Classifying Users through Keystroke Dynamics

Ioannis Tsimperidis, Georgios Peikos, Avi Arampatzis

Abstract The billions of users connected to the Internet together with the anonymity that each of them can have behind a computer is a source of many risks, such as financial fraud and seduction of minors. Most methods that have been proposed to remove this anonymity are either intrusive, or violate privacy, or expensive. We propose the recognition of certain characteristics of an unknown user through keystroke dynamics, which is the way a person is typing. The evaluation of the method consists of three stages: the acquisition of keystroke dynamics data from 110 volunteers during the daily use of their device, the extraction and selection of keystroke dynamics features based on their information gain, and the testing of user characteristics recognition by training five well-known machine learning models. Experimental results show that it is possible to identify the age group, the handedness, and the educational level of an unknown user with an accuracy of 87.6%, 97.0%, and 84.3%, respectively.

Key words: Keystroke dynamics, user characteristic classification, data mining, feature selection, information gain, digital forensics.

1 Introduction

Today there are more than 4 billion Internet users in the world who use online services in order to communicate, entertain, educate, work, etc. The way we talk over the Internet with someone else differs radically from the way we do it in person.

Ioannis Tsimperidis
Democritus University of Thrace, 67100 Xanthi, e-mail: itsimper@ee.duth.gr

Georgios Peikos
Democritus University of Thrace, 67100 Xanthi, e-mail: georpeik1@ee.duth.gr

Avi Arampatzis
Democritus University of Thrace, 67100 Xanthi, e-mail: avi@ee.duth.gr

Most of the time we do not see the face of our interlocutor, we do not hear his/her voice, and in general, the stimuli that give us information about who is and what his/her intentions are, cease to exist. In addition, we have to consider that often a user is talking to someone completely unknown and that kids participate in these conversations, especially in social networks. It is easily understood that these lurk many dangers, such as financial fraud, seduction of minors, anonymous threats, etc.

One solution to this problem is to know some characteristics of the user we are talking to, such as gender, age, and so on. There are several proposals for achieving these, such as that of Cheung and She [4] who tried to recognize the gender of users from the images generated by their mobile devices and shared in social networks. Arroju et al. [1] try to determine the gender and age of Twitter users based on the contents of their tweets. Although in most cases gender and/or age of users are sought, there are also methods in the literature trying to discover other characteristics, such as the work of Seneviratne et al. [11] where it is attempted to determine, among others, the religion and the spoken languages of unknown users.

All the aforementioned approaches show some limitations. For example, some of them require special equipment, such as special cameras or keyboards, or can only be applied if the target user has an account in some social network, while some others use features derived from the use of certain language, and therefore are incapable of dealing with the multilinguality of today's Internet. In contrary, methods based on keystroke dynamics features are free from such limitations. This is because the only device needed is the common QWERTY keyboard, and furthermore, these methods are language independent since the features derive mainly from how the user uses the keyboard rather than the words he/she writes in a specific language. Finally, data can be collected non-obtrusively, preserving also user privacy contentwise. The keystroke dynamics features used can be categorized into temporal and non-temporal and are described in detail in [14].

The present study is not yet another work on user authentication, as is the case with most studies on keystroke dynamics, but an attempt to classify users according to some inherent or acquired characteristics of them, namely age, handedness, and educational level. The rest of the paper is organized as follows. Section 2 lists the related works in user classification through keystroke dynamics. Section 3 describes the phases of the method followed. Section 4 presents and comments on the classification results obtained by using five well-known machine learning models, i.e. the support vector machine (SVM) witn polynomial kernel, the logistic regression (LR), the Bayes classifier (NB), the Bayesian network classifier (BNC), and the radial basis function network (RBFN). Finally, Section 5 concludes the paper.

2 Related Work

Although most research in keystroke dynamics has as its object user authentication, there are some published papers in user classification. Once again, the characteristic sought in most cases is users' gender, followed by age. For example, Buriro et al. [3]

tried to estimate user's gender, age, and handedness. They defined 3 age groups (teenagers, adults, and senior users) and used several machine learning models for classification. The best results were 87.9% and 95.5% accuracy in age and handedness classification, respectively. Random Forest was the most successful classifier among 7 others, in the work of Roy et al. [10]. They divided users into two classes, kids and adults, and used three fixed text datasets. Finally, using an Ant Colony Optimization technique they achieved accuracy of 92.2%.

Studies with more age classes are those of Tsimperidis et al. [15] who divided the users into 4 groups, and Pentel [9] into 6 groups. The former study used 120 down-down digram latencies as features, and with a dataset of 239 logfiles presented 66.1% accuracy coming from MLP combined with a boosting algorithm, while the latter study with data from more than 7,000 users, each of which was recorded for about 320 keystrokes, and 134 keystroke dynamics features in total, reached 61.6% accuracy using Random Forest.

Handedness is a human characteristic that has been extensively researched in terms of economy, sociology, biology, criminology, etc. [6]. In the field of user classification through keystroke dynamics, Brizan et al. [2] collected data from 329 users, and their experimental results showed an F-score of 0.223 for the left-hand class with a baseline of 0.1. Shen et al. [12] exploited a dataset created by 51 users and extracted keystroke durations and digram latencies as features, and achieved an accuracy of 87.75%. Another approach is that of Pentel [8] who collected data from 504 users through an electronic questionnaire. Despite the small number of keystrokes per user in dataset, they managed to present high performance with F-score of 0.995. Similarly, in the work of Shute et al. [13], 65 volunteers were recorded in the same laptop. The authors split the keyboard into six segments, and then fed the features, which were keystroke durations only, in 3 classifiers resulting in an accuracy of 94.5%.

User classification studies based on how users use the keyboard are quite rare. There may be no other published work in seeking age and handedness of unknown users other than those mentioned. In fact, we have not found any paper referring to user classification according to educational level, which is one of the main focuses of our work.

3 Method

Our methodology consists of three consecutive phases. In the first phase, we collected free-text data from volunteers who agreed to participate in the experiment of extracting real-life keystroke dynamics features. In the second phase, we ran a feature selection algorithm to sort the features according to their contained information. In the third phase, the age, the handedness, and the educational level of an unknown user are sought by training and hyperparameter-tuning five well-known machine learning algorithms, namely SVM, LR, NB, BNC, and RBFN.

3.1 Keystroke Dynamics Dataset

Keystrokes dynamics datasets can be created by recording users either in fixed- or in free-text. The term "fixed-text" refers to the typing of a specific text usually in some closed environment, while "free-text" indicates the recording of volunteer during the typical daily use of his/her computer. In this work, the free-text approach is followed as it integrates with the subject's regular typing activities better and is less intrusive.

To create a suitable dataset, a free text keylogger named IRecU, which can be installed on any Microsoft Windows-based devices, was designed and developed. In each of the volunteers who participated in this project, IRecU was installed on their personal computer and it was possible to record their typing at anytime, anywhere they wanted to work, and using any application, gathering data from thousands of keystrokes, in order to get the best possible approximation of the actual use of their computer.

After two recording periods totaling 18.5 months, 110 users were recorded forming a dataset of 362 log files. In each file there are some metadata with the characteristics of the user being recorded, while keystrokes were written in the form:

```
78,2018-03-19,45743645,"dn"
79,2018-03-19,45743769,"dn"
78,2018-03-19,45743785,"up"
79,2018-03-19,45743879,"up"
```

In each record, which is a user's action on the keyboard, there are four fields separated by commas. The first field represents the virtual key code of the key used, the second indicates the date the action took place in the yyyy-mm-dd format, the third is the elapsed time since the beginning of that day (12:00am) in milliseconds, and the forth is the action, "dn" for key-press and "up" for key-release.

Each log file contains data from about 3,500 keystrokes. Demographics of the dataset that are of interest to this research are shown in Table 1. As it can be seen, the dataset is unbalanced in each of the characteristics being studied. However, it is evident that with regards to age and educational level each class is adequately represented, while with regards to handedness the dataset is as unbalanced as it should be, since the ratio of right- to left-handers is approximately 9:1 [5].

3.2 Feature Extraction and Feature Selection

In order to keep the complexity low, among the hundreds of thousands of available keystroke dynamics features we considered the most frequently-used, namely the keystroke durations and down-down digram latencies. For the feature extraction we developed a software application, named ISqueezeU, which reads the files created by IRecU and extracts the desired features.

Table 1 Number of volunteers and logfiles per age, dominant hand, and educational level.

Characteristic	Class	Volunteers		Log Files	
		#	%	#	%
Age	18-25	23	20.9	71	19.6
	26-35	37	33.6	129	35.6
	36-45	37	33.6	117	32.3
	46+	13	11.9	45	12.5
Handedness	Right-handers	98	89.1	322	88.9
	Left-handers	9	8.2	31	8.6
	Ambidextrous	3	2.7	9	2.5
Educational Level	ISCED-3	21	19.1	62	17.1
(According UNESCO)	ISCED-4	7	6.4	23	6.4
	ISCED-5	13	11.8	49	13.5
	ISCED-6	36	32.7	120	33.2
	ISCED-7-8	33	30.0	108	29.8

Although we chose to extract a small part of the available features, their number is n^2+n, with n being the number of keyboard keys, which is a large number that can lead to systems with high time complexity. Therefore, a feature selection procedure is needed.

Of the thousands of features, there must be selected those which are most capable of distinguishing users according to the studied characteristics. A method to do this is by calculating the information gain (*IG*) of each feature *f*, which is the measure that illustrates the ability of that feature to reduce the entropy of a system *x*. It is expressed as:

$$IG(x,f) = H(x) - H(x|f) = -\sum_{i=1}^{m} P(x_i) \cdot lnP(x_i) - \frac{1}{N} \sum_{j=1}^{k} n_j \cdot H(x_j) \,, \qquad (1)$$

In Equation 1, $H(x)$ is the entropy of the system x, and $H(x|f)$ is calculated by splitting the dataset into groups according to the value of the particular feature f. These two terms are analyzed into sums of products, as shown in the equation, where m is the length of vector x, which in the classification problem is the number of classes, and $P(x_i)$ is the probability of class x_i. Also, N is the number of instances of the initial dataset, k is the number of groups that the initial dataset was split, n_j is the number of instances of the j-th group, and $H(x_j)$ is the entropy of the j-th group.

This procedure is also described in the work of Menzies et al. [7] and if applied to every extracted feature in our classification problems, then a list with the amount of information that every feature carries will emerge. In Table 2, the first 5 features are ranked with the highest *IG* for age, handedness, and educational level classification problems, where each of them is represented by the virtual key code of the keys that compose it.

Table 2 Keystroke dynamics features with the highest IG in the three classification problems.

	Age				Handedness				Educational Level		
#	Feature	Key(s)	IG	#	Feature	Key(s)	IG	#	Feature	Key(s)	IG
1	69	E	0.1519	1	69	E	0.0832	1	76	L	0.1801
2	69-82	E-R	0.1016	2	65	A	0.0782	2	32	(space)	0.1727
3	80	P	0.0868	3	79	O	0.0723	3	80	P	0.1354
4	32	(space)	0.0867	4	82	R	0.0618	4	77	M	0.1319
5	68	D	0.0862	5	82-65	R-A	0.0602	5	65-32	A-(space)	0.1294

3.3 Experimental Procedure and Validation of Models

The feature selection procedure indicated 715, 230, and 727 features with non-zero information gain for the age, handedness, and educational level classification problems, respectively. Since we try to predict user characteristics with high precision, we decided to take advantage of any feature that carries some information, and thus all those with non-zero *IG* were used.

Several machine learning models were tested which presented low accuracy and/or too long training times. The five models which presented the best performance in terms of accuracy and time complexity were SVM, LR, NB, BNC, RBFN. Therefore, the results of these models will be presented.

The purpose of model validation is to ensure that the implementations of the models are correct and work as they should. There are many techniques that can be utilized to verify a model and several of them were adopted to validate the five models used in this work.

Firstly, to assess the performance of the five models fairly, we use the well-known 10-folds cross-validation, which divides the data into 10 disjoint parts, uses 9 of them for training and the remaining one for testing, in a round-robin fashion. Secondly, to evaluate the effectiveness of the feature selection procedure we additionally use F-score, as a combined measurement of precision and recall, because accuracy alone cannot fully give the picture of the overall performance of a model when classes are imbalanced. Finally, to assess the ranking ability of the classifiers we use the area under the ROC curve (AUC) or ROC index.

4 Experiments and Results

For each user characteristic contemplated in this paper and for each of the five mentioned models, a large number of experiments of multiclass classification were conducted in Weka to find the values of classifiers' hyperparameters that lead to the best performance, in terms of accuracy (Acc.), time complexity (TBM - Time to Build Model), F-score (F1), and ROC index (AUC).

The results after hyperparameter tuning are shown in Table 3.

Table 3 Performance of the five models in the classification problems.

Model	Age				Handedness				Educational Level			
	Acc.	TBM	F1	AUC	Acc.	TBM	F1	AUC	Acc.	TBM	F1	AUC
SVM	77.1%	0.22	0.767	0.868	94.8%	0.05	0.943	0.824	77.9%	0.22	0.778	0.897
LR	73.5%	4.87	0.734	0.871	95.6%	0.80	0.952	0.970	72.9%	6.89	0.729	0.896
NB	69.9%	0.02	0.694	0.842	89.0%	$<$0.01	0.879	0.621	68.0%	0.02	0.670	0.850
BNC	71.8%	2.47	0.717	0.903	96.1%	0.03	0.959	0.969	63.8%	0.09	0.640	0.861
RBFN	87.6%	5.39	0.875	0.940	97.0%	0.56	0.971	0.961	84.3%	6.98	0.842	0.893

From Table 3, it can be seen that RBFN outperforms in terms of accuracy and F-score all other models in every classification problem examined in this work. In regards to AUC, RBFN has the best performance in age classification and similar value with LR and BNC in handedness classification, and with SVM and LR in educational level classification. However, the fastest model in each experiment proved to be NB followed by BNC and SVM, on the contrary, it has disadvantage in accuracy, F-score, and AUC in almost every case. RBFN and LR have the longest training time, but they are not prohibitive for their use as they are (without condensation method, reducing the dimensionality, etc.). In conclusion, it seems that the RBFN model is the most suitable for user classification according, mainly because it correctly predicts the age group, handedness, and educational level of an unknown user with 87.6%, 97.0%, and 84.3%, respectively.

5 Conclusion

Often, full anonymity on the Internet can make it difficult for users to access useful services, or even worse, be the advantage of malicious users. Existing methods that achieve user characteristics recognition require specific data, or are intrusive, or violate privacy. On the contrary, keystroke dynamics provide a non-intrusive low-cost method using data coming only from the way users use the keyboard.

This study presents a process in which the most suitable keystroke dynamics features are selected to identify age, handedness, and educational level of an unknown user. To accomplish the objective, a new keystroke dynamic dataset was created from recording users during the daily usage of their devices. Then, a feature selection procedure was followed and several machine learning models were tested to show that it is possible to recognize the aforementioned three characteristics of an unknown Internet user with accuracy of 87.6%, 97.0%, and 84.3%, respectively, using only a few hundred features and in a short time of model training.

Having the ability to recognize some characteristics of an unknown user who types a certain piece of text has significant value in digital forensics, targeted advertisement, and facilitating users.

Possible extensions of this research are, firstly, the combination of the results from the various models and their fusion using the theory of Dempster-Shafer. Sec-

ondly, the extension of the existing dataset and its balancing with undersampling, in order to re-conduct experiments and re-confirm our conclusions. Thirdly, the assessment of a model whether it does better or worse between consecutive classes or classes that are far apart (such as age groups) may provide useful directions on how to improve effectiveness. Finally, the implementation of an adaptive system that will utilize data related to the way users type and predict their characteristics. The system will modify its parameters according to new data, with possible changes in the way user types, so as to improve its ability to predict.

References

1. Arroju, M., Hassan, A., Farnadi, G. (2015). Age, gender and personality recognition using tweets in a multilingual setting. In: Proceedings of 6th Conference and Labs of the Evaluation Forum: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Toulouse, France. pp. 23-31
2. Brizan, D.G., Goodkind, A., Koch, P., Balagani, K., Phoha, VV., Rosenberg, A.: Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. Int. J. Hum.-Comput. Stud. **82**,57-68 (2015)
3. Buriro, A., Akhtar, Z., Crispo, B., Del Frari, F. (2016). Age, gender and operating-hand estimation on smart mobile devices. In: Proceedings of 2016 International Conference of the Biometrics Special Interest Group. Darmstadt, Germany. pp. 273-280
4. Cheung, M., She, J.: An analytic system for user gender identification through user shared images. ACM Trans. Multimed. Comput., Commun., and Appl. **13(3)**, 30:1-30:20 (2017)
5. Dragovic, M., Hammond, G.: A classification of handedness using the Annett Hand Preference Questionnaire. Br. J. Psychol. **98(3)**, 375-387 (2007)
6. Goodman, J.: The wages of sinistrality: Handedness, brain structure, and human capital accumulation. J. Econ. Perspect. **28(4)**, 193-212 (2014)
7. Menzies, T., Greenwald, J., Frank, A:. Data mining static code attributes to learn defect predictors. IEEE Trans. Softw. Eng. **33(1)**, 2-13 (2007)
8. Pentel, A. (2017). High precision handedness detection based on short input keystroke dynamics. In: Proceedings of 8th International Conference on Information, Intelligence, Systems & Applications. Larnaca, Cyprus. pp. 1-5
9. Pentel, A.: Predicting user age by keystroke dynamics. In: Silhavy, R. (ed). Artificial Intelligence and Algorithms in Intelligent Systems, pp. 336-343. Switzerland: Springer International Publishing (2018)
10. Roy, S., Roy, R., Sinha, D.D.: ACO-Random forest approach to protect the kids from Internet threats through keystroke. Int. J. Eng. and Technol. **9(3S)**, 279-285 (2017)
11. Seneviratne, S., Seneviratne, A., Mohapatra, P., Mahanti, A.: Predicting user traits from a snapshot of apps installed on a smartphone. ACM SIGMOBILE Mob. Comput. and Commun. Rev. **18(2)**, 1-8 (2014)
12. Shen, C., Xu, H., Wang, H., Guan, X. (2016). Handedness Recognition through Keystroke-Typing Behavior in Computer Forensics Analysis. In: Proceedings of 2016 IEEE Trustcom/BigDataSE/ISPA. Tianjin, China. pp. 1054-1060
13. Shute, S., Ko, R.K.L., Chaisiri, S. (2017). Attribution using keyboard row based behavioural biometrics for handedness recognition. In: Proceedings of 2017 IEEE Trustcom/BigDataSE/ICESS. Sydney, NSW, Australia. pp. 1131-1138
14. Tsimperidis, I., Arampatzis, A., Karakos, A.: Keystroke dynamics features for gender recognition. Digit. Investig. **24**, 4-10 (2018)
15. Tsimperidis, I., Rostami, S., Katos, V.: Age detection through keystroke dynamics from user authentication failures. Int. J. Digit. Crime and Forensics **9(1)**, 1-16 (2017)