user profile, intruder detection,
mouse activity,
user verification

Malgorzata PALYS[1], Tomasz Emanuel WESOŁOWSKI[1]

FEATURES REDUCTION FOR COMPUTER USER PROFILING BASED ON MOUSE ACTIVITY

The article is related to the computer systems security and to the problem of detecting the masqueraders, intruders who pretend to be legitimate users of computer systems, in particular. The research concerns computer user verification based on analysis of the mouse activity in computer system. The article presents an improvement of user verification method by introducing a new method of user profiling. The new user profiling method is based on the reduced number of features without the loss of information is introduced. The presented method of user profiling allows to simplify and speed up a user's activity data analysis and is not causing the deterioration of intrusion detection. Additionally, a method of aggregating mouse activity basic events into a higher level events is described. This article presents the preliminary research and conclusions.

1. INTRODUCTION

The computer systems are present in every aspect of human life. One of the most important role of computer systems is being a part of a medical health care and rescue system. They provide services used in management of health care facilities, they allow fast and accurate analysis of medical data. What is even more important, medical facilities deal with data of a sensitive nature (both the personal data of patients and their medical records) which by law should be especially protected against intruders attack. Within various kinds of cyber-intruders attacks [2], [6] intruders who pretend to be a legitimate user (masqueraders) constitute the greatest threat [7], [8]. Masqueraders (in most cases insiders) use the opportunity to overtake the role of a user already authorized in the system, who temporarily left his workspace.

In most cases user authorization is performed by simple username/password method which can be insufficient. To increase the security level more advanced methods are necessary and biometrical methods are one of the most efficient security measures. By means of biometrics users are verified by their individual physical (face or fingerprint recognition) or behavioral (signature [5] or walking style analysis) characteristics. Methods based on computer systems peripherals (like keyboard or mouse) also are considered when analyzing human behavioral characteristics. Unfortunately the analysis of the keyboard use as a continuous process [1], [4], [10] is very difficult due to the security issues related to the possibility of personal data

[1] University of Silesia, Institute of Computer Science, Będzińska 39, 41-200 Sosnowiec, Poland
e-mail: {malgorzata.palys, tomasz.wesolowski}@us.edu.pl

(usernames, passwords, PIN codes) leakage. Users activity connected to computer mouse is free of these issues and furthermore does not generate any extra costs associated with the purchase of sensors. Presented in this paper method of recording and analyzing the activity of a user is simple and effective. For the above reasons the introduced profiling method can be easily implemented into the intrusion detection systems.

The paper presents a new method of computer user profiling for the intrusion detection system based on mouse activity analysis. The higher level mouse event used in the analysis is also described. The introduced method reduces by means of statistics the set of features necessary to obtain user's profile without causing the deterioration of intrusion detection. Reducing the number of features simplifies the data analysis and user profiling process and decreases the time and complexity of intrusion detection.

1.1. DATA SET DESCRIPTION

The data set consisting of users' activity data used in the study had been created by using a dedicated software created by the authors. The registration of user's activity data is performed automatically and continuously without involving a user. For the purpose of the study the activity of 10 users was recorded till the moment of writing this article but the data set is continuously extended by adding the activity logs of further users. The data are captured on the fly and saved in the text files on the ongoing basis.

Each activity log file starts with the information on screen resolution set in user's computer system. The following lines contain the description of events representing users activity appearing in the computer system connected to the use of both computer mouse and keyboard. Each line represents a single event and it starts with an event *id* describing the type of an event followed by the timestamp of an event and the additional data related to it. The possible *id* values for mouse activity events are: M – mouse move, L/l – left mouse button down/up, R/r – right mouse button down/up, S – mouse scroll. Additional data section of mouse events consists of two coordinates (x, y) describing the position of the mouse cursor on the screen at the moment of an event. The negative values of coordinates are possible when working with multiple screens.

2. DATA PROCESSING AND ANALYSIS

The dedicated activity registration software records an activity of a computer user when using the computer mouse and/or keyboard. In the first stage of data analysis some data preprocessing is necessary.

At first the text logs are filtered and only the information about the screen resolution and mouse events are accepted for further analysis. All the events related to a use of a keyboard or indicating a switching of the window are omitted. Still, after filtering the number of recorded events is very high. In the study it was noticed, that the software recorded even 60000 mouse events per hour and most probably for some advanced users it could be more in the future (Fig. 1). The recorded data are raw, they consist only of a timestamp and coordinates. To allow a statistical analysis of the data set it is necessary to change its form. In [9] it has been proposed to aggregate the basic mouse events into the higher level event (HL-Event). HL-Events aggregate all the basic mouse events (move, scroll) represented by timestamp and coordinates, that took place between two mouse button click events. Based on such a subset of data later the characteristics are determined. This is why the sequence of each user's mouse events was reorganized into a set of HL-Events.

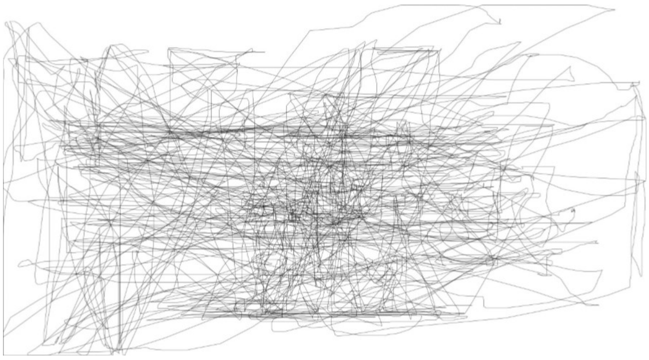

Fig. 1. The trace of a mouse cursor recorded during one hour of activity.

2.1. FEATURES OF MOUSE ACTIVITY

The introduced method is based on the approach presented in [3], [9] where a method of user profiling based on the 25 features extracted from a users' activity data for intrusion detection was introduced.

The high number of features causes a long time of an analysis and profiling. Because the aim of the study is to develop an intrusion detection methods able to detect masqueraders in a real time - based on a few minutes of their malicious activity - it seems reasonable to simplify the activity analysis process in order to shorten a time of profiling. One way to achieve this is to reduce a number of analyzed features but only while maintaining or improving the performance of intrusion detection.

3. REDUCING NUMBER OF FEATURES

Taking into consideration the research and mouse activity features previously described in [3], [9] the goal is to select only the most important of all the introduced features in order to simplify the analysis of a user behavior. Eventually, the research should result in developing a computer user profiling method working in an on-line mode – allowing to analyze the activity of a user in a real time. For this reason, the less features taken for an analysis the better.

To limit the number of features to these important ones, it is necessary to use a strategy for creating the feature space, called Feature Selection. The Feature Selection is a process of searching for such a subset of features that will eventually deliver the best efficiency of solving a given problem.

For the presented in this paper study as a feature selection method, the multiple regression was selected. Its purpose is to present a relationship between several independent variables and a dependent variable. The regression helps to answer the question of how an individual independent variables affect a dependent variable. In general, a multiple regression equation has the following form:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n, \tag{1}$$

where:
$\hat{Y}$ - a dependent variable,
X_i - an i-th independent variable representing an i-th feature, $i = 1, \ldots, n$,
b_i – a coefficient of an i-th independent variable, $i = 1, \ldots, n$,
b_0 – a constant term,
n – a number of features.

In case of the presented study the mouse activity features constitute independent variables. A dependent variable $\hat{Y}$ represents the distance value taken for the k-NN classifier in an intrusion detection [9] calculated as an Euclidean distance between a tested sample (feature vector based on a particular HL-Event) and a legitimate user's profile. The determination of coefficients b_i, $i = 0, \ldots, n$ values was carried out by a least squares method. For this purpose, a use of StatSoft Statistica software was made.

4. EXPERIMENTS AND RESULTS

First, the experiments were performed for the same parameters of user profiling and intrusion detection as in [9] in order to verify if the new proposed user profiling method could replace the former profile based on all 25 features without the loss of accuracy of intrusion detection. In previous work [9] an Euclidean distance E_j between a user's profile and a j-th tested sample was calculated. As the result of experiments a reduced set of 13 features have been established: the length, the filling ratio, the filling-distance ratio, the average velocity, the correction-velocity ratio, the average vertical velocity, the jerk, the trajectory correctness, the braking, the approaching, the Trajectory Center of Mass (TCM), the Scattering Coefficient (SC) and the Velocity Curvature (VC). For the reduced set of features the general form of the multiple regression equation is as follows:

$$\begin{aligned} \hat{Y} &= b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8 + b_9X_9 + \\ &+ b_{10}X_{10} + b_{11}X_{11} + b_{12}X_{12} + b_{13}X_{13}, \end{aligned} \tag{2}$$

where:
$\hat{Y}$ – the profile variable,
X_1 – the length,
X_2 – the filling ratio,
X_3 – the filling-distance ratio,
X_4 – the average velocity,
X_5 – the correction-velocity ratio,
X_6 – the average vertical velocity,
X_7 – the jerk,
X_8 – the trajectory correctness,
X_9 – the braking,
X_{10} – the approaching,
X_{11} – the TCM,
X_{12} – the SC,
X_{13} – the VC.

After determining a reduced set of features the experiments were performed in order to determine the optimal number of feature vectors (HL-Events) which make up a user's profile. As a goodness of fit measure the coefficient of determination R^2 was used (3).

$$R^2 = 1 - \frac{\sum\limits_{j=1}^{k} \left(E_j - \hat{Y}_j\right)^2}{\sum\limits_{j=1}^{k} \left(E_j - \bar{E}\right)^2}, \quad R^2 \in [0, 1], \tag{3}$$

where:
$\hat{Y}$ – a characteristic of a j-th sample (feature vector) calculated by the introduced regression

equation, $j = 1, \ldots, k$,
E_j – an Euclidean distance between the profile and a j-th tested sample derived from the previous experiments [9], $j = 1, \ldots, k$,
$\bar{E}$ – an average of E_j values in a tested set of feature vectors,

$$\bar{E} = \frac{1}{k} \sum_{j=1}^{k} E_j, \tag{4}$$

k – the number of tested samples (feature vectors), k = 2000.

The results of these experiments are presented in Table 1.

Table 1. Determining the optimal number of feature vectors constituting a profile of a user.

HL-Events/Profile	R^2
100	0.593
200	0.978
300	0.964
400	0.998
500	0.998
600	0.998
700	0.998
800	0.998
900	0.998
1000	0.998

The results can be presented also in a graphical form (Fig. 2). It can be noticed that after reaching the number of 400 HL-Events in a profile the coefficient of determination R^2 does not change significantly with a growing number of HL-Events. Based on this observation the optimal number of feature vectors for building a user's profile has been designated as 400.

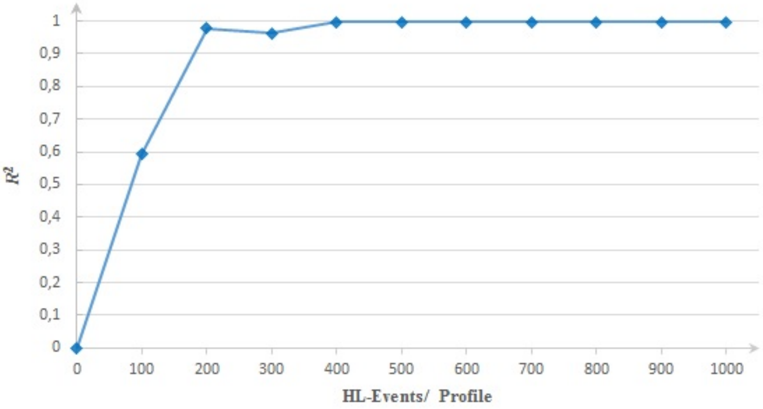

Fig. 2. The coefficient of determination R^2 versus number of HL-Events/Profile.

4.1. MULTIPLE REGRESSION EQUATION

It should be noted that taking a minimum of 400 feature vectors, a regression equation can be determined, which gives the coefficient of determination $R^2 = 0.998$ what can be interpreted as a very high goodness of fit (99.8%). Accordingly, for an exemplary user, taking into account the reduced set of thirteen features, the multiple regression equation was determined:

$$
\begin{aligned}
\hat{Y} &= 0.85673 - 2.46489 \cdot X_1 + 2.8942 \cdot X_2 - 0.57794 \cdot X_3 - 1.0917 \cdot X_4 + \\
&+ 2.25417 \cdot X_5 + 1.77867 \cdot X_6 + 0.82027 \cdot X_7 + 0.95887 \cdot X_8 - 0.55724 \cdot X_9 - \\
&- 0.99827 \cdot X_{10} - 0.17526 \cdot X_{11} + 1.00005 \cdot X_{12} + 0.99176 \cdot X_{13},
\end{aligned} \tag{5}
$$

where:
$\hat{Y}$ – the profile variable,
$X_1, ..., X_{13}$ – the set of features described above (2).

Such an equation has to be determined for each user. The above equation describes in the study a user's profile. In particular, the coefficients and the constant term constitute a profile of a user. The value of $\hat{Y}$ calculated for a tested sample (feature vector) is equivalent to the Euclidean distance between the tested sample and the previously introduced profile of a user [9].

Hereby, the goal of the research has been reached. The new method of computer user profiling introduced in this paper allows to create a profile of a user based on less features (the number of necessary features was reduced by almost 50%) and it does not cause the loss of intrusion detection accuracy. The intrusion detection experiments performed for the new introduced method based on the multiple regression equation $\hat{Y}$ and the reduced set of 13 features derived the same results as the experiments performed on the previous profiling method based on an Euclidean distance E [9] and on the full set of 25 features (Table 2).

Table 2. Comparison of use profiling methods.

	Method E	Method $\hat{Y}$
Number of features	25	13
Intrusion detection EER [%]	17.88	17.88

5. CONCLUSIONS

The goal of the research on user's mouse activity is to develop a user profiling method that would allow to detect an intruder in short time by analyzing a short time activity on the fly. This type of detection is especially desirable in environments where due to a temporary absence of a legitimate users at the workplace an unauthorized person could access a protected resource. Such an environment is for example nurses' duty desk at the hospital. When an emergency takes place the medical staff starts the emergency procedures leaving their computer systems temporarily unattended.

The new introduced user profiling method is based on almost 50% lower number of features as the one previously presented in [9], it is also less complex in calculations. Therefore, the proposed method allows to simplify and to accelerate the mouse activity analysis process. At the same time it does not cause the deterioration of intrusion detection by preserving the accuracy (the average EER was kept on the level of 17.88%).

The article presents the preliminary research and conclusions however the results are promising. The research was based on the previously conducted study but some observations were made that most probably will result in further improvements. The reduced set of 13 features was determined on the basis of pre-defined parameters. The arithmetic mean used to create the profile is very sensitive and a large impact on its value has the noise in the activity data. To reduce the influence of outliers a different model for creating a profile using, among others, the median or dominant could be proposed together with different filtering methods.

In the future it is planned to perform research aimed at optimizing the method of profiling and intrusion detection.

BIBLIOGRAPHY

[1] BANERJEE S. P., WOODARD D. L. Biometric authentication and identification using keystroke dynamics: A survey. Journal of Pattern Recognition Research, 2012, Vol. 7. pp. 116–139.

[2] DENNING D. E. Cyberspace attacks and countermeasures. 1997. In Internet Besieged D. E. Denning and P. J. Denning (eds), ACM Press, New York, pp. 29–55.

[3] FEHER C., ELOVICI Y., MOSKOVITCH R., ROKACH L., SCHCLAR A. User identity verification via mouse dynamics. Information Sciences, 2012, Vol. 201. pp. 19–36.

[4] KUDŁACIK P., PORWIK P., WESOŁOWSKI T. Fuzzy approach for intrusion detection based on user's commands. Soft Computing, Springer-Verlag Berlin Heidelberg, 2015, doi: 10.1007/s00500-015-1669-6.

[5] PALYS M., DOROZ R., PORWIK P. On-line signature recognition based on an analysis of dynamic feature. In: IEEE Int. Conference on Biometrics and Kansei Engineering (ICBAKE 2013), Tokyo Metropolitan University Akihabara, 2013. pp. 103–107.

[6] PUSARA M., BRODLEY C. E. User re-authentication via mouse movements. In: Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security, 2004. pp. 1–8.

[7] SALEM M. B., HERSHKOP S., STOLFO S. J. A survey of insider attack detection research. Insider Attack and Cyber Security: Beyond the Hacker, Advances in Information Security, Springer, 2008, Vol. 39. pp. 69–90.

[8] SCHONLAU M., ET AL. Computer intrusion: detecting masquerades. Statistical Science, 2001, Vol. 16. pp. 58–74.

[9] WESOŁOWSKI T., PALYS M., KUDŁACIK P. Computer user verification based on mouse activity analysis. Studies in Computational Intelligence, Springer International Publishing, 2015, Vol. 598. pp. 61–70.

[10] WESOŁOWSKI T. E., PORWIK P. Keystroke data classification for computer user profiling and verification. Computational Collective Intelligence, LNAI, Springer International Publishing, 2015, Vol. 9330. pp. 588–597.