Applied Soft Computing 36 (2015) 408–418

Contents lists available at ScienceDirect

Applied Soft Computing

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

An uncertainty-managing batch relevance-based approach to network anomaly detection

Gianni D'angelo [a], Francesco Palmieri [b,*], Massimo Ficco [c], Salvatore Rampone [a]

[a] Department of Sciences and Technologies, Sannio University, Via dei Mulini 59A, I-82100 Benevento, Italy
[b] Department of Computer Science, University of Salerno, Via Giovanni Paolo II, 132, I-84084 Fisciano, SA, Italy
[c] Department of Industrial and Information Engineering, Second University of Naples, Via Roma 29, I-81031 Aversa, CE, Italy

ARTICLE INFO

Article history:
Received 22 October 2014
Received in revised form 23 June 2015
Accepted 22 July 2015
Available online 3 August 2015

Keywords:
Network anomaly detection
Machine learning
Supervised classification
Fuzzy-based techniques
Inductive inference

ABSTRACT

The main aim in network anomaly detection is effectively spotting hostile events within the traffic pattern associated to network operations, by distinguishing them from normal activities. This can be only accomplished by acquiring the a-priori knowledge about any kind of hostile behavior that can potentially affect the network (that is quite impossible for practical reasons) or, more easily, by building a model that is general enough to describe the normal network behavior and detect the violations from it. Earlier detection frameworks were only able to distinguish already known phenomena within traffic data by using pre-trained models based on matching specific events on pre-classified chains of traffic patterns. Alternatively, more recent statistics-based approaches were able to detect outliers respect to a statistic idealization of normal network behavior. Clearly, while the former approach is not able to detect previously unknown phenomena (zero-day attacks) the latter one has limited effectiveness since it cannot be aware of anomalous behaviors that do not generate significant changes in traffic volumes. Machine learning allows the development of adaptive, non-parametric detection strategies that are based on "understanding" the network dynamics by acquiring through a proper training phase a more precise knowledge about normal or anomalous phenomena in order to classify and handle in a more effective way any kind of behavior that can be observed on the network. Accordingly, we present a new anomaly detection strategy based on supervised machine learning, and more precisely on a batch relevance-based fuzzyfied learning algorithm, known as U-BRAIN, aiming at understanding through inductive inference the specific laws and rules governing normal or abnormal network traffic, in order to reliably model its operating dynamics. The inferred rules can be applied in real time on online network traffic. This proposal appears to be promising both in terms of identification accuracy and robustness/flexibility when coping with uncertainty in the detection/classification process, as verified through extensive evaluation experiments.

1. Introduction

Together with the astonishing deployment of network technologies and the consequent increment in traffic volumes, the importance of network misuse detection and prevention frameworks is proportionally growing in almost all the modern organizations, in order to protect the most strategic resources from both external and internal threats. In this scenario, the task of identifying and categorizing network anomalies essentially consists in determining all the circumstances in which the network traffic pattern deviates from its normal behavior, that in turn depends on multiple elements and considerations associated to the activities taking place every day on the network.

However, the main difficulty related to a really effective detection is associated to the continuous evolution of anomalous phenomena, due to the emergence of new previously unknown attacks, so that achieving a precise, stable and exhaustive definition of anomalous behavior, encompassing all the possible hostile events that can occur on a real network, is practically impossible. Nevertheless, detection systems must not be limited by the a priori knowledge of a specific set of anomalous traffic templates or be conditioned by a large number of complex operating parameters (e.g., traffic statistic distributions and alarm thresholds), and hence have to be able to recognize and directly classify any previously

* Corresponding author. Tel.: +39 089227710.
E-mail addresses: dangelo@unisannio.it (G. D'angelo), fpalmieri@unisa.it (F. Palmieri), massimo.ficco@unina2.it (M. Ficco), rampone@unisannio.it (S. Rampone).

http://dx.doi.org/10.1016/j.asoc.2015.07.029
1568-4946/

unknown phenomenon that can be experienced on the network. As a consequence, the ultimate goal of modern anomaly detection systems is behaving in a adaptive way in order to flag in "real-time", all the deviations from a model that is built dynamically and in an incremental way by capturing the concept of normality in network operations according to a learning-by-example strategy. These new systems, overcoming the known limitations of the more traditional ones based on pattern detection and statistical analysis, are empowered by flexible machine learning techniques.

Accordingly, we propose a novel anomaly detection strategy, particularly suitable for IP networks, based on supervised machine learning, and more specifically on a batch relevance-based fuzzy-fied learning algorithm known as U-BRAIN.

This strategy aims at understanding the processes that originate the traffic data, by deriving the specific laws and rules governing it, in order to reliably model its underlying dynamics. This is accomplished by performing inductive inference (or better, generalization) on traffic observations, based on some empirical pre-classified "experiential" (training) data, representing incomplete information about the occurrence of specific phenomena that describe normal or anomalous network activities. In addition, the adopted learning scheme allows a certain degree of uncertainty in the whole detection process making the resulting framework more solid and flexible in managing the large variety and complexity of real traffic phenomena. Then the inferred rules can be applied in real time on online network traffic.

We evaluated the effectiveness of the presented detection framework within a widely known test case scenario, in order to make the achieved results comparable with those of other proposal available in literature. These results demonstrated a quite satisfactory identification accuracy by placing our strategy among the most promising state-of-the-art proposals.

2. Background and related work

Network anomaly detection has gained a great attention in security research with about 40 years of experiences available in literature. The first approach to automatic detection has been proposed in [1], followed by a large number of contributions exploring many other solutions and proposals [2–4].

The earliest and more traditional detection approaches, mainly aiming at spotting intrusion activities, work by matching specific traffic patterns, gathered from the packets under observation, against a list of predefined signatures, each associated to a known attack or hostile/anomalous behavior. Some well-known examples are SNORT [5] and BRO [6]. While ensuring very good response times and a quite satisfactory degree of effectiveness in case of previously known menaces, these approaches are almost totally clueless in presence of new (zero-day) attacks, or when, due to minor modifications in its behavior, an already known attack does not closely match the associated signatures. In both the cases new up-to-date signatures must be generated and added to the list as soon as more detailed information about the hostile behavior become available. Unfortunately, this implies human intervention, and hence too much time to ensure real-time response.

Other very common detection systems are based on a statistical idealization of the network behavior and process the traffic observations through statistical-analysis techniques by flagging the outliers as anomalous events. The most significant examples are NIDES (Next-Generation Intrusion Detection Expert System) [7], an hybrid system providing a statistical analysis engine, and SPADE (Statistical Packet Anomaly Detection Engine) [8] a statistical detection system based on determining anomaly scores, available as a plug-in for SNORT. However, while straightforward and robust in their formulation (they do not require prior knowledge of the security menaces nor need packet inspection), statistic detection approaches may result too simplistic in their basic assumptions and hence scarcely reliable in their results. In fact, being based only on the statistical properties of the involved traffic flows, these approaches are too sensitive to the normality assumption, and really effective only against specific phenomena that imply significant variations in the statistical properties of the network traffic (Volume-based Attacks). More precisely, such detection techniques have to be based on extremely accurate statistical distributions that describe the traffic under observation. Unfortunately, modeling real network traffic, typically characterized by an inhomogeneous usage pattern, by using only pure statistical methods, may result in a poor choice in terms of real effectiveness. Furthermore, these solutions cannot be aware of hostile activities that only affect the packet contents (such as stack smashing or other kind of malicious code exploiting system/services vulnerabilities) or explicitly conceived to be undistinguishable from regular user activities (e.g., low-rate DoS attacks).

As an alternative that may reveal extremely effective in coping with the above challenges, machine learning provides fully automated detection capabilities, by allowing a system to learn by example what are the anomalous events occurring in the observed traffic. It also allows improving the detection performance over time with experience, as more and more examples (or training data), describing normal or anomalous behaviors, are provided in its knowledge base. In this way, the detection function, that is essentially a binary classifier working on the normal and anomalous traffic classes, is inferred from the aforementioned training data. Such data consist of a set of pre-classified traffic samples. In supervised learning, each sample is a pair consisting of an input object (typically a vector of traffic features) and a desired output value (the class value) also called the supervisory signal. The inferred classifier should assign the right output class value to any valid input sample. This implies that the learning paradigm should be reasonably capable to perform generalization from the knowledge contained in the training data to previously unseen situations.

The use of machine learning in anomaly detection, with the development of generalization capabilities from past experiences for classifying future data as normal or anomalous has been exploited in many proposals [9], based on neural networks [10,11], SVMs [12] and data mining techniques [13,14]. These approaches can be further subdivided into generative or discriminative. A typical generative approach (e.g., [15]) constructs a model by starting only from normal training examples, and then evaluate several test instances in order to appreciate how well they fit such model. As an example, the ideas presented in [16] explore different machine learning techniques to construct detection models from past behavior. On the other hand, discriminative techniques (e.g., [12]), attempt to understand the difference between the normal and anomalous instance classes. A learning approach for reproducing packet level alerts for anomaly detection at the flow level has been presented in [17]. Several approaches rely on clustering techniques, such as ADWICE [18], performing unsupervised detection based on a fast incremental clustering technique. K-Means+ID3 [19], instead, is a supervised learning approach combining k-Means clustering and the ID3 decision trees in order to classify anomalous and normal network activities. Regarding the use of tree-based structures, the DGSOT+SVM [20] scheme is an iterative approach leveraging the dynamic generation of self-organizing hierarchical trees together with SVMs to be trained on the tree nodes, where support vectors are used as the basic knowledge to control the tree growth. Also non-linear analysis, combined with recurrence quantification techniques [21] has been used to construct a flexible detection approach based on understanding the most hidden traffic dynamics by simultaneously observing the traffic behavior on multiple time scales.