Effective Anomaly Detection Using Hidden-Semi Markov Model

KhaingShweWutyi, Mie Mie Su Thwin
University of Computer Studies Mandalay
wutyi.pan@gmail.com

Abstract

Anomaly detection studies the normal behavior of the monitored system and then looks out for any difference in it to detect anomalies or attacks. It is able to detect new attacks as any attack is assumed to be different from normal activity. It sometimes sets false alarms because it erroneously classifies the normal user behaviors as attacks.Different techniques have been used for anomaly detector generation.In this paper, we would like to propose Hidden-Semi Markov Model (HSMM) as it is introduced in intrusion detection for several years. Based on this HSMM, an algorithm of anomaly detection is presented in this paper, which computes the distance between the processes monitored by intrusion detection system and the perfect normal processes. In this algorithm, we use the average information entropy (AIE) of fixed-length observed sequence as the anomaly detection metric based on maximum entropy principle (MEP). To improve accuracy, the segmental K-means algorithm is applied as training algorithm for the HSMM. By comparing the accurate rate with the experimental results of previous research, it shows that our method can perform a more accurate detection.

Keywords: *Intrusion detection, Anomaly detection, Hidden semi-Markov model (HSMM), Maximum entropy principle (MEP), Segmental K-means algorithm*

1. Introduction

With the enormous growth of computer networks usage and the huge increase in the number of applications running on top of it, network security is becoming increasingly more important. As it is shown in [1], all the computer systems suffer from security vulnerabilities which are both technically difficult and economically costly to be solved by the manufacturers.

Therefore, the role of Intrusion Detection Systems (IDSs), as special-purpose devices to detect anomalies and attacks in the network, is becoming more important. The research in the intrusion detection field has been mostly focused on anomaly-based and misuse-based detection techniques for a long time. While misuse-based detection is generally favored in commercial products due to its predictability and high accuracy, in academic research anomaly detection is typically conceived as a more powerful method due to its theoretical potential for addressing novel attacks.

Conducting a thorough analysis of the recent research trend in anomaly detection, one will encounter several machine learning methods reported to have a very high detection rate of 98% while keeping the false alarm rate at 1% [2]. However, when we look at the state of the art IDS solutions and commercial tools, there is few products using anomaly detection approaches, and practitioners still think that it is not a mature technology yet. To find the reason of this contrast, we studied the details of the research done in anomaly detection and considered various aspects such as learning and detection approaches, training data sets, testing data sets, and evaluation methods. Our study shows that there are some inherent problems like redundant records in the train data set of KDDCUP'99 data set [3], which is widely used as one of the few publicly available data sets for network-based anomaly detection systems.

The new version of KDD data set, NSL-KDD is publicly available now. Although, the data set still suffers from some of the problems discussed by McHugh [4] and may not be a perfect representative of existing real networks, because of the lack of public data sets for network-based IDSs.

The research analysis for anomaly detection fully based on several machine learning methods on various training and testing dataset [2]. Firstly, our study analyze the inherent problems in KDDcup 99 dataset and we found the better solution that our study should base on NSL-KDD dataset for finding accuracy in intrusion detection.

Hidden Markov model (HMM) has been introduced into intrusion detection field for many years and has achieved many satisfying but the major weakness of HMM lies in its high false rejection rate (FRR) and false acceptance rate (FAR). The inherent duration probability density of a state in HMM is exponential, which may be inappropriate for the modeling of audit data of computer systems. We can handle this problem well by developing a hidden semi-

Markov model (HSMM) for the normal behavior of computer systems.

In this paper, we present a novel anomaly detection approach for intrusion detection based on HSMM. The approach described here applies machine learning techniques to learn the normal behavior of a particular program in order to detect aberrations. By implementing detection at the software process level, multiple, diverse, and overlapping detectors can be embedded within the software infrastructure to provide system-wide coverage.

The rest of the paper is structured as follows: section 2 present some related work based on intrusion detection research. Section 3 explains detailed description of the attacks present in NSL-KDD dataset and analysis of dataset on various data mining techniques and machine learning techniques.Section 4 constructs a hidden semi-Markov model for normal behavior of computer system, and proposes an anomaly detection algorithm based on this model. Finally, we give our conclusion and future work in Section 5.

2. Related Work

The inherent problem of KDD dataset leads to new version of NSL KDD dataset that are mentioned in [6, 7]. It is very difficult to signify existing original networks, but still it can be applied as an effective benchmark data set for researchers to compare different intrusion detection methods [4]. In [7] they have conducted a statistical analysis on this data set and found two important issues which highly affect the performance of evaluated system, and results in very poor evaluation of anomaly detection approaches. To solve these issues, they proposed a new dataset, NSL-KDD, which consists of only selected records form the complete KDD dataset and does not suffer from any of the mentioned shortcomings.

In [5] they use k-mean clustering technique on NSLKDD dataset to find the accuracy forintrusion detection. Shilpa et.al [8] used principal component analysis on NSL KDD dataset for feature selection and dimension reduction technique for analysis on anomaly detection. Generally, Data mining and machine learning technology has been widely applied in network intrusion detection and prevention system by discovering user behavior patterns from the network traffic data.

A hidden Markov model is a doubly embedded stochastic process with an underlying stochastic process that is not observable (it is hidden), but can only be observed through another set of stochastic process that produces the sequence of observations [12]. HMM is a useful tool to model sequence symbols. The states of HMM represent some unobservable conditions of the system being modeled. In each state, there is a certain probability of producing any of the observable system outputs and a separate probability indicating the likely next states. An HMM can be described using its characteristic parameters. The further information about these parameters can be found in [12].

In previous research, hidden Markov model (HMM) has been applied in intrusion detection systems, but it has a major weakness: the inherent duration probability density of a state in HMM is exponential, which may be inappropriate for the modeling of audit data of computer system. That's why we develop Hidden Semi Markov Model to reduce this weakness.

3.Dataset Description

The statistical analysis showed that there are important issues in the data set which highly affects the performance of the systems, and results in a very poor estimation of anomaly detection approaches. To solve these issues, a new data set as, NSL-KDD [6] is proposed. The advantage of NSL KDD dataset are-

- No redundant records in the train set, so the classifier will not produce any biased result
- No duplicate record in the test set which have better reduction rates.
- The number of selected records from each difficult level group is inversely proportional to the percentage of records in the original KDD data set.

Table 1. Attacks in Testing Dataset

Attacks in Dataset	Attack Type (37)
DOS	Back,Land,Neptune,Pod,Smurf, Teardrop,Mailbomb,Processtable,Udpstor m,Apache2,Worm
Probe	Satan,IPsweep,Nmap,Portsweep,Mscan,Saint
R2L	Guess_password,Ftp_write,Imap,Phf,Multi hop,Warezmaster,Xlock,Xsnoop,Snmpguess,Snmpgetattack,Httptunnel, Sendmail, Named
U2R	Buffer_overflow,Loadmodule,Root kit,Perl ,Sqlattack,Xterm,Ps

The training dataset is made up of 21 different attacks out of the 37 present in the test dataset. The known attack types are those present in the training dataset while the novel attacks are the additional

attacks in the test dataset i.e. not available in the training datasets. The attack types are grouped into four categories: DoS, Probe, U2R and R2L. Table 1 shows the major attacks in both training and testing dataset [5].

3.1.Dataset Analysis by Data Mining Techniques and Machine Learning Techniques

Using data processing techniques, it perceive and extrapolate knowledge that may scale back the probabilities of fraud detection [9], improve audit reactions to potential business changes, and make sure that risks area unit managed in exceedingly a lot of timely and active manner. Additionally to employing a specific data processing tool, internal auditors will choose between ranges of knowledge mining techniques. The foremost unremarkably used techniques embody artificial neural networks, decision trees, and nearest-neighbor methodology. Each of the techniques are analyzed the knowledge in numerous ways:

- Artificial neural networks are unit non-linear, predictive models that learn through training. Though they're powerful predictive modeling techniques. The auditors will simply use them is reviewing records to spot fraud and fraud-like actions, they're higher utilized in things wherever they will be used and reused, like reviewing MasterCard transactions each month to envision for anomalies.
- Decision trees are unit arborous structures that represent decision sets. These choices generate rules that are used to classify data.
- The nearest-neighbor methodology classifies knowledge set records supported similar data in an exceedingly historical dataset. Auditors will use this approach to outline a document that's fascinating to them and raise the system to go looking for similar things.

Each of these approaches has both advantages and disadvantages that need to be considered prior to their use. Neural networks, which are difficult to implement, require all input and resultant output to be expressed numerically, thus needing some sort of interpretation. The decision tree technique is the most commonly used methodology, because it is simple and straightforward to implement and the nearest-neighbor method relies more on linking similar items. A good way to apply advanced data mining techniques is to have a flexible and interactive data mining tool that extract, import, and analyze the data. On integrating data mining with warehouse it simplifies mining result.

Irrespective of good anomaly detection methods are used, the problems such as high false alarm rates is difficult in finding proper features, and high performance requirements still exist. Therefore, if we are able to mix the advantages of learning schemes in machine learning methods, according to their characteristics in the problem domain, then the combined approach can be used as an efficient means for detecting anomalous attacks. Some of the classification algorithm that most commonly used to classify the dataset are SVM, J48, Random forest, CART and Navie- Bayes [10].

4. Hidden Semi Markov Model

Because of the weakness in Hidden Markov Model, we adopt hidden semi-Markov model (HSMM) for intrusion detection, which is introduced in the following.

A semi-Markov HMM (more properly called a hidden semi-Markov model, or HSMM) is similar to HMM except that each state in HSMM can emit a sequence of observations [13].

Because of this difference, the duration probability density of a state in HSMM can be an arbitrary distribution. An HSMM can be described as

$\lambda = (N, M, V, A, B, \pi)$ where

- N is the size of $\varphi = \{0, 1, \dots, N-1\}$, which is the state space of hidden semi-Markov chainH_t, t = 1, 2, 3, …;
- $V = \{V_0, V_1, \dots, V_{M-1}\}$ is visible symbols;
- M is the number of all visible symbols;
- $B = \{b_i(k)\}$, $i \varepsilon \Phi$, $1 \le k \le M$, is the distribution of visible symbols V;
- $A = [a_{ij}]_{NxN}$is the distribution of state transfer probabilities;
- $\pi = \{\pi_0, \pi_1, \dots, \pi_{N-1}\}$ is the initial distribution;
- O_t,t = 1, 2, … , T, $O_t \varepsilon$ V is visible symbol sequence;
- T is the number of observed visible symbol.

4.1. Detection Algorithm Based on HSMM

From maximum entropy principle (MEP) [14], we know that when a computer system is running in normal state, the audit data it generates contains less information than that it generates when running in anomaly state. Namely, the information entropy of anomaly state is larger than that of normal state, so the information entropy can act as the metric in anomaly detection.

But when the length of visible symbol sequence increases, the information entropy of visible symbol opsense to compare the value of information entropy

among the same-length sequences. In order to use entropy metric on variable-length symbol sequences, we compute the average information entropy (AIE) of visible symbol sequences, and use it as the metric to distinguish between normal behavior and anomaly behavior. Let E(N) be the average information entropy (AIE) of visible symbol sequences, we can get:

$$E(N) = \frac{-\sum_{i=1}^{N} In\ Pi\{O|\lambda\}}{N} \quad (1)$$

For convenience of on-line detection, we can use the following iterative algorithm:

$$E(N) = \frac{-\sum_{i=1}^{N} In\ Pi\{O|\lambda\}}{N} = \frac{N-1}{N} E(N-1) - \frac{In\ PN\{O|\lambda\}}{N} \quad (2)$$

Initial value is $E(1) = -InP1\{O|\lambda\}$.

4.2. Detection Based on Classifier

The data in NSL-KDD dataset is either labeled as normal or as one of the 24 different kinds of attack. These 24 attacks can be grouped into four classes: Probe, DoS, R2L, and U2R. The effectiveness of the algorithm is performed in weka tool [11]. It is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [12]. WEKA consists of four application namely Explorer, Experimenter, Knowledge flow, Simple Command Line Interface and also Java interface. The experimental steps are as follows-

1. Select and preprocess the dataset.
2. Run the classifier algorithm and detection algorithm based on HSMM.
3. Compare the classifier result and detection algorithm result.

The first step is to perform discretization as preprocess. Discretization is the process of turning numeric attributes into nominal attributes. The main benefit is that some classifiers can only take nominal attributes as input, not numeric attributes. Another advantage is that some classifiers that can take numeric attributes can achieve improved accuracy if the data is discretized prior to learning. From 41 attribute we have filtered to 13 feature vectors by using CFS subset technique to get an optimum selection from complete dataset for training as well as for testing experiments. Table II shows the test accuracy that achieved by using the six algorithms for the full dimension data and also after the feature reduction.

Table 2.Test Accuracy for different classes of attack

Class Name & Test Classification Algorithm	Normal		DOS		Probe		U2R		R2L	
	With 41 fea:	With 15 fea:	With 41 fea:	With 15 fea:	With 41 fea:	With 15 fea:	With 41 fea:	With 15 fea:	With 41 fea:	With 15 fea:
	Accuracy									
Random Forest	**9.1**	**9.8**	**8.7**	**9.1**	**7.6**	**8.9**	**7.5**	**8.7**	**6.8**	**7.9**
J48	**8.9**	**7.5**	**2.4**	**8.3**	**0.2**	**6.0**	**3.9**	**5.5**	**7.6**	**8.9**
SVM	**8.1**	**8.9**	**7.8**	**8.6**	**0.7**	**1.3**	**3.7**	**5.9**	**1.8**	**3.9**
CART	**8.9**	**1.9**	**2.7**	**9.5**	**2.1**	**5.4**	**3.1**	**0.7**	**0.8**	**9.0**
Navie Bayes	**0.3**	**5.9**	**2.7**	**5.0**	**0.9**	**5.1**	**0.7**	**4.3**	**9.8**	**1.1**
HSMM	**0.0**	**2.0**	**0.2**	**4.5**	**1.2**	**0.0**	**9.8**	**0.1**	**0.1**	**0.7**

5. Conclusion and Future Work

In this paper, hidden semi-Markov model is introduced into intrusion detection systems. We present an algorithm of anomaly detection based on HSMM, which computes the distance between the processes monitored by intrusion detection system and the perfect normal processes. In this algorithm, based on maximum entropy principle (MEP), we introduce the concept of average information entropy (AIE), which is used as detection metric via analyzing variable-length observed symbol sequences. To improve accuracy, the segmental K-means algorithm is applied as training algorithm for the HSMM. Experimental results show that this approach is not only valuable in theory, but also can be effectively applied to monitoring realtime computer systems.

We have also analyzed the NSLKDD dataset that solves some of the issues of KDD cup99 data. The analysis shows that NSL KDD dataset is very ideal for comparing different intrusion detection models. Using all the 41 features in the dataset toevaluate the intrusive patterns may leads to time consuming and it also reduce performance degradation of the system. Some of the features in the dataset are redundant and irrelevant for the process. The experiment has been carried out with different classification algorithms for the dataset with and without feature reduction and it's clear that Random Forest shows a high test accuracy compared to all other algorithms in both the cases. So in the case of reduced feature set this analysis shows that Random Forest is speeding up the training and the testing methods for intrusion detection that is very essential for the network application with a high speed and even providing utmost testing accuracy. In future

we can try to improve the Random Forest algorithm to build an efficient intrusion detection system.

References

[1] C. E. Landwehr, A. R. Bull, J. P. McDermott, and W. S. Choi, "A taxonomy of computer programSecurity flaws," ACM Computer. Surv. vol. 26, no. 3, pp. 211–254, 1994.

[2] M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03), pp. 172– 179, 2003.

[3] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddc up 99/kddcup99.html, October 2007. IJERTV2IS120804

[4] J. McHugh, "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, vol. 3, no. 4, pp. 262–294, 2000.

[5] Vipin Kumar, HimadriChauhan, DheerajPanwar, "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume3, Issue-4, September 2013.

[6] "Nsl-kdd data set for network-based intrusion detection systems." Available on: http://nsl.cs.unb.ca/KDD/NSL-KDD.html, March 2009.

[7] MahbodTavallaee, EbrahimBagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", In the Proc. Of the IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009), pp. 1-6, 2009.

[8] Shilpalakhina, Sini Joseph and Bhupendraverma, "Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD", International Journal of Engineering Science and Technology, Vol. 2(6), 2010, 17901799.

[9] Lei Li, De-Zhang Yang, Fang-Cheng Shen, "A Novel Rule-based Intrusion detection System Using Data andMining", In the Proc. Of 3IEEE International Conference on Computer Science and Information Technology, pp. 169-172, 2010.

[10] Xindong Wu, Vipin Kumar, J. Ross Quinlan, JoydeepGhosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, " Top 10 Algorithms in Data Mining", KnowlInfSyst (2008) 14:1–37,DOI 10.1007/s10115-007-0114-2

[11] Weka– Data Mining Machine Learning Software. http://www.cs.waikato.ac.nz/ml/weka/

[12] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (1989) 257–286.

[13] Murphy, Hidden semi-Markov models (HSMMs), June 2002, <http://www.ai.mit.edu/~muphyk>.

[14] E.T. Jaynes, Information theory and statistical mechanics, Physical Review 106 (4) (1957) 620–630.