Developing a Keystroke Dynamics Based Agent Using Rough Sets

Kenneth Revett[1], Sérgio Tenreiro de Magalhães[2], and Henrique M. D. Santos[2]

[1] *University of Westminster*
Harrow School of Computer Science
London, UK HA1 3TP
revettk@westminster.ac.uk

[2] *Universidade do Minho*
Department of Information Systems
Campus de Azurem
4800-058 Guimaraes, Portugal
{psmagalhaes, hsantos} @dsi.uminho.pt

Abstract. Software based biometrics, utilising keystroke dynamics has been proposed as a cost effective means of enhancing computer access security. Keystroke dynamics has been successfully employed as a means of identifying legitimate/illegitimate login attempts based on the typing style of the login entry. In this paper, we collected keystroke dynamics data in the form of digraphs from a series of users entering a specific login ID. We wished to determine if there were any particular patterns in the typing styles that would indicate whether a login attempt was legitimate or not using rough sets. Our analysis produced a sensitivity of 98%, specificity of 94% and an overall accuracy of 97% with respect to detecting intruders. In addition, our results indicate that typing speed and particular digraph combinations were the main determinants with respect to automated detection of system attacks.

1. Introduction

Keystroke dynamics was first introduced in the early 1980s as a method for identifying the individuality of a given sequence of characters entered through a traditional computer keyboard [1]. Keystroke dynamics originated from studies of the typing patterns exhibited by users when entering text into a computer using a standard keyboard. Researchers in the field focused on the keystroke pattern, in terms of keystroke duration and keystroke latencies. Evidence from preliminary studies indicated that typing patterns were sufficiently unique as to be easily distinguishable from one another, much like a person's written signature [1,2]. Efforts focused on acquiring keystroke attributes based on the dynamic aspects of user input. The results from these preliminary studies have formed the basis for a software-based enhancement to login security. The basic idea is to extract characteristic signatures from a particular user's entry of a login ID – and use this information along with the login ID in deciding whether a login attempt is legitimate. If the typing characteristics of the owner of a login ID could be ascertained, then any differences in typing patterns associated with a particular login attempt *may* be the result of a fraudulent attempt to use those details. Thus, the notion of a software based biometric security enhancement system was born. Indeed, there are commercial systems such as BioPassword that have made use of this basic premise [12].

Deterministic algorithms have been applied to keystroke dynamics since the late 70's. In 1980 Gaines [1] presented a report of his work to study the typing patterns of seven professional typists. The small number of volunteers and the fact that the algorithm is deduced from their data and not tested in other people later, results on a lower confidence on the false acceptance ratio (FAR) and false rejection ratio (FRR) values presented. But the method used to establish a pattern was a breakthrough: a study of the time spent to type the same two letters, when together in the text. In 1997 Monrose and Rubin use the Euclidean Distance and probabilistic calculations based on the assumption that the latency times for one-digraph exhibits a Normal Distribution [5]. Later, in 2000, the same authors presented a Bayesian similarity based metric algorithm for identification of attackers [6]. In 2005 Magalhães and Santos [3] presented an improvement of the Joyce and Gupta's algorithm, while Revett and Khan [9] presented evidence of the existence of a set of procedures (typing rhythms, length of the password, *etc.*) that can enhance the precision of these algorithms. In this study, we employ a rough sets based classifier in order to determine which attributes in the input signature are important to the identification of a legitimate owner of a login ID sequence.

The rough set theory, proposed by Pawlak [8,9], is an attempt to propose a formal framework for the automated transformation of data into knowledge. It is based on the idea that any inexact concept (for example, a class label) can be approximated from below and from above using an indiscernibility relationship (generated by information about objects). Pawlak [8] points out that one of the most important and fundamental notions to the rough set philosophy is the need to discover redundancy and dependencies between features. Since then this philosophy has been used successfully in several tasks as, for example, construction of rule based classification schemes, identification and evaluation of data dependencies, information-preserving data reduction [7,10]. In this work, we utilised an implementation of rough sets (Rosetta – see ref 11) in order to determine if a set of rules could be generated that could provide sufficient discriminatory capacity to automatically determine if a user was an intruder. In this study, we asked 100 volunteers to enter a login ID. A small sample of the volunteers was designated as the rightful owner of the login ID. They were instructed to enter it into our system with full knowledge that they were designated as the owners and were instructed to enter their login ID with the same characteristics every time (on average 50 entries). The rest of the volunteers were instructed to enter the login ID as many time as possible over a 7-day period. We recorded specific keystroke dynamics (e.g. digraph times) and then used Rosetta to extract a rule base from this data. The next section describes in detail the experimental method employed in this study, followed by a results section and lastly a brief discussion of this work.

2. Methods

In this study, we asked users (approximately 100) to enter a passphrase (Login ID) that consisted of a string of 14 characters ('ensouspopulare'), which is composed of three words in Portugese, through an Intranet based portal. Please note that all subjects that participated in this study were native Portugese speakers. A subset of the users (10) were designated as the owner of this passphrase and was asked to enter the passphrase on numerous occasions (approximately 50). The entries were collected over a 7-day period to ensure that we acquired a robust sampling of the variations of the input style for passphrase entry. For each passphrase entry we collected all of the digraphs, the time elapsing between successive (3 in total), the total time spent entering the passphrase, and the half-way time

Table 1. This table presents a sample of 5 legitimate users ('1' in the Legit? column) and 5 illegitimate users (with a '0' in the Legit? column). All other values in the table are the digraph times in mS. Please note that there are 5 additional attributes not included in this table for the sake of presentation clarity. The additional attributes are: W1 (first word), W2 (second word), W3 (third word), WH (half the total time), and TT (total time). The TX headings in this table represents the digraph number. Legit refers to whether the entry was made by the designated owner

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	Legit?
281	344	297	218	375	266	328	266	234	313	515	282	281	1
343	266	875	297	250	719	593	250	235	312	281	282	250	1
375	359	250	328	328	469	406	282	265	344	359	344	359	1
250	328	266	234	375	328	516	266	234	297	312	297	235	1
391	250	578	297	250	328	297	265	282	312	594	265	438	1
390	344	266	312	297	313	375	312	266	531	547	453	235	0
546	625	297	344	360	343	641	313	296	344	469	500	219	0
344	359	266	266	312	266	344	265	266	312	266	438	234	0
531	501	843	344	344	453	656	297	750	344	453	328	297	0
390	344	297	281	297	313	453	312	266	391	390	532	265	0

point. These formed the objects in our decision table, which included a binary decision class based on whether the entry was from the legitimate user or not.

Our rough sets software, Rosetta, has a limitation of 500 objects, so we split the decision table into legitimate and illegitimate users (approximately 500 of each) and randomly selected 250 objects from each decision class. We repeated this process 10 times, and report the average results when applicable in this paper. We then discretised the attributes (except for the decision attribute) using an entropy/MDL algorithm. We then split the decision table up in a 70:30 split (legitimate and non-legitimate entries respectively). We generated reducts using the Dynamic Reduct option, exhaustive RSES algorithm. We then generated decision rules that were then applied to the testing set. Since the critical factor in this study is the information content of the rules, we were interested in yielding a rule set with minimal cardinality, while obviously maintaining high accuracy levels. To achieve this aim, we filtered the rules based on support since the initial rule set contained over 74,000 rules – too large to be of practical use. In the next section, we describe the key results that were obtained in this study.

3. Results

In Table 1 we present a sample of the objects in the decision table, which for the sake of clarity does not present the values for the word lengths, total time and the halfway time. We then discretised the entire decision table using the entropy/MDL option in Rosetta, on all attributes except for the decision class. We then split the decision table into a 70:30 split, which we used for training and testing purposes respectively. We then generated dynamic reducts (using the Exhaustive calculation RSES) option in Rosetta. Lastly, we generated rules from the reducts – in order to minimise the redundancy in the resultant rule set. Without any filtering, 74,392 rules were generated. Since the primary goal of this study was to determine if a set of rules could be generated that would allow a software based biometric system to distinguish legitimate from non-legitimate users, to make the system computationally tractable. If Table 2 below, we present data on the relationship between the number of rules (filtered on support) and the classification accuracy.

Table 2. Results from high-pass filtering of the rules based on support. We excluded all rules that had a support less then the specified filter threshold. Note that the accuracy was reduced by just over 2%, but the number of rules was reduced to 0.6% of the default value

Filter Threshold (based on Support)	Accuracy	Number of Rules
<= 0	99.1%	74,392
<= 4	97.8%	2,401
<= 10	97.5%	604
<= 20	96.8%	452

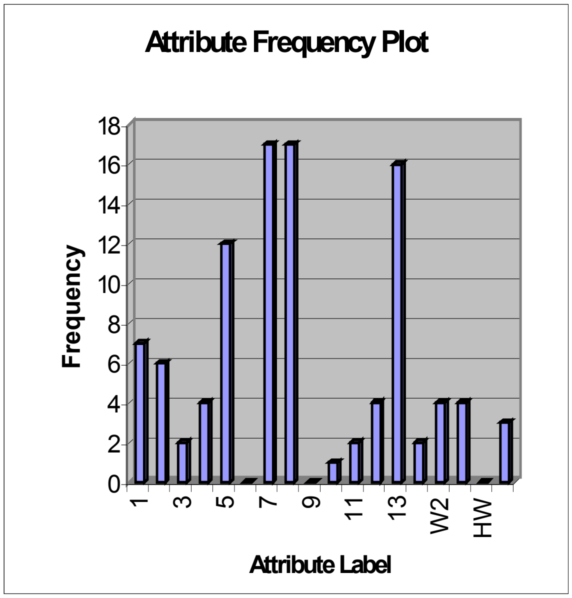
The accuracy of the classification task (with maximal filtering) – segregating legitimate from non-legitimate users was approximately 97%. Table 3 below presents a randomly selected confusion matrix that presents the key summary statistics regarding the classification accuracy of the resulting classifier.

Table 3. A sample confusion matrix for a randomly selected application of the rule set generated using rough sets. The top entry in the 3rd column is the sensitivity, the value below that is the specificity. The entry at the bottom of column two is the positive predictive value (PPV), the last entry in column three is the predictive negative value (PNV) and the lower right hand corner is the overall classification accuracy

Outcomes	0	1	
0	74	3	0.96
1	2	71	0.97
	0.97	0.96	**0.97**

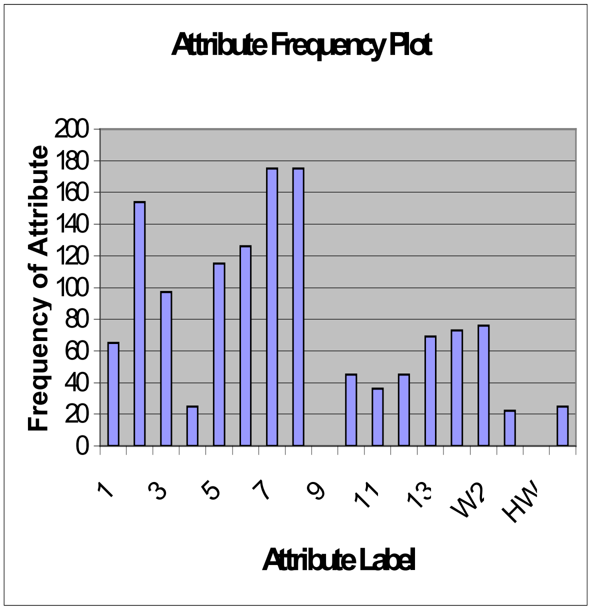
The primary result of this study was the rule set that was used to distinguish a legitimate from an illegitimate login attempt. The primary attributes used in this study were digraphs – the amount of time required to depress two keys (in this study keys on a standard PC keyboard). We collected all digraphs (13 in all), plus the time taken for each word in the login ID, the total time and the half-way time point for entering the login ID. We present summary statistics in Figure 1 below, which depicts a frequency plot of the occurrences of the various digraphs that were found in the resulting rule set.

Figure 1. Frequency plot for all attributes from the rules (17 in total) corresponding to the legitimate login entries – please note that there are a total of 392 instances of all 17 rules for legitimate login attempts



Additionally, we examined the attributes to determine if any were more representative of the rule set than others. We found that attributes 7 & 8 occurred in 100% of the rules, 7,8 & 13 occurred in 94.6% of the rules, and attribute set 5,7,8, & 13 occurred in 72.4% of all instances of the rules (392 instances of 17 rules). This key result indicates that a subset of the attributes, primarily 5,7,8, & 13 are the most frequent occurring attributes and may therefore serve as a signature for a legitimate login attempt, for this particular login. We performed this same analysis on the illegitimate login attempts, which we summary with regards to the attribute frequency in Figure 2. The analysis of the non-legitimate login rules is not as straightforward as for the legitimate login rules. For one, there are many more of them – 175 versus 17 for the legitimate login attempt rules (this excludes the non-deterministic rule set consisting of 260

Figure 2. Attribute label versus frequency for illegitimate login attempts. Note that the total number of unique rules for the deterministic non-legitimate access classification was 175. The values indicate attribute in the total rule base



rules). In addition, the average rule length increased from 5 attributes to 8. Even with these differences, we can account for 65% of the data by focusing on attributes 2,3,5,6,7, & 8 – a reduction of 6/16 attributes (63.5%). Lastly, there were a significant number of non-deterministic rules – which were not able to map attribute values to specific decision classes. There were a total of 260 of such rules – and their examination proved to be quite useful – as they highlight bordering cases between the decision classes. Specifically, we found that in many instances, the same attributes that were significant in the crisp rule set were mapped to different decision classes. After careful, inspection, we found that the difference was based on the magnitude of the attribute – which in this decision table – represents the digraph time. For the non-legitimate login attempts, all digraphs were on the low end of the discretisation range. For the legitimate login attempts, this trend generally held as well, accept for digraphs 5 & 7. It was found unanimously (see Table 4 for details) that for the legitimate user, the digraph values for attributes 5 & 7 were sufficient to distinguish the login attempt in virtually 100% of the cases. That is, the typing speed – reflected in the digraph values was sufficient to distinguish a valid from invalid login attempt, when combined with a

Table 4. A random sample of 6 rules (generated filtering on support >= 20). Note that there is a mixture of deterministic (with a single decision '1' or '0') and non-deterministic rules with two decisions: '1' and '0'. The '*' refers to either 0 if it appears on the left of a tuple, or the maximal value following discretisation if it appears on the right end of a tuple. All rules are generated in conjunctive normal form from discretised data

Rule	Decision
T2([*, 391)) AND T3([*, 399)) AND T5([*, 238)) AND T6([*, 282)) AND T7([*,274)) AND T8([*,235)) AND T12([*,368)) AND T13([*,317) AND W1([*,704)) =>	0
T1([*, 391)) AND T2([*, 269)) AND T3([*, 399)) AND T4([*,274)) AND T5([*,238)) AND T7([*,274)) AND T8([*,235)) AND T13([*,317)) =>	0
T2([*, 269)) AND T3([*, 399)) AND T5([*, 238)) AND T6([*, 282)) AND T7([*,274)) AND T8([*,235)) AND TT([*,4204)) =>	0 and 1
T2([*, 269)) AND T4([*, 274)) AND T5([*, 238)) AND T6([*, 282)) AND T7([*,274)) AND T8([*,235)) AND T13([*,317)) AND W1([*,704)) =>	0 and 1
T3([*,399)) AND T5([246, 289)) AND T7([274,*)) AND T8([*,235)) AND T12([*,368)) AND T13([*,317)) =>	1
T5([246-289)) AND T7([274, *) AND T8([*, 235)) AND T11([*, 430)) AND T12([*,368)) AND T13([*,317)) =>	1

specific digraph pattern. In this particular case, the combination of typing speed for digraphs 5 and 7 were sufficient to discriminate between legitimate owners and attackers/non-legitimate owners of the login ID.

4. Discussion

In this pilot study, we used rough sets to mine a small database of keystroke based biometric data – using only digraph times. The purpose was to develop an approach to developing a situated agent that could be used to determine whether a login attempt was legitimate or not. Using a reasonable sized dataset, we generated a decision table by including the correct decision class (legitimate or non-legitimate owner). Our methodology based on rough sets was able to predict with a high degree of accuracy whether the attempt was legitimate or not based on the decision rules that we generated from rough sets (97% or more classification accuracy). The most interesting result from this study indicates that the digraph times and specific digraphs (see Table 4 for details of the rules) were sufficient to determine whether a user was legitimate. As can be seen in Table 4, the decision class '1' – the non-legitimate owner took the least amount of time in entering the characters of their login ID compared with that of an non-legitimate owner. The results of this study corroborate our previous work [4] - but in this study, we used the keystroke dynamics of a series of owners of a given login ID/passphrase. In addition to typing speed, there appears to be unique digraphs that are sufficient to distinguish the actual owner versus and imposter – the essence of keystroke dynamics. This implies that instead of using all of the digraphs in a signature for verification, we may only require a subset of them – depending on the particular login ID characteristics of the owner. This reduction in the number of attributes that must be stored and searched through reduces the computational load of the verification system. The use of rules generated from rough sets based classifiers can be enhanced by the addition of more attributes into the decision table. With these encouraging results, we are expanding our analysis using much larger datasets, both in terms of the number of objects, but also by the inclusion of additional attributes. We hope to discover what attributes are critical for particular login Ids in order to tailor the system so that it can emphasise those keystroke dynamic features that are indicative of the legitimate owner. For instance, in addition to individual digraphs associated with particular keys, we also investigated obtaining composite attributes such as the total time and half time for the entry of the login ID. Although these attributes did not appear significantly in the rule set, there was clearly a trend for these higher order attributes to segregate across different class decision boundaries. We will continue to explore the addition of higher order attributes into our decision table in order to help increase the classification accuracy of our biometrics based security enhancement system. In particular, we can explore the use of association rules and other rule based systems and compare them with the rough sets approach used in this work.

5. References

[1] Gaines, R. et al. Authentication by keystroke timing: Some preliminary results. Rand Report R-256-NSF. (1980) Rand Corp.

[2] Joyce, R. and Gupta, G.. Identity authorization based on keystroke latencies. *Communications of the ACM*. Vol. 33(2), (1990) pp 168-176.

[3] Magalhães, S. T. and Santos, H. D., 2005, An improved statistical keystroke dynamics algorithm, *Proceedings of the IADIS MCCSIS 2005*.

[4] Magalhães, S. T. and Revett K. Password Secured Sites – Stepping Forward with Keystroke Dynamics, *International Conference on Next Generation Web Service Practises*, Seoul, Korea, 2005.

[5] Monrose, F. and Rubin, A. D., 1997. Authentication via Keystroke Dynamics. *Proceedings of the Fourth ACM Conference on Computer and Communication Security*. Zurich, Switzerland.

[6] Monrose, F. and Rubin, A. D., 2000. Keystroke Dynamics as a Biometric for Authentication. *Future Generation Computing Systems (FGCS) Journal: Security on the Web*.

[7] Pawlak, Z. Rough Sets, International Journal of Computer and Information Sciences, 11, (1982) pp. 341-356.

[8] Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data. Kluwer (1991).

[9] Revett, K. and Khan, A., 2005, Enhancing login security using keystroke hardening and keyboard gridding, *Proceedings of the IADIS MCCSIS 2005*, pp 471-475.

[10] Slezak, D.: Approximate Entropy Reducts. Fundamenta Informaticae (2002).

[11] Rosetta: http://www.idi.ntnu.no/~aleks/rosetta

[12] BioPassword: http://www.biopassword.com/bp2/welcome.asp