Expert Systems With Applications 54 (2016) 29–47

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Unsupervised learning clustering and self-organized agents applied to help network management

Luiz Fernando Carvalho[a,*], Sylvio Barbon Jr.[a], Leonardo de Souza Mendes[b], Mario Lemes Proença Jr.[a]

[a] Computer Science Department, State University of Londrina (UEL), Rodovia Celso Garcia Cid, PR 445 Km 380, Campus Universitário, Londrina 86051-980, Brazil.

[b] Department of Communications (DECOM), School of Electrical and Computer Engineering, University of Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, Barão Geraldo, Campinas 13083-970, Brazil

ARTICLE INFO

Keywords:
Ant Colony Optimization
Traffic characterization
Network management
Unsupervised learning
Anomaly detection
Self-organized agents

ABSTRACT

Traffic monitoring and anomaly detection are essential activities for computer network management, since they provide relevant information about its current performance and contribute to network control. Although there are several studies in this area, diagnosis and resolution of anomalies are still challenging issues. From an expert system point of view, current solutions have not been sufficient to meet the requirements demanded for use in large-scale network environments, and thus a significant portion of budgets on the workforce are spent to network management. Based on this context, the focus of this paper consists of the development of a system able to proactively monitor the network and detect anomalous events, reducing manual intervention and the probability of errors in decision-making, regarding network management. The proposed approach characterizes the normal pattern of the network traffic and detects anomalous behavior, outage events and attacks by deviations from this pattern. For this purpose, an unsupervised learning methodology is used to extract features of traffic through IP flows attributes, collected from a network structure. Aiming to improve its efficiency, a modification of the Ant Colony Optimization metaheuristic is proposed, which through self-organized agents optimizes the analysis of multidimensional flows attributes and allows it to be completed in time to mitigate the impact on large-scale networks. In addition to notify the network manager about the anomalies, the system provides necessary information to identify and take action against them. The resulting detection system was tested with real and simulated data, achieving high detection rates while the false alarm rate remains low.

1. Introduction

Networking services have become increasingly indispensable due to the constant development of communication technologies. This advancement has ensured the provision of resource and mechanisms such as file sharing, remote access, search, people interaction and others. However, to meet the growing demand for new services quality, networks have turned into complex systems composed of heterogeneous elements which operate at high speed and constantly interact with each other (Proença, Coppelmans, Bottoli, & Souza Mendes, 2006). The networks growth made their management by human operators a complicated task, resulting in the need for automation of administrative functions.

Network operators are often faced with unusual events, some of which may exhibit malicious behavior. Using large scale computer networks, their monitoring and control are hampered. In order to maintain the reliability and availability characteristics, the administrator must ensure a precise traffic analysis to promote diagnosis which assists decision-making. Nonetheless, detection and classification of anomalies is a challenging research field, since the traffic monitoring becomes more difficult every day, requiring proactive methods to detect these activities in advance, mitigating its effects and ensuring the network proper functioning (Bhuyan, Bhattacharyya, & Kalita, 2014; Kim, Lee, & Kim, 2014).

In the literature, methods of detecting anomalies are divided into two groups. Traditionally, in signature-based detection, the systems have a database provided by human experts with description of anomalies characteristics which can be detected (Bhuyan et al., 2014). The second category comprises the profile-based detection. This approach creates a normal network profile taking into account its historical traffic. The detection is performed based on

[*] Corresponding author. Tel.: +55 43 33714534; fax: +55 43 84558489.
E-mail addresses: luizfcarvalhoo@gmail.com (L.F. Carvalho), barbon@uel.br (S. Barbon Jr.), lmendes@decom.fee.unicamp.br (L.de S. Mendes), proenca@uel.br (M.L. Proença Jr.).

http://dx.doi.org/10.1016/j.eswa.2016.01.032
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

inferences of significant changes in network operations indicators, which are not consistent with the estimates defined in the normal behavior profile. Although there may be higher false alarms rates, profile-based systems are more promising, since they can detect unknown anomalies, due to its flexibility.

This paper presents a system able to automatically detect security risks and anomalies in order to help administrators in the network management. This paper seeks to build an approach which identifies varied set of network anomalies and attacks with high accuracy and low false alarm rate. According to Molnar and Moczar (2011), many previous works are focused on analysis of a single traffic attribute for anomaly detection, making it unfeasible for recognition of complex attacks. In this paper, we tackle this limitation by developing a seven-dimensional analysis of IP network traffic. For this purpose, it is believed that greater precision is achieved by using both information regarding the traffic volume, e.g., bits, packets and amount of flows transmitted per second as well as the distribution of the attributes contained in the packet header fields (source and destination IP addresses and ports of origin and destination).

The details level of the proposed monitoring is related to the analysis of flows using IP network management protocols such as NetFlow, sFlow and IPFIX. Thus, source and destination IP addresses, source and destination ports, types of used protocols and other traffic properties make it possible to identify patterns of behavior between hosts, in order to identify problems, anomalies and attacks. This management approach is an alternative to the traditional model provided by SNMP (Simple Network Management Protocol) objects.

The proposed system uses the profile-based methodology, being divided into two steps. The first one is the network traffic characterization based on volume features and also the distribution of attributes IP addresses and ports. For this to occur, we introduce the concept of DSNSF (Digital Signature of Network Segment using Flow Analysis), which is responsible for describing the normal behavior profile of network to the analyzed attributes. For DSNSF creation, a modification of the Ant Colony Optimization metaheuristic is used in order to optimize the extraction of behavior patterns of traffic through an unsupervised learning mechanism.

The second step corresponds to the detection of events which cause significant changes traffic in relation to normal behavior, previously characterized. This approach uses Adaptive Dynamic Time Warping (ADTW), a modification from the traditional pattern matching method for this purpose. It provides accurate anomaly detection, recognizing shifted behavior between the DSNSF and real traffic series through time alignment, enabling improved analysis of sudden events and those that occur along the time.

The use of DSNSF along with anomalies recognition provides significant information to the solution of attacks or failures. Furthermore, by using the DSNSF it is possible to constantly monitor the traffic, due to its ability to provide information about the use of resources utilized by network services. This information assists the network administrator in quick decision making, promoting the network reliability and availability of the services offered by it.

The presented system was subjected to evaluation using data from a real network environment. In addition to this data source, anomalies were injected to the real traffic to verify the effectiveness of ACODS in detecting many types of security threats. Complementing the review, a comparison with the traditional clustering algorithm K-means used for DSNSF generating is performed. The presented system is also compared in terms of the detection rate and computational cost with two recent methods found in the literature, PCADS and HWDS.

Based on this scenario, the main contributions of the paper are summarized in the following four points:

- ACODS enables automation in both network monitoring and detection of anomalous traffic events. The principal advantage of this methodology is the automatization of manual management activities, in order to ensure greater accuracy and reduce errors from human intervention.
- Contrary to widely used static thresholds, ADTW allows a more flexible analysis between the series which describe the behavior pattern and the current traffic, avoiding legitimate traffic to be misclassified as anomalous, so that network administrator is not overloaded with false notifications.
- Performance tests considering real traffic traces were performed aiming to validate the effectiveness of the proposed system. Real-world data have been rapidly changing over time. Thus, to reflect the state of non-stationary data in near real-time, a computational model constructed in the training process should be continuously updated whenever new traffic pattern is observed. ACODS uses unsupervised learning in the training stage, eliminating the need for labeled data or network administrator surveillance.
- The administrator is supplied with reports on network resources utilization at the exactly moment when the anomaly is detected. This information – when, who (IP addresses) and how (ports and protocols) – is the basis to build a solution to the problem.

The remainder of this paper is organized as follows. Section 2 shows the related work. Section 3 presents the proposed anomaly detection system. Results and performance of the system are described and discussed in Section 4. Finally, Section 5 concludes the paper.

2. Related work

Efficient identification and categorization of the various network anomalies types is not simple and several issues have to be highlighted. Firstly, the diagnosis of abnormalities is becoming more complicated since anomalous behavior patterns should be recognized from an increasing amount of multidimensional data traffic. Such data may contain noise either derived from its collect process or cross-traffic burstiness which aggregates all measurement links. Secondly, the anomaly detection in large-scale networks implies difficulties both in cost of continuous monitoring as in processing multiple traffic attributes accompanied by the lack of automated tools for near real-time detection (Marnerides, Schaeffer-Filho, & Mauthe, 2014).

Currently, network-monitoring applications generally require occasionally manual tuning parameters of internal anomaly detection approach, which administrators are neither interested nor feel comfortable in doing. To address the tuning challenge, Liu et al. (2015) advocate for using supervised machine learning techniques. The proposed method called Opprentice learns, i.e., extract traffic features from labeled data, capturing the domain knowledge from the network operators, just as an apprentice of the operators. Then the features and the labels are used to train a random forests model in order to select appropriate detector-parameters combinations that satisfy administrator accuracy preference. Also, regarding machine learning, Elhag, Fernández, Bawakid, Alshomrani and Herrera (2015) present linguistic fuzzy association rule mining classifier. This approach uses labeled data (KDD CUP' 99 dataset) and divide-and-conquer learning scheme to obtain a better separability between a normal activity and the different attack types. Using the same dataset, Support Vector Machine (SVM) with three types of kernel (Linear, polynomial and RBF) is used to detect web applications attacks in Alazab, Hobbs, Abawajy, Khraisat, and Alazab (2014). In this method, signature-based detection techniques are used to recognize known attacks, while anomaly-based detection