms by means of the achievement of smoother borderline for the rules of the final system; (2) a higher interpretability of the obtained rule set for the better understanding of the working procedure of the system.
2. A pairwise learning approach is applied for addressing the classification of the multiple classes, i.e. normal behaviour and all intrusion alarms. By following a divide and conquer scheme, we are able to learn a better suited discrimination function for each pair of classes, thus improving the overall classification.
3. To the best of our knowledge, this is the first research paper that combines GFS and OVO for the IDS problem. Furthermore, the baseline classifier used, i.e. FARC-HD, excel from the algorithms of the state-of-the-art as it is well-suited for high dimensional problems.
4. Finally, the goodness of this new methodology is shown by means of its high performance when contrasted versus several GFS algorithms developed for IDS, and with the C4.5 decision tree. We must stress the good behaviour of our approach especially for the minority classes.

In order to carry out the study, this manuscript is organized as follows. First, Section 2 introduces the preliminary concepts for this paper, i.e. the context of IDS, some basic notions on FRBCSs and the related work. Next, Section 3 introduces our proposal for the development of a combined approach between GFS and pairwise learning for the improvement on misuse detection. Then, the experimental framework including the features of the KDDCUP'99 dataset, metrics of performance, algorithms for comparison and their parameters, are presented in Section 4. The analysis of the results is shown throughout Section 5. Finally, Section 6 summarizes and concludes the work.

2. Preliminaries: Intrusion Detection Systems and fuzzy rule based classification systems

Prior to the description of our proposal, we must introduce some preliminary concepts which will help to understand the context of this work and the features of the solution which is to be develop. According to this, we will first present a brief review on IDS (Section 2.1), and then we will recall some basic concepts on FRBCSs (Section 2.2). Finally, Section 2.3 merges the former topics, and study those works on IDS related to fuzzy systems in general and GFS in particular.

2.1. Intrusion Detection Systems

Any information system should accomplish three main principles for guarantee a correct access to the data, namely