Detecting masquerades with principal component analysis based on cross frequency weights

Wei Wang, Sylvain Gombault

Detecting Masquerades with Principal Component Analysis Based on Cross Frequency Weights

Wei Wang and Sylvain Gombault

RSM Department, GET/ENST Bretagne
2, rue de la Châtaigneraie
35576 Cesson Sévigné, France
wei.wang.email@gmail.com,sylvain.gombault@enst-bretagne.fr

Abstract. In this paper, several cross frequency weights are used for extracting attributes of audit events. Principal Component Analysis (PCA) are then employed to discover the interrelationships and dependencies among features in a large number of variables and also to reduce the high dimensionality of these variables. Command data are used in the experiments for masquerade detection and the results demonstrate the effectiveness and efficiency of the method.

1 Introduction

Masquerades are people who impersonate other people on a computer [1]. Masquerading can be a serious threat to the security of computer systems. However, user behavior varies widely and masquerades are usually difficult to be detected. Extracting important features from user behavioral data, therefore, is crucial for effective masquerade and intrusion detection.

Almost all the existing work in anomaly masquerade and intrusion detection considered two probabilistic attributes of activities in computer systems, namely, the transition attributes (e.g., [2-4]) and the frequency attributes (e.g.,[5-9]) of audit data. Schonlau et al. [1] attempted to detect masquerades by building normal user behavioral models using truncated command sequences. Experiments with six masquerade detection techniques [1]: Bayes one-step Markov, Hybrid multi-step Markov, IPAM, Sequence-Match, Compression and Uniqueness, were performed and compared. The first five methods are mainly based on the transition information of user command data.

There are many issues to be resolved in masquerade and intrusion detection. First, a computer system in daily operation can produce massive data streams. Fast processing of typically high dimensional audit data is thus essential for a practical Intrusion Detection System (IDS) so that actions for response can be taken as soon as possible. However, intrusion detection methods considering the transition attributes of audit data usually require much time to train the models and to detect intrusions. Intrusion detection methods only taking account

of frequency information can improve some real-time performance but cannot achieve good detection accuracy [7] and this is not enough for a practical IDS.

Second, most of current masquerade and intrusion detection methods utilize the first-order statistics for detection (e.g.,[5]). However, many masquerades and intrusions in information systems manifest themselves by the correlation changes. The detection methods ignoring the coherent relations and dependencies of variables usually resulted in high false positive rates.

In this paper, we propose a masquerade and anomaly intrusion detection method with Principal Component Analysis (PCA) based on frequency weights that not only consider the frequency information of events in each sequence of audit data, but also consider the distribution of the event in the whole data. The weights are originally from information retrieval and were known as *tfidf* (term frequency - inverse document frequency). Plain Term Frequency (TF), Mean *tfidf* (*Mtfidf*) and LOG *tfidf* (*LOGtfidf*) are used in this paper for feature transformation. PCA is then used for masquerade detection based on the cross frequency weights. PCA is employed to discover the interrelationships and dependencies among features in a large number of variables by using the second-order statistics in the variable of audit data. Moreover, PCA condenses the valuable information in a large number of variables into a smaller set of dimensions so that the data can be largely reduced for real-time masquerade and intrusion detection.

2 Detecting masquerades with Principal Component Analysis Based on Cross Frequency Weights

2.1 Feature transformation with frequency weight schemes

Thefrequency weight schemes are described below and the notation and terminology used in this paper are listed in Tab. 1.

Table 1. The notation and terminology

n	Total number of sequences in the observation data set
m	Total number of distinct events in the observation data set
f_{ij}	Frequency of event i in sequence j
n_i	Number of times that event i appears in the observation data set
s_i	Number of sequences containing event i
X	Test sequences
T	Training sequences in the observation data set

1. **TF (Plain Term Frequency)**: Nearly all the current frequency based intrusion detection methods used Plain Term Frequency (TF) for feature transformation [5-9]. It can be defined as $tf_{ij} = f_{ij}$.
2. **Mtfidf (Mean term frequency - inverse document frequency)**: *Mtfidf* has been widely used in information retrieval and we use this scheme for masquerade and intrusion detection. It is defined as $Mtfidf_{ij} = f_{ij} \times \log \frac{n}{s_i}$.

3. **LOGtfidf (LOG term frequency - inverse document frequency)**: *LOGtfidf* is defined as $LOGfidf_{ij} = \log(0.5 + f_{ij}) \times \log \frac{n}{s_i}$.

2.2 PCA based masquerade and intrusion detection

Given a set of observations (sequences) be $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}$, suppose each observation is represented by a row vector of length m. The average observation is defined as $\mu = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i}$. Observation deviation from the average is defined as $\mathbf{\Phi} = \mathbf{x_i} - \mu$. The sample covariance matrix of the data set is defined as $C = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - \mu)(\mathbf{x_i} - \mu)^{\mathbf{T}} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{\Phi_i}\phi_{\mathbf{i}}^{\mathbf{T}}$. The covariance matrix C considers the first and second-order statistic of variables in audit data. Suppose $(\lambda_1, \mathbf{u_1}), (\lambda_{\mathbf{2}}, \mathbf{u_2}), \cdots, (\lambda_{\mathbf{m}}, \mathbf{u_m})$ are m eigenvalue-eigenvector pairs of the sample covariance matrix C. We choose k eigenvectors having the largest eigenvalues. Often there will be just a few large eigenvalues, and this implies that k is the inherent dimensionality of the subspace governing the "signal" while the remaining $(m-k)$ dimensions generally contain noise [10]. The dimensionality of the subspace k can be determined by $\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{m}\lambda_i} \geq \alpha$ [10], where α is the ratio of variation in the subspace to the total variation in the original space. We form a $m \times n$ matrix $\mathbf{U}$ whose columns consist of the k eigenvectors. The representation of the data by principal components consists of projecting the data onto the k-dimensional subspace according to the rules $\mathbf{y_i} = \mathbf{U^T}(\mathbf{x_i} - \mu) = \mathbf{U^T}\phi_{\mathbf{i}}$ [10].

A test data vector $\mathbf{x_i}$ that represents a test sequence of data can be projected onto the k-dimensional subspace according to the rules. The distance between the test data vector and its reconstruction in the subspace is simply the distance between the mean-adjusted input data vector $\mathbf{\Phi} = \mathbf{x} - \mu$ and $\phi_{\mathbf{f}} = \mathbf{U}\mathbf{U^T}(\mathbf{x} - \mu) = \mathbf{U}\mathbf{y}$. If the test data vector $\mathbf{x}$ is normal, the test data vector and its reconstruction will be very similar and the distance between them will be very small [10]. As PCA seeks a projection that best represents the data in a least-square sense, we use the squared Euclidean distance to measure the distance between these two vectors $\varepsilon = \|\phi - \phi_{\mathbf{f}}\|$.

In anomaly detection, ε are characterized as *anomaly indexes*. If ε is above a predetermined threshold, then the test data $\mathbf{x}$ is classified as normal. Otherwise it is treated as anomalous.

3 Testing results

3.1 Data set

The command data sets collected by Schonlau et al. [1] are used in our experiments for masquerade detection. The command data consists of user names and the associated command sequences (without arguments). 50 users are included with 15000 consecutive commands for each user, divided into 150 blocks of 100 commands. The first 50 blocks are uncontaminated and used as training data. Starting at block 51 and onward, some masquerading command blocks, randomly drawn from outside of the 50 users, are inserted into the command sequences of

the 50 users. The goal is to correctly detect the masquerading blocks in the user community. The data are available at http://www.schonlau.net/intrusion.html.

3.2 Testing results

In the experiments, we first convert each block of data into a feature vector based on the three weights. PCA is then used for masquerade detection. We use the same threshold for all the users and there is no updating during the training and detection steps in our experiments. α was set as 99.99% in the experiments. Receiver Operating Characteristic (ROC) curves are used to evaluate the masquerade detection performance. The ROC curve is the plot of Detection Rates (DR)against False Alarm Rates (FAR).

For evaluating the performance of different weights, we plot ROC curves of the results shown in Fig.1 base on PCA method with plain TF, Mtfidf and LOGtfidf weights. It is easily observed from the figure that LOGtfidf and Mtfidf are much better than TF in terms of detection accuracy. In details, LOGtfidf is little better than Mtfidf. Using the cross frequency weights significantly improves the detection accuracy.

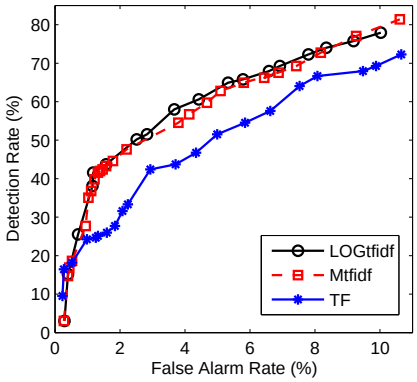


Fig. 1. ROC curves for PCA method with various different frequency weights

4 Results comparison

To facilitate comparison, we also used Chi-square distance test (also called as X^2 test) for masquerade and intrusion detection as it is typical for considering the first-order statistic of audit data.

4.1 Competitive approaches

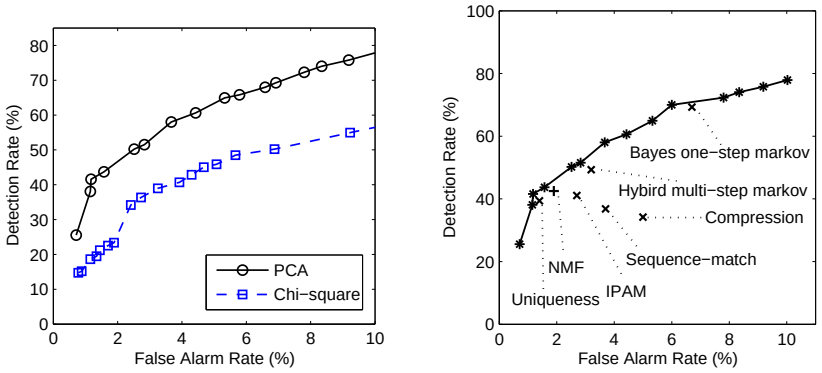
1. **Chi-square distance test**: For a given test vector $\mathbf{x}$, the X^2 test statistic is given by $X^2 = \sum_{i=1}^{m} \frac{x_i - \bar{t}_i}{\bar{t}_i}$, where x_i is the i-th variable in the test vector $\mathbf{x}$

and $\bar{t}_i$ is the averaged i-th variable of all the training vectors. The distance of a test vector $\mathbf{x}$ from the center of the normal data population can be measured by X^2 test and are considered as *anomaly index* for the test vector. When the m variables are independent and m is large (e.g., greater than 30), the X^2 statistic follows approximately a normal distribution according to the central limit theorem [10]. We compute the mean and standard deviation of the X^2 population as $\overline{X}^2$ and $\sigma_{\overline{X}^2}$ and set a threshold based on a zone of some combinations of $\overline{X}^2$ and $\sigma_{\overline{X}^2}$, e.g., $[\overline{X}^2 - \beta\sigma_{\overline{X}^2}, \overline{X}^2 + \beta\sigma_{\overline{X}^2}]$, where β is a variable parameter. For a test sequence $\mathbf{x}$, if its *anomaly index* is outside of the zone, it is then classified as abnormal.

2. **Existing approaches in the literature**: As mentioned in Section 1, Schonlau et al. [1] used six methods to detect masquerades based on the same data sets. We also used NMF for masquerade detection.

4.2 Results comparison

From Fig.2 (a), it is observed that PCA method outperforms the Chi-square test method using the same *LOGtfidf* weight for masquerade detection. This may show that considering the first and second-order statistic of audit data can improve detection accuracy compared to only using that of first-order statistic. Fig.2 (b) shows the results for PCA method and Chi-square method by using the *LOGtfidf* weight along with the results from another 7 methods in [1,6]. It is observed that PCA method achieves the best results among the other 7 methods. By using the same PCA method, the *LOGtifidf* weight improves the detection results with 23.6% than plain frequency TF based on the same data set. PCA method improves the detection results with 29.3% than Chi-square test.



(a) (b)

Fig. 2. (a): ROC curves for PCA and X^2 test method using *LOGtfidf* weight; (b): ROC curves for PCA method using *LOGtfidf* weight along with the results from other 7 methods

5 Conclusion and future work

Using the cross frequency weights are essentially more effective than only using the plain frequency attributes of audit data for masquerade intrusion. In addition, the computation cost of cross frequency weights is almost as the same as that of the plain frequency and is low overhead. In this way, the detection accuracy can improve a lot while the computation expense almost does not increase so that an effective IDS can be developed for real-time detection. The PCA method using the first and second-order of audit data can also improve the detection accuracy compare to only using that of the first-order statistic.

We have tested our method on command data for masquerade detection. The testing results are promising in terms of detection accuracy and computation expense. We believe that our work has some contributions to the masquerade and intrusion detection areas as well as IT service management. In the future work, we plan to implement the method onto practical network environments for real-time detection. Finding more effective weights for extracting valuable features of audit data and combining the frequency attributes with the transition information of audit data are also being investigated to achieve lower false alarm and missing alarm rates.

References

1. M. Schonlau, W. Dumouchel, W.-H, Ju, et al., "Computer Intrusion: Detecting Masquerades". Statistical Science, vol.16, no.1, pp.58-74, 2001.
2. D. Y. Yeung and Y. Ding, "Host-based intrusion detection using dynamic and static behavioral models". Pattern Recognition, vol.36, no.1, pp.229-243, 2003.
3. S. Forrest, S. A. Hofmeyr, A. Somayaji et al., "A sense of self for Unix processes". In Proceedings of the 1996 IEEE Symposium on Research in Security and Privacy, pp. 120-128. IEEE Computer Society Press, 1996.
4. W. Wang, X. Guan, and X. Zhang, "Modeling program behaviors by hidden Markov models for intrusion detection". Proceedings of the Third IEEE International Conference on Machine Learning and Cybernetics, pp. 2830-2835, 2004.
5. Y. Liao and V. R. Vemuri, "Using text categorization techniques for intrusion detection". In 11th USENIX Security Symposium, pp. 51-59, 2002.
6. W. Wang, X. Guan, and X. Zhang, "Profiling program and user behaviors for anomaly intrusion detection based on non-negative matrix factorization". In Proceedings of 43rd IEEE Conference on Decision and Control, pp. 99-104, 2004.
7. W. Wang, X. Guan, X. Zhang, and L. Yang, "Profiling program behavior for anomaly intrusion detection based on the transition and frequency property of computer audit data". Computers & Security, Elsevier, vol. 25, no 7, pp. 539-550, 2006.
8. W. Wang, X. Guan and X. Zhang, "A Novel Intrusion Detection Method Based on Principal Component Analysis in Computer Security". 2004 IEEE Symposium on Neural Networks, Dalian, China. LNCS 3174. pp. 657-662. Aug 2004.
9. W. Wang, S. Gombault, "Distances measures for anomaly intrusion detection". To appear in 2007 international conference on security and management, Las Vegas, NV, USA. June 2007.
10. R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification. China Machine Press, Beijing, 2nd edition edition, 2004.