A MODEL FOR USER AUTHENTICATION BASED ON MANNER OF KEYSTROKE AND PRINCIPAL COMPONENT ANALYSIS

YANG WANG[1], GUANG-YU DU[1], FU-XIONG SUN[2]

[1]School of Electronic Information, Wuhan University, Wuhan 430072, China
[2] School of Information, Zhongnan University of economics and law, Wuhan 430060, China
E-MAIL: wangyang_wh@126.com

Abstract:

As information and network security is exposing to increasing evil threats significantly, authentication is now playing a more and more important role in security defense system. In this paper, a model that authenticates and identifies access to computer systems and networks by applying keystroke manner is put forward. In the model, legitimate user's typing manner such as durations of keystrokes and latencies between keystrokes are collected during training step, and several features are extracted from the keystroke data. Then user's manner template of keystroke is built by principal component analysis. While identifying, some principal component scores of user data are used to judge the user's validity. The procedure of authentication is invisible to user. The results of related experiment show that the model has a good discerning ability.

Keywords:

Information security; authentication; keystroke; principal component analysis

1. Introduction

Because of the widely application in almost all technical, industrial, and business fields, computers and network are playing increasingly vital roles in modern society. Security of information systems has become more important than ever before.

User passwords are the most primary means of users authentication to computers since the introduction of access controls in computer systems. However, duo to its simplicity, it is vulnerable to hacker attacks. Researches have shown that users are likely to employ passwords that can be broken by an exhaustive search of a relatively small subset of all possible passwords.

In order to avoid the disadvantage of passwords, the biometrics technology is applied in users authentication to strengthen the security of computer and networks. Biometrics usually means the physical trait or behavior characteristic. Some biometrics technologies used for identification-based systems include hand geometry, iris pattern recognition, face recognition, blood vessel patterns in the retina and handprint, fingerprint and voiceprint, and handwritten signatures [1][2][3][4]. Unfortunately, these technologies are too expensive to apply widely or easy to fool. To provide an inexpensive and more reliable method, some researches employ biometrics technology of keystroke dynamics in password typing, which costs low for there is no additional hardware needed for getting the keystroke data. Keystroke dynamics utilizes the habit and rhythm in which each user types his password. It measures the keystroke rhythm of a user in order to build a template that can identify the valid user. When a user types the word, it will compare the manner of the user to the template and decide whether the user is the valid user or an intruder.

However, theses methods just identifies the manner of the user when he types his password once, therefore intruder who disguises himself as a valid user can know when the identification process is running. Moreover, the manner template of a user built by these methods is related to the password of user. In this paper, a model for user authentication based on manner of keystroke and principal component analysis (PCA) is presented, which authenticates user and builds manner template without limiting the characters of password or time of identification. The subsequent sections are organized as follows. Section2 describes the architecture of the model; Section 3 describes the data sampling and the feature extracting; Section 4 presents the user authentication method based on principal component analysis; Section 5 describes experiment of user authentication and results. Conclusions and future work are given in the last section.

2. Model of The User Authentication

The architecture of authentication model based on manner of keystroke and principal component analysis is shown in Figure 1.

1-4244-0060-0/06/$20.00 ©2006 IEEE

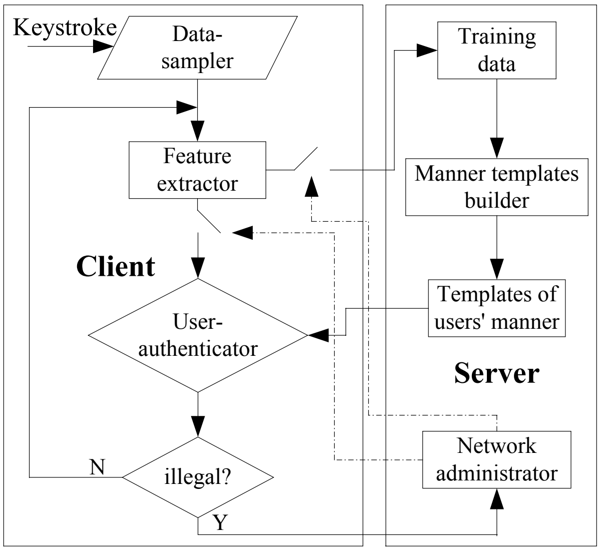

Figure 1. Architecture of the authentication model

As shown in Figure 1, the authentication model is consisted of clients and server. The clients are located at the different computers in network. The server is located on the management workstation of the network administrator.

The five main components of the authentication model are Data-sampler, Feature-extractor, User-authenticator, Manner-templates -builder and Network-administrator.

Data-sampler is responsible for capturing the stroke data of keyboard. It is a software plugin that runs in the background while user strikes keyboard.

Feature-extractor is used to analyze and transform volumes of raw keystroke data collected by the Data-sampler. First, it reassembles those data by means of the window technique. Then, it computes the statistical features of those data in the window. Finally, it produces a behavior vector by statistical features of the data in the window. The behavior vector is in a format that the User-authenticator, Manner-templates-builder can recognize.

Manner-templates-builder applies principal component analysis (PCA) to extract the manner of keystroke from training data. The method of manner extracting will be put forward in the section 4. In the builder, the manner templates extracted from training data are stored into a database for authentication.

User-authenticator identifies the behavior vector and generates judgment. It compares the incoming behavior vector with the template of user's manner. When the result of manner compare is "illegal", the User-authenticator will give an alarm to Network-administrator.

Network-administrator will make decision according to predefined strategy when User-authenticator alarms. And Network-administrator controls the work mode of each client. When training, it stores vectors from Feature-extractor to database and then start Manner-templates-builder to obtain manner templates of users. When authenticating, it monitors the results of User-authenticator from client.

In the authentication model, Data-sampler and the User-authenticator run in the background, and when authentication is started by the Network-administrator is unknowable to users, which can increase the invisibility and the security of authentication.

3. Data Sampling and Feature Extracting

3.1. Data Sampling

To collect data, a program coded in Visual C++ is located at client computer in the background. When user strikes keyboard, Data-sampler collects the typing sequences of keystrokes. For any stroke during the typing sequence, Data-sampler records its timestamp of key press time and key release time. A typing sequence example of string "stroke" is shown in Figure 2.

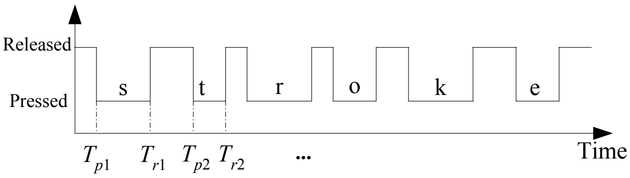

Figure 2. Timestamp of typing sequence

In Figure 2, the crest of square wave represents that a key is pressed while the peak of square wave represents that a key is released.

In our model, Data-sampler collects two kinds of keystroke data which are represented by two combinations of timestamp:

1) Key press time, namely, m_p, which is the duration time from one key being pressed to the key being released. In Figure 2, the key press time of character 's' is $m_{p1}=T_{r1}-T_{p1}$.
2) Latency of key press time, namely, m_l, which is the duration time from one key being pressed to the next key being pressed. In Figure 2, latency of key press time of character 's' and 't' is $m_{l1}=T_{p2}-T_{p1}$.

Both key press time and latency of key press time present keystroke manner of user. Key press time can reflect the strength of user's keystroke while Latency of key press time can reflect the speed of it.

Data-sampler records key press time and latency of key press time and stores them to array M_P and array M_L respectively. However, occasionally, some application software will cause user to press key or not press any key for a long time, which will result in large values in key press time and latency of key press time. These values are outliers which are random and will disturb the building of keystroke manner of user, so they must be removed from M_p and M_l by predefined thresholds.

3.2. Features

Partition M_P and M_L are to k intervals and the number of elements in each interval is equal. If there are N elements in M_P and M_L and the number of elements in each interval is n, then the numbers of intervals is k=INT(N/n). Each interval is defined as M_{Xi} (X=(P,L), i=1,2…k). Compute statistical values of each interval and regard each value as a feature of keystroke manner of user, then a behavior vector is extracted.

3.2.1. Distribution Features

These features represent distribution of key press time and latency of key press time in intervals M_{Xi} (X=(P,L)), which indicates the probability of key press time value and latency of key press time value of in M_{Xi}. In our model, key press time value and latency of key press time value are divided into 5 levels. For an element m_{ij},(j=1,2,…,n) in an interval M_{Xi}, its value belongs to level S (S=1,2, …, 5). Define probability density functions $P_X(S)$ of time value as follows:

$$P_X(S) = \frac{n_S}{n} \quad (X=(P,L)) \tag{1}$$

where n_S is the number of elements whose values belong to level S in the interval.

Because probability density functions are stable according to different users, they can be used as features to show the characteristic of user keystroke. The mathematical expectation E_{Xi} and Variance V_{Xi} of an interval M_{Xi} are also used as features.

$$E_{Xi} = \frac{1}{n}\sum_{j=1}^{n} m_{ij} \tag{2}$$

$$V_{Xi} = \frac{1}{n}\sum_{j=1}^{n} (m_{ij} - E_{Xi}) \tag{3}$$

where i=1,2,...,k;j=1,2,...,n;X=(P,L).

3.2.2. Difference Features

Given an element m_{ij} in an interval M_{Xi}, define the difference of m_{ij} and m_{ik} (k=j+d) is $\Delta_{xi}(m_{ij})$ as follows:

$$\Delta_{xi}(m_{ij}) = |\ m_{ij} - m_{ik}\ | \tag{4}$$

Difference value is divided into m levels. Define probability density functions $P\Delta_X(m)$ of difference value as $P_X(m)$=n_m/n, where X=(P,L) and n_m is the number of elements whose difference values belong to level m in the interval. Define three difference features, Contrast C_X, Second order moment M_X and entropy E_X as follows:

$$C_X = \sum_{i=1}^{m} i^2 P\Delta_X(i) \tag{5}$$

$$M_X = \sum_{i=1}^{m} (P\Delta_X(i))^2 \tag{6}$$

$$E_X = \sum_{i=1}^{m} P\Delta_X(i) \lg P\Delta_X(i) \tag{7}$$

Difference features can reflect the dynamic characteristic of keystroke manner.

Based on these two kinds of features, keystroke data of user can be turned into behavior vectors shown in Table 1.

Table 1. Data of behavior vectors

Feature	Vectors			
	No.1	No.2	No.3	…
P_{P1}	0.006667	0.013333	0.013333	…
P_{P2}	0.366667	0.306667	0.353333	…
P_{P3}	0.433333	0.460000	0.433333	…
P_{P4}	0.160000	0.186667	0.160000	…
P_{P5}	0.040000	0.040000	0.046667	…
P_{L1}	0.620000	0.560000	0.506667	…
P_{L2}	0.306667	0.273333	0.313333	…
P_{L3}	0.053333	0.106667	0.100000	…
P_{L4}	0.013333	0.040000	0.060000	…
P_{L5}	0.013333	0.026667	0.026667	…
E_{Pi}	0.139113	0.144364	0.140536	…
V_{Pi}	0.049788	0.048871	0.050702	…
E_{Li}	0.191854	0.244815	0.260238	…
V_{Li}	0.152613	0.203649	0.207102	…
C_P	2.006667	2.006667	2.013333	…
M_P	0.263556	0.250400	0.251022	…
E_P	0.099531	0.095247	0.095056	…
C_L	1.793333	1.853333	1.920000	…
M_L	0.316222	0.248133	0.230711	…
E_L	0.116012	0.093542	0.087845	…

Behavior vectors of each user can describe the keystroke manner of the user, so the model will build manner templates and identify user through those vectors.

4. Authenticator based on PCA

Because the keystroke manner of an intruder is different to valid user, the behavior vectors of the intruder can be regarded as outliers relative to the vectors of valid user. In our model, we treat the process of user authentication as outlier detection.

As an outlier detection method, principal component analysis [5] has been applied in many fields [6] [7]. The principal component based approach to outlier detection has some advantages. It doesn't have any distributional assumption, and it can reduce the dimensionality of vectors without losing valuable information, which will increase the speed of outlier detection on the basis of accuracy.

4.1. Building manner templates

PCA can explain the variance and covariance structure of a set of features through a few new variables called principal components which are linear combinations of the original features. The original features are converted to principal components by a PCA matrix. When applying PCA on keystroke data, the PCA matrix can give the distribution of behavior vectors, which reflects the keystroke manner of user.

Consider a multivariate $X=(X_1, X_2, \ldots, X_p)^T$ with the covariance matrix S, where X_i is a feature, i=1,2,…,n. Let $(\lambda_1,e_1),(\lambda_2,e_2),\ldots,(\lambda_p,e_p)$ be the p eigenvalue-eigenvector pairs of the covariance matrix S, where$\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_p\geq0$,then $U=(e_1, e_2,\ldots, e_p)^T$ is the PCA matrix of X.

The *ith* principal component is $Y_i=e_i^TX$. The principal components are uncorrelated to each other. The variance of each principal component Y_i is λ_i, so the variance of *ith* Y_i is the *ith* highest. The total variation in all the principal components is equal to the total variation in the original features.

When building manner template of a user, consider a sample matrix of a valid user behavior vectors $V=(V_1^T, V_2^T, \ldots, V_n^T)^T$, where $V_i=(v_{i1}, v_{i2},\ldots, v_{ip})^T$, i=1,2,…,n. As shown in section 3.2, in our model the number of features is 20, therefore p=20.

Because the covariance matrix S of features cannot be known directly, to archive the PCA matrix U of V, S should be estimated from the sample covariance matrix S_V.

$$S_V = \sum_{i=1}^{n}(V_i-\overline{V})(V_i-\overline{V})^T/(n-1) \tag{8}$$

where $\overline{V}=\sum_{i=1}^{n}V_i/n$.

The PCA matrix of the user keystroke vectors is composed of the p eigenvectors of the covariance matrix S_V. However, when some features are in a much bigger magnitude than others, the first few principal components will be combined mainly by those features. For this reason, it is better to carried out PCA of sample vectors on the sample correlation matrix $R_V=(r_{ij})_{p\times p}$.

where $r_{ij}=\dfrac{s_{Vij}}{\sqrt{s_{Vii}s_{Vjj}}}(i,j=1,2,\cdots,p)$.

If $(\lambda_1,e_1),(\lambda_2,e_2),\ldots,(\lambda_p,e_p)$ are the p eigenvalue-eigenvector pairs of the R_V, then the $U=(e_1, e_2,\ldots, e_p)^T$ is the PCA matrix of keystroke data. Through U, the principal component scores $Y=(y_1,y_2,\ldots,y_p)^T$ of a vector $T==(t_1,t_2,\ldots,t_p)$ can be computed as follows:

$$Y=UT' \tag{9}$$

Where T' is called the standardized vector of T, $T'=(t_1',t_2',\ldots,t_p'),t_i'=(t_i-\overline{V}_i)/\sqrt{s_{Vii}}$, i=(1,2,…,p). And y_i is called the score of vector T on the *ith* principal component.

4.2. Identification by PCA based outlier detection

When identifying a user, judge whether the behavior vector T of the user are outliers relative to the vectors of the valid user by principal component scores Y of T according to equation (9).

For any $r < p$, the first r principal components give the best representation of the data in r dimensions, in the sense that the total amount of variation explained by these components is maximum in the set of all r dimensional representations [8]. If the first $r <$ p principal components provide an adequate representation of the data, then the remaining principal components are often interpreted as representing the near-constant linear relations in the data [8]. The variation of the corresponding principal component scores will be small. Because vectors that are outliers usually arise from consistent perturbations in the original features, they are typically incompatible with the orientation of the training data. As a consequence, the principal component scores of outliers on the last few principal components are especially larger than those of training vectors, so the last few principal components are useful in multivariate outlier analysis. Therefore, in our model, we identify behavior vector by its last few principal component scores, which are represented by ξ that is the sum of squares of the last few standardized scores.

$$\xi=\sum_{i=r+1}^{p}y_i^2/\lambda_i \tag{10}$$

where r can satisfy $\sum_{i=1}^{r}\lambda_i/\sum_{i=1}^{p}\lambda_i\geq95\%$.

Once the ξ of a behavior vector of a user's keystroke data is larger than a predefined threshold θ, the authenticator will decide that the user is an intruder and give an alarm. Else, the authenticator will decide that the user is valid.

5. Experiment

To prove that the valid user's keystroke manner is identical and stable, and to show the validity of our model, an experiment of user authentication based on our model is given here. In the experiment, 10 hours consecutive keystroke data of 16 users are collected by Data-sampler, and then these data are converted to behavior vectors.

For authentication, these 16 users are divided into two groups. The first 8 users are assumed as valid users in one group while the others are defined as invalid users in another group.

The behavior vectors of valid users are divided into two datasets. One of them is treated as the training dataset and the other is included in the test dataset. The behavior vectors of invalid users are all included in the test dataset.

When training, perform principal components analysis on training dataset. During the identification, decide whether each vector in the test dataset is an outlier relative to a valid user's training vectors based on PCA. An outlier is judged as an intrusion to get access of a valid user.

The result of identification is presented by parameters shown in Table 2.

Table 2. Parameters describing the identification result

	Identification	
Actural	valid	invalid
valid	true positive(TP)	false negative(FN)
invalid	false positive(FP)	true negative(TN)

To measure the accuracy of identification, two parameters, α and β, are proposed in our experiment. Where α=TP/(TP+FN), β=TN/(FP+TN).

The results of identification of test dataset are shown in Table 3.

Table 3. The results of Identification

UserID	α	β	UserID	α	β
1	85.1%	86.4%	5	86.6%	85.3%
2	83.8%	85.3%	6	70.1%	67.8%
3	85.5%	84.7%	7	85.7%	86.5%
4	67.3%	69.3%	8	84.6%	85.2%

Table 3 illustrates that the identification of users have good accuracy, which means the valid user's manner is identical and stable. However, the identification of user 4 and 6 is not good as others. After analysis, we find that the train data of user 4 and 6 are much less than those of other valid users. It means that the discerning ability of our method is unstable while the training data are not enough, which is the defect that lowers the effectiveness of the authentication model.

6. Conclusions

In this paper, a model for user authentication based on manner of keystroke and PCA is put forward. The authentication in our model is invisible to users and the keystroke manner used in the model is not limited to that of the user's password. Experiment shows that discerning ability of our method is good. However, the model in this paper requires large training dataset to guarantee the effectiveness of authentication. Therefore, increasing the discerning ability based on small training dataset is our future work.

References

[1] Monrose.F,Reiter.M.K,Wetzel.S,"Password hardening based on keystroke dynamics", Proceedings of the 6th ACM Conference on Computer and Communication Security, pp. 26-32, November 1999.

[2] Changshui Zhang, Yanhua Sun, "AR model for keystroker verification", Proceeding of 2000 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2887-2890, October 2000.

[3] Haider. S. Abbas. A. Zaidi. A.K, "A multi-technique approach for user identification through keystroke dynamics", 2000 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1336-1341, October 2000.

[4] Robinson J A,Liang V M,Chambers J.A.M, "Computer User Verification Using Login String Keystroke Dynamics", IEEE Transactions on Systems, Man and Cybernetics,Vol 28, No 2, pp. 63-70, 1998.

[5] Johnson. R.A, Wichern. D.W, "Applied Multivariate Statistical Analysis," 4th Ed., Prentice-Hall, NJ, 1998.

[6] Lalor G.C, Zhang C, "Multivariate outlier detection and remediation in geochemical databases", The Science of the Total Environment, Vol 281, No 1, pp. 99-109, December 2001.

[7] Jackson D.A, Chen Y, "Robust principal component analysis and outlier detection with ecological data", Environmetrics, Vol 15, No 2, pp. 129- 139, 2004.

[8] Fan J.C, Mei C.L, "Data analysis", Science Press, Beijing, 2000.