Journal of Network and Computer Applications 41 (2014) 473–487

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Review

Information theoretic feature space slicing for statistical anomaly detection

Ayesha Binte Ashfaq [a,*], Sajjad Rizvi [b], Mobin Javed [c], Syed Ali Khayam [d], Muhammad Qasim Ali [e], Ehab Al-Shaer [e]

[a] Department of Computing, School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Pakistan
[b] Michigan State University, East Lansing, MI, USA
[c] University of California, Berkeley, CA, USA
[d] PLUMgrid Inc., 440 North Wolfe Rd., Sunnyvale, CA, USA
[e] University of North Carolina, Charlotte, USA

ARTICLE INFO

Article history:
Received 26 January 2013
Received in revised form
9 November 2013
Accepted 14 January 2014
Available online 22 January 2014

Keywords:
Feature slicing
Conditional entropy
Information content
Clustering
Statistical anomaly detection

ABSTRACT

Anomaly detection accuracy has been a serious limitation in commercial ADS deployments. A main reason for this limitation is the expectation that an ADS should achieve very high accuracy while having extremely low computational complexity. The constraint of low computational cost has recently been relaxed with the emergence of cheap high-performance platforms (e.g., multi-core, GPU, SCC, etc.). Moreover, current ADSs perform anomaly detection on aggregate feature spaces, with large volumes of benign and close-to-benign feature instances that overwhelm the feature space and hence yield low accuracies. In this paper, we ask and address the following question: *Can the accuracy of an ADS be improved if we slice ADS feature space at the cost of higher computational resource utilization?* We first observe that existing ADSs are not designed to exploit better computational platforms to achieve higher accuracies. To mitigate this problem, we identify the fundamental accuracy limiting factors for statistical network and host-based ADSs. We then show that these bottlenecks can be alleviated by our proposed feature space slicing framework. Our framework slices a statistical ADS' feature space into multiple disjoint subspaces and then performs anomaly detection separately on each subspace by utilizing more computational resources. We propose generic information-theoretic methods for feature space slicing and for determining the appropriate number of subspaces for *any* statistical ADS. Performance evaluation on three independently-collected attack datasets and multiple ID algorithms shows that the enhanced ADSs are able to achieve dramatic improvements in detection (up to 75%) and false alarm (up to 99%) rates.

© 2014 Elsevier Ltd. All rights reserved.

Contents

1. Introduction . . . 474
2. Datasets and anomaly detectors . . . 475
3. Terminology . . . 475
4. Motivation and analysis . . . 475
   4.1. What are the fundamental accuracy limiting factors in statistical ADS design? . . . 476
      4.1.1. Averaging out . . . 476
      4.1.2. Noise . . . 477
      4.1.3. Discussion . . . 478
   4.2. How can the accuracy limiting factors be mitigated by slicing an ADS's feature space? . . . 478
      4.2.1. Maximum entropy detector's feature slicer . . . 478
      4.2.2. PCA-based subspace method's feature slicer . . . 478
      4.2.3. Discussion . . . 479
5. Information-theoretic feature space slicing . . . 479

* Corresponding author: Tel.: +923235095427.
E-mail address: ayesha.ashfaq@seecs.edu.pk (A.B. Ashfaq).

1084-8045/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jnca.2014.01.002

5.1. System-level design of a feature space slicer . . . 480
5.2. How should statistical similarity be defined? . . . 481
5.2.1. Probability and information gain-based feature space slicing (PFS, IGFS) . . . 481
5.2.2. Information content based feature space slicing (ICFS) . . . 481
5.2.3. Discussion . . . 482
5.3. How many subspaces should a feature space be sliced into? . . . 482
5.3.1. Existing methods . . . 482
5.3.2. Conditional entropy based method . . . 483
5.4. Accuracy evaluation . . . 483
6. Limitations of feature space slicing . . . 484
6.1. Higher computation resource requirement . . . 484
6.2. Accuracy . . . 484
6.3. Evasion . . . 484
6.4. Non-statistical ADSs . . . 484
6.5. ADSs with dependence across feature classes . . . 484
7. Conclusion . . . 484
Appendix A. Evaluation datasets . . . 484
A.1. Network traffic datasets . . . 484
A.1.1. LBNL dataset . . . 484
A.1.2. Endpoint dataset . . . 485
A.2. Host-based system calls datasets . . . 485
Appendix B. Network and host-based ADSs . . . 485
B.1. TRW-CB (Schechter et al., 2004) . . . 485
B.2. NIDES (Next-Generation Intrusion Detection Expert System) . . . 486
B.3. KL detector (Khayam et al., 2008) . . . 486
B.4. Kalman filter based detection (Soule et al., 2005) . . . 486
B.5. SVM on bags of system calls (Kang et al., 2005) . . . 486
B.6. Host-based KL detector (Lee and Xiang, 2001) . . . 486
References . . . 486

1. Introduction

In 2003, Gartner Inc. reported Anomaly Detection Systems (ADSs) as a market failure and predicted that "ADSs will be obsolete by 2005" (Stiennon, 2003; Gartner Inc., 2003). Seven years later, we know this prediction to be overly pessimistic as ADS products have seen significant commercial deployments. Nevertheless, there is widespread consensus that ADS technology has not been as disruptive as was originally anticipated. Despite its pessimistic projections, Gartner did rightly predict that the following reasons (among others) will limit ADSs' commercial appeal: (a) low accuracies (low detection rates, high false alarm rates), and (b) inability to monitor high-speed traffic. These two factors represent an interesting trade-off because an ADS is supposed to build a highly robust traffic model (to achieve high accuracy), but is expected to do so at an extremely low computational cost (to allow real-time deployment). Commercial ADSs try to find the right balance of accuracy and computational complexity, but generally end up leaning too heavily towards one end of this spectrum.[1]

Many cheap high-performance platforms-ranging from low-end multi-core processors (Intel Core™2 Duo Processor; AMD Athlon™II X2 Dual-Core Processor) to massively parallel Graphical Processing Units (GPUs) (nVIDIA Motherboard GPUs; Intel Larrabee Integrated Graphics Accelerator) and Single-chip Cloud Computers (SCCs) have emerged to cater for the speed and power requirements of compute-intensive applications. These high-performance platforms are now being proposed to speed up network security devices (Vasiliadis et al., 2008; Aldwairi et al., 2005; Piyachon and Luo, 2006; Yi et al., 2007). As these off-the-shelf high-performance platforms become pervasive, the computational resource constraint on ADSs is becoming increasingly less critical. However, the inherent design of current ADSs does not allow them to exploit the parallelism available in emerging hardware. Though, if multiple instances of an ADS (treated as a black box) are deployed on multiple processing cores (one ADS per core) for anomaly detection, it can in turn enable ADSs to exploit the strengths of off-the-shelf high performance platforms. Consequently, an important question arises: Can the accuracy of an existing ADS be improved if more computational resources are available for its deployment? If the answer to this question is in affirmative then a subsequent question follows: Can a *generic* method be developed that can allow an existing statistical ADS to enhance its accuracy at the expense of more computational resources?

In this paper, we address the above questions in the context of statistical ADSs. A statistical ADS flags traffic anomalies by measuring perturbations in a probabilistic model of network or host traffic. Our performance evaluation in Ashfaq et al. (2008) showed that statistical ADSs offer the most promising accuracies among existing research ADSs e.g rule-based ADSs, volumetric ADSs, etc. We identify two fundamental factors that limit the accuracy of a statistical ADS: (1) *averaging out*: in which high volumes of benign feature instances inundate malicious perturbations and (2) *noise*: which is introduced by close-to-benign malicious feature instances (e.g., successful portscans, attacks on common ports/servers, etc.) To mitigate these factors, we propose that an ADS *slices* statistically similar feature instances into disjoint subspaces at its input and then performs anomaly detection on each subspace separately. Such a method localizes the averaging out and noise artifacts to a few subspaces while ADS accuracy is enhanced in the remaining subspaces, albeit at the cost of higher computational resources.

To achieve accurate feature slicing, we propose a novel method which quantifies each feature instance based on its information

[1] For instance, some ADSs run offline but claim to have near 0% false alarm rates (FireEye Malware Protect System (MPS)), while other products run at wirespeeds but do not provide accuracy guarantees (Cisco Anomaly Guard Module Homepage; Arbor Networks PeakFlow Homepage; Endace NinjaBox Homepage).

Download English Version:

https://daneshyari.com/en/article/457245

Download Persian Version:

https://daneshyari.com/article/457245

Daneshyari.com