A COMPARISON OF ARTIFICIAL NEURAL NETWORKS AND CLUSTER ANALYSIS FOR TYPING BIOMETRICS AUTHENTICATION

Leenesh Kumar MAISURIA, Cheng Soon ONG & Weng Kin LAI <lai@mimos.my>
Information Management and Computational Intelligence Research Group
MIMOS Berhad.
Technology Park Malaysia
57000 Kuala Lumpur,
MALAYSIA

Abstract

Password authentication is the most commonly used identification system in today's computer world. Its security can be enhanced using typing biometrics as a transparent layer of user authentication. Our research focuses on using the time period between keystrokes as the measure of the individual's typing pattern.

The typing pattern of a particular individual can be represented by the weights of a fully trained Multi Layer Perceptron (MLP). Alternatively, each user's typing pattern can be viewed as a cluster of measurements that can be differentiated from clusters of other users.

Keywords: typing biometrics, artificial neural networks, cluster analysis, Multi Layer Perceptrons, K-means clustering, Binary Classification.

1.0 Introduction

The conventional method for user authentication is a password known to the user only. There is no security in the use of passwords if the password is known to an imposter. Hence, to enhance user authentication we may replace passwords with biometric identification of the user e.g. Voice or fingerprint recognition. This may be feasible in terms of cost, when we need to secure highly sensitive material, but for common usage the cost of buying new hardware may be difficult to justify.

Typing biometrics uses the typing patterns of a user as a means of user authentication. The advantage of the typing biometrics based authentication system is that it can be developed as a software program and distributed to other users via the Internet. Hence there is minimal cost to the user. It can be also be used to secure restricted sites on the Internet.

A user authentication system has two issues to consider:

- Recognition of the authorised user
- Rejection of the impostor

Our current research focuses on the means of improving the recognition rates of the authentic user. The methods that are being researched are clustering techniques and Artificial Neural Networks , in conjunction with data processing to improve the identification rates.

1.1 Multiple Layer Perceptron

The MLP is a subset of Artificial Neural networks (ANN) which is modelled on the biological brain to imitate its processing power [1]. This is schematically represented in figure 1.

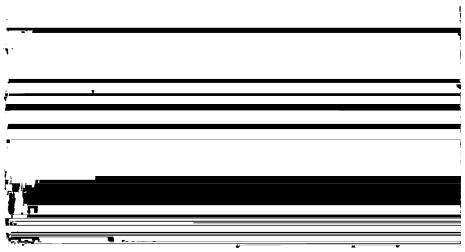

Figure 1: A schematic representation of the Multiple Layer Perceptron.

During training, the weights between the input-hidden layers and the hidden-output layers were modified using

the Hebbian Learning Rule. The network was trained until the weights stabilized.

Once trained, new patterns were fed into the system and compared to decide if the new patterns were similar to the patterns that had been used to train the network.

The input data to the MLP system were between 0 and 1. The schematic representation of the conversion of the patterns from the timings to the required format is shown in figure 2. This conversion was chosen after careful analysis of the collected data.

Statistical analysis showed that 90% of the pattern times were between 6000 and 42000 clock cycles. Hence this range was used for the conversion; and the data that fell below 6000 clock cycles and above 42000 clock cycles were converted to 0s and 1s respectively.

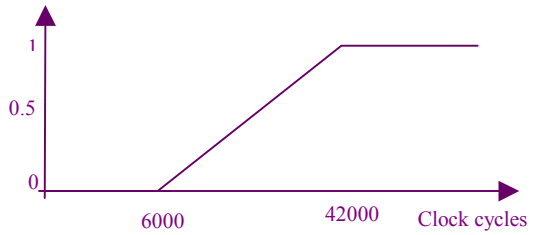

Figure 2: A schematic representation describing the conversion of the individual timings, in each of the typing patterns, to the range of [0-1].

1.2 Cluster Analysis

Clustering is the technique of grouping together pattern vectors that in some sense belong together because of similar characteristics. It seeks to organise information about variables so that relatively homogeneous groups, or "*clusters*," can be formed. The clusters formed with this family of techniques should be highly internally homogenous (members are similar to one another) and highly externally heterogeneous (members are not like members of other clusters).

The modified K-means cluster algorithm is a special case of the partitional cluster algorithm [2, 3]. The basic idea is to start with a random initial partition and iteratively assign patterns to clusters so as to reduce the clustering criterion. Our modified algorithm differs from the basic K-means [2, 4] in that is utilizes different metric spaces for its allocation function, representation function and clustering criteria [3].

1.2.1 Modified K-means Cluster Algorithm

The problem of clustering can be formally stated as follows. Given p patterns in an n dimensional metric space, determine a partition of patterns into K groups, or clusters. For the purposes of this experiment the number of clusters is set to two. The modified K-means cluster algorithm is as follows [2, 3, 4 and 5].

1. Select an initial partition with 2 clusters by randomly selecting two points as the centroids of their respective clusters.
2. Generate a new partition by assigning each pattern to a cluster based on the allocation function. A point is allocated to a cluster if it is closer to the centroid of the cluster. The distance is measured using the particular metric being considered.
3. Compute the new cluster identities (centroids) based on the representation scheme, which is given by Equation 1. This is the weighted average of the reciprocal distances.
4. Repeat steps 2 and 3 until a near optimum value of the cluster criterion is found. The cluster criterion is described in Equation 2 and 3. It is the square error criterion modified to allow different metric spaces.
5. Repeat steps 2 to 4 until cluster membership stabilizes.
6. Perform steps 1 to 5 for 100 repetitions and record the resulting clusters. The cluster configuration with the most occurrences is considered to be the correct partition of the data set.

Eight metrics were investigated in order to explore the effect of different metric spaces for classifying the typing patterns [3]. Some of the metrics was of the correlation type, e.g. Correlation Coefficient and Kendall's Correlation Coefficient. Others are more intuitive measures of distance, e.g. Euclidean and City Block. There were also non-linear metrics, e.g. Minkowsky, Camberra, Chebyshev and Quadratic. The most successful metric was found to be the Camberra Metric, which is shown in Equation 4.

Equation 1

$$m^{(K)} = \frac{\sum_{i=1}^{nK} \frac{x_i^{(K)}}{d(x_i^{(K)}, m(K))}}{\sum_{i=1}^{nK} \frac{1}{d(x_i^{(K)}, m^{(K)})}}$$

Equation 2

$$e_K{}^2 = \sum_{i=1}^{nK} \left(d(x_i{}^{(K)}, m^{(K)})\right)^2$$

Equation 3

$$E_K{}^2 = \sum_{i=1}^{K} \left(e_i{}^2\right)$$

Equation 4

$$d(X_p, X_q) = \sum_{i=1}^{n} \frac{|x_{pi} - x_{qi}|}{|x_{pi} + x_{qi}|}$$

2.0 Experimental Set-up

2.1 Data Collection

20 subjects provided data over 3 sittings.

At the first sitting each subject keyed-in his/her password 60 times (3 sets of 20 patterns[1] each).

At the second sitting each of the subjects provided 3 sets of data using the passwords of 3 other subjects, which were classified as "uninformed imposter".
Before the third sitting, each of the subjects was required to observe the typing patterns of the subjects, that they had been allocated to, and attempt to imitate their typing patterns. At the third sitting each of the subjects keyed-in the passwords of the 3 subjects whom they had observed. These were termed "informed imposters". Each of the sittings was 2-3 weeks apart. This is shown in figure 3.

2.2 Classification using Multi Layer Perceptron

The data from the first sitting was used to train the MLP and record the patterns of the authentic users. To record the characteristic patterns of the subjects, 10 patterns were used from the second set of their first sitting. This was because the user would have been familiar with the password by then but has yet to suffer a deterioration due to boredom or fatigue. The second phase tested the ease with which an imposter was able to imitate a given user by just knowing the password of that user.

[1] One pattern refers to each entry of the password, containing the timing differences between the subsequent keystrokes.

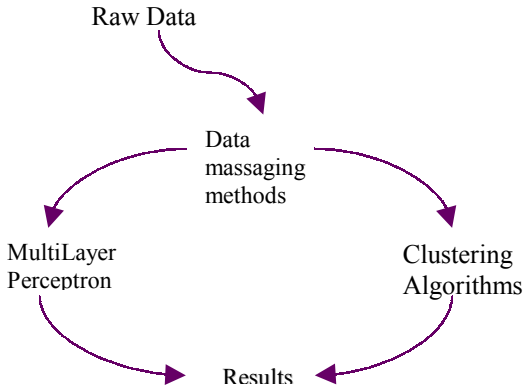

Figure 3: This is a simple representation of the experimental set-up.

In each of the phases, each pattern was analyzed without any modifications (O) and without the last interval. (NL). It had been observed that some users tended to type in their patterns and wait for an inconsistent amount of time before pressing enter. The data massage was applied to the patterns to see if it improved the recognition rates of the authentic user.

2.3 Classification using Cluster Analysis

The second set of data from the first sitting of the authentic user was selected to be the typing template for the reasons above. For each classification test, a point was chosen from either an authorized user file or an impostor file. This pattern vector was then added to the template file and the modified K-means cluster algorithm was applied to it. If the test point was found to be an outlier, then the particular sample was classified as an impostor, if not then the sample was classified as an authorized user. An outlier is defined as a cluster with only one data point in it.

3.0 Results

The results for each person was collected and the average rate of correct classification for both methods and both representations were calculated. There were large differences in the rates for different users. For example, some users managed to achieve 100% acceptance at later attempts of their own password, but others achieved less than 50% acceptance. This observation applies to impostor rejection rates as well.

The data used to generate the graph of informed impostor rejection in figure 4 and 5, was only calculated from the imposters who had made attempts to practice the typing patterns of their allocated authentic users.

3.1 Authentic user recognition and Impostor rejection using MLP

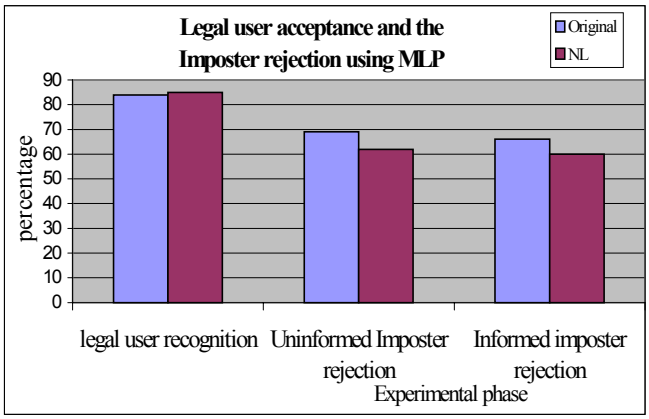

Figure 4: The results of legal user acceptance and impostor rejection using MLP.

The average acceptance rate for the legal user was found to be 84% and 85% , when the data was analysed as original and NL respectively.
From the graph, it can be seen that the impostors were able to better immitate the legal users after having observed legal users keying in their respective passwords.

3.2 Authentic user recognition and Impostor rejection using clustering Algorithms

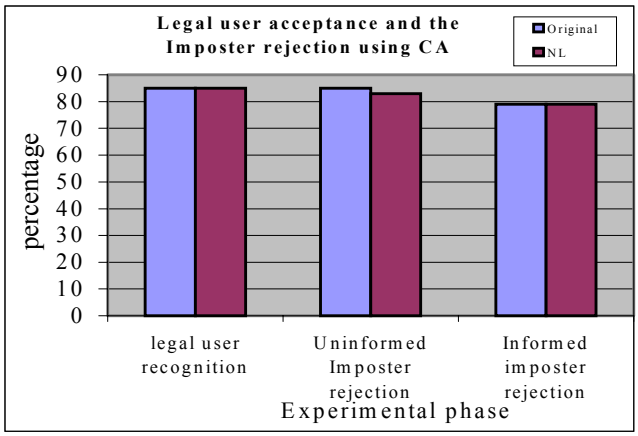

Figure 5: The results of legal user recognition, and the informed and uninformed impostor rejection using cluster analysis.

For cluster analysis, we found that there was a slight decrease in the rejection rate of the impostor when the subject observed the authorised user's typing pattern.

4.0 Discussion

Research so far has suggested that on its own, typing biometrics can not totally replace the accuracy of user identification using passwords. This is because of the variations in one's typing patterns introduced by factors such as one's mental situation e.g. tiredness, or physical injury. However, such factors will also plague many other biometric systems. These effects will not be so pronounced for users on the same PC.

From the analysis of the data, it was found that the last interval had a significantly larger standard deviation than the other measurements. This agreed with the observation that users tended to wait for a variable period before pressing the "Return" key. As shown in figures 4 and 5, the user recognition rates improved by 1% for both MLP and K-means after removing the last interval.

Comparing the results of the ignorant versus informed impostor, the informed user shows a higher success rate. i.e. better imitation of a given password. Another benefit of the new representation is that the degree of deterioration of rejection rates decreases, as shown in table 1. This refers to the improved ability of the new representation to recognise impostors who have practiced typing the authentic password the correct way.

	MLP	K-means
Original	3%	6%
No Last	2%	4%

Table 1: Deterioration of rejection rates of the impostor after observation of typing style.

Both, the MLP and K-means, produced authorised user acceptance rates, which were within 1% of each other for both the representations. However, the identification of impostors using MLP, at 69% rejection rate, needs to be improved, as the K-means achieved an average of 85% rejection. Our future work will focus on improvements to this imposter recognition rate.

Since the passwords used were not the users' actual passwords (so that their privacy and security was not compromised), there may have been variations in typing style for the same user.

5.0 Conclusions

Typing biometrics has the potential to be an additional and transparent layer of user authentication. The comparison of two classification methods, one statistical, the other of soft computing, provides an insight into this binary classification problem. Approximately four thousand sets of typing patterns were collected and tested using the above mentioned methods.

There was a large variation in the latency of the last keystroke, corresponding to the "Return" key. By removing the last data measurement, the authorized user acceptance rate improved. Using the same technique, a smaller drop in the detection of impostors was achieved.

In general, both methods were comparable in classifying the authorized user correctly. However, the modified K-means had a 16% higher impostor rejection rate for the uninformed impostor. This was probably due to the difficulty in selecting a proper bad training set for the MLP.

References

[1] S. Haykin, "Neural Networks : A Comprehensive Foundation", *Prentice Hall*, 1994.

[2] Eric Backer, "Computer Assisted Reasoning in Cluster Analysis", *Prentice Hall*, 1995.

[3] R. S. Michalski, R. E. Stepp, E. Diday, "A recent advance in Data Analysis: Clustering Objects into Classes Characterised by Conjunctive Concepts", *North Holland Publishing Company*, 1981.

[4] Richard O. Duda, Peter E Hart, " Pattern Classification and Scene Analysis", *John Wiley & Sons, Inc.*, 1973.

[5] M.R. Anderberg, "Cluster Analysis for Applications", *Academic Press*, 1973

Biography

Leenesh Maisuria is currently completing his fourth year of a Bachelor of Science/Bachelor of Engineering (Biomedical Engineering) at The University of Western Australia.
He was an assistant researcher at MIMOS Berhad, attached to the Computational Intelligence Group. His research interests are Biomedical engineering applications.

Cheng Soon Ong has a B.Sc. (majoring in computer science and mathematics) from The University of Sydney and is currently completing his final year of a Bachelor of Electrical Engineering (Information Systems) at the University of Sydney.
He was an assistant researcher at MIMOS Berhad, attached to the Computational Intelligence Group. His current research interests are machine learning and optimization problems.

Weng Kin Lai holds a PhD in Engineering from the University of Auckland, and has a MSc (Electronics) from Queen's University of Belfast (UK)
He is currently leading a research team in Computational Intelligence with MIMOS Berhad. His current research interests include evolutionary computation, artificial neural networks and fuzzy logic.
Dr Lai is a member of the IEEE Neural Networks Society.