The New Multilayer Ensemble Classifier for Verifying Users Based on Keystroke Dynamics

Rafal Doroz(✉), Piotr Porwik, and Hossein Safaverdi

Institute of Computer Science, University of Silesia,
Ul. Bedzinska 39, 41-200 Katowice, Sosnowiec, Poland
{rafal.doroz,piotr.porwik,hossein.safaverdi}@us.edu.pl
http://www.biometrics.us.edu.pl, http://www.zsk.us.edu.pl

Abstract. In this work we proposed the new multilayer ensemble classifier which can be applied in many domains, especially in the biometric systems. Proposed classifier works on database which comprises data from keystroke dynamics. Such kind of data allows us to recognize computer users who use password. It is a typical case among the users every day work. Obtained results confirm that proposed multilayer ensemble classifier gives the high security level. For this reason our method can be used to protect computer resources against forgers and imposters.

Keywords: Keystroke dynamics · Ensemble classifiers · Biometrics

1 Introduction

The increasing use of internet and computers makes that in practice our work can be observed by illegitimate users. Today internet service is widely available, hence many, even remotely, users have access to computers and resources [7]. For example, almost everyday we need to check our bank account, e-mail account or we have to fill out online forms, so we need to use our personal information. There are several ways to protect the sensitive information. Biometric techniques are, where we analysis voice samples, fingerprint, iris, lip prints [18,19], digital signature, gait, etc. allow us to increase the security of the computer system. Keystroke dynamic is one of the important features in biometrics because it does not need any additional devices to install on computer and it could be fully implemented by means of the software [8], [12–14], [17]. A password is also one of the best solutions to increase the computer system security [4]. Although the password has a lot of benefits, in some cases it raises the danger. When someone illegally gets the password, the system cannot correctly recognize the person, then impostor can obtain unauthorized access to computer resources.

One way to overcome this problem is behavioral biometric of keystroke dynamics [4]. The keystroke rhythms of a user are measured to develop a unique biometric template of the user's typing pattern for future authentication [7].

© Springer International Publishing Switzerland 2015
M. Núñez et al. (Eds.): ICCCI 2015, Part II, LNCS 9330, pp. 598–605, 2015.
DOI: 10.1007/978-3-319-24306-1_58

Keystroke dynamics allows to measure timing patterns which could be used to recognize individuals. In this technique time durations between two keystrokes and pressure on the key can be measured [8], [12–14]. It allows us to build a unique pattern for the individuals. Keystroke dynamics is developed [1], [9,10] and accuracy of this biometric is improved. In our paper we used a machine learning technique to improve the recognition of authorized (legitimated) and unauthorized (illegitimated) users [9–11].

2 Ensemble Classifier

To make a pattern for each user, each of them types by means of the computer keyboard the password "try-mbs" 10 times [5,6]. In practice password is represented by the vector $\mathbf{v}_j^k = [\,156\ 188\ 266\ 375\ 343\ 219\ 203\ k\,]$, where numbers indicate time between two consecutive pressed keys, k is a label of a given user, and $j = 1, ..., 10$ is number of pattern. Mentioned above vectors will be supplied to an input of classifiers. By means of computer keyboard, user registers the password which was mentioned above, the password consists of eight letters and it ends by label. Such data forms a vector that is parallelly supplied to the inputs of four classifiers (see Fig. 1).

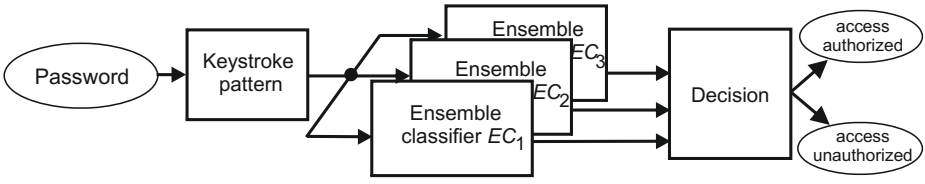

Fig. 1. The general structure of proposed classifier devoted to computer user's recognition.

It could be seen that proposed classifier consists of three ensemble-based sub-classifiers. The main part of a system is ensemble of the classifiers EC_i, $i = 1, ..., 3$ which consists of four single classifiers: c_1,c_2,c_3 and c_4. The ensemble classifiers EC_i work paralelly and each of them has the same structure.

In practice different types of classifiers can be applied, but we propose the following classifiers: Kstar(c_1), BayesNet(c_2), LibSVM(c_3) and HoeffdingTree(c_4).

Kstar: is an instance-based classifier. The classifier works on the training instance dataset. And for classification some similarity functions are used. It differs from other instance-based learners in that it uses entropy as distance function. The fundamental assumption of such a classifier is that similar instances will have similar classifications.

BayesNet: Bayesian networks (BNs), belong to the family of probabilistic models. This classifier is used to represent knowledge about an uncertain domain.

In particular, each node in this network represents a random variable, while the connections between network nodes represent probabilistic dependencies between corresponding random variables. These conditional dependencies are often estimated by using known statistical methods. Hence, BNs combine principles from network, and probability theory, as well as computer science, and statistics.

LibSVM: implements algorithm for kernelized support vector machines (SVMs). This classifier utilized well known support vector classification. It is special library which allow us to accelerate computation.

HoeffdingTree: Hoeffding tree is an incremental, anytime decision tree induction algorithm that is capable of learning from massive data streams, assuming that the distribution generating examples does not change over time. Hoeffding trees exploit the fact that a small sample can often be enough to choose an optimal splitting attribute. This idea is supported mathematically by the Hoeffding bound, which quantifies the number of observations needed to estimate some statistics within a prescribed precision [3].

This selection follows from the fact that mentioned classifiers give the best accuracy level compared to other classifiers. Each the i'th classifier c_i on its output calculates probability $P_i \in [0, 1]$ that a given individual is legitimated or not. The EC_i structure is presented in Fig. 2. The local IF THEN rule creates a local EC_i ensemble decision.

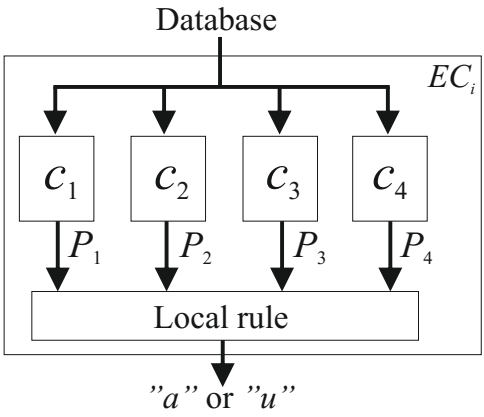

Fig. 2. Single ensemble classifier structure (one out of three).

Let S_g be a total probability generated by all classifiers. Let S_f be a probability that a given user password is forged. For such assumptions, we have:

$$S_g = \sum_{i=1}^{4} P_i, \tag{1}$$

$$S_f = \sum_{i=1}^{4} (1 - P_i), \tag{2}$$

where P_i is probability of legitimate user password produced by the i'th classifier.

The classifiers c_1,c_2,c_3 and c_4 are single classifiers and local rule produce results of classification - password was authorized ("a") or not ("u"). For such assumption, the local rule can be formulated as follow:

$$\begin{array}{ll} if & S_g > S_f \; then \text{ user is legitimated } (\text{“}a\text{”}) \\ otherwise & \text{user is illegitimated } (\text{“}u\text{”}) \end{array} \tag{3}$$

3 Classifier in the Training Mode

Learning set for a given user consists of 10 genuine and 10 forged passwords [5,6]. In practice password is changing to a vector form, which has been shown in the previous section. Our database comprises only original passwords [5,6], therefore forged examples have to be formed from the genuine passwords of other users.

Let $O^k = \left\{\mathbf{v}_1^k, \mathbf{v}_2^k, ..., \mathbf{v}_{10}^k\right\}$ be a set of genuine passwords of the person k. The database W includes passwords of 100 users and each of them typed the same password 10 times, therefore $W = \{O^k\}$, $k = 1, ..., 100$. Each ensemble classifier EC_i, $i = 1, ..., 3$ (see Fig. 2) is learned separately by means of the three kinds of learning sets DS_1^k, DS_2^k and DS_3^k, respectively. For such assumptions the forged ($F^a \subset W$) and genuine ($O^k \subset W$) passwords of legitimated user k are marked below as follows:

$$\begin{array}{lll} DS_1^k = O^k \cup F^a, & DS_2^k = O^k \cup F^b, & DS_3^k = O^k \cup F^c \\ & a \neq b \neq c \neq k \quad \text{and} \quad \text{k} = 1, ..., 100 & \end{array} \tag{4}$$

It should be noticed that password F^a is a genuine password of the person a but it is treated as a forged password for the person k.

The classifiers c_1,c_2,c_3 and c_4 work in supervised mode. After training, classifier can be switched into verified mode.

4 Classifier in the Verify Mode

Unlike the other approaches [1], [4], [16] instead of mixing biometric features we used different machine learning algorithms only. As was explained above, the global classifier has a multi-layer structure which brightly follows from Fig. 1.

During verification procedure three ensembles of classifiers form the answer in the majority voting scheme, which is presented synthetically in Fig. 3. The voting scheme uses IF THEN rule to build ultimate decision whether the password of a given user is authorized ("a") or unauthorized ("u") [9–11]:

$$\begin{array}{ll} if & \bigcup_{i=1}^{3} \{a\}_{EC_i} > \bigcup_{i=1}^{3} \{u\}_{EC_i} \; then \text{ user is legitimated } (\text{“}a\text{”}) \\ otherwise & \text{user is illegitimated } (\text{“}u\text{”}) \end{array} \tag{5}$$

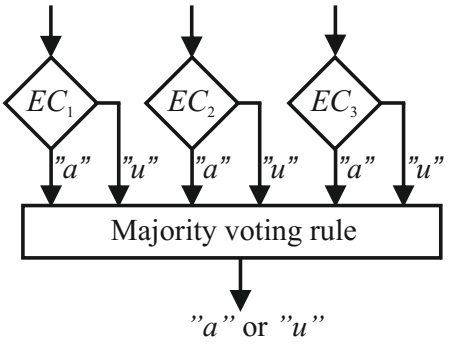

Fig. 3. Ensemble of classifiers work in the password verification mode.

5 Results Obtained

Experiments carried out allow us comparing the obtained results with other methods. In the consecutive experiments, the best structure of classifiers has been established. The results have been gathered in Table 1. In this experiment instead of ensemble, various single classifiers have been checked. Table 1 presents accuracy of the best five single classifiers from many others which were tested.

Table 1. Computer users recognizing accuracy of the best single classifiers.

Classifier	Accuracy [%]
NaiveBayesUpdatable	84.85
RBFClassifier	84.56
NaiveBayesSimple	84.25
HoeffdingTree	84.22
NaiveBayes	82.98

In the next investigation we built a different ensemble classifiers. Instead of single classifier, the one-layer ensemble with different members (c_1,c_2,c_3 and c_4) have been tested. Table 2 presents the accuracy results for the best two one-layer ensemble classifiers.

From Table 2 follows that accuracy of ensemble classifiers is still not enough in professional biometric systems. Such results should be improved. To realize this task we proposed the new type of multilayer ensemble classifiers. The general structure of these ensembles have been already presented in Fig 1. In the long time investigation carried out, we found the best classifier members of ensemble classifiers. Members of the three ensemble classifiers have been presented in Table 3 as well as the best accuracy which such classifiers reached.

Additionally we built confusion matrix for the ensemble which worked in the computer users verification mode. It presented in Table 4. Confusion matrix has

Table 2. Different ensemble classifiers (like as in Fig 2) and their accuracy for the users recognition by means of keystroke dynamics.

Ensemble Classifiers		Accuracy [%]
c_1	BayesNet	
c_2	NaiveBayesSimple	
c_3	IBK	88.7
c_4	RandomForest	
c_1	Kstar	
c_2	HoeffdingTree	
c_3	NaiveForest	87.7
c_4	IsolationForest	

Table 3. The best selection of the ensemble classifier members (like as in Fig 3) for keyboard dynamics - based biometric system.

Members of ensemble of ensemble classifiers		Accuracy [%]
c_1	Kstar	
c_2	BayesNet	
c_3	LibSVM	98.4
c_4	HoeffdingTree	

Table 4. Confusion matrix for the best ensemble classifier selection.

TP = 971	FN = 29
FP = 3	TN = 997

Table 5. Accuracy comparison with different methods.

Method	Accuracy [%]	Database	Used feature(s)	Nr of users
Proposed method	98.40	try4-mbs	latency	100
Loy [5]	96.14	try4-mbs	latency	100
Monrose [7]	92.14	private	variuos	63
Guven [2]	95.00	private	variuos	12
Sung [15]	98.13	practice	variuos	100

been built for ensemble with members from Table 3. Our results were compared with other investigations, which is presented in Table 5. Form this table follows that proposed multilayer classifier structure gives the best accuracy level compared to state of the art proposition announced in literature. It means that in some cases biometric systems can be simply improved.

6 Conclusions

From obtained results follow that new type of multilayer classifier gives the highest accuracy level compared to single classifier approaches and for other single layer ensemble classifier structures. Accuracy of 98.4% is promising result and can be treated as very high biometric factor. Hence our method can be included

into the professional biometric systems based on the keystroke dynamics. The newest literature announced that it is possible to obtain the highest accuracy level compare to our proposition, but it needed more sophisticated method what causes that system works slowly.

In future we will try to investigate other ensemble classifiers with various members.

Acknowledgments. This work was supported by the Polish National Science Centre under the grant no. DEC-2013/09/B/ST6/02264.

References

1. Doroz, R., Porwik, P.: Handwritten signature recognition with adaptive selection of behavioral features. In: Chaki, N., Cortesi, A. (eds.) Computer Information Systems – Analysis and Technologies. CCIS, vol. 245, pp. 128–136. Springer, Heidelberg (2011)
2. Guven, A., Sogukpinar, I.: Understanding users' keystroke patterns for computer access security. Computers Security **22**(8), 695–706 (2003)
3. Kirkby R.: Improving Hoeffding Trees, PhD thesis, University of Waikato (2007)
4. Kang, P., Cho, S: Keystroke dynamics-based user authentication using long and free text strings from various input devices, Information Sciences **308**, 72–93 (2015)
5. Loy, C.C., Lim, C.P., Lai, W.K.: Pressure-based typing biometrics user authentication using the fuzzy ARTMAP neural network. In: International Conference on Neural Information Processing, Taiwan (2005)
6. Loy, C.C., Lai, W.K., Lim, C.P: Keystroke patterns classification using the ARTMAP-FD neural network. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Taiwan (2007)
7. Monrose, F., Rubin, A.D.: Keystroke dynamics as a biometric for authentication. Future Generation Computer Systems **16**(4), 351–359 (2000)
8. Panasiuk, P., Saeed, K.: Influence of database quality on the results of keystroke dynamics algorithms. In: Chaki, N., Cortesi, A. (eds.) CISIM 2011. CCIS, vol. 245, pp. 105–112. Springer, Heidelberg (2011)
9. Porwik, P., Doroz, R., Wrobel, K.: A new signature similarity measure. In: World Congress on Nature Biologically Inspired Computing, NaBIC 2009, pp. 1022–1027. IEEE (2009)
10. Porwik, P., Doroz, R., Orczyk, T.: The k-NN classifier and self-adaptive Hotelling data reduction technique in handwritten signatures recognition. Pattern Analysis and Applications, 1–19 (2015)
11. Porwik, P., Doroz, R.: Self-adaptive biometric classifier working on the reduced dataset. In: Polycarpou, M., de Carvalho, A.C.P.L.F., Pan, J.-S., Woźniak, M., Quintian, H., Corchado, E. (eds.) HAIS 2014. LNCS, vol. 8480, pp. 377–388. Springer, Heidelberg (2014)
12. Rybnik, M., Panasiuk, P., Saeed, K.: User authentication with keystroke dynamics using fixed text. In: International Conference on Biometrics and Kansei Engineering, ICBAKE 2009, pp. 70–75. IEEE (2009)
13. Rybnik, M., Panasiuk, P., Saeed, K., Rogowski, M.: Advances in the keystroke dynamics: the practical impact of database quality. In: Cortesi, A., Chaki, N., Saeed, K., Wierzchoń, S. (eds.) CISIM 2012. LNCS, vol. 7564, pp. 203–214. Springer, Heidelberg (2012)

14. Rybnik, M., Tabedzki, M., Saeed, K.: A keystroke dynamics based system for user identification. In: 7th Computer Information Systems and Industrial Management Applications, CISIM 2008, pp. 225–230. IEEE (2008)
15. Sung, K., Cho, S.: GA SVM Wrapper Ensemble for Keystroke Dynamics Authentication Department of Industrial Engineering, Seoul National University, San 56–1, Shillim-dong, Kwanak-gu, Seoul, 151–744, Korea (2005)
16. SZS, I., Cherrier, E., Rosenberger, C., Bours, P.: Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords. Computers Security **45**, 147–155 (2014)
17. Teh, P.S., Teoh, A.B.J., Tee, C., Ong, T.S.: Keystroke dynamics in password authentication enhancement. Expert Systems with Applications **37**(12), 8618–8627 (2010)
18. Wrobel, K., Doroz, R., Palys, M.: A method of lip print recognition based on sections comparison. In: IEEE Int. Conference on Biometrics and Kansei Engineering, pp. 47–52. Tokyo Metropolitan University Akihabara, Tokyo, Japan (2013)
19. Wrobel, K., Doroz, R., Palys, M.: Lip print recognition method using bifurcations analysis. In: Nguyen, N.T., Trawiński, B., Kosala, R. (eds.) ACIIDS 2015. LNCS, vol. 9012, pp. 72–81. Springer, Heidelberg (2015)