A Novel Intrusion Detection Method Based on Principle Component Analysis in Computer Security*

Wei Wang[1], Xiaohong Guan[1, 2], and Xiangliang Zhang[3]

[1] SKLMS (State Key Laboratory for Manufacturing Systems Engineering) and Research Center for Networked Systems and Information Security,
Xi'an Jiaotong University, 710049 Xi'an, China
{wwang, xhguan}@sei.xjtu.edu.cn
[2] Center for Intelligent and Networked Systems,
Tsinghua University, 100084 Beijing, China
[3] Department of electronic science and technology,
Xi'an Jiaotong University, 710049 Xi'an, China
zhangxl@mailei.xjtu.edu.cn

Abstract. Intrusion detection is an important technique in the defense-in-depth network security framework and a hot topic in computer security in recent years. In this paper, a new intrusion detection method based on Principle Component Analysis (PCA) with low overhead and high efficiency is presented. System call data and command sequences data are used as information sources to validate the proposed method. The frequencies of individual system calls in a trace and individual commands in a data block are computed and then data column vectors which represent the traces and blocks of the data are formed as data input. PCA is applied to reduce the high dimensional data vectors and distance between a vector and its projection onto the subspace reduced is used for anomaly detection. Experimental results show that the proposed method is promising in terms of detection accuracy, computational expense and implementation for real-time intrusion detection.

1 Introduction

Intrusion detection system (IDS) is an important component of the defense-in-depth or layered network security mechanisms [1]. In general, the techniques for intrusion detection fall into two major categories depending on the modeling methods used: misuse detection and anomaly detection. Since anomaly detection can be effective against new attacks, it has become a hot topic in research of computer security.

Many types of data can be used for anomaly detection, such as Unix commands, audit events, keystroke, system calls, and network packages, etc. Early studies [2, 3] on anomaly detection mainly focus on learning normal system or user behaviors from monitored system log or accounting log data. Examples of the information derived

* The research in this paper was supported in part by the National Outstanding Young Investigator Grant (6970025), National Natural Science Foundation (60243001) and 863 High Tech Development Plan (2001AA140213) of China.

F. Yin, J. Wang, and C. Guo (Eds.): ISNN 2004, LNCS 3174, pp. 657–662, 2004.

from these logs are: CPU usage, time of login, duration of user session, names of files accessed, etc. In recent years, many research in anomaly detection focus on learning normal program behavior. In 1996, Forrest et al. introduced a simple anomaly detection method based on monitoring the system calls issued by active, privileged processes [4]. This work was extended by various methods. Lee et al. used data mining approach to study a sample of system call data to characterize sequences occurring in normal data by a small set of rules [5]. Warrender et al. proposed Hidden Markov Model (HMM) method for modeling and evaluating invisible events based on system calls [6].

In practice, a protected computer system could produce massive data streams, for example, during the experiments of capturing the system calls on the *sendmail*, only 112 messages produced a combined trace length of over 1.5 million system calls [4]. Therefore, processing the high dimensional audit data in real time for online intrusion detection would be computationally expensive.

Principle Component Analysis (PCA, also called Karhunen-Loeve transform) is one of the most wildly used dimension reduction techniques for data analysis and compression in practice. In this paper, we discuss a novel intrusion detection method based on PCA, by which intrusion detection can be employed in a lower dimensional subspace and the computational complexity can be significantly reduced. Two types of data are used to verify the proposed method and the testing results show that the method is efficient and effective.

2 The Proposed Intrusion Detection Method Based on PCA

Suppose an observation dataset is divided into m blocks by a fixed length (e.g. divided consecutively by 100 in the command data) or by an appointed scheme (e.g. separated by processes in system call data), and there are totally n unique elements in the dataset, the observed data can be expressed by m vectors with each vector containing n observations. A $n \times m$ Matrix X, where each element X_{ij} stands for the frequency of i-th individual element occurs in the j-th block, is then constructed.

Given a training set of data vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$, the average vector $\boldsymbol{\mu}$ and each mean-adjusted vector can be computed. m eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{u}_1), (\lambda_2, \mathbf{u}_2), \cdots, (\lambda_m, \mathbf{u}_m)$ of the sample covariance matrix of vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$ can be also computed [7, 8].

Several eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k$ ($k << m$) forming the $n \times k$ matrix U, which can be used to represent the distribution of the original data, is decided by experiences. Any data vector of the training set can be represented by linear combination of the k eigenvectors such that the dimensions of the data are reduced.

Given a test data vector $\mathbf{t}$, it can be projected onto the k-dimensional subspace according to the rules [7]

$$\mathbf{y} = U^T(\mathbf{t} - \boldsymbol{\mu}) \tag{1}$$

The distance between the test data vector and its projection onto the subspace is simply the distance between the mean-adjusted input data vector $\mathbf{\Phi} = \mathbf{t} - \boldsymbol{\mu}$ and

$$\mathbf{\Phi}_f = U\mathbf{y} \tag{2}$$

If the test data vector is normal, the vector and its projection would be very similar and the distance between them would be very small and near to zero [9]. Based on this property, normal program and user behaviors are profiled for anomaly detection. In this paper, three measures, squared Euclidean distance, Cosine distance and Signal-to-Noise Ratio (SNR) measure, are applied to map the distance or similarity of this two vectors for comparison of the experimental results.

Squared Euclidean distance, Cosine distance and SNR measure, are defined respectively by the following rules for anomaly detection

$$\varepsilon_1 = \left\| \mathbf{\Phi} - \mathbf{\Phi}_f \right\|^2 \tag{3}$$

$$\varepsilon_2 = \frac{\mathbf{\Phi}^T \mathbf{\Phi}_f}{\left\| \mathbf{\Phi} \right\| \left\| \mathbf{\Phi}_f \right\|} \tag{4}$$

$$\varepsilon_3 = 10 \log \left(\frac{\left\| \mathbf{\Phi} \right\|^2}{\left\| \mathbf{\Phi} - \mathbf{\Phi}_f \right\|^2} \right) \tag{5}$$

In the procedure of anomaly detection, ε_1, ε_2 and ε_3 are considered as *detection index*. If either ε_1, ε_2 is below or ε_3 is above a predetermined threshold, the test data $\mathbf{t}$ is then classified as normal, otherwise as anomalous.

3 Experiments

3.1 Experiments on System Call Data

The first data used in the experiments is from the data set collected by Warrender and Forrest [6]. Since a great number of traces are included in the *lpr* data, we use MIT *lpr* data to validate the proposed method in this paper. The data set is available for downloading at http://www.cs.unm.edu/~immsec/. The procedures of generating the data are also described in the website. Each trace of the *lpr* data is the list of system calls issued by a single process from the beginning of its execution to the end. The data set includes 2703 traces of the normal data and 1001 traces of the intrusion data. We use the former 600 traces of the normal data and the former 300 traces of the intrusion data in the experiments. The data descriptions in the experiments are shown as Table 1.

Table 1. Data descriptions in the experiments

Number of system calls	Number of unique system calls	Number of normal traces	Number of intrusion traces
842,279	41	600	300

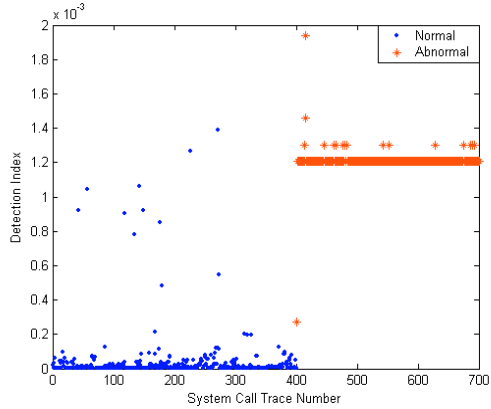

Fig. 1. Experimental results on the MIT *lpr* system call data. The y-axis represents the *detection index* and x-axis represents the system call trace number. The star (*) stands for abnormal data and dot (•) for normal data

Using PCA for intrusion detection, we can get good testing results. Figure 1 shows the experimental results using squared Euclidean distance measure with the former 200 traces of data for training and other 700 traces for testing. It is observed that abnormal data can be easily distinguished from normal data.

Using different number of the traces in the normal data for training and different distance or similarity measures for anomaly detection, we can get different detection rate and false alarm rate. We use principle component percentage of the total variation as 99.9% in the experiments and the results are summarized as table 2.

Table 2. False Alarm Rate (FAR) and Detection Rate (DR) with different conditions

Number of the normal training traces	Squared Euclidean distance measure		Cosine distance measure		SNR measure	
	FAR	DR	FAR	DR	FAR	DR
100	3.4%	100%	12.4%	100%	12.4%	100%
200	2.75%	100%	10.25%	100%	10.25%	100%
300	3%	100%	8%	100%	8%	100%
400	4%	100%	7%	100%	7%	100%

From table 2, we observe that the experimental results are the best for low alarm rate when the number of the training traces is 200 with squared Euclidean distance

measure. Another observation is that the squared Euclidean distance is better than Cosine distance measure and SNR measure for intrusion detection.

3.2 Experiments on Unix Command Data

To further investigate the performance of intrusion detection using the proposed method, we use another data set which comes from a UNIX server at AT&T's Shannon Research Laboratory. User names and the associated command sequences make up the data. Fifty users are included with 15000 commands for each user, divided into 150 blocks of 100 commands. The first 50 blocks are uncontaminated while some masquerading command blocks are inserted into the command sequences of the 50 users starting at block 51 and onward. The goal is to correctly detect the masquerading blocks in the user community. The data are available at http://www.schonlau.net/intrusion.html, see [3] for more about the data descriptions. We revised the data and reconstructed them in the experiments. We selected two data sets of two users from the user community. The first 50 data blocks of the first user are used for training and other data, which contain 100 data blocks of the first user considered as normal and 150 blocks of the second user as abnormal, are used for testing. User 5 and user 32 are selected in the experiments.

We used principle component percentage of the total variation as 99.9% and squared Euclidean distance for anomaly detection in the experiments. Experimental results are shown as Fig.3.

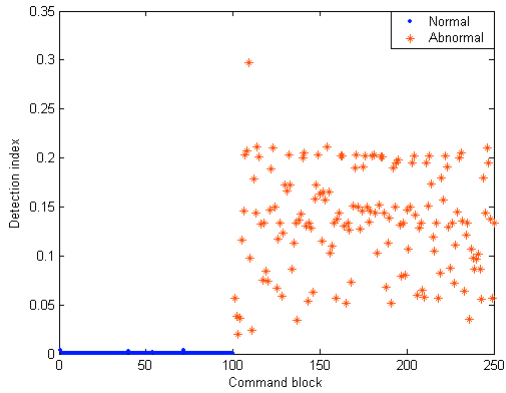

Fig. 2. The experimental results of the combining data of user 5 and user 32. All the data blocks of user 5 and 32 are uncontaminated, therefore the first 100 data blocks from user 5 are treated as normal (•) and blocks 101~250 from user 32 are considered as abnormal (*)

From Fig.2, it is easily observed that the abnormal data can be 100% distinguished from the normal data without false alarm by using PCA.

4 Conclusion

In this paper, a new intrusion detection method based on PCA is proposed. Instead of considering the transition information of the system calls or commands, the new method takes into account those of frequency property. Since there is no need to consider each system call in each trace or command in each block, the computational cost of the proposed method is low and suitable for real-time intrusion detection. Data found in intrusion detection problem are often high dimensional in nature. By using the proposed method, the high dimensional data can be greatly reduced by projecting them onto a lower dimensional subspace for intrusion detection so that the complexity of the detecting algorithm is significantly reduced.

The method is implemented and tested on the system call data from University of New Mexico and the Unix command data from AT&T Research lab. Experiment results show that the method is promising in terms of detection accuracy, computational expense and implementation for real-time intrusion detection.

Further research is in progress to mix the frequencies property with the transition information of system calls and commands so that lower false alarms and missing alarms can be achieved.

References

1. Lee, W., Xiang, D.: Information-Theoretic Measures for Anomaly Detection. Proceedings of the IEEE Symposium on Security and Privacy, IEEE Computer Society Press, Oakland, CA (2001) 130-143
2. Anderson, D., Frivold, T., Valdes, A.: Next-Generation intrusion Detection Expert System (NIDES): A Summary. Technical Report SRI-CSL-95-07, Computer Science Laboratory, SRI International, Menlo Park, California (1995)
3. Schonlau, M., Theus, M.: Detecting Masquerades in Intrusion Detection Based on Unpopular Commands. Information Processing Letters, Vol. 76 (2000) 33-38
4. Forrest, S., Hofmeyr, S.A., Somayaji, A., Longstaff, T.A.: A Sense of Self for Unix Processes. Proceedings of the IEEE Symposium on Security and Privacy, IEEE Computer Society Press, Oakland, CA (1996) 120-128
5. Lee, W., Stolfo, S.: Data Mining Approaches for Intrusion Detection. Proceedings of the 7th USENIX Security Symposium, Usenix Association, San Antonio, Texas (1998) 79-94
6. Warrender, C., Forrest, S., Pearlmutter, B.: Detecting Intrusions Using System Calls: Alternative Data Models. Proceedings of the IEEE Symposium on Security and Privacy, IEEE Computer Society Press, Oakland, CA (1999) 133-145
7. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2nd Edn. China Machine Press, Beijing (2004) 568-570
8. Jolliffe, I.T.: Principal Component Analysis. 2nd Edn. Springer-Verlag, New York (2002)
9. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience. Vol. 3, No. 1 (1991) 71-86