INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN: 2278 – 0211 (Online)

Machine Learning For Cloud Computing Intrution Detection

Ms. Vijitha.kondiparthi
Pursuing the masters degree inSoftware Engineering ,
Anurag Engineering College, JNT university , India
Ms. Greeshmananda.V
Pursuing the masters degree in Software Engineering ,
Anurag Engineering College, JNT university, India

Abstract:

The term "cloud computing" has emerged as a major ICT trend and has been acknowledged by respected industry survey organizations as a key technology and market development theme for the industry and ICT users in 2010. In Cloud Computing intrusion detection research, one popular strategy for finding attacks is monitoring a cloud's activity for anomalies: deviations from profiles of normality previously learned from benign traffic typically identified using tools borrowed from the machine learning community. However, despite extensive academic research one finds a striking gap in terms of actual deployments of such systems: compared with other intrusion detection approaches, machine learning is rarely employed in operational "real world" settings. We examine the differences between the cloud computing intrusion detection problem and other areas where machine learning regularly finds much more success. Our main claim is that the task of finding attacks is fundamentally different from these other applications, making it significantly harder for the intrusion detection community to employ machine learning effectively. We support this claim by identifying challenges particular to cloud computing intrusion detection, and provide a set of guidelines meant to strengthen future research on anomaly detection.

Key words: *anomaly detection; machine learning; intrusion detection; cloud security, profile.*

1. Introduction

"Cloud computing" is essentially composed of a large-scale distributed and virtual machine computing infrastructure. This new paradigm delivers a large pool of virtual and dynamically scalable resources including computational power, storage, hardware platforms and applications to users via Internet technologies. Private and public organizations alike can make use of such cloud systems and services while many advantages may be derived when migrating all or some information ser-vices to the cloud computing environment. Examples of these benefits include increases in flexibility and budgetary savings through minimization of hardware and software investments.

Cloud computing is internet based computing where virtual shared servers provide software, infrastructure, platform, devices and other resources and hosting to customer as a service on pay-as you-use basis.

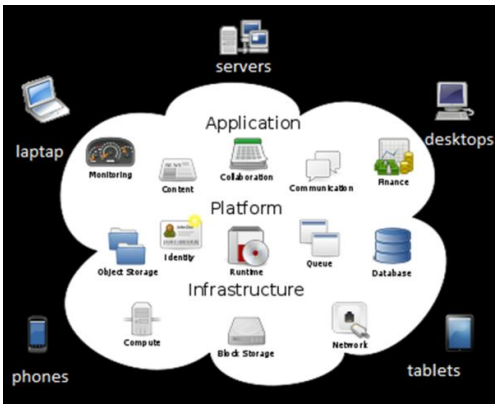

Figure 1

2.Cloud Computing

Traditionally, cloud computing intrusion detection systems (CIDS) are broadly classified based on the style of detection they are using: systems relying on misuse-detection monitor activity with precise descriptions of known malicious behavior, while anomaly-detection systems have a notion of normal activity and flag deviations from that profile. Both approaches have been extensively studied by the research community for many years. However, in terms of actual deployments, we observe a striking imbalance: in operational settings, of these two main classes we find almost exclusively only misuse detectors in use—most commonly in the form of signature systems that scan cloud's traffic for characteristic byte sequences. This situation is somewhat striking when considering the success that machine-learning—which frequently forms the basis for anomaly-detection—sees in many other areas of computer science, where it often results

in large-scale deployments in the commercial world. Examples from other domains include product recommendations systems such as used by Amazon [4] and Netflix [5]; optical character recognition systems (e.g., [6], [7]); natural
language translation [8]; and also spam detection, as an
example closer to home [9].

In this paper we set out to examine the differences between the intrusion detection domain and other areas where machine learning is used with more success. Our main claim is that the task of finding attacks is fundamentally different from other applications, making it significantly harder for the intrusion detection community to employ machine learning effectively. Believe that a significant part of the problem already originates in the premise, found in virtually any relevant textbook, that anomaly detection is suitable for finding novelattacks; Argue that this premise does not hold with the generality commonly implied. Rather, the strength of machine-learning tools is finding activity that is similar to something previously seen, without the need however to precisely describe that activity up front (as misuse detection must).

3.Challenges Of Cloud Computing

The challenges of cloud computing includes:

- A very high cost of errors;
- lack of training data;
- a semantic gap between results and their operational interpretation;
- enormous variability in input data; and
- fundamental difficulties for conducting sound evaluation.

While these challenges may not be surprising for those who have been working in the domain for some time, they can be easily lost on newcomers. To address them, we deem it crucial for any effective deployment to acquire deep, semantic insight into a system's capabilities and limitations, rather than treating the system as a black box as unfortunately often seen. We stress that we do not consider machine-learning an inappropriate tool for intrusion detection. Its use requires care, however: the more crisply one can define the context in which it operates, the better promise the results may hold. Likewise, the better we understand the semantics of the detection process, the more operationally relevant the system will be. Consequently, we also present a set of

guidelines meant to strengthen future intrusion detection research. Throughout the discussion, we frame our mindset around on the goal of using an anomaly detection system effectively in the "real world", i.e., in large-scale, operational environments. We focus on cloud computing intrusion detection as that is our main area of expertise, though we believe that similar arguments hold for host-based systems. For ease of exposition we will use the term anomaly detection somewhat narrowly to refer to detection approaches that rely primarily on machine-learning. By "machine-learning" we mean algorithms that are first trained with reference input to "learn" its specifics (either supervised or unsupervised), to then be deployed on previously unseen input for the actual detection process. While our terminology is deliberately a bit vague, we believe it captures what many in the field intuitively associate with the term "anomaly detection".
We structure the remainder of the paper as follows. In Section II, we begin with a brief discussion of machine learning as it has been applied to intrusion detection in the past. We then in Section III identify the specific challenges machine learning faces in our domain. In Section IV we present recommendations that we hope will help to strengthen future research, and we briefly summarize in Section V.

4. Machine Learning In Intrution Detection

Anomaly detection systems find deviations from expected behavior. Based on a notion of normal activity, they report deviations from that profile as alerts. The basic assumption underlying any anomaly detection system—malicious activity exhibits characteristics not observed for normal usage— was first introduced by Denning in her seminal work on the host-based IDES system [10] in 1987. To capture normal activity, IDES used a combination of statistical metrics and profiles. Since then, many more approaches have been pursued. Often, they borrow schemes from the machine learning community, such as information theory [12], neural networks [13], support vector machines [14], genetic algorithms [15], artificial immune systems [16], and many more. In our discussion, we focus on anomaly detection systems that utilize such machine learning approaches. Chandola et al. provide a survey of anomaly detection in [17], including other areas where similar approaches are used, such as monitoring credit card spending patterns for fraudulent activity. While in such applications one is also looking for outliers, the data tends to be much more structured. For example, the space for representing credit card transactions is of relatively low dimensionality and semantically much more well-defined than network traffic [18]. Anomaly detection

approaches must grapple with a set of well-recognized problems [19]: he detectors tend to generate numerous false positives; attack-free data for training is hard to find; and attackers can evade detection by gradually teaching a system to accept malicious activity as benign. Our discussion in this paper aims to develop a different general point: that much of the difficulty with anomaly detection systems stems from using tools borrowed from the machine learning community in inappropriate ways. Compared to the extensive body of research, anomaly detection has not obtained much traction in the "real world". Those systems found in operational deployment are most commonly based on statistical profiles of heavily aggregated traffic [21]. While highly helpful, such devices operate with a much more specific focus than with the generality that research papers often envision. We see this situation as suggestive that many anomaly detection systems from the academic world do not live up to the requirements of operational settings.

5.Challenges Of Using Machine Learning

In the following identify the unique challenges anomaly detection faces when operating on cloud traffic. We note that our examples from other domains are primarily for illustration, as there is of course a continuous spectrum for many of the properties discussed (e.g., spam detection faces a similarly adversarial environment as intrusion detection does). We also note that we are cloud security researchers, not experts on machine-learning, and thus we argue mostly at an intuitive level rather than attempting to frame our statements in the formalisms employed for machine learning. However, based on discussions with colleagues who work with machine learning on a daily basis, we believe these intuitive arguments match well with what a more formal analysis would yield.

5.1.Outlier Detection

Fundamentally, machine-learning algorithms excel much better at finding similarities than at identifying activity that does not belong there: the classic machine learning application is a classification problem, rather than discovering meaningful outliers as required by an anomaly detection system [22]. Consider product recommendation systems such as that used by Amazon [4]: it employs collaborative altering; matching each of a user's purchased (or positively rated) items with other similar products, where similarity is determined by products that tend is bought together. In some sense, outlier detection is also a classification problem: there are two classes, "normal" and "not

normal", and the objective is determining which of the two more likely matches an observation. However, a basic rule of machine-learning is that one needs to train a system with specimens of all classes, and, crucially, the number of representatives found in the training set for each class should be large [23].

6.2.High Cost of Errors

In intrusion detection, the relative cost of any misclassification is extremely high compared to many other machine learning applications. A false positive requires spending expensive analyst time examining the reported incident only to eventually determine that it reflects benign underlying activity. As argued by Axelsson, even a very small rate of false positives can quickly render a CIDS (Cloud Intrusion Detection System) unusable [24]. False negatives, on the other hand, have the potential to cause serious damage to an organization: even a single compromised system can seriously undermine the integrity of the IT infrastructure.

6.3.Semantic Gap

Anomaly detection systems face a key challenge of transferring their results into actionable reports for the network operator. In many studies, we observe a lack of this crucial final step, which we term the semantic gap. Unfortunately, in the intrusion detection community we find a tendency to limit the evaluation of anomaly detection systems to an assessment of a system's capability to reliably identify deviations from the normal profile. While doing so indeed comprises an important ingredient for a sound study, the next step then needs to interpret the results from an operator's point of view—"What does it mean?"

Answering this question goes to the heart of the difference between findings "abnormal activity" and "attacks". Those familiar with anomaly detection are usually the first to acknowledge that such systems are not targeting to identify malicious behavior but just report what has not been seen before, whether benign or not.

When addressing the semantic gap, one consideration is the incorporation of local security policies. While often neglected in academic research, a fundamental observation about operational networks is the degree to which they differ: many security constraints are a site-specific property. Activity that is fine in an academic setting can be banned in

an cloud enterprise network; and even inside a single organization, department policies can differ widely. Thus, it is crucial for a CIDS to accommodate such differences.

6.4.Diversity of Network Traffic

Network traffic often exhibits much more diversity than people intuitively expect, which leads to misconceptions about what anomaly detection technology can realistically achieve in operational environments. Even within a single network, the network's most basic characteristics—such as bandwidth, duration of connections, and application mix—can exhibit immense variability, rendering them unpredictable over short time intervals (seconds to hours).

Finally, we note that traffic diversity is not restricted to packet-level features, but extends to application-layer information as well, both in terms of syntactic and semantic variability.

7.Recommendations For Using

7.1.Machine Learning

In light of the points developed above, we now formulate guidelines that we hope will help to strengthen future research on anomaly detection. We note that we view these guidelines as touchstones rather than as firm rules; there is certainly room for further discussion within the wider intrusion detection community. If we could give only one recommendation on how to improve the state of anomaly detection research, it would be: Understand what the system is doing.

7.2.Understanding the Threat Model

Before starting to develop an anomaly detector, one needs to consider the anticipated threat model, as that establishes the framework for choosing trade-offs. Questions to address include:

7.2.1.<u>What Kind Of Environment Does The System Target?</u>

Operation in a small cloud network faces very different challenges than for a large enterprise or backbone network; academic environments impose different requirements than commercial enterprises.

7.2.2.What Do Miss Attacks Cost?

Possible answers ranges from "very little" to "lethal." A site's determination will depend on its security demands as well as on other deployed attack detectors.

7.2.3.Keeping The Scope Narrow

It is crucial to have a clear picture of what problem a system targets: what specifically are the attacks to be detected? The more narrowly one can define the target activity; the better one can tailor a detector to its specifics and reduce the potential for misclassifications. Of course machine-learning is not a "silver bullet" guaranteed to appropriately match a particular detection task. Thus, after identifying the activity to report, the next step is a neutral assessment of what constitutes the right sort of tool for the task; in some cases it will be an anomaly detector, but in others a rule-based approach might hold more promise. A common pitfall is starting with the premise to use machine learning (or, worse, a particular machine-learning approach) and then looking for a problem to solve.

7.3.Reducing the Costs

Anecdotally, the number one complaint about anomaly detection systems is the excessive number of false positives they commonly report.

7.4.Evaluation

- Working with data: The single most important step for sound evaluation concerns obtaining appropriate data to work with. The "gold standard" here is obtaining access to a dataset containing real network traffic from as large an environment as possible; and ideally multiple of these from different networks. Work with actual traffic greatly strengthens a study, as the evaluation can then demonstrate how well the system should work in practice. In our experience, the best way to obtain such data is to provide a clear benefit in return to the network's operators; either, ideally, by research that aims to directly help to improve operations, or by exchanging the access for work on an unrelated area of importance to the operators.
- Understanding results: The most important aspect of interpreting results is to understand their origins. A sound evaluation frequently requires relating input

and output on a very low-level. Researchers need to manually examine false positives. If when doing so one cannot determine why the system incorrectly reported a particular instance, this indicates a lack of insight into the anomaly detection system's operation.

8.Conclusion

The security of cloud computing is a new research area requiring more input from both the academic and industrial communities. Our work examines the surprising imbalance between the extensive amounts of research on machine learning-based anomaly detection pursued in the academic intrusion detection community, versus the lack of operationaldeployments of such systems. We argue that this discrepancy stems in large part from specifics of the problem domain that make it significantly harder to apply machine learning effectively than in many other areas of computer science where such schemes are used with greater success.

To overcome the challenges of cloud computing, provide a set of guidelines for applying machine learning to cloud intrusion detection. In particular, we argue for the importance of obtaining insight into the operation of an anomaly detection system in terms of its capabilities and limitations from an operational point of view. It is crucial to acknowledge that the nature of the domain is such that one can always find schemes that yield marginally better curves than anything else has for a specific given setting. Such results however do not contribute to the progress of the field without any semantic understanding of the gain. We hope for this discussion to contribute to strengthening future research on anomaly detection by pinpointing the fundamental challenges it faces. We stress that we do not consider our discussion as final, and we look forward to then intrusion detection community engaging in an ongoing dialog on this topic.

9.Acknowledgment

The making of the paper needed co-operation and guidance of a number of people. We therefore consider it our prime duty to thank all those who had helped us for making it successful. It is our immense pleasure to express our gratitude to Mr. K. Ananda kumar (Associate professor) as a guide who provided us constructive and positive feedback during the preparation of this paper. Last but not least, we are thankful to our friends and library staff members whose encouragement and suggestion helped us to complete our seminar. We are also thankful to our parents.

7.Reference

1. R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina, "Controlling data in the cloud: outsourcing computation without outsourcing control," in Proceedings of the 2009 ACM workshop on Cloud computing security, pp. 85–90, ACM, 2009.
2. R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599–616, 2009.
3. K. Pemmaraju, "Leaders in the cloud: Identifying the business value of cloud computing for customers and vendors," Mar 2010. available at: http://www.sandhill.com/opinion/editorial.php?id=296.
4. G. Linden, B. Smith, and J. York, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76–80, 2003.
5. J. Bennett, S. Lanning, and N. Netflix, "The Netflix Prize," in Proc. KDD Cup and Workshop, 2007.
6. L. Vincent, "Google Book Search: Document Understanding on a Massive Scale," 2007.
7. R. Smith, "An Overview of the Tesseract OCR Engine," in Proc. International Conference on Document Analysis and Recognition, 2007.
8. P. Graham, "A Plan for Spam," in Hackers & Painters. O'Reilly, 2004.
9. D. E. Denning, "An Intrusion-Detection Model," IEEE Transactions on Software Engineering, vol. 13, no. 2, pp. 222–232, 1987.
10. H. S. Javitz and A. Valdes, "The NIDES Statistical Component: Description and Justification," SRI International, Tech. Rep., 1993.
11. W. Lee and D. Xiang, "Information-Theoretic Measures for Anomaly Detection," in Proc. IEEE Symposium on Security and Privacy, 2001.
12. Z. Zhang, J. Li, C. Manikopoulos, J. Jorgenson, and J. Ucles, "HIDE: a Hierarchical Network Intrusion Detection System Using Statistical Preprocessing and Neural Network Classification," in Proc. IEEE Workshop on Information Assurance and Security, 2001.
13. W. Hu, Y. Liao, and V. R. Vemuri, "Robust Anomaly Detection Using Support Vector Machines," in Proc. International Conference on Machine Learning, 2003.

14. C. Sinclair, L. Pierce, and S. Matzner, “An Application of Machine Learning to Network Intrusion Detection,” in Proc. Computer Security Applications Conference, 1999.
15. S. A. Hofmeyr, “An Immunological Model of Distributed Detection and its Application to Computer Security,” Ph.D. dissertation, University of New Mexico, 1999.
16. V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” University of Minnesota, Tech. Rep., 2007.
17. R. J. Bolton and D. J. Hand, “Statistical Fraud Detection: A Review,” Statistical Science, vol. 17, no. 3, 2002.
18. C. Gates and C. Taylor, “Challenging the Anomaly Detection Paradigm: A Provocative Discussion,” in Proc. Workshop on New Security Paradigms, 2007.
19. “Peakflow SP,” http://www.arbornetworks.com/en/ peakflow-sp.html.
20. “StealthWatch,” http://www.lancope.com/products/.
21. I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques (2nd edition). Morgan Kaufmann, 2005.
22. R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification (2nd edition). Wiley Interscience, 2001.
23. V. Paxson, “Bro: A System for Detecting Network Intruders in Real-Time,” Computer Networks, vol. 31, no. 23–24, pp. 2435–2463, 1999.
24. P. Gill, M. Arlitt, Z. Li, and A. Mahanti, “YouTube Traffic Characterization: A View From the Edge,” in Proc. ACM SIGCOMM Internet Measurement Conference, 2008.
25. Sebastian Roschke, Feng Cheng, Christoph Meinel,“ Intrusion Detection in the Cloud", Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009.
26. Chi-Chun Lo, Chun-Chieh Huang, Joy Ku, “A Cooperative Intrusion Detection System Framework for Cloud Computing Networks", 39th International Conference on Parallel Processing Workshops, 2010.
27. Andreas Haeberlen,“ An Efficient Intrusion Detection Model Based on Fast Inductive Learning”, Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007.

28. Richard Chow, Philippe Golle, Markus Jakobsson, "Controlling Data in the Cloud: Outsourcing Computation without Outsourcing Control", ACM Computer and Communications Security Workshop, CCSW 09, November 13, 2009.
29. Kleber, schulter, "Intrusion Detection for Grid and Cloud Computing", IEEE Journal: IT Professional, 19 July 2010.