The Use of Free-Text Keystroke Dynamic and Single-Layer Perceptron in Users Classification

Chatthip Sutthaluk[1], Sukree Sinthupinyo[2]
Department of Computer Science
Faculty of Science and Technology, Thammasat University,
Pathumthani, Thailand
ithip_302@yahoo.com[1], sukree@cs.tu.ac.th[2]

***Abstract*—Keystroke Dynamic is one of biometric measurement techniques, in which sequences of time intervals from one keypress to another keypress are used to identify users. Nowadays, there are a number of methods used to authenticate users by measuring time interval of typing. This paper presents a different aspect of free-text keystroke dynamic which can be used to classify users by Single-Layer Perceptron (SLP). This method not only helps in identify person who is producing sequence of keystrokes, but also with the results obtained from Neural Network, it can reveal some specific characteristics of that person. Finally, we compared the accuracy obtained from our approach to other methods by 10 fold cross validation technique. The experimental results show that SLP yields the better results and is more comprehensible than other methods compared in this research.**

I. Introduction

Basically, individual has its own natures in typing characters via a keyboard. Some people type slowly while some people type fast. Some type by using both hands on a keyboard but some use only two index finger to type. Therefore, keystroke rhythms produced by each person are different. These sequences of keystroke can reveal typing characteristics of individuals varying in term of their identities.

In this research, free-text typing rhythms produced by a user are used as input data of a Neural Network. This biometric measurement is based on time interval between keypress, which are collected while a user is typing in password or other free text. Then, these collected data are fed to Neural Networks to classify users and determine their specific identities. In this paper, we employ the Single-Layer Perceptron technique to identify users existing in a system. Our proposed idea not only can classify users, but also can reveal the typing characteristics of users. We ran experiments to evaluate three different machine learning techniques and found that the SLP method yields better result. Moreover the weights and structure of SLP can show some typing characteristics of each unique user.

II. Authentication

A. Keystroke Dynamic

Bergadano, Gunetti and Picardi [1] proposed that keystroke dynamic is a method of User Authentication based on a typing verification of two measuring rates; False Acceptance Rates (FAR) or "False Positives" and False Rejection Rate (FRR) or "False Negatives" from two computing formulas:

$$\%\ FRR = \frac{FA}{N} * 100$$

$$\%\ FAR = \frac{FP}{N} * 100$$

where, FA is number of False Alarms,
N is number of total samples, and
FP is number of False Positive.

Bleha et al. [2] combined a decision process of Bayes Classifier with a pattern recognition technique. Obaidat and Sadoun [3] presented techniques which used key hold time combined with interkey for the identification process when users type their login strings. Joyce and Gupta [4] used the method of keyboard latency information capture while they are typing string in login process. Monrose and Rubin [5] applied the concept presented by Joyce and Gupta [4] to authenticate and recognize user.

Haider et al. [6] described a multi-technique approach rather than a statistical approach alone. They combined fuzzy logic, neural networks and statistical techniques to verify users.

TABLE I

SAMPLE TIME MEASURING OF USER'S TYPING PHRASE "ตามหนังสือที่"

Typing	Phrase: "ตามหนังสือที่"											
	ต-า	า-ม	ม-ห	ห-น	น-ั	ั -ง	ง-ส	ส-ื	ื-อ	อ-ท	ท-ี	ี - '
1st Typing	390	231	1642	351	380	241	340	611	200	1542	431	301
2nd Typing	551	200	1643	330	291	590	621	471	401	821	390	230
3rd Typing	1062	811	1052	290	361	651	410	421	851	1472	691	301
4th Typing	781	401	1232	260	321	690	491	491	451	691	410	331
5th Typing	401	260	1412	962	941	611	581	480	301	1883	400	210

In Robinson et al. [7], the keystroke and the key hold intervals were used as input data classified by Minimum Intra-Class Distance (MICD). Gunetti and Picardi [8] proposed a technique which uses keystroke rhythms-based measurement of free text, in which typing of free text samples is compared to verify personal identity. The measurement was based on two typing rates of FAR and FRR. To classify users, Generic Distance Measure is applied along with created classification rules using R Measure and A Measure. Rao [9] used keystroke dynamic as biometric measured by Key press and Key hold time, and the pattern recognition chosen for a classification, including Naïve Bayes Classifier, Nearest Neighbor Classifier and Support Vector Machine.

B. Neural Network in Keystroke Dynamic

Biometric technique employs different approaches for user authentication. One of them is measuring keystroke latencies and using them to authenticate users. This method is interesting because it can verify users at the time they are typing in login string and password before login process and free text after login process. This technique identifies a user through pattern recognition by co-measuring key hold time and key press interval.

Artificial Neural Network has been used in dynamic keystroke in several approaches, e.g. fuzzy ARTMAP, Radial Basis Function Networks (RBFN), and Learning Vector Quantization (LVQ) [3]. Furthermore, some works have been done by using both neural network and other classical pattern algorithms, e.g. the Backpropagation with sigmoid transfer function, hybrid sum-of-product (HSOF), sum-of-product (SOF), potential function and Bayes' rule algorithm, which are applied together to measure a level of efficiency improvement. Haider, Abbas and Zaidi [6] proposed an enhanced classification by selecting fuzzy logic and three-layer feed forward network implemented by the Backpropagation algorithm. Herein, keystroke interval data is included in measurement and compared in terms of error rate. All classifiers are combined together to verify the error rate again.

III. METHODOLOGY

As described in the previous section, when a user types sentences or repeated words, such as user's login and password, their keystrokes can be used for authentication. In this research, SLP is applied to find a relationship between the time interval of key typing and the users themselves. Our approach is described in details as follows.

A. Keystroke Interval Time

The keystroke interval time is measured while a user is typing a free-text sentence. A unit of time used in our experiment is in millisecond. These values will finally be used as input data for the SLP.

Table 1 shows an example of phrases which we collected in our experiments. An often typed phrase in users' everyday work can be divided into a number of character bigrams. Each bigram's time interval is measured and used as one input attribute for the SLP.

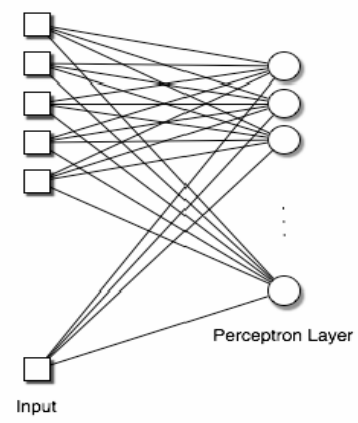

Fig. 1. Structure of our SLP.

B. Structure of Neural Network

Artificial Neural Network is a computational model based on biological neurons, which is aimed at improving computers' learning ability as well as to enable them to practice, and apply knowledge and skills to solve their own problems. This concept has been widely investigated in several works. Many of them invent different models of Artificial Neural Network to apply with their own works. This concept could be helpful for them from such a simple decision to a complex task.

Our network consists of only one layer of perceptrons. The number of perceptrons in the layer is equal to the number of users in a system. Each perceptron represent a particular user. The input of each perceptron comes from the time interval of a bi-gram key press in each sentence. The reason that we select SLP in this work is that we would like to know the user's typing characteristic which is hidden in the weight and structure of the network. Hence, the final weight of our SLP can reveal this desired knowledge.

The perceptrons consists of two ordinary units, i.e. the summation unit and activation unit. In training process both units involve in adjusting weights of the input links. Albeit in recognition process, the activation unit is removed. The node with highest summation of dot product of input vector and weight vector is selected

IV. Experiment

A. Experimental Results

In our experiment, we used the well-known machine learning tool, WEKA [10], to compare the performance of SLP against two other machine learning techniques, namely Decision Tree Learning (DTL) using C4.5 algorithm [11] and Multi-Layer Perceptron (MLP), by using ten-fold cross validation method.

The structure of MLP in our experiment contains one hidden layer and one output layer. Number of hidden nodes is simply equal to number of input nodes. Each input node represents the time interval between each bigram, same as in SLP. Output node also represents each user.

We first collected the keystroke time interval from 15 users who are working at the Office of Don Muang RTAF Base Commander, the Royal Thai Air Force.

Each user was asked to type ten often used phrases, as

TABLE II

Units for Magnetic Properties

ID	Phrases
P1	ตามหนังสือที่
P2	รายละเอียดตามความทราบแล้วนั้น
P3	สน.ผบ.ดม.ตรวจสอบและพิจารณาแล้ว เห็นสมควรให้
P4	จึงเรียนมาเพื่ออนุมัติตามข้อ ๓ ให้ต่อไป
P5	จึงเรียนมาเพื่อโปรดพิจารณาดำเนินการให้ต่อไป
P6	จึงเรียนมาเพื่อพิจารณา หากเห็นชอบด้วย ขอได้ลงชื่อในร่างคำสั่ง ฯ ที่แนบให้ต่อไป
P7	จึงเสนอมาเพื่อพิจารณา หากเห็นชอบด้วย ขอได้ลงชื่อในหนังสือเสนอ กพ.ทอ.ให้ต่อไป
P8	กระผมพิจารณาแล้ว เห็นสมควรลงชื่อในร่างคำสั่ง ฯ ที่แนบให้ต่อไป
P9	กพ.บก.สน.ผบ.ดม.ดำเนินการในส่วนที่เกี่ยวข้อง และเก็บรวบรวมไว้เป็นหลักฐานต่อไป
P10	สำนักงานผู้บังคับทหารอากาศดอนเมือง

TABLE III

Experimental Results. The Asterisk (*) Denotes the Significance Level 0.1 when compared with SLP.

Phrases	Accuracy		
	SLP	MLP	DTL
P1	75.00±19.25	64.17±22.92	80.83±27.23
P2	64.17±20.43	75.00±15.21	69.17±32.88
P3	89.33±14.70	83.83±18.02	64.67±20.49*
P4	79.50±17.32	77.17±20.43	68.83±27.90
P5	69.50±15.44	72.50±15.44	60.67±23.50
P6	74.67±29.92	76.33±29.08	68.17±19.66
P7	73.33±18.84	73.33±18.84	46.67±27.83*
P8	76.00±18.51	80.33±18.84	69.17±22.72
P9	97.50±7.91	97.50±7.91	46.67±30.23*
P10	70.00±25.52	73.33±15.61	61.67±16.76
Average	76.90±10.30	77.35±10.30	63.65±11.99

TABLE IV

THE WEIGHTS OBTAINED FROM SLP TRAINING OF THE PERCEPTRONS REPRESENTING U1 AND U3

User	Phrase: "ตามหนังสือที่"											
	ต-า	า-ม	ม-ห	ห-น	น-ั	ั -ง	ง-ส	ส-ื	ื-อ	อ-ท	ท-ี	ี - '
U1	-1.76	**6.77**	-1.71	1.01	2.06	-2.47	0.30	0.69	-0.31	-0.58	-0.35	0.16
U3	-0.73	2.42	-0.98	-0.11	**-7.15**	-1.63	3.36	0.60	-0.42	-0.88	**-6.41**	3.12

shown in Table 2, ten times for each phrase. Next we identified each user based on its keystroke time interval. Each Machine Learning technique compared in this experiment constructed ten hypotheses; one hypothesis was for one phrase. Finally, we analyzed the results and extract knowledge from our SLP.

B. Knowledge Discovered from SLP

In previous section, we can see from Table III that the accuracies obtained from SLP and MLP are not significantly different. Because structure of SLP is simpler than that of MLP. Hence we can easier extract some kind of knowledge from the SLP's structure.

We show the weight structure of one of perceptron trained from our experiment in Table IV. We can see from the table that the weight of input bigram "า-ม" of the user U1 is highest which means that if the time interval between pressing key "า" and "ม" is relatively high, it is high probability of classify as the user U1. Moreover, for the user U3, the weight of input bigram "น-ั" and "ท-ี" are large negative values which means that the time interval of this bigram should be very low for the user U3.

V. CONCLUSIONS AND FUTURE WORK

We have presented an idea of our ongoing research in discriminating users using their keystroke latencies. As shown in the previous section, the experimental results show that the Neural Networks yield more accurate classification results than those from the Decision Tree method. The percent accuracies from both structure of Neural Networks, namely SLP and MLP, are not significantly different but SLP's structure is much simpler than MLP. So that we can extract knowledge from SLP as shown in Section Knowledge Discovered from SLP.

We can see that the weights obtained from SLP show some particular characteristic of users in keystroke latency. We plan to employ other clustering techniques to extract these characteristics from keystroke latency pattern, which can be used to identify and represent an anonymous user.

REFERENCES

[1] F. Bergadano, D. Gunetti and C. Picardi. "User Authentication through Keystroke Dynamics" ACM Trans. Information and System security, Vol. 5, no. 4, 2002, pp. 367 – 397

[2] S. Bleha, C. Slivinsky and B. Hussien. "Computer – Access Security System Using Keystroke Dynamic", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 12, No. 12, December 1990, pp. 1217 – 1222.

[3] M. S. Obaidat and B. Sadoun. "Verification of Computer Users Using Keystroke Dynamics", IEEE Transaction on Systems, Man, and Cybernetics, Part B. Vol. 27, Issue 2, Apr 1997, pp. 261 – 269

[4] R. Joyce and G. Gupta. "Identity Authentication Based on Keystroke Latencies", Communications of the ACM, Vol. 33, pp. 168 – 176, Feb. 1990

[5] F. Monrose and A. Rubin. "Authentication via Keystroke Dynamic", Proceedings of the Fourth ACM Conference on Computer and Communication Security. Zurich, Switzerland, 1997, pp. 48 – 56

[6] S. Haider, A. Abbas, and A. K. Zaidi. "A Multi–Technique Approach for User Identification Through Keystroke Dynamic", IEEE International Conference on Systems, Man, and Cybernetics. Volume 2., Nashville, TN. USA, 2000, pp. 1336 – 1341

[7] J. A. Robinson, V. M. Liang, J. A. Michael Chambers, and C. L. Mackenzie. "Computer User Verification Using Login String Keystroke Dynamic", IEEE Transaction on Systems, Man, and Cybernetics – Part A: Systems and Humans, Volume 28, No. 2, Mar 1998, pp. 236 – 241

[8] D. Gunetti and C. Picardi. "Keystroke Analysis of Free Text", ACM Transaction on Information and System Security, Vol. 8, No. 3, August 2005, pp. 312 - 347

[9] B. Rao. "Continuous Keystroke Biometric System", Media Arts and Technology, September 2005

[10] H. Witten and E. Frank. "Data Mining: Practical machine learning tools and techniques (Second Edition)", Morgan Kaufmann, San Francisco, 2005.

[11] J. R. Quinlan. "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.