Expert Systems with Applications 39 (2012) 1822–1829

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection

Carlos A. Catania [a,*], Facundo Bromberg [b], Carlos García Garino [a,c]

[a] ITIC, Universidad Nacional de Cuyo, Mendoza, Argentina
[b] Dept. Sistemas de Información, FRM – UTN, Mendoza, Argentina
[c] Facultad de Ingeniería, Universidad Nacional de Cuyo, Mendoza, Argentina

ARTICLE INFO

Keywords:
Anomaly detection
Intrusion detection systems
SVM
Labeling

ABSTRACT

In the past years, several support vector machines (SVM) novelty detection approaches have been applied on the network intrusion detection field. The main advantage of these approaches is that they can characterize normal traffic even when trained with datasets containing not only normal traffic but also a number of attacks. Unfortunately, these algorithms seem to be accurate only when the normal traffic vastly outnumbers the number of attacks present in the dataset. A situation which can not be always hold.

This work presents an approach for autonomous labeling of normal traffic as a way of dealing with situations where class distribution does not present the imbalance required for SVM algorithms. In this case, the autonomous labeling process is made by SNORT, a misuse-based intrusion detection system. Experiments conducted on the 1998 DARPA dataset show that the use of the proposed autonomous labeling approach not only outperforms existing SVM alternatives but also, under some attack distributions, obtains improvements over SNORT itself.

1. Introduction

In the past years network security has become a serious problem. In the early years of the Internet, the set of network protocols that supported it worked reasonable well. However as the Internet grew, underlying security faults in those protocols were observed. Security faults in protocols such as ARP, TCP, TELNET, SMTP and FTP have caused most of known attacks against network data confidentiality, authenticity and availability. Currently most of these problems have been fixed, however new ways to develop attacks are discovered everyday.

Network managers must be well prepared in order to prevent network attacks, e.g., being informed about new vulnerabilities. For several years, intrusion detection systems (**IDS**) provided an invaluable help to network managers, becoming an integral part of any network security package.

In the intrusion detection field two different approaches can be observed: misuse detection and anomaly detection (Mukherjee, Heberline, & Levitt, 1994). The main idea behind misuse detection is to represent attacks in a form of a pattern or a signature in such a way that even variations of these attacks can be detected. Based on these signatures, this approach detects attacks through a large set of rules describing every known attack (Tsai, Hsu, Lin, & Lin, 2009; Wu & Yen, 2009). The main disadvantage of the signature based approach is its difficulty for detecting unknown attacks. The main goal of the anomaly detection approach is to build a statistical model for describing normal traffic. Then, any deviation from this model can be considered an anomaly, and recognized as an attack. Notice that when this approach is used, it is theoretically possible to detect unknown attacks, although in some cases, this approach can lead to a high false attack rate. This ability to detect unknown attacks has been the cause of the increasing interest in developing new techniques to build models based on normal traffic behavior in the past years.

The anomaly detection approach has been a very active research topic inside the machine learning community and it has been the subject of many articles over the past years. One of the most successful approaches is based on the idea of collecting data only from network normal operation. Then, based on this data describing normality, any deviation would be considered an anomaly. Different techniques were proposed for characterizing the concept of normality (Catania & García Garino, 2008; Hofmeyr, Forrest, & Somayaji, 1998; Lee & Stolfo, 1998). In practice, however, it is difficult to obtain clean data to implement these approaches. Verifying that no attacks are present in the training data can be an extremely hard task, and for large samples this is simply infeasible. On the other hand, if the data containing attacks is assumed attack free, intrusions similar to the ones present in the training

* Corresponding author.
E-mail address: ccatania@itu.uncu.edu.ar (C.A. Catania).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.068

data will be accepted as normal patterns, resulting in inaccurate models and consequently, an increment in the number of misdetections.

Recently, different authors proposed the use of unsupervised algorithms for dealing with datasets presenting not only normal traffic but also a considerable number of attacks (Eskin, Arnold, Prerau, Portnoy, & Stolfo, 2002; Feng, Wu, Wu, Xiong, & Zhou, 2005; Laskov, Schafer, & Kotenko, 2004). This situation could be considered more suitable than using datasets with only normal traffic instances. In this sense, SVM for novelty detection (Schölkopf, Platt, Shawe-taylor, Smola, & Williamson, 2001; Tax & Duin, 1999) was proposed as an alternative approach with a significant success rate.

Unfortunately, as noticed by Eskin et al. (2002), SVM for novelty detection works under the assumption that the number of normal traffic instances vastly outnumbers the number of anomalies. Eskin suggests datasets with a proportion of at least 98.5% of normal traffic.

To the best of the authors knowledge, there is no study which confirms the number of attacks laying under such low proportion. Informal observations of real traffic however, show that it is possible to find periods of time where the number of attacks presents in traffic could easily outnumber normal traffic instances. This situation can be observed in commonly used datasets for intrusion detection evaluation such as the 1998 DARPA dataset (Lippmann et al., 2000). This dataset was provided by DARPA to the machine learning community in the context of the 1999 KDD Cup for evaluating different IDS approaches. Since its publication DARPA dataset has been widely used by many IDS researchers over the years. Interestingly, the 1998 DARPA class distribution does not exhibit the required imbalance. Moreover, the percentage of attacks present in the dataset is around 50%. Certainly, under these situations algorithms such as SVM for novelty detection could suffer considerable performance loss.

To deal with these imbalanced class distribution situations a novel approach is proposed. The idea is to provide a strategy for autonomous labeling only normal traffic, following the hypothesis that using an autonomous labeling tool may help reducing the presence of attacks in the traffic instances used for training, and consequently improving the performance of SVM for novelty detection. In this work, SNORT (Roesch, 1999), a very well known misuse signature-based IDS system, is proposed as a strategy for autonomous labeling normal traffic.

The rest of the work is organized as follows: in Section 2 main characteristics of SVM for novelty detection are briefly discussed, together with its application to the traffic network detection field. Then, in Section 3, a new approach for autonomous labeling normal traffic is presented. In Section 4 a set of experiments is conducted on the 1998 DARPA dataset in order to evaluate the performance of the different approaches. Finally, conclusions and future work are provided in Section 5.

2. SVM for novelty detection

Since its introduction in the mid-1990s (Boser, Guyon, & Vapnik, 1992; Cortes & Vapnik, 1995; Vapnik, 1998), The SVM algorithm has been widely used, being the subject of many articles on classification and other pattern recognition problems (Lee & Verri, 2002).

SVM approach for classification differ from other classification algorithms by three important properties. First, its formulation presents an important theoretical result, proving that the generalization error is minimized when the *margin* is maximized, where the margin is defined as the distance of the solution hyperplane to its closest point (Vapnik, 1998). This property is unique to SVM and is one of its main advantages when compared to other classification algorithms. Another important property is that the search for the maximal margin is a convex (quadratic) optimization problem, i.e., with only one minima, resulting in an efficient learning stage. In most cases, the input data points are not separable by the separation surface, so a standard approach (first introduced for the Perceptron algorithm of Rosenblatt (1958)), is to project the data points to higher dimension *feature* space. That usually affects the generalization error. However, for SVM, it can be proven (Vapnik, 1998) that for the maximal margin, the generalization error is still minimal, regardless of the dimension of the feature space. Finally, the formulation of the optimization problem (as shown in the next section for SVM for novelty detection) can be expressed solely in terms of the dot product between the feature vectors (denoted its *kernel*), which further reduces the computational complexity by permiting an efficient pre-computation of these quantities.

SVM for novelty detection is a generalization of the core SVM ideas for classification problems. Traditional SVM approaches for classification uses as input training data consisting of a mixture of data labeled by two classes. In the intrusion detection problem this would consist of data labeled both as *attack* and *non-attack*. The model constructed by these approaches discriminates the input space in two infinite regions, one per class, using a hyperplane as a separation surface. In contrast, the main idea in SVM for novelty detection (Schölkopf et al., 2001; Tax & Duin, 1999) is to use as input a description of only the *normal* class of objects (*non-attack* in IDS), assuming the rest as *anomalies* (in our problem, the *attacks*). The model constructed by this approach discriminates the input space in a finite region containing the normal objects, while all the rest of the (infinite) space is assumed to contain the anomalies.

The SVM for novelty detection variants appear in the literature of intrusion detection with different names, which could lead to some confusion. In some cases they are referenced as SVM one-class algorithms. SVM for non supervised learning is another widely used name by some authors. Although, all of these names describe important characteristics of this kind of algorithms, in this work the term SVM for novelty detection will be preferred.

Two major approaches were proposed for generalizing SVM to the problem of novelty detection. One approach, proposed by Tax and Duin (1999), is based on the idea of finding a hypersphere with center $\mathbf{c}$ and minimal radius R containing the *normal* data, discriminating all other data not in the sphere as *anomalies*. As in standard SVM approaches, the discriminating surface (the sphere), as well as the data, may be mapped into a higher dimension feature space by a kernel function (see more details in next section). Another approach proposed by Schölkopf et al. (2001) tries to separate the normal data points from the anomalies by finding the hyperplane that is maximally distant from the origin. When a RBF kernel is used, it was shown that the two approaches converge to the same solution (Campbell, 2000). In this work the Tax's approach is preferred, which is explained in more detail below. For a description of Schölkopf's hyperplane formulation the reader is referred to Schölkopf et al. (2001).

2.1. SVM based on the hypersphere formulation

The sphere formulation has an intuitive geometric idea: the normal data $\{\mathbf{x}_i, i = 1,\ldots,N\}$ can be concisely described by a sphere, of center $\mathbf{c}$ and radius R, first projecting the data to some high-dimensional feature space by the mapping Φ, obtaining the projected set of points $\{\Phi(\mathbf{x}_i), i = 1,\ldots,N\}$, and assuming the projected *normal* points lie within the sphere. A graphical example of this can be observed in Fig. 1. Non-separability of the training data in the feature space can be addressed by introducing slack variables