2012

Applying Feature Selection to Reduce Variability in Keystroke Dynamics Data for Authentication Systems

Mark Abernethy
Murdoch University

Shri Rai
Murdoch University

DOI: 10.4225/75/57a841bcbefaf
Originally published in the Proceedings of the 13th Australian Information Warfare and Security Conference, Novotel Langley Hotel, Perth, Western Australia, 3rd-5th December, 2012

APPLYING FEATURE SELECTION TO REDUCE VARIABILITY IN KEYSTROKE DYNAMICS DATA FOR AUTHENTICATION SYSTEMS

Mark Abernethy and Shri Rai
Murdoch University, Perth, Western Australia
mark.abernethy@murdoch.edu.au
s.rai@murdoch.edu.au

Abstract

Authentication systems enable the verification of claimed identity. Password-based authentication systems are ubiquitous even though such systems are amenable to numerous attack vectors and are therefore responsible for a large number of security breaches.

Biometrics has been increasingly researched and used as an alternative to password-based systems. There are a number of alternative biometric characteristics that can be used for authentication purposes, each with different positive and negative implementation factors. Achieving a successful authentication performance requires effective data processing.

This study investigated the use of keystroke dynamics for authentication purposes. A feature selection process, based on normality statistics, was applied to reduce the variability associated with keystroke dynamics raw data.

Artificial Neural Networks were used for classification, and results were calculated as the false acceptance rate (FAR) and the false rejection rate (FRR). Experimental results returned an average FAR of 0.02766 and an average FRR of 0.0862, which were at least comparable with other research efforts in this field.

Keywords
Authentication, Verification, Biometrics, Keystroke Dynamics, Artificial Neural Networks.

INTRODUCTION

The traditional and most commonly used authentication procedure requires legitimate users to have an account on a computer system, and requires them to supply a username and password pair for authentication. The username provides a naming or labelling for an identity; the password (*verification token*) is used to verify a claimed identity (Joyce and Gupta, 1990; Garfinkel et al., 2003; Zhang et al., 2009).

However, user-defined passwords are typically low in entropy (little randomness); high entropy passwords are difficult for users to remember (Schneier, 2000). Thus in common use, passwords are easy to guess or vulnerable to numerous attack vectors. Even when users choose higher entropy passwords, it is common for them to be written down and left in the vicinity of their computer (for easy remembrance); this makes them vulnerable to loss or theft. Other candidate verification tokens are keys and ID cards, which are also easily lost or stolen.

Therefore, token-based authentication procedures continue to be some of the most widely exploited methods for compromising computer systems, and remain a significant cause of system security breaches (Garfinkel et al., 2003). Additionally, there is no guarantee that knowledge or possession of the verification token truly confirms identity (Monrose and Rubin, 2000); it merely verifies presence of the token (Schneier, 2000).

Biometrics is being increasingly investigated as an alternative to token based procedures. Biometrics is discussed in next section and includes a review of research related to keystroke dynamics. Subsequent sections present the experimental method, results and conclusion.

BIOMETRICS

Biometric authentication concerns the use of the physical traits and behavioural characteristics that make each individual unique, and encompasses any personal characteristic that can be used to uniquely verify a person's identity (Monrose and Rubin, 2000). Familiar examples of biometric characteristics are a person's signature, fingerprints, retinal/iris patterns, voice/speech patterns, and facial image patterns. Less familiar examples are hand geometry, gait recognition, and keystroke dynamics.

Biometric authentication systems essentially operate as a pattern recognition system (Jain et al., 2004). That is, features of a biometric characteristic determine a distinct pattern for each person. Authentication is granted or denied on the basis of recognition of this pattern, in data supplied during authentication.

Biometric technologies are increasingly automated (Wayman et al., 2005) and involve (Liu and Silverman, 2001) capture of biological data; extraction of uniquely identifiable features; processing these features into a format that can be stored and later retrieved (the *registered template*); comparison of the registered template with a template processed from a sample provided at the time of authentication (the *query sample*).

To incorporate biometrics into an authentication system, selecting the particular characteristic/s to utilize becomes important. These characteristics should (ideally) meet the following requirements (Matyas Jr and Riha, 2000; Jain et al., 2004) universality; uniqueness; permanence; collectability; performance; acceptability; circumvention.

The following two performance variables are used in biometric research:

- False Acceptance Rate (**FAR**): the number of false positives (Type I errors) divided by the number of samples used to test for Type I errors.

- False Rejection Rate (**FRR**): the number of false negatives (Type II errors) divided by the number of samples used to test for Type II errors..

Keystroke Dynamics

Keystroke dynamics–a behavioural biometric characteristic–involves analyzing a computer user's habitual typing pattern when interacting with a computer keyboard (Monrose and Rubin, 2000). Typing involves predominantly subconscious control of finger movement (for those who type regularly); it incorporates movement characteristics that are different between individuals and consistent over time (Gaines et al., 1980). The habitual nature of typing facilitates the development of a *typing signature* that is distinguishable enough between people to be used for authentication purposes.

Examination of the times between keystroke events reveals a definite pattern for each person (Gaines et al., 1980). *Digraphs* describe pairs of typed characters; they are analyzed because they are the most elemental typing unit (as opposed to analyzing the typing of a complete word, sentence, or paragraph). *Digraph time* or *keystroke latency* is a metric calculated from the time the first key is depressed to the time the second key is depressed. This was the first metric used for digraph analysis; a vector of such metrics can be determined for each person, which described a unique pattern for that person. However, this metric provides a relatively course granularity for use in discerning a typing pattern.

Since Brown and Rogers (1993), the most commonly used metrics in keystroke dynamics research are the *keystroke duration* and *digraph latency.* Keystroke duration is the total time a key is depressed. Digraph latency is the time between two successive keystrokes in typing any particular digraph.

The typing signature gained from habitually typing a known password or passphrase exhibits less variability than that gained from typing general text. However, both exhibit a higher degree of variability than some other biometric characteristics. The next section demonstrates that keystroke dynamics can still confidently be used, provided the appropriate issues are carefully considered.

Related research

Early researchers in keystroke dynamics used various typed samples of prose. Gaines et al. (1980) had 6 participants provide 3 samples totalling 818 words; Umphress and Williams (1985) had 17 participants provide 2 samples totalling 1,700 characters; Leggett and Williams (1988) had 36 participants provide 2 samples totalling 537 characters; Joyce and Gupta (1990) used 6 participants to represent legitimate users and 27 to represent impostors; each participant provided 13 and 30 samples (of 32 characters) in 2 collection sessions.

All four research efforts utilized the one metric (keystroke latency), with Umphress and Williams (1985) and Leggett and Williams (1988) capturing keystroke event times at a 10-millisecond resolution, and Gaines et al. (1980) captured keystroke event times at a 1-millisecond resolution.

The research methodology utilized deterministic statistical calculations to determine the performance variables used to present results. Gaines et al. (1980) reported a FAR of 0.0 and a FRR of 0.04; Umphress and Williams (1985) reported a FAR of 0.0588 and a FRR of 0.1176; Leggett and Williams (1988) reported a FAR of 0.05 and a FRR of 0.055; Joyce and Gupta (1990) reported a FAR of 0.0025 and a FRR 0.1636.

Though the results seem impressive, there were validity concerns due to the small number of participants, the small number of samples supplied by the participants, the course granularity of the capture resolution, the use of only one metric, and the basic statistical calculations used in the methodology. However, the investigations provided a valuable foundation for successive studies.

Brown and Rogers (1993) conducted the earliest investigation utilizing Artificial Neural Networks (ANNs) as a pattern classifier for keystroke dynamics data. They had 46 participants to represent legitimate users and 15 to represent impostors; each participant provided 41 and 30 samples (of 8 characters). Keystroke event times were captured at a 1-millisecond resolution. They were the first to design their experiment using the keystroke duration and digraph latency metrics.

Obaidat and Sadoun (1997) conducted an extensive experiment to compare various classifiers for pattern recognition; they used 5 statistical pattern recognition techniques, and 8 different ANNs. Fifteen participants provided a total of 9,210 samples each of 7 characters. Cho et al (2000) conducted data collection for their experiment online. They recruited 21 participants to represent legitimate users and 15 to represent impostors, who provided 275 and 75 samples (of 8 characters). Abernethy et al. (2004) recruited 50 participants who provided 50 samples (of 32 characters). All three experiments captured keystroke event times at a 1-millisecond resolution, and used the keystroke duration and digraph latency metrics.

Brown and Rodgers (1993), Cho et al. (2000), and Abernethy et al. (2004) used the multi-layer perceptron (MLP) ANN for classification. Brown and Rodgers (1993) forced the FAR to 0.0 and achieved a FRR of 0.115. Cho et al. (2000) reported a FAR of 0.0 and a FRR of 0.01; Abernethy et al. (2004) achieved a FAR of 0.0119 and a FRR of 0.108. For consistency of comparison, the results achieved by Obaidat and Sadoun (1997) for the MLP were a FAR of 0.0 and a FRR of 0.0005.

Revett et al. (2007) conducted an experiment using the probabilistic neural network. They recruited 20 participants (legitimate users) and 30 (impostors) who provided 13 and 30 samples (of between 6 to 15 characters). Keystroke event times were captured at a 1-millisecond resolution, and 8 primary and secondary metrics were calculated from the captured keystroke times. Results were presented as an average FAR/FRR of 0.039.

From the review, it is evident that experiments involving keystroke dynamics have achieved acceptable accuracy. Also, the use of ANN classifiers demonstrates better accuracy than other classification methods. Further research was needed to see the effect of a much larger number of participants as well as samples per participant. It was expected that the variability evident in the data would increase and a process for reducing this would be needed.

RESEARCH METHOD

Ninety participants were recruited from the author's institution. The recruitment criterion was that participants typed on a standard computer keyboard on a regular basis. Participants provided typed samples of a 20-character string. The composition of the string was based on following character combinations: 'io', 'in', 'no', 'on', 'ul', 'il', 'ly' (recommended by Gaines et al., 1980). The equivalent left hand character combinations are: 'ew', 'eb', 'bw', 'wb', 'rs', 'es', 'st'. Some of these combinations were employed to form a 4-word phrase that was as sensible as possible. Sensibility was important, so participants could more easily learn to type the phrase habitually. The derived phrase was "*lyles best lino sets*".

The following constraints were incorporated into a capture program: no short cuts were allowed (i.e. using copy and paste); correct spelling was required; case sensitivity was imposed; an upper limit of 750 milliseconds, for time intervals between keystroke events, was imposed. These were enforced to ensure consistency, and data integrity, for all typed samples.

For each participant, a brief data collection session was held each week over 8 weeks. Twenty samples were collected each session, with each sample entered one after the other. Samples from the first week were treated as familiarization and were not used in the experiment; so only samples collected in the last 7 sessions were utilized. All data collection sessions were supervised.

Time values associated with keystroke events were captured at a 1-millisecond resolution. The recorded values were the key press and key release event times for each character typed, including the 'Enter' key–required for calculating the digraph latency for the last character typed. Therefore, each participant's raw data file contained 140 correctly typed samples consisting of 40 key press and release time values, and the key press and release time values corresponding to the 'Enter' key.

The keystroke duration and digraph latency metrics were calculated from the raw data files. The outcome of the metrics extraction process was a 40-element array of 20 metric pairs. All 140 samples in each participants data file were thus processed, and the resultant metrics recorded.

Feature Selection

Twenty metrics from the available 40 per sample were selected for the experiment, based on research by Obaidat and Sadoun (1997). This required a feature selection method that reduced the level of noise in the data.

For the purpose of identifying variables responsible for data noise, statistical measures that can estimate a distribution's coincidence with a normal (Gaussian) distribution (or deviation from it) are the normality, kurtosis, and skewness coefficients, and the standard deviation (The Northwestern University Medical School, 2007). These statistics help to identify extreme values in the distribution (i.e. values at the tails of the normal distribution curve). The higher the variable frequencies at the tails, the more noise is evident in the data. The SPSS software was utilized to calculate these statistics for the 40 metrics across all 140 samples per participant. Utilizing the determined statistics, the 'best' 20 metrics were selected for each participant (Abernethy, 2011).

Once the requisite metrics were identified, extracted and recorded, they were normalized according to the min/max method (Indovina et al., 2003).

Final Analysis Procedure

Artificial Neural Networks (ANNs) were used to classify participants typing patterns. Carefully and intelligently guiding the ANN training process can improve its capability to accurately discern patterns in noisy data (Wong et al., 2005). By identifying metrics responsible for data noise and selecting those less affected (as described in the previous section), the ANN is better able to discern patterns in the data. The data files used in the final analysis procedure were those obtained by the feature selection process.

Participants' data files were randomly allocated to the training or non-training group, and subsequently referred to as training group members' files and non-training group members' files respectively. As all 140 samples in each of the non-training group members' files were available for testing purposes, it was considered that 40 data files would be sufficient for this purpose. So, 50 participants' data files were assigned to the training group, and the remaining 40 participants' data files were assigned to the non-training group. The experiment then proceeded in two phases.

ANN Training Phase

One ANN was trained per training group member, and a training input file generated as follows:

- 30 samples (of the 140) were randomly chosen from that member's metric data file, for the *positive* training case. Positive training case samples are those that the ANN is intended to recognize.
- 1 sample was randomly chosen from each of the other training group member's data files, for the *negative* training case. There were 49 such members; thus 49 samples. Negative training case samples are those that the ANN is intended to not recognize.

Therefore, the 50 training group members' training input files consisted of 79 samples each. With 40 samples removed and used for training purposes, 100 samples remained per training group member for testing purposes.

A cross validation file (to assist the training process) was generated for each training group member as follows:

- 10 samples were randomly chosen from the same file from which the 30 training samples were chosen (excluding the training samples).

Therefore, the 50 training group members' validation files consisted of 10 samples each.

The objective of the training phase was to obtain a registered template (for each training group member) associated with their training input file. The back propagation ANN architecture was used as a pattern classifier. On completion of training, the weights of the trained ANNs were used as registered templates and subsequently used during the testing phase.

ANN Testing Phase

A testing input file was generated for each training group member as follows:

- 100 samples (not used in training/validation) for the member being tested were used for the *positive* testing case.
- 100 unused samples from each of the other training group members were used for the *negative* testing case.
- 140 samples from each of the non-training group members were used for the *negative* testing case.

Positive case testing examines whether the ANN has correctly recognized samples belonging to the member that it has been trained to recognize (samples not seen during training). Non-recognition of any positive case sample is a false negative (Type II error). Negative case testing examines whether the ANN has correctly rejected samples belonging to someone other than the member it has been trained to recognize. Incorrect recognition of any negative case sample is a false positive (Type I error).

Therefore, the 50 training group members' testing input files consisted of 10,600 samples each (100 positive case samples, plus 100 negative case samples from each of the other 49 training group members, plus 140 negative case samples from each of the 40 non-training group members).

RESULTS

The classification outcome for an authentication system involving biometrics is the likelihood that two biometric samples either belong to the same individual or to different individuals (Maltoni et al., 2003). This necessitates a subjective decision based on whether or not the predicted outcome should be accepted or rejected. This determination may be termed 'the final classification decision', and typically involves the use of a 'decision threshold' applied to the predicted outcome (Abernethy, 2011).

The individual results, presented in Table 1, indicate that the lowest FAR (column 4) was 0 attained by member 85, whilst members 16 and 52 attained the next lowest FAR of 0.00009524. This meant that for member 85 no impostor samples (out of 10,500) were incorrectly accepted, and that for members 16 and 52 one impostor sample each was incorrectly accepted. Other low FAR scores were 0.00047619 and 0.00085714, attained by members 18 and 27 respectively. These could be considered extremely good FAR results.

FAR scores of 0.15019048, 0.144, 0.126, and 0.10628571 (attained by members 43, 74, 3, and 12 respectively) were the highest of all FAR scores. Other high FAR scores were 0.08647619 and 0.08104762, attained by members 49 and 29 respectively. These would be considered unacceptably high FAR results.

The lowest FRR of 0 (Table 1 column 3) was attained by member 85, whilst members 40 and 52 shared a rate of 0.01. This meant that for member 85 no genuine samples (out of 100) were incorrectly rejected, and for members 40 and 52 one genuine sample each was incorrectly rejected. Other low FRR scores were 0.02 and 0.03, attained by members 45 and 18 respectively. These could be considered good FRR results.

Training group members 61 and 74 score the highest FRR of 0.22. This meant that 22 genuine samples were incorrectly rejected. Other high FRR scores were 0.21 and 0.19, attained by members 3 and 43 respectively. These would be considered unacceptably high FRR results.

The last two rows of Table 1 provide the average and standard deviation FAR and FRR figures for all training group members. The average FAR was 0.02766095 with a standard deviation of 0.03806474. This meant that on average, there were approximately 3 in 100 impostor samples incorrectly accepted. The average FRR was 0.0862 with a standard deviation of 0.0515. This meant that on average, there were approximately 9 in 100 genuine samples incorrectly rejected.

CONCLUSION

The results demonstrate that the experiment performed well compared to other research efforts. The methodology aimed to improve verification accuracy by applying feature selection to keystroke dynamics data. The experiment involved more participants than most studies reviewed, and there were generally many more samples provided by participants (to enable thorough positive and negative case testing).

The experiment confirmed findings of previous research that keystroke dynamics can provide an acceptable alternative to traditional authentication methods, provided variability existent in raw data can be reduced.

The feature selection method (using normality statistics) was successful for many participants, and performed comparably to most previous studies–though the results were not as accurate when compared with other biometric characteristics. However, there may be other feature selection methods that could return improved results.

PARTICIPANT	FALSE REJECTION RATE	FALSE ACCEPTANCE RATE
1	0.06	0.0156
2	0.09	0.0653
3	0.21	0.126
5	0.05	0.0047
7	0.04	0.0025
9	0.07	0.052
12	0.12	0.1063
14	0.06	0.0134
16	0.05	0.0001
18	0.03	0.0005
20	0.06	0.0101
21	0.08	0.0174
23	0.09	0.0141
24	0.09	0.0695
25	0.09	0.0014
27	0.05	0.0009
29	0.13	0.0810
32	0.1	0.0010
34	0.04	0.0012
36	0.05	0.0117
38	0.04	0.0054
40	0.01	0.0032
41	0.09	0.0103
43	0.19	0.1502
45	0.02	0.0219
46	0.05	0.012
47	0.08	0.0217
49	0.06	0.0865
52	0.01	0.0001
54	0.08	0.0115
56	0.13	0.0055
58	0.13	0.0510
60	0.07	0.0011
61	0.22	0.0087
63	0.11	0.0608
65	0.15	0.004
67	0.14	0.0292
68	0.1	0.0184
69	0.06	0.0119
72	0.05	0.0027
74	0.22	0.144
76	0.06	0.0138
78	0.08	0.0110
80	0.09	0.0230
81	0.14	0.0134
83	0.07	0.0044
85	0	0
87	0.12	0.0270
89	0.05	0.028
90	0.13	0.0074
AVERAGE	0.0862	0.0277
STD DEV.	0.0515	0.0381

1. Testing Phase Results

REFERENCES

Abernethy, M. (2011). User Authentication Incorporating Feature Level Data Fusion Of Multiple Biometric Characteristics. PhD thesis, Information Technology, Murdoch University, Australia.

Abernethy, M., Rai, S. M., and Khan, M. S. (2004). User Authentication Using Keystroke Dynamics and Artificial Neural Networks. In Proceedings of the 5th Australian Information Warfare and Security Conference. Perth Western Australia.

Brown, M. and Rogers, S. J. (1993). User Identification via Keystroke Characteristics of Typed Names Using Neural Networks. International Journal of Man-Machine Studies, 39(6): 999-1014.

Cho, S., Chigeun, H., Han, D. H., and Kim, H.-I. (2000). Web-Based Keystroke Dynamics Identity Verification Using Neural Networks. Journal of Organizational Computing and Electronic Commerce, 10(4):295-307.

Gaines, R. S., Lisowski, W., Press, S. J., and Shapiro, N. (1980). Authentication by Keystroke Timing: Some Preliminary Results. Technical report, Rand Corporation. Report number: R-2526-NSF.

Garfinkel, S., Spafford, G., and Schwartz, A. (2003). Practical Unix & Internet Security. O'Reilly & Associates, third edition.

Indovina, M., Uludag, U., Snelick, R., Mink, A., and Jain, A. (2003). Multimodal Biometric Authentication Methods: A COTS Approach. In Proceedings of the Workshop On Multimodal User Authentication (MMUA), pages 99-106, Santa Barbara, California.

Jain, A. K., Ross, A., and Prabhakar, S. (2004). An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 14(1):4-20.

Joyce, R. and Gupta, G. (1990). Identity Authentication Based on Keystroke Latencies. Communications of the ACM, 33(2):168-176.

Leggett, J. and Williams, G. (1988). Verifying Identity via Keystroke Characteristics. International Journal of Man-Machine Studies., 28(1):67-76.

Liu, S. and Silverman, M. (2001). A Practical Guide to Biometric Security Technology. IT Pro (IEEE), 1(1):27-32.

Maltoni, D., Maio, D., Jain, A. K., and Prabhakar, S. (2003). Handbook of Fingerprint Recognition. Springer.

Matyas Jr, V. and Riha, Z. (2000). Biometric Authentication Systems. Technical report, Ecom-Monitor. Available: http://www.ecom-monitor.com/papers/biometricsTR2000.pdf.

Monrose, F. and Rubin, A. D. (2000). Keystroke Dynamics as a Biometric for Authentication. Future Generation Computer Systems, 16(4): 351-359.

Obaidat, M. S. and Sadoun, B. (1997). Verification of Computer Users Using Keystroke Dynamics. IEEE Transactions On Systems, Man, And Cybernetics-Part B: Cybernetics, 27(2): 261-269.

Revett, K., Gorunescu, F., Gorunescu, M., Ene, M., Magahaes, S., and Santos, H. (2007). A Machine Learning Approach to Keystroke Dynamics Based User Authentication. International Journal of Electronic Security and Digital Forensics, 1(1): 55-70.

Schneier, B. (2000). Secrets & Lies: Digital Security in a Networked World. John Wiley and Sons, Inc.

The Northwestern University Medical School (2007). Examining Normality Test Results. PROPHET: StatGuide. Available: http://www.basic.northwestern.edu/statguidefiles/n-dist_exam_res.html.

Umphress, D. and Williams, G. (1985). Identity Verification Through Keyboard Characteristics. International Journal of Man-Machine Studies, 23(3): 263-273.

Wayman, J., Jain, A., Maltoni, D., and Maio, D. (2005). Biometric Systems: Technology, Design and Performance Evaluation, Chapter 1: An Introduction to Biometric Authentications Systems, pages 1-20. Springer-Verlag, first edition.

Wong, K. W., Fung, C. C., Gedeon, T. D., and Ong, Y. S. (2005). Neural Network Applications In Information Technology And Web Publishing, chapter 22. Generalization Of Neural Networks For Intelligent Data Analysis, pages 304-317. Borneo Publishing, Sarawak, Malaysia.

Zhang, J., Luo, X., Akkaldevi, S., and Ziegelmeyer, J. (2009). Improving Multiple-Password Recall: An Empirical Study. European Journal Of Information Systems, pages 1-12.