Person Identification Based on Keystroke Dynamics: Demo and Open Challenge

Krisztian Buza

Brain Imaging Center, Research Center for Natural Sciences
Hungarian Academy of Sciences, Budapest, Hungary
buza@biointelligence.hu
http://www.biointelligence.hu

Abstract. Person identification based on keystroke dynamics is a challenging task with applications in various domains ranging from online education to internet banking. State-of-the-art solutions for this task are based on machine learning. In this paper, we present our solution which is based on dynamic time warping (DTW) and ECkNN, a recent hubness-aware regression technique. We performed initial evaluation on a dataset containing 200 typing sessions and we show that the proposed approach outperforms popular time-series classifiers. Additionally, we point out that we integrated the proposed approach into a Python-based web server which allows to demonstrate real-world applications of the proposed person identification technique. Furthermore, in order to motivate research in this domain, we announce an open challenge.

Keywords: person identification, keystroke dynamics, hubness-aware regression, challenge

1 Introduction

Conventional techniques for person identification range from passwords to biometric identification, such as fingerprints, iris-patterns, electroencephalograph-based and electrocardiograph-based person identification [6, 8]. Online services, such as online banking or online courses, require cheap, widely accessible and reliable person identification techniques. It was shown that the dynamics of typing is characteristic to particular users, and users are hardly able to mimic the typing dynamics of others [10].

Although the dynamics of typing, e.g. the time series of the duration of keystrokes, is characteristic to users, it is obvious that even the same user can not always type with the *exactly* same dynamics. This is illustrated in Fig. 1. The figure shows the durations of the first 25 keystrokes in case of typing the same text by two different users. Each of the users typed the same text two times. The time series of *user1* are shown in the left of the figure, while the time series of *user2* are shown in the right of the figure. As one can see, the time series of the same user are more similar to each other than the time series of different users.

In: S. España, M. Ivanović, M. Savić (eds.): Proceedings of the CAiSE'16 Forum at the 28th International Conference on Advanced Information Systems Engineering, Ljubljana, Slovenia, 13-17.6.2016, published at http://ceur-ws.org

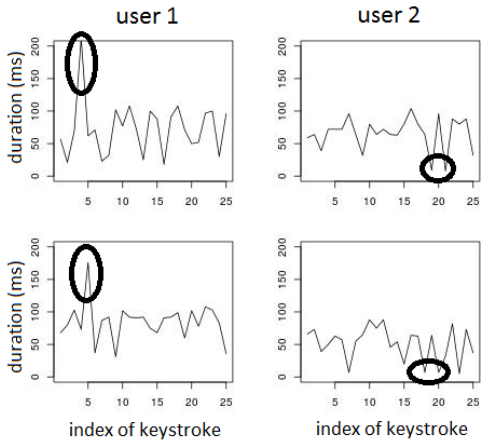

Fig. 1. The duration of 25 consecutive keystrokes in case of two different users (in the left and right) when typing the same text.

In particular, a peak (i.e., an exceptionally long keystroke) close to the fifth position is characteristic to *user1*, whereas exceptionally short keystrokes close to position twenty are characteristic to *user2*. In case if we consider a large set of users, such as millions of students participating in online education, it may be difficult and time-consuming for human experts to identify patterns that are able to reliably distinguishing users from each other. Therefore, approaches based on machine learning are required for user identification based on typing patterns.

In our solution, we consider the task of person identification based on the dynamics of typing as a time-series classification problem, for which various approaches have been introduced ranging from neural networks [11, 18] over Hidden Markov Models [7] to support vector machines [5] and Bayesian networks [2]. However, the 1-nearest neighbour (1NN) classifier with dynamic time warping (DTW) as distance measure was shown to be an extremely competitive classifier, outperforming many complex models, such as neural networks, Hidden Markov Models or super-kernel fusion scheme [19].

Although the empirical evidence is also justified by theoretical results [3, 4], one of the recently observed shortcomings of nearest neighbour models is their suboptimal performance in the presence of bad hubs [12, 17]. Informally, we say that an instance x is a *bad hub*, if x appears as a nearest neighbour of surprisingly many other instances, but x belongs to a class which is different from the class of those instances that have x as their nearest neighbour. With *hubness*, we refer to the presence of bad hubs, a phenomenon that has been observed in various datasets, including time series datasets [15]. For a more formal definition of bad hubs, we refer to [15], in which hubness-aware classifiers are surveyed and applied to the classification of time series. Here, we only note that we made

similar observations for time series representing keystroke dynamics, i.e., bad hubs are present in such data as well.

As the aforementioned studies show, bad hubs are responsible for surprisingly large fraction of the total classification error of nearest neighbour classifiers, therefore, reduction of the detrimental effect of bad hubs can substantially improve the accuracy of time-series classification. Consequently, we base our solution on ECkNN, one of the recent hubness-aware machine learning techniques. In Section 4 we present the results of our initial evaluation on a dataset containing 200 typing sessions and we show that the proposed approach outperforms popular time-series classifiers. Furthermore, we point out that we integrated the proposed approach into a Python-based web server which allows to demonstrate real-world applications of the proposed technique. Moreover, in order to motivate research in this domain, we announce an open challenge.

2 Nearest neighbour regression with error correction

Nearest neighbour regression with error correction (ECkNN) is a hubness-aware extension of the kNN regression. By design it is suitable to various types of data, e.g. vector data, time series, etc., given that an appropriate distance measure between the instances of the dataset is available. As we work with time-series data describing the dynamics of typing, the instances are time series in our case. Next, we describe ECkNN in more detail.

In its training phase, ECkNN implements error correction on the training data. In particular, the corrected label $y_c(x)$ of an instance x is defined as

$$y_c(x) = \begin{cases} \frac{1}{|\mathcal{I}_x|} \sum\limits_{x_i \in \mathcal{I}_x} y(x_i) & \text{if } |\mathcal{I}_x| \geq 1 \\ y(x), & \text{otherwise} \end{cases}, \tag{1}$$

where $\mathcal{I}_x$ denotes the set of those training instances that have x as one of their k-nearest neighbours, $|\mathcal{I}_x|$ is the size of the set $\mathcal{I}_x$ and $y(x)$ is the original (i.e., uncorrected) label of instance x. When ECkNN is applied to predict labels for new instances, it performs k-nearest neighbour regression using the corrected labels. That is: for a new instance x^*, ECkNN searches for the k-nearest neighbours of x^* among the training instances and outputs the average of the corrected labels of the neighbours as the estimated label of x^*. For more details about ECkNN we refer to [1].

As mentioned previously, the dynamics of typing is captured by time series data. In order to use ECkNN with time series data, it is necessary to determine the nearest neighbours of time series. For this purpose, we use Dynamic Time Warping (DTW), an extraordinarily popular time series distance measure that is robust to elongations and noise [13, 15].

3 Person identification based on typing patterns

We base our solution on the wide-spread "classification-via-regression" approach. In particular, we use ECkNN regression for the person identification task in the following way: for each pair of users (u, v) we train a separate model. While doing that, we associate time series describing the typing dynamics of user u with label 0. Similarly, the time series of user v are associated with label 1. When a new time series is presented to the trained model, the model outputs a continuous value between 0 and 1 (bounds are inclusive). Values close to 0 (or 1, resp.) indicate that, according to the model, the new time series is more likely to represent the typing dynamics of user u (or v, respectively).

Usage of pairwise models, i.e., models that distinguish between two users, is consistent with the circumstances under which the person identification problem arise in real-world applications: for example, in case of online exams and online banking, the user claims an identity and the task is to decide if the claimed identity matches the user's true identity. Assume that the claimed identity is u^*, and other users of the system are denoted by $u_i, 1 \leq i \leq n$. In this case, for each user u_i (except u^*), we decide if the typing pattern is more consistent with the typing patterns of u^* than u_i. These n decisions can be implemented in parallel if the number of users is high and several computational units are available.[1]

At each of the above decisions, a simple decision threshold of 0.5 could be applied, i.e., given a model trained to distinguish between the typing patterns of users u and v, if the model outputs less than 0.5 when a new time series is presented to the model, then the decision is u, otherwise the decision is v. However, the simple threshold of 0.5 may be suboptimal, therefore, we learn the threshold in the following way: once the model is trained, we present the time series of the training set to the model and obtain the output of the model for the training time series. Then, we determine the threshold that gives the highest accuracy on the training data.

4 Initial experimental evaluation

Typing Dynamics Data We collected time series data describing the dynamics of typing, or *typing patterns* for short. In the initial evaluation, we used the data from four different users, each of them donated approximately 50 typing patterns, resulting in a collection of 200 typing patterns in total. In each of the typing sessions, the users were asked to type the same short text of a few sentences. In particular, the users were asked to type the following text based on the English Wikipedia page about Neil Armstrong:

[1] As there might be users with similar typing dynamics, depending on the costs of different types of errors, in order to successfully authenticate user u^* we might allow a few of these pairwise decisions to "fail" in the sense that the model outputs that the typing pattern is more likely to belong to user u_i instead of u*. Concretely, in case if we allow t of the pairwise decisions to "fail" in the above sense, we say that we set the value of the *tolerance parameter* to t.

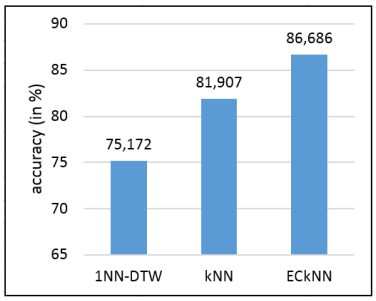

Fig. 2. Accuracy of our approach (ECkNN) and the baselines.

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 both (i) the time between consecutive keystrokes and (ii) the duration of each keystroke, i.e., the time between pressing and releasing a key. We used a self-made JavaScript application, a PHP script and a Python script to capture the aforementioned time series, save and preprocess the data. Note that due to typing errors, the length of the typing patterns (and the corresponding time series) varies slightly.

Experimental protocol In order to simulate the scenario in which users provide few typing patterns when they register into a system, we used the first five typing patterns per user as training data. The remaining typing patterns were used as test data in order to evaluate the system.

We trained models to distinguish two users: i.e, we trained a separate model for each pair of users. For each model, we measured the accuracy, i.e., the ratio of correctly classified instances. We report accuracies averaged over all the pairs of users. We used the binomial test suggested by Salzberg [14] to judge if the differences between the models are statistically significant or not.

We measured the accuracy of our approach, denoted as ECkNN, and the baselines in case of (i) using only the time series of the times between consecutive keystrokes, (ii) using only the time series of the duration of keystrokes, and (iii) using both of the aforementioned two time series. In the latter case, we combined the output of the two models used in the first two cases by averaging their outputs.

We used the publicly available ECkNN-implementation from the PyHubs library.[2] We set $k = 5$ for ECkNN which is in accordance with other works on hubness-aware machine learning [1, 16].

As described in Section I, 1NN-DTW was reported as an extremely competitive time series classifier, therefore we used it as one of the baselines. Addi-

[2] `http://biointelligence.hu/pyhubs`

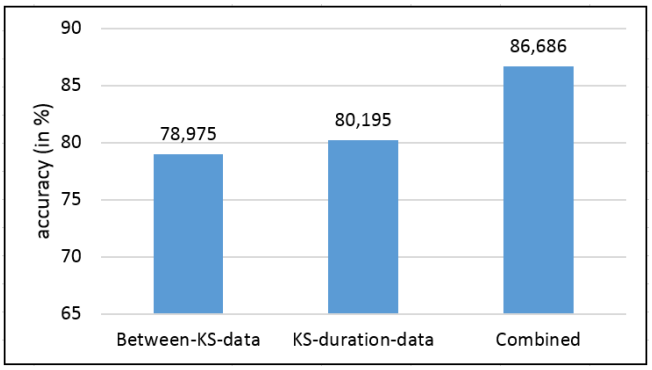

Fig. 3. Accuracy of our approach for various types of data.

tionally, we performed experiments with k-nearest neighbour regression (kNN) with DTW and k=5 and a decision threshold learned on the training data (see Section 3 for the details of the "classification-via-regression" approach and how we learned the decision threshold).

Experimental results Fig. 2 and Fig. 3 summarize the results of our experiments. Fig. 2 shows classification accuracy of the combined models that use both types of time series. We observed that ECkNN outperforms both baselines statistically significantly according to binomial tests at significance level of $p = 0.001$. In Fig. 3, we examine the performance of both types of time series in more detail: keypress duration time-series seems to be more informative than the times between consecutive keystrokes. Most importantly, the combination of both types of information leads to the best performance out of the examined cases.

5 Demo and open challenge

In order to demonstrate applications of the proposed approach in real-world systems, we integrated it into a Python-based web server.

In order to motivate further research in this domain and to allow the evaluation of various models and preprocessing pipelines, we published raw data describing the dynamics of typing and announce an open challenge. The description of the challenge and the dataset is available at:

`http://www.biointelligence.hu/typing-challenge/index.php`

The data used for the challenge contains more than 500 typing sessions. It was collected from 12 users, therefore, the associated identification task is inherently more challenging than the task we considered for the initial evaluation. With "open challenge" we mean that the challenge shall run and submissions of new solutions shall be possible as long as there is interest both from the side of the scientific community and the organizers.

The performance of the proposed approach and the baselines is available on the leaderboard as "ECKNN", "1NN" and "KNN". In the "Person Authentication" task (Task 1), we set the tolerance parameter to $t = 2$, whereas in the "Person Identification" task (Task 2) we used the majority vote of all the pairwise decision models.

6 Conclusions and outlook

In this paper, we considered the task of person identification based on the dynamics of typing. We presented initial results of our experiments with ECkNN, a hubness-aware regression technique. We compared the results of ECkNN with 1NN-DTW and kNN regression which are highly competitive baselines.

We integrated the proposed approach into a Python-based web server which aims to demonstrate the applicability in real-world systems. In order to encourage large scale evaluation of various machine learning techniques and future research, we announced an open challenge.

In case of person identification based on keystroke dynamics, a relatively small set of reliably labeled data is given, e.g. the typing patterns that were recorded when the user registered to the system. However, substantially more unlabeled data (or "weakly labeled" data, under the assumption that the user claims an identity when she tries to log in) may be collected during the usage of the system. The presence of unlabeled data was not taken into account in our work, but it might be exploited using semi-supervised machine learning techniques, such as the SUCCESS approach [9] that was designed for time series classification.

As hubness-aware approaches performed well for the person identification based on the dynamics of typing, we envision that similar techniques may be applied to person identification based on electroencephalograph (EEG) and electrocardiograph (ECG) signals as well.

Acknowledgments. We thank Ladislav Peška for implementing the submission system used for the challenge. We thank Dóra Neubrandt for her help with the implementation of the server demonstrating person identification based on keystroke dynamics. This research was performed within the framework of the grant of the Hungarian Scientific Research Fund – OTKA PD 111710. This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

1. Buza, K., Nanopoulos, A., Nagy, G.: Nearest neighbor regression in the presence of bad hubs. Knowledge-Based Systems 86, 250–260 (2015)
2. Buza, K., Schmidt-Thieme, L.: Motif-based classification of time series with bayesian networks and svms. In: Advances in Data Analysis, Data Handling and Business Intelligence, pp. 105–114. Springer (2009)

3. Chen, G.H., Nikolov, S., Shah, D.: A latent source model for nonparametric time series classification. In: Advances in Neural Information Processing Systems. pp. 1088–1096 (2013)
4. Devroye, L., Györfi, L., Lugosi, G.: A probabilistic theory of pattern recognition, vol. 31. Springer Science & Business Media (2013)
5. Eads, D.R., Hill, D., Davis, S., Perkins, S.J., Ma, J., Porter, R.B., Theiler, J.P.: Genetic algorithms and support vector machines for time series classification. In: International Symposium on Optical Science and Technology. pp. 74–85. International Society for Optics and Photonics (2002)
6. Gargiulo, F., Fratini, A., Sansone, M., Sansone, C.: Subject identification via ecg fiducial-based systems: Influence of the type of qt interval correction. Computer methods and programs in biomedicine 121(3), 127–136 (2015)
7. 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. pp. 309–316. AUAI Press (2004)
8. Marcel, S., Del Millan, J.R.: Person authentication using brainwaves (eeg) and maximum a posteriori model adaptation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(4), 743–752 (2007)
9. Marussy, K., Buza, K.: Success: a new approach for semi-supervised classification of time-series. In: Artificial Intelligence and Soft Computing. pp. 437–447. Springer (2013)
10. Monrose, F., Rubin, A.D.: Keystroke dynamics as a biometric for authentication. Future Generation computer systems 16(4), 351–359 (2000)
11. Nanopoulos, A., Alcock, R., Manolopoulos, Y.: Feature-based classification of time-series data. International Journal of Computer Research 10(3), 49–61 (2001)
12. 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)
13. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. Acoustics, Speech and Signal Processing, IEEE Transactions on 26(1), 43–49 (1978)
14. Salzberg, S.L.: On comparing classifiers: Pitfalls to avoid and a recommended approach. Data mining and knowledge discovery 1(3), 317–328 (1997)
15. Tomašev, N., Buza, K., Marussy, K., Kis, P.B.: Hubness-aware classification, instance selection and feature construction: Survey and extensions to time-series. In: Feature selection for data and pattern recognition, pp. 231–262. Springer (2015)
16. Tomašev, N., Mladenić, D.: Nearest neighbor voting in high-dimensional data: Learning from past occurrences. In: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on. pp. 1215–1218. IEEE (2011)
17. Tomašev, N., Radovanovic, M., Mladenic, D., Ivanovic, M.: A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian knn,. In: Proc. 20th ACM Int. Conf. on Information and Knowledge Management (CIKM). pp. 2173–2176 (2011)
18. 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. In: Signals, Systems and Computers, 2001. Conference Record of the Thirty-Fifth Asilomar Conference on. vol. 2, pp. 911–915. IEEE (2001)
19. Xi, X., Keogh, E., Shelton, C., Wei, L., Ratanamahatana, C.A.: Fast time series classification using numerosity reduction. In: Proceedings of the 23rd international conference on Machine learning. pp. 1033–1040. ACM (2006)