A new Data Mining-based Approach to Improving the Quality of Alerts in Intrusion Detection Systems

Hadi Barani Baravati [1], Javad Hosseinkhani[2], Solmaz Keikhaee[2], Meisam Ostad Hossein Khayat [2], and Malek Havasi[2]

Department of Computer, Iranshahr Branch, Islamic Azad University, Iranshahr, Iran[1]
Department of Computer, Damavand Branch, Islamic Azad University, Damavand, Iran[2]

Summary
Data mining is about finding insights which are statistically reliable, unknown previously, and actionable from data. This data must be available, relevant, adequate, and clean. Also, the data mining problem must be well-defined, cannot be solved by query and reporting tools, and guided by a data mining process model Thus it is essential to use different security tools in order to protect computer systems and networks. Among these tools, Intrusion Detection Systems (IDSs) are one of the components of Defense-in-depth. One major drawback of IDSs is the generation of a huge number of alerts, most of which are false, redundant, or unimportant. Among different remedy approaches, many researchers proposed the use of data mining. Most of the research done in this area could not address the problems completely. Also, most of them suffer from human dependency and offline functionality. In this research, an online approach is proposed in order to manage alerts issued by IDSs. The proposed approach is able to process alerts produced by heterogeneous IDS systems. The approach is evaluated using DARPA 1999 dataset and Shahid Rajaee Port Complex dataset. Evaluation results show that the proposed approach can reduce the number of alerts by 94.32%, effectively improving alert management process. Because of the utilization of ensemble methodology and ideal algorithms in the proposed methodology, it can advise network security specialist the talk about of the monitored network within an online manner.

Key words:
Web Data Mining, Quality of Alerts, Data Mining, Intrusion Detection.

1. Introduction

Intrusion detection corresponds to a collection of techniques that are being used to identify disorders against pcs and network infrastructures. As the cost of the information processing and Internet accessibility falls, more and more organizations are becoming vulnerable to a wide variety of cyber threats. According to a recent survey by CERT/CC [1], the rate of cyber-attacks has been more than doubling every year in recent times. Therefore, it has become increasingly important to make our information systems, especially those used for critical functions in the military and commercial sectors, resistant to and tolerant of such attacks. The most widely deployed methods for detecting cyber terrorist attacks and protecting against cyber terrorism employ signature-based detection techniques. Such methods can only detect previously known attacks that have a corresponding signature, since the signature database has to be manually revised for each new type of attack that is discovered. These limitations have led to an increasing interest in intrusion detection techniques based on data mining [2, 3, 4, 5, and 6].

Data mining established intrusion recognition techniques generally belong to 1 of 2 categories; misuse diagnosis and anomaly detection. In misuse detection, each instance in a data set is labeled as ‘normal’ or ‘intrusive’ and a learning algorithm is trained over the labeled data. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks, as long as they have been labeled appropriately. Research in misuse detection has focused mainly on classification of network intrusions using various standard data mining algorithms [2, 4, 5, and 6], rare class predictive models, association rules [7, 8] and cost sensitive modeling. Unlike signature based intrusion detection systems, models of misuse are created automatically, and can be more sophisticated and precise than manually created signatures. A key advantage of misuse detection techniques is their high degree of accuracy in detecting known attacks and their variations. Their obvious drawback is the inability to detect attacks whose instances have not yet been observed [23]. The term fraud here refers to the abuse of a profit organization’s system without necessarily leading to direct legal consequences. In a competitive environment, fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe. Fraud detection, being part of the overall fraud control, automates and helps reduce the manual parts of a screening/checking process. This area has become one of the most established industry/government data mining applications. It is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option

Manuscript received August 5, 2017
Manuscript revised August 20, 2017

is to tease out possible evidences of fraud from the available data using mathematical algorithms.

2. Learning from Rare Classes

In misuse detection related problems, standard data mining techniques are not applicable due to several specific details that include dealing with skewed class distribution, learning from data streams and labeling network connections. The problem of skewed class distribution in the network intrusion detection is very apparent since intrusion as a class of interest is much smaller i.e. rarer than the class representing normal network behavior [21, 22]. In such scenarios when the normal behavior may typically represent 98-99% of the entire population a trivial classifier that labels everything with the majority class can achieve 98-99% accuracy. It is apparent that in this case classification accuracy is not sufficient as a standard performance measure.

3. Related Works of Anomaly Detection Techniques

Most research in supervised anomaly detection can be considered as performing generative modeling. These approaches attempt to build some kind of a model over the normal data and then check to see how well new data fits into that model. An approach for modeling normal sequences using look ahead pairs and contiguous sequences is presented in [13]. A statistical method for ranking each sequence by comparing how often the sequence is known to occur in normal traces with how often it is expected to occur in intrusions is presented in [14]. One approach uses a prediction model obtained by training decision trees over normal data [9], while others use neural networks to obtain the model [15] or non-stationary models [16] to detect novel attacks. Lane and Brodley [17] performed anomaly detection on unlabeled data by looking at user profiles and comparing the activity during an intrusion to the activity during normal use. Similar approach of creating user profiles using semi-incremental techniques was also used in [18]. Barbara used pseudo-Bayes estimators to enhance detection of novel attacks while reducing the false alarm rate as much as possible [10]. A technique developed at SRI in the EMERALD system [11] uses historical records as its normal training data. It then compares distributions of new data to the distributions obtained from those historical records and differences between the distributions indicate an intrusion. Recent works such as [19] and [20] estimate parameters of a probabilistic model over the normal data and compute how well new data fits into the model.

The proposed approach is able to process alerts produced by heterogeneous IDS systems. The approach is evaluated using DARPA 1999 dataset and Shahid Rajaee Port Complex dataset. Evaluation results show that the proposed approach can reduce the number of alerts by 94.32%, effectively improving alert management process. Because of the use of ensemble approach and optimal algorithms in the proposed approach, it can inform network security specialist the state of the monitored network in an online manner.

4. Proposed Work and Experiments

This research has been applied the proposed detection schemes to 1999 DARPA Intrusion Detection Evaluation Data [12] as well as to the real network data from the Web. The DARPA'99 data contains two types: training data and test data. The training data consists of 7 weeks of network-based attacks inserted in the normal background data.

Attacks in training data are labeled. The test data contained 2 weeks of network-based attacks and normal background data. 7 weeks of data resulted in about 5 million connection records. Although DARPA'99 evaluation represents a significant advance in the field of intrusion detection, there are many unresolved issues associated with its design and execution. In his critique of DARPA evaluation, starting from usage of synthetic simulated data for the background (normal data) and using attacks implemented via scripts and programs collected from a variety of sources. In addition, it is known that the background data contains none of the background noise (packet storms, strange fragments …) that characterize real data. However, in the lack of better benchmarks, vast amount of the research is based on the experiments performed on this data. The evaluation of any intrusion detection algorithm on real network data is extremely difficult mainly due to the high cost of obtaining proper labeling of network connections. Figure 1 shows that Architecture proposed approach for managing Alerts. The main reason for this procedure is to associate new constructed features with the connection records from "list files" and to create more informative data set for learning.

Since the amount of available data is huge (e.g. some days have several million connection records), this work has sampled sequences of normal connection records in order to create the normal data set that had the same distribution as the original data set of normal connections. This Research has used this normal data set for training our

anomaly detection schemes, and then examined how well the attacks may be detected using the proposed schemes.

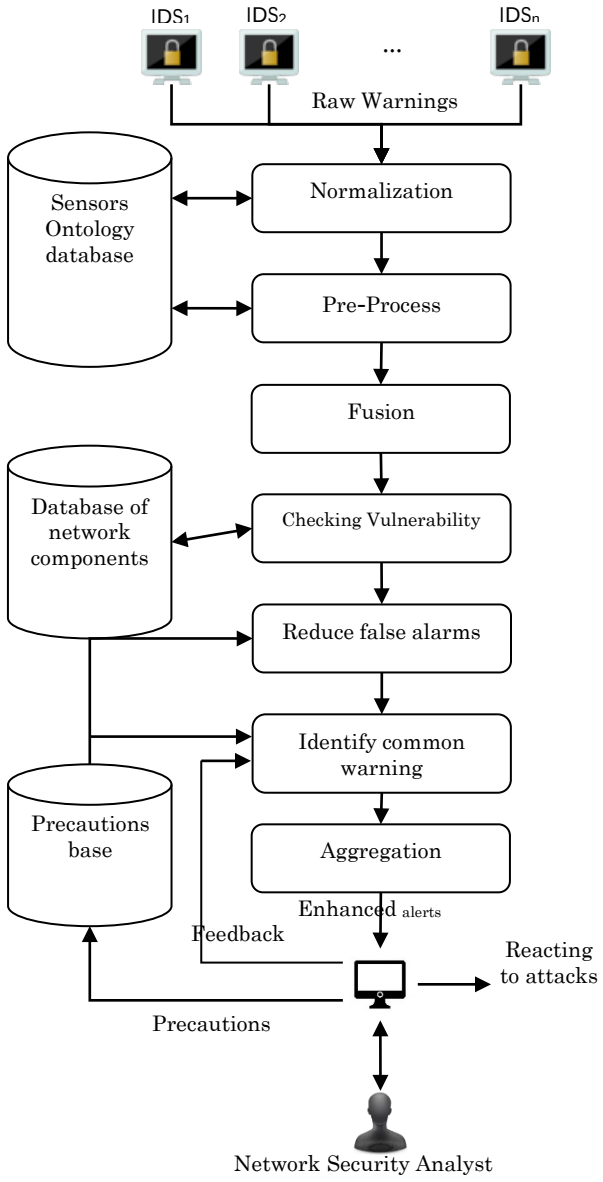

Fig. 1 Architecture Proposed Approach for Managing Alerts.

Algorithm 1 shows that Sub-component code Normalization.

```
Normalize(rawAlert) {
  alert = new Alert();
  alert.AID =
  AttackNamesDB.GetAttackName(rawAlert.Nam
  e, rawAlert.Sensor.Type);
  foreach(attribute in Alert.Attributes) {
    alert.attribute.Value =
    NormalizationDB.GetStandardAttributeValue(
    attribute, rawAlert);
  }
  send alert to next component;
}
```

Algorithm 1. Sub-Component Code Normalization

Table 1: Evaluating the Results of the Proposed Approach in Data Collection Darpa 1999

The fifth week	The fourth week	The third week	The second week	The first week	
6090	260	82	7512	104	*alerts*() #
4385	219	82	464	104	*FP* # *alerts*()
12	10	9	12	18	*output* # *alerts*()
2	5	9	6	18	*output* # *FP alerts*()
%99.80	%96.15	%89.02	%99.84	%82.69	(%) *RR*
%99.95	%97.72	%89.02	%98.71	%82.69	(%) *FPRR*

Table 1 report on additional metrics for evaluating the results of the proposed approach in data collection DARPA 1999. The table shows that results of the proposed approach at the dataset DARPA 1999 on bursty attacks

5. Conclusion and Future Works

Many anomaly diagnosis techniques intended for revealing circle intrusions tend to be offered with this report. To aid applicability involving anomaly diagnosis techniques, an activity intended for extracting practical statistical written content based along with temporal attributes is additionally applied. Experimental results performed on DARPA 98 data set indicate that the most successful anomaly detection techniques were able to achieve the detection rate of 74% for attacks involving multiple connections and detection rate of 56% for more complex single connection attacks,

while keeping the false alarm rate at 2%. When the false alarm rate is increased to 4%, the achieved detection rate reaches 89% for bursty attacks and perfect 100% for single-connection attacks. Computed ROC curves indicate that the most promising technique for detecting intrusions in DARPA'99 data is the LOF approach. In addition, when performing experiments or real network data, the LOF approach was very successful in picking several very interesting novel attacks. Considering the DARPA'99 data, performed experiments also demonstrate that for different types of attacks, different anomaly detection schemes were more successful than others. For example, the unsupervised SVMs were very promising in detecting new intrusions since they had very high detection rate but very high false alarm rate too. Therefore, future work is needed in order to keep high detection rate while lowering the false alarm rate.

Our long-term goal is to develop an overall framework for defending against attacks and threats to computer systems. Although our developed techniques are promising in detecting various types of intrusions they are still preliminary in nature. Data generated from network traffic monitoring tends to have very high volume, dimensionality and heterogeneity, making the performance of serial data mining algorithms unacceptable for on-line analysis. Therefore, development of new anomaly detection algorithms that can take advantage of high performance computers is a key component of this project. According to our preliminary results on real network data, there is a significant non-overlap of our anomaly detection algorithms with the SNORT intrusion detection system, which implies that they could be combined in order to increase coverage. The approach is looked at utilizing DARPA 1999 dataset and Shahid Rajaee port intricate dataset. Evaluation outcomes display that the suggested approach may reduce the amount of alerts by simply 94. 32%, properly increasing notify management procedure. Because of the by using attire approach and optimum algorithms inside suggested approach, it might enlighten circle safety measures practitioner their state on the monitored circle in an on the internet approach.

Acknowledgments

This research is funded by the Iranshahr Branch in Islamic Azad University, Iranshahr, Iran. The authors would like to thank the Research Management Centre of Islamic Azad University-Iranshahr Branch and cooperation including Lecturers and other individuals who are either directly or indirectly involved in this project.

References

[1] Agarwal, A., Johri, S., Agarwal, A., Tyagi, V. & Kumar, A. (2012). "Multi Agent Based Approach For Network Intrusion Detection Using Data Mining Concept". Journal of Global Research in Computer Science, 3(3), 29-32.

[2] Agrawal, R., Imielinski, T. & Swami, A. (1993). "Mining association rules between sets of items in large databases". Paper presented at the Proceedings of the 1993 ACM SIGMOD international conference on Management of data, Washington, D.C., USA.

[3] Al-Mamory, S. O. & Zhang, H. (2010). "New data mining technique to enhance IDS alarms quality". Journal in Computer Virology, 6(1), 43-55.

[4] Alpaydin, E. (2010). "Introduction to Machine Learning" (2nd ed.). Cambridge, Massachusetts, London, England: The MIT Press.

[5] Anderson, J. P. (1980). "Computer security threat monitoring and surveillance": Technical report, James P. Anderson Company, Fort Washington, Pennsylvania.

[6] Bace, R. & Mell, P. (2001). "NIST special publication on intrusion detection systems": DTIC Document.

[7] Balthrop, J., Forrest, S. & Glickman, M. R. (2002). "Revisiting LISYS: parameters and normal behavior". Paper presented at the Evolutionary Computation. CEC '02. Proceedings of the 2002 Congress on.

[8] Bloedorn, E., Christiansen, A. D., Hill, W., Skorupka, C., Talbot, L. M. & Tivel, J. (2001)."Data mining for network intrusion detection: How to get started": MITRE Technical Report.

[9] Brugger, S. T. (2004). "Data mining methods for network intrusion detection". University of California at Davis. Retrieved 7/24/2013, from http://neuro.bstu.by/ai/To-dom/My_research/failed%201%20subitem/For-research/D-mining/Anomaly-D/Intrusion-detection/brugger-dmnid.pdf

[10] BugTraq. Retrieved 7/24/2013, from http://www.securityfocus.com/archive/1

[11] Cheng-Yuan, H., Yuan-Cheng, L., Chen, I. W., Fu-Yu, W. & Wei-Hsuan, T. (2012). "Statistical analysis of false positives and false negatives from real traffic with intrusion detection/prevention systems". Communications Magazine, IEEE, 50(3), 146-154.

[12] Vulnerabilities, CVE Common. "Exposures (2014)." Common Vulnerabilities and Exposures (2014).

[13] DARPA. (1998). "MIT Lincoln Laboratory: Communications & Information Technology". Retrieved 7/24/2013, from http://www.ll.mit.edu/mission/communications/ist/index.html

[14] Denning, D. E. (1987). "An Intrusion-Detection Model". Software Engineering, IEEE Transactions on, SE-13(2), 222-232.

[15] Dongre, S. S. & Wankhade, K. K. (2012). "Intrusion Detection System Using New Ensemble Boosting Approach". International Journal of Modeling and Optimization, 2(a).

[16] EMERALD. (1996). "Event Monitoring Enabling Responses to Anomalous Live Disturbances (EMERALD)". Retrieved 7/24/2013, from http://www.sdl.sri.com/projects/emerald/

[17] Mahoney, M. V. & Chan, P. K. (2003). "An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection". In G. Vigna, C. Kruegel & E. Jonsson (Eds.), Recent Advances in Intrusion Detection 2820, 220-237: Springer Berlin Heidelberg.
[18] McHugh, J. (2000). "Testing Intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory". ACM Transactions on Information and System Security, 3(4), 262-294.
[19] Modi, C., Patel, D., Borisaniya, B., Patel, H., Patel, A. & Rajarajan, M. (2013). "A survey of intrusion detection techniques in Cloud". Journal of Network and Computer Applications, 36(1), 42-57.
[20] Mohammed, R. G. & Awadelkarim, A. M. (2011). "Design and Implementation of a Data Mining-Based Network Intrusion Detection Scheme". Asian Journal of Information Technology, 10(4), 136-141.
[21] Ektefa, Mohammadreza, Sara Memar, Fatimah Sidi, and Lilly Suriani Affendey. "Intrusion detection using data mining techniques." In Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on, pp. 200-203. IEEE, 2010.
[22] Nmap. "Nmap - Free Security Scanner For Network Exploration & Security Audits". Retrieved 7/24/2013, from http://nmap.org/
[23] Okamoto, T. (2011). "An artificial intelligence membrane to detect network intrusion". Artificial Life and Robotics, 16(1), 44-47.