Projection-based Person Identification

Dora Neubrandt[1], Krisztian Buza[2]

[1] Department of Computer Science and Information Theory,
Budapest University of Technology and Economics
dori@biointelligence.hu
[2] Knowledge Discovery and Machine Learning,
Rheinische Friedrich-Wilhelms-Universität Bonn, Germany,
chrisbuza@yahoo.com
http://www.biointelligence.hu/typing.html

Abstract. The increasing interest in person identification based on keystroke dynamics can be attributed to several factors. First of all, it is a cheap and widely applicable technique, whereas online services such as internet banking or online tax declaration require reliable person identification methods. Furthermore, there are various attack techniques against the existing identification methods, thus combining the existing methods with new person identification methods could improve the reliability of the identification. Recent research shows that person identification based on machine learning using keystroke dynamics data works surprisingly well. This is because the dynamics of typing is characteristic to users and a user is hardly able to mimic the dynamics of typing of another user. In this paper, we propose to use a projection-based classification technique for the task of person identification based on keystroke dynamics.

Keywords: keystroke dynamics, machine learning, person identification, PROCESS

1 Introduction

Person identification became an important research topic as it may be used in various online services ranging from internet banking to online tax declarations. Conventional person identification methods use, for example, passwords or biometric features such as fingerprints or iris patterns. Despite the fact that solely password-based identification systems are very vulnerable to various threats, most online services' identification procedures rely on passwords.

With the dynamics of typing we mean a time series, in which each of the values correspond to the duration of a keystroke. The dynamics of typing has been shown to be characteristic to individuals [1, 2], and people are hardly able to mimic others' typing dynamics. Furthermore, person identification based on keystroke dynamics is cheap and it may be implemented with a standard keyboard (i.e., it does not necessarily require special hardware). Therefore, it is widely applicable and suitable for the task of online person identification.

Our goal is to develop an accurate keystroke dynamics-based person identification technique. We consider the task of *person identification based on keystroke dynamics* as a time-series classification problem for which various techniques have been developed recently[3], one of them is a projection-based approach, PROCESS [3], which has only been applied to the classification of electroencephalograph (EEG) signals previously. Despite the promising results, PROCESS has not been applied to person identification based on keystroke dynamics yet. This paper aims to close this gap.

In particular, we propose to use a projection-based classification method for the task of person identification based on keystroke dynamics. We performed experiments on publicly available real-world typing dynamics data. Our results show that our approach outperformed other state-of-the-art classifiers.

The reminder of the paper is organized as follows: Section 2 gives a brief overview of the most important related works. In Section 3 we describe the proposed projection-based classification of keystroke dynamics data. Section 4 presents our experimental results, while we draw conclusions in Section 5.

2 Related Work

Person identification based on keystroke dynamics can be considered as a classification task, for which various classifier may be applied, such as neural networks [4], data evolution methods [5] and classifiers based on Kolmogorov-Smirnov test [6]. While ensemble techniques [7, 8] have been found to be powerful for various classification tasks, their recent applications include person identification based on keystroke dynamics data [9].

In this paper we consider the task of person identification based on keystroke dynamics as a time-series classification problem, for which nearest neighbor classifiers with dynamic time warping (DTW) distance [10] have been shown to be competitive [11, 12] with various, more complex models such as neural networks [13], Hidden Markov Models [14] or "super-kernel fusion scheme". While these empirical results are supported by studies focusing on the theoretical aspects of classification [15, 16], a few instances, called *bad hubs* have been shown to be responsible for a surprisingly large fraction of the classification error of the nearest neighbor classifier [17] which motivated the development of hubness-aware classifiers [18] and regressors [19] and their application to the person identification task [20].

On the other hand, in the case of electroencephalograph (EEG) signals, a projection-based classifier, PROCESS [3] outperformed several hubness-aware classifiers, such as k-Nearest-Neighbor with Hubness-aware Weighting (HWkNN), Naive Hubness Bayesian k-Nearest Neighbor (NHBNN) or Hubness Information k-Nearest Neighbor (HIkNN), both in terms of accuracy and computational time. Despite these promising results, none of the aforementioned works used projection-based classifiers, such as PROCESS, for the task of person identifica-

[3] See Section 2 for an overview of related works.

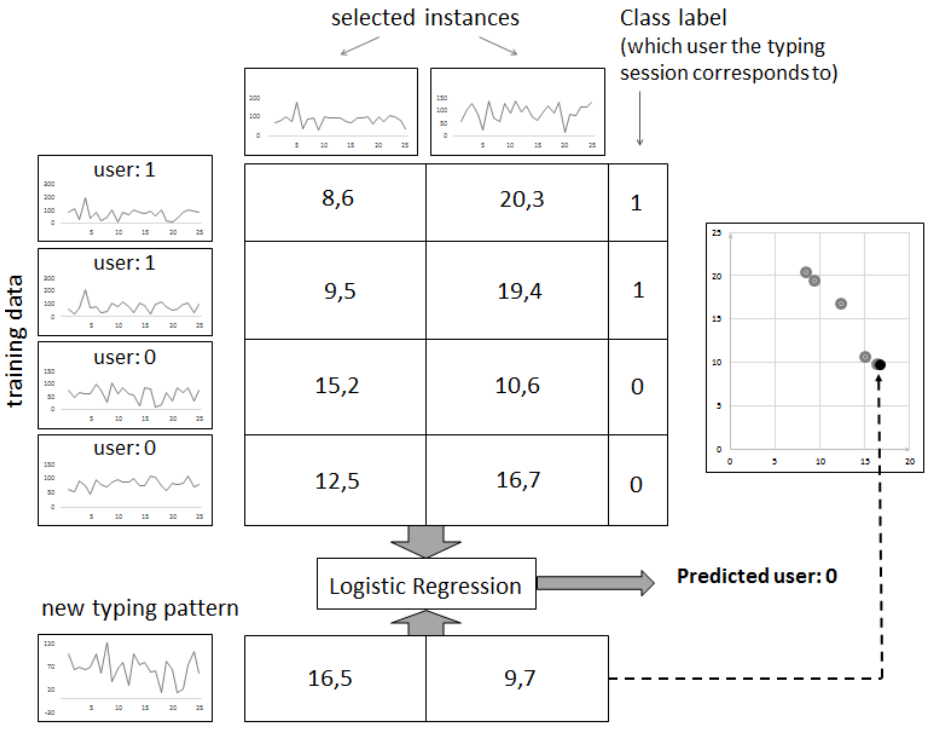

Fig. 1. Projection-based classification of keystroke dynamics data.

tion based on keystroke dynamics previously. In this paper, we aim to close this gap by proposing to use a projection-based classifier for this task.

3 Projection-Based Classification of Keystroke Dynamics Data

As mentioned above, with keystroke dynamics data we mean time series, in which each of the values corresponds to the duration of a keystroke. Thus in our case the instances correspond to time series describing typing dynamics. In particular, each typing session corresponds to a time series.

First, we project time series into a vector space. Subsequently, we train a logistic regression classifier using the projected instances which can be applied to classify new instances. This is illustrated in Fig. 1. Next we describe the details of this approach.

First, we select a subset of the training instances. Then for each training instance, we calculate its distance from the selected instances. Thus we map the time series into a d-dimensional vector space where d is the number of se-

Algorithm 1 Projection of an instance

Require: instance x, set of reference instances $\mathcal{S}$
Ensure: Projected instance p
1: **for** $i = 1 \dots d$ **do**
2: $\quad p[i] \leftarrow$ DTW distance of x and the j-th instance of $\mathcal{S}$
3: **end for**

Algorithm 2 Projection-based classification – Training

Require: Training data $\mathcal{D}$, integer d
Ensure: Trained classifier
1: $\mathcal{S} \leftarrow$ select d instances from $\mathcal{D}$
2: **for** $i = 1 \dots |\mathcal{D}|$ **do**
3: $\quad \mathcal{P}[i] \leftarrow$ project the i-th instance of $\mathcal{D}$ with Alg. 1 using $\mathcal{S}$ as reference instances
4: **end for**
5: $\mathcal{L} \leftarrow$ logistic regression classifier trained on $\mathcal{P}$

lected instances. For the calculation of distances, we use Dynamic Time Warping (DTW), see e.g. [18] for the detailed description of DTW. Next we train a logistic regression classifier using the projected training data. Alg. 1 shows the pseudocode of projection, while Alg. 2 summarizes the training procedure.

If we want to classify a new time series, we project it into the same vector space by calculating its distance from the selected instances, and then we use the previously trained classifier (logistic regression) to predict the class label of the new instance, see Alg. 3.

We note that the standard logistic regression is a binary classifier, i.e., it assumes that only two class labels (usually coded as "0" and "1") are present. However, it can be applied to multiclass classification tasks (such as person identification) if the multiclass task is reduced to a set of binary classification tasks. There are various techniques for that, such as one-vs-rest, one-vs-one or error correcting output coding, see e.g. [21] for details. In state-of-the-art implementations of logistic regression (such as the one in the *scikit-learn* machine learning library[4]) this issue is handled internally, i.e., the logistic regression classifier implements the interface of a multiclass classifier.

[4] `http://scikit-learn.org`

Algorithm 3 Classification of a new instance

Require: Selected instances $\mathcal{S}$, instance x to be classified
Ensure: Predicted class label of instance x
1: $p \leftarrow$ project the x with Alg. 1 using $\mathcal{S}$ as reference instances
2: $\mathcal{L}$.predict_label(p)

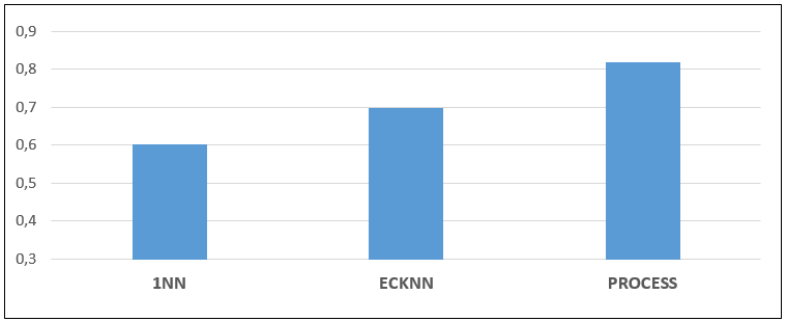

Fig. 2. Accuracy of person identification in case of the proposed projection-based approach, PROCESS, and the baselines.

4 Experiments

In this section we describe the experimental evaluation of projection-based person identification.

We collected time series describing keystroke dynamics, or *typing patterns* for short, from 12 different users over several months, resulting in a collection of 548 typing patterns in total. In each of the typing sessions, the users were asked to type the following short text based on the English Wikipedia page about Neil Armstrong:

That's one small step for a man, one giant leap for mankind. Armstrong prepared his famous epigram on his own. In a post-flight press conference, he said that he decided on the words just prior to leaving the lunar module.

In each typing session, we measured the duration of each keystroke, i.e., the time between pressing and releasing a key. Thus each typing pattern is a time series of keystroke duration values corresponding to a typing session. We used a self-made JavaScript application and a PHP script to capture the aforementioned time series and to save the data. Note that due to typing errors, the length of typing patterns may vary from session to session.

For the evaluation of our approach, we considered the task of *person identification*, wherein a set of training time series and some "new" (i.e., test) time series are given, and the task is to decide, which user belongs to the "new" (i.e., test) time series.[5]

From the collected data, we used the first 5 typing pattern from each user as training data, and the rest is used as test data. This corresponds to the assump-

[5] We note that the task of *person identification* is different from *person authentication*, in case of which the user claims an identify and the system has to decide whether the true identity matches the claimed identity. The task of person identification is inherently more challenging compared with the person authentication task, therefore, we decided to evaluate the proposed approach in context of person identification.

Predicted labels

Actual labels	1	2	3	4	5	6	7	8	9	10	11	12
1	35	0	4	0	1	0	1	4	0	0	0	0
2	4	37	0	0	0	0	1	0	0	0	1	0
3	0	0	43	1	0	1	0	2	0	0	0	0
4	2	0	15	15	0	0	0	9	1	2	1	0
5	0	0	0	0	45	0	0	0	0	0	0	0
6	0	0	0	0	0	46	0	0	0	0	0	0
7	0	11	0	0	0	0	47	0	0	0	1	1
8	0	0	11	1	0	0	0	32	0	1	2	0
9	0	0	0	1	0	0	0	0	13	0	0	0
10	0	0	0	0	0	0	0	0	0	35	0	0
11	0	6	0	0	0	0	0	0	0	0	40	0
12	0	0	1	0	0	0	2	1	0	0	1	10

Fig. 3. Confusion matrix of projection-based classification.

tion that the system is trained on a small set of typing patterns, such as the typing data recorded when the user registers to the system. We measured the quality of the models in terms of accuracy, i.e., the proportion of correctly classified typing patterns. The data, together with a web-based evaluation system, is publicly available at `http://www.biointelligence.hu/typing-challenge/task2/` .

We implemented the projection-based approach in Python. We used all the training instances as "selected" instances, i.e., we calculated the distance of each training and test instance from all the training instances and mapped the data into a $12 \times 5 = 60$ dimensional vector space. After the projection, we trained a logistic regression classifier from the *sklearn* machine learning library on the projected training instances, and used this classifier to label the test instances.

For the reasons mentioned in Section 2, we chose 1NN and k-Nearest-Neighbor with Error Correction (ECkNN) as baselines, see [20] for the detailed description of these models.

As can be seen in Fig. 2, our results show that the proposed projection-based approach, PROCESS, outperformed both baselines (1NN and ECkNN).

Fig. 3 shows the confusion matrix of the projection-based classification. It can be seen that in the majority of the cases the predicted label equals to the actual

label. However, we can notice that there are some users who have very similar typing patterns, so the predictions for their typing patterns may be confused – for example: user 3 and user 4.

4.1 Relevance to real-world applications

Based on the results of the above classification experiments, we aimed to approximate what additional security can be achieved using person identification based on keystroke dynamics in real-world applications. As an example, we consider credit card transactions on the internet. Usually, all the data required to perform a transaction is available on the bank card. We consider a hypothetical system which additionally checks the user's keystroke dynamics. There are two possible errors in such a system: (i) an illegitimate user is allowed to perform a transaction because his/her typing pattern is recognized erroneously as the typing pattern of the owner of the bank card, or (ii) a legitimate user is not allowed to perform a transaction because his/her typing pattern is not recognized properly.

Under the assumption that the classification results are representative for the application, using *per-user* confusion matrices, we estimate the probability of a legitimate user not being allowed to perform a transaction to be approximately 20%. On the other hand, the system is expected to recognize 98% of the illegitimate transactions which may be considered as a reasonably high level of additional security for the case if the bank card gets lost or stolen.[6] [7]

5 Conclusions and Outlook

In this paper we proposed to use a projection-based approach for the task of person identification based on keystroke dynamics. We evaluated our model on publicly available real-world typing data, and compared it with state-of-the-art classifiers. Our results show that our approach outperforms the baselines on the task of person identification and it may provide additional security in applications.

[6] We note that they were calculated under the assumption that the observations in the classification experiment are representative to real-world application scenarios. This includes (but it is not limited to) the assumption of a *naive attacker*. That is, we did *not* assume an "intelligent" attacker who would try to record and/or imitate the dynamics of the legitimate user. Instead, a *naive attacker* was assumed who simply steals a bank card (or the information printed on the card) and tries to use it for internet-based transactions without paying attention to imitate the owner's dynamics of typing.

[7] We also note that there is a trade-off between the aforementioned two types of error and, if required, recognition systems may be tuned in order to decrease one of them, while the other type of error may increase. Though in principle, such tuning is possible in case of projection-based classification as well (for example, based on the continuous output of logistic regression), this is left for future work.

Acknowledgment

D. Neubrandt was supported by the "Új Nemzeti Kiválóság Program" (ÚNKP-16-1-1) of the "Emberi Erőforrások Minisztériuma".

References

1. Antal, M., Szabó, L. Z., László, I.: Keystroke dynamics on android platform. Procedia Technology 19, 820–826 (2015)
2. Monrose, F., Rubin, A.D.: Keystroke dynamics as a biometric for authentication. Future Generation Computer Systems 16(4), 351–359 (2000)
3. Buza, K., Koller, J., K. Marussy: PROCESS: Projection-Based Classification of Electroencephalograph Signals. Lecture Notes in Computer Science 9120, 91–100 (2015)
4. Wong, F.W.M.H., Supian, A.S.M., Ismail, A.F., Kin, L.W., Soon, O.C.: Enhanced user authentication through typing biometrics with artificial neural networks and k-nearest neighbor algorithm. 35th IEEE Asilomar Conference on Signals, Systems and Computers, vol. 2, 911–915 (2001)
5. Kozierkiewicz-Hetmanska, A., Marciniak, A., Pietranik, A.: Data Evolution Method in the Procedure of User Authentication Using Keystroke Dynamics. Computational Collective Intelligence, Lecture Notes in Computer Science 9875, 379–387 (2016)
6. Ceffer A., Levendovszky J.: Kolmogorov-Smirnov test for keystroke dynamics based user authentication. 17th IEEE International Symposium on Computational Intelligence and Informatics, DOI: 10.1109/CINTI.2016.7846387 (2016)
7. Wozniak, M., Jackowski, K.: Fusers Based on Classifier Response and Discriminant Function – Comparative Study. Lecture Notes in Computer Science 5271, 361–368 (2008)
8. Kurzynski, M., Wozniak, M.: Combining classifiers under probabilistic models: experimental comparative analysis of methods. Expert Systems 29(4), 374–393 (2012)
9. Doroz, R., Porwik, P., Safaverdi, H.: The New Multilayer Ensemble Classifier for Verifying Users Based on Keystroke Dynamics. Computational Collective Intelligence, Lecture Notes in Computer Science 9330, 598–605 (2015)
10. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 26(1), 43–49 (1978)
11. Xi, X., Keogh, E., Shelton, C., Wei, L., Ratanamahatana, C.A.: Fast time series classification using numerosity reduction. Proceedings of the 23rd ACM International Conference on Machine Learning, 1033–1040 (2006)
12. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment 1(2), 1542–1552 (2008)
13. Nanopoulos, A., Alcock, R., Manolopoulos, Y.: Feature-based classification of time-series data. International Journal of Computer Research, 10(3), 49–61 (2001)
14. Kim, S., Smyth, P., Luther S.: Modeling waveform shapes with random effects segmental Hidden Markov Models. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 309–316 (2004)

15. Chen, G.H., Nikolov, S., Shah, D.: A latent source model for nonparametric time series classification. Advances in Neural Information Processing Systems 26, 1088–1096 (2013)
16. Devroye, L., Györfi, L., Lugosi, G.: A probabilistic theory of pattern recognition, Springer Science & Business Media (1996)
17. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data., The Journal of Machine Learning Research 11, 2487–2531 (2010)
18. Nenad, T., Buza, K., Marussy, K., B. Kis, P.: Hubness-aware classification, instance selection and feature construction: Survey and extensions to time-series. Feature selection for data and pattern recognition, 231–262 (2015)
19. Buza, K., Nanopoulos, A., Nagy, G.: Nearest neighbor regression in the presence of bad hubs. Knowledge-Based Systems, 86, 250–260 (2015)
20. Buza, K.: Person Identification Based on Keystroke Dynamics: Demo and Open Challenge. CAiSE Forum 2016, 28th International Conference on Advanced Information Systems Engineering (2016)
21. Witten, I.H., Frank, E., Hall, M., Pal, C.: Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann (2016)