Computer Networks 000 (2016) 1–9

Contents lists available at ScienceDirect

Computer Networks

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

Learning combination of anomaly detectors for security domain

Martin Grill[a,b,*], Tomáš Pevný[a,b]

[a] Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Republic
[b] Cisco Systems, Inc., United States

ARTICLE INFO

Article history:
Received 27 November 2015
Revised 10 May 2016
Accepted 29 May 2016
Available online xxx

Keywords:
Anomaly detection
Ensemble systems
Positive unlabeled data
Accuracy at top

ABSTRACT

This paper presents a novel technique of finding a convex combination of outputs of anomaly detectors maximizing the accuracy in τ-quantile of most anomalous samples. Such an approach better reflects the needs in the security domain in which subsequent analysis of alarms is costly and can be done only on a small number of alarms. An extensive experimental evaluation and comparison to prior art on real network data using sets of anomaly detectors of two existing intrusion detection systems shows that the proposed method not only outperforms prior art, it is also more robust to noise in training data labels, which is another important feature for deployment in practice.

1. Introduction

Increasing numbers of attacks against computing infrastructure and the critical importance of the infrastructure for enterprises drives the need to deploy progressively more sophisticated defense solutions to protect network assets. An essential component of the defense are Intrusion Detection Systems (IDS) [1] searching for evidence of ongoing malicious activities (network attacks) in network traffic crossing the defense perimeter.

Many intrusion detection systems are implemented as ensembles of relatively simple, yet heterogeneous detectors [2,3], where some of them can be specialized to particular types of intrusions, whereas others can be general anomaly detectors capable of detecting previously unseen attacks at the expense of higher false alarm rates. Such a setup has multiple advantages, including faster processing of the data stream, lower complexity of the detectors, and simpler inclusion of domain knowledge into the system. The main drawback is that combining outputs of individual detectors is a non-trivial problem. Although a vast prior art on the problem exists [4–6], we believe that peculiarities of the security domain, namely a highly imbalanced ratio of non-alarm and alarm samples in the data, lack of accurately labeled datasets, and the need of extremely low false positive rates, call for a tailored solution.

The rationale behind the above specifics is that from the user perspective each raised alarm needs to be thoroughly investigated, which is expensive and can be done only for a small number of them. Hence reporting high numbers of false positives renders any intrusion detection system useless (recall that most of the samples are legitimate). Note that using a supervised method to learn the combination may bring the expense of lower generalization, but according to our experience completely unsupervised approaches rarely have false positive rate low enough to be usable in practice. Moreover, anomaly detectors and their features are usually selected based on the experience of the designer, which is a kind of proxy for labels and surely not guaranteed to be complete.

Obtaining labeled data in security domains and in network intrusion detection especially can be difficult, time consuming, and expensive. Besides, labeled data frequently contains errors in labels of different sorts, for example some alerts might be missed and labeled as legitimate samples, or even worse, all samples of alerts of certain types might be missed and labeled as legitimate.

The above concerns motivated the main goals and contributions of this paper, which are a method of finding a convex combination of outputs of a fixed set of anomaly detectors maximizing the number of true alarms in τ-fraction of most anomalous connections (samples)[1] and an experimental study of the effect of different types of label noise in the training data on the accuracy of combinations obtained by different methods to better understand their advantages and drawbacks. Conducted experiments revealed that the proposed method is not only better than the state of the art, but also more robust with respect to various kinds of noise in labels we can expect in intrusion detection domains.

If the proposed method requires labeled data, one can ask why not use them to train a classifier and sidestep the use of anomaly

* Corresponding author.
E-mail addresses: magrill@cisco.com (M. Grill), tpevny@cisco.com (T. Pevný).

[1] Since the experimental evaluation is performed with network intrusion detection systems, the terms sample and connection are used interchangeably.

http://dx.doi.org/10.1016/j.comnet.2016.05.021
1389-1286/

detectors? The most important reason to favor anomaly detectors is that network traffic discussed in this paper is very non-stationary and anomaly detectors are good at coping with this aspect, as they can constantly update their models (see [7–9] for a review).

This paper is organized as follows: The next section formally defines the problem and presents the proposed solution. Section 3 reviews related work and algorithms that we evaluate in the experimental section. The experimental Section 4 compares the proposed solution with existing methods using sets of anomaly detectors from two different network intrusion detection systems operating on two different data sources.

2. Proposed method

Prior art in combining detectors and anomaly detectors in particular is vast [4,10], nevertheless we feel that security domains requires a tailored solution because of its prominent requirement of extremely low false positive rate. We assume that the network operator observers connections (samples) from an unknown distribution $P_o = \pi P_a + (1-\pi)P_b$ with P_a/P_b being distributions of malicious/background samples and $\pi \in [0, 1]$. The network operator uses set of m anomaly detectors on samples $\mathcal{H}_m = \{h_k : \mathcal{X} \mapsto [0,1]\}_{k=1}^m$ (w.l.o.g. it is assumed that zero means the sample is legitimate and one means the sample is malicious) and wishes to have a convex combination of anomaly detectors $\alpha = (\alpha_1, \ldots, \alpha_m)$ that would maximize the number of alarms in top τ quantile of the distribution of the combined anomaly scores. For purposes of this paper it is safe to assume that each connection (sample) is described by m-dimensional vector (an output of m anomaly detectors), which implies that distributions P_o, P_a, and P_b are defined on the m-dimensional Euclidean space.The requirements on detectors having their image in the interval [0, 1] and learning a convex combination instead of a linear one are to improve interpretability of the results as discussed in [11], but can be dropped. The same work also presents a general approach to scale the output of any anomaly detector to the interval [0, 1] reviewed in Appendix A.

With respect to the above, networks operator's goal can be written as

$$\arg\min_{\alpha\in\mathbb{R}^m} R(H_\alpha) = \underbrace{\mathbb{E}_{x\sim P_b}\left[\mathbb{1}(\alpha^{\mathrm{T}}h(x) \geq q_{\alpha,\tau})\right]}_{R^{\mathrm{fp}}(H_\alpha)} + \underbrace{\mathbb{E}_{x\sim P_a}\left[\mathbb{1}(\alpha^{\mathrm{T}}h(x) < q_{\alpha,\tau})\right]}_{R^{\mathrm{fn}}(H_\alpha)}, \tag{1}$$

subject to

$$\begin{aligned} H_\alpha(x) &= \sum_{k=1}^m \alpha_k h_k(x) = \alpha^{\mathrm{T}}h(x), \\ \mathbf{1}^{\mathrm{T}}\alpha &= 1, \\ \alpha_i &\geq 0,\ \forall i \in \{1,\ldots,m\}, \end{aligned} \tag{2}$$

where the first term in (1) is the false alarm rate, the second term is the false negative rate, and finally $q_{\alpha,\tau}$ is a τ-quantile of observed distribution of ensemble's output $\{\alpha^{\mathrm{T}}h(x)|x \in P_o\}$. The minimized term (1) captures the accuracy of a particular convex combination in top τ-quantile of its distribution, which is the goal.

In theory it would be sufficient if (1) minimizes either only the false positive rate R^{fp} or only the false negative rate R^{fn}, because each of them together with constraints (2) implies minimization of the other. But including both terms increases the robustness with respect to noise on labels, since the error and its gradient are estimated from larger number of samples implying their better estimates. This is demonstrated in Appendix B, where the combination of anomaly detectors was found by optimizing either only false positive rate or only false negative rate under constraints (2). The experiments have confirmed that optimizing the proposed (1) is indeed more robust to error in labels, which are almost inevitable in security domains. In the rest of this section we show, how to find a good solution in practice using adaptation of the method of Boyd et al. [12].

First, the true loss function (1) cannot be used in practice, since the true probability distributions P_a and P_b are not known. Therefore the expectations are replaced by their empirical estimates calculated from some labeled data used for learning the weight vector α. Below the $\mathcal{S} = \mathcal{S}_a \cup \mathcal{S}_b$ denotes the set of available samples with $\mathcal{S}_b$ being the set of background (legitimate) samples and $\mathcal{S}_a$ the set of malicious samples. The empirical estimate of (1) is therefore

$$\hat{R}(H_\alpha) = \frac{1}{|\mathcal{S}_b|}\sum_{x\in\mathcal{S}_b}\mathbb{1}\left[\alpha^{\mathrm{T}}h(x) \geq \hat{q}_{\alpha,\tau}\right] + \frac{1}{|\mathcal{S}_a|}\sum_{x\in\mathcal{S}_a}\mathbb{1}\left[\alpha^{\mathrm{T}}h(x) < \hat{q}_{\alpha,\tau}\right], \tag{3}$$

where $\hat{q}_{\alpha,\tau}$ is an empirical estimate of the true quantile $q_{\alpha,\tau}$ defined as

$$\hat{q}_{\alpha,\tau} = \arg\max_{\omega} \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}}\left[\mathbb{1}(\alpha^{\mathrm{T}}h(x) \leq \omega)\right] \leq \tau. \tag{4}$$

Since the empirical loss function (3) is neither convex nor smooth, finding the optimal solution is an NP-complete problem. A usual approach is to replace indicator function $\mathbb{1}$ with a convex surrogate, for example an exponential used in this work.[2] This substitution leads to the following optimization problem

$$\begin{aligned} \arg\min_{\alpha} \quad & \frac{1}{|\mathcal{S}_b|}\sum_{x\in\mathcal{S}_b}\exp\left(\alpha^{\mathrm{T}}h(x) - \hat{q}_{\alpha,\tau}\right) \\ & + \frac{1}{|\mathcal{S}_a|}\sum_{x\in\mathcal{S}_a}\exp\left(\hat{q}_{\alpha,\tau} - \alpha^{\mathrm{T}}h(x)\right) \\ \text{subject to} \quad & \mathbf{1}^{\mathrm{T}}\alpha = 1, \\ & \alpha_i \geq 0,\ \forall i \in \{1,\ldots,l\}, \\ & \hat{q}_{\alpha,\tau} \text{ is a } \tau\text{-quantile defined in (4).} \end{aligned} \tag{5}$$

where the optimized term (further denoted as $\hat{R}_{\exp}(H_{\boldsymbol{\alpha}})$) is an upper bound of the empirical loss function $\hat{R}(H_{\boldsymbol{\alpha}})$ defined in Eq. (3).

Nevertheless the last problem is still hard to solve, as it is not convex. Boyd et al. [12] showed how to find a good solution in polynomial time using series of convex problems. However his algorithm does not guarantee finding the global minimum, and the computational complexity prevents it from being used on problems with millions of samples. We therefore propose to solve (5) by a simple gradient algorithm summarized in Algorithm 1, which albeit not reaching the global minimum performs well, according to our experiments. In each step the current solution $\boldsymbol{\alpha_k}$ is updated by subtracting a small multiple of the gradient of (5), which is decreasing in each step to ensure convergence. The $\boldsymbol{\alpha_k}$ is then truncated to satisfy the constraints, and finally the estimate of the quantile $\hat{q}_{\alpha,\tau}$ is updated. The algorithm may find sub-optimal solutions but the experiments in Section 4 show that the solutions found are in most of the cases better than the ones of the state-of-the-art methods. Additionally, detailed discussion about the differences between the solution found by Boyd et al. and the one found by the proposed algorithm can be found in Appendix C.

The combination of detectors found by the above algorithm is optimized with respect to the *known* malware, by which we understand the malware whose samples are present in the training set

[2] The chosen convex surrogate does not have a significant impact on the solution and can be replaced by the reader's favorite choice, e.g. logistic, hinge, truncated square, etc.

Please cite this article as: M. Grill, T. Pevný, Learning combination of anomaly detectors for security domain, Computer Networks (2016), http://dx.doi.org/10.1016/j.comnet.2016.05.021