Systematization of Metrics in Intrusion Detection Systems

Yufan Huang, Xiaofan He and Huaiyu Dai
Department of ECE
North Carolina State University, USA
Email: {yhuang20,xhe6,hdai@ncsu.edu}

ABSTRACT

Intrusion detection assumes paramount importance in this information era due to its capability of providing security protection to information systems. In addition to advancing the specific intrusion detection techniques, substantial efforts have been devoted to the taxonomy of existing IDSs, mostly focusing on the methodology, audit source and architecture aspects. The employed metric is another decisive factor of IDS performance, yet a systematized understanding in this aspect is still lacking. As an initial effort towards this objective, a categorization of IDS metrics is proposed in this work, where existing IDS metrics are divided into four types - information theoretic, probabilistic, proximity-based, and reliability-based metrics. Simulation studies of several intrusion detection algorithms that match the proposed categorization are also conducted based on the KDD'99 dataset.

Keywords

Intrusion Detection Systems, KDD'99, Systematization

1. INTRODUCTION

While bringing unprecedented convenience to our lives, information systems are vulnerable to various forms of adversarial intrusions, leaving their security an ever-present concern. As an effective approach to provide security protection to information systems through monitoring the system/network usage for malicious activities and policy violations, intrusion detection systems (IDSs) have assumed paramount importance in this information era and thus have received extensive research efforts in the past two decades (e.g., [1] and the references therein).

In literature, substantial efforts have also been devoted to the taxonomy of IDSs [2]. However, they mainly focus on methodology, audit source, and architecture; while systematization of IDSs in terms of the employed metric that introduces quantifiable effectiveness to the intrusion detection process is equally important but less investigated. In this work, a categorization of IDS metrics is proposed as an initial effort towards this objective. Particularly, existing IDS metrics are divided into four categories: 1) information theoretic metrics which quantify the uncertainty in the observed data using information theoretical measures, 2) probabilistic metrics that are built upon quantities from probability and detection theory, 3) proximity-based metrics that concern about the closeness between different data points, and 4) reliability-based metrics which weight observation quality and trustworthiness according to a set of pre-specified rules. Each of these four types is illustrated with corresponding IDSs in literature, and finer categorization within each type is also provided whenever possible. Moreover, numerical studies of several representative algorithms using these metrics are conducted over the KDD'99 dataset [3].

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

Copyright is held by the owner/author(s).
HotSoS '15 Apr 21-22, 2015, Urbana, IL, USA
ACM 978-1-4503-3376-4/15/04.
http://dx.doi.org/10.1145/2746194.2746222.

2. TAXONOMY OF IDS METRICS

This section presents the proposed categorization of existing IDS metrics (information theoretical, probabilistic, proximity-based, and reliability-based), and discusses existing intrusion detection metrics belonging to these categories.

2.1 Information Theoretical IDS Metrics

Information theoretical metrics quantify the uncertainty in the observed data using information theoretical measures. The performance of many IDSs rely crucially on a proper feature selection, and the information-theoretic metrics (e.g., entropy and information gain) have been proven to be effective for this purpose. In addition, more involved information theoretic metrics built upon compression algorithms have also been used in existing IDSs. For example, the compression-based dissimilarity [4], which describes the ratio of the codeword lengths with and without the testing sample included, is used as a metric for intrusion detection. The information density, which is defined as the ratio between the extra bits needed for the sampling data by the grammar-based lossless compression algorithm and the length of the sampling data, is used as intrusion detection metric in [5].

2.2 Probabilistic IDS Metrics

This type of IDS metrics are often built upon quantities based on probability and detection theory. Occurrence frequency based IDS metrics are the simplest of this type which simply count the number of appearances of certain data of interest. The inverse of occurrence frequency of each testing data point in the normal database is used as an anomaly score in [6], and containing many testing data

points with high scores will be treated as a sign of intrusion in a testing dataset. In addition to the occurrence frequency, likelihood and posterior are another two fundamental probabilistic quantities that have been widely used to build IDS metrics. Taking the logarithm of the likelihood, the resulting log-likelihood of a testing data point with respect to (w.r.t.) a trained probabilistic model is used as the metric for anomaly detection in [7]. Moreover, there are also a family of probabilistic IDS metrics built upon the celebrated Markov models. In [8], the likelihood of a testing sequence w.r.t. the trained Markov model in the normal phase is used to detect intrusion.

2.3 Proximity-based IDS Metrics

There are also a notable amount of IDS metrics concerning the closeness between different data points, and these IDS metrics can be categorized as proximity-based IDS metrics. In existing IDSs, various forms of proximities have been considered. The simplest form of proximity would be the distance from a (vector) point to another object, which can be a (set of) point(s) in the normal space, or the normal space itself. Also, the proximity can be measured by the distance between two distributions. In [9], intrusion is detected by computing a similarity score between the probability density functions (PDF) of the normal dataset and the testing dataset. Several different similarity metrics can be included, such the Kolmogorov-Smirnov (KS) statistic and a deviation from the mean statistic. Metrics that measure the proximity between two data sequences have also been employed by existing IDSs. The clustering based IDSs discussed in [10] use the length of the longest common subsequence, average distance to other samples, distance to the nearest neighbor, the edit distance and the hamming distance as metrics for intrusion detection.

2.4 Reliability-based IDS Metrics

Reliability-based metrics weight observation quality and trustworthiness according to a set of rules that are usually specified by human experts or learned through data mining. In [11], consistency w.r.t. data dependency rules is used to identify malicious access to a database, and the mismatch rate of a testing sequence of system calls w.r.t. the patterns recorded during training phase is used as a metric for intrusion detection. In [12], consistency w.r.t. trained fuzzy rules is used to detect abnormal behaviors.

3. SIMULATIONS AND CONCLUSIONS

The ID3 and C4.5, Naive Bayes (NB) and Tree Augmented Naive (TAN) Bayes, K-mean (K) and Y-mean (Y) clustering algorithms that correspond to the proposed information theoretic, probabilistic, and proximity based metrics, respectively, are examined over the KDD'99 dataset, and the corresponding results are presented in Table 1.

In this work, based on a survey of existing IDSs, a categorization of IDS metrics is proposed where IDS metrics are divided into four types - information theoretic, probabilistic, proximity-based, and reliability-based metrics. Six intrusion detection algorithms are chosen to match these metric categories and our simulation results show that these metrics and the corresponding algorithms provide similar performance over the KDD'99 dataset.

Table 1: Detection Accuracy

%	ID3	C4.5	NB	TAN	K	Y
Normal	92.79	96.33	98.13	98.56	97.56	89.89
DoS	96.03	97.02	95.74	97.10	97.31	91.17
R2L	2.24	4.58	0.56	3.15	6.43	5.19
U2R	3.07	1.75	3.51	7.02	29.82	10.96
Probe	86.46	80.82	72.25	78.90	87.54	75.25

4. REFERENCES

[1] T. F. Lunt, "A survey of intrusion detection techniques," *Computers and Security*, vol. 12, no. 4, pp. 405–418, 1993.

[2] H. Debar, M. Dacier, and A. Wespi, "Towards a taxonomy of intrusion-detection systems," *Computer Networks*, vol. 31, no. 8, pp. 805–822, 1999.

[3] KDD99 data set. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[4] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards parameter-free data mining," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2004, pp. 206–215.

[5] N. Wang, J. Han, and J. Fang, "An anomaly detection algorithm based on lossless compression," in *2012 IEEE 7th International Conference on Networking, Architecture and Storage (NAS).* IEEE, 2012, pp. 31–38.

[6] T. Lane and C. E. Brodley, "Temporal sequence learning and data reduction for anomaly detection," *ACM Transactions on Information and System Security (TISSEC)*, vol. 2, no. 3, pp. 295–331, 1999.

[7] S. Staniford, J. A. Hoagland, and J. M. McAlerney, "Practical automated detection of stealthy portscans," *Journal of Computer Security*, vol. 10, no. 1, pp. 105–136, 2002.

[8] N. Ye *et al.*, "A Markov chain model of temporal behavior for anomaly detection," in *Proceedings of the 2000 IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*, vol. 166. West Point, NY, 2000, p. 169.

[9] J. Li and C. Manikopoulos, "Early statistical anomaly intrusion detection of DoS attacks using mib traffic parameters," in *Information Assurance Workshop, 2003. IEEE Systems, Man and Cybernetics Society.* IEEE, 2003, pp. 53–59.

[10] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 823–839, 2012.

[11] Y. Hu and B. Panda, "A data mining approach for database intrusion detection," in *Proceedings of the 2004 ACM symposium on Applied computing.* ACM, 2004, pp. 711–716.

[12] A. Abraham, R. Jain, J. Thomas, and S. Y. Han, "D-scids: Distributed soft computing intrusion detection system," *Journal of Network and Computer Applications*, vol. 30, no. 1, pp. 81–98, 2007.