World Automation Congress

Seventh International Symposium on Manufacturing with Applications

Maui, Hawaii
June 11-16, 2000

Enhanced Password Authentication Through Typing Biometrics With The K-Means Clustering Algorithm

Cheng Soon Ong and Weng Kin Lai

Enhanced Password Authentication through Typing Biometrics with the K-Means Clustering Algorithm

***Cheng Soon* ONG**

School of Electrical and Information Engineering
University of Sydney,
AUSTRALIA.

***Weng Kin* LAI**, *MIEEE*

MIMOS Berhad,
Technology Park Malaysia,
57000 Kuala Lumpur,
MALAYSIA.

lai@mimos.my

ABSTRACT

Password authentication is the most commonly used identification system in today's computer world. It can be enhanced using typing biometrics as a secondary check. Our research focuses on using the time period between keystrokes as the measure of the typing pattern. Each user's typing pattern can be viewed as a cluster of measurements that can be differentiated from clusters of other users. The effect of different metric spaces for cluster analysis was also investigated. The measurements of the keystroke latency are analysed using the modified K-means clustering algorithm to classify the users.

KEYWORDS: *Computer security, typing biometrics, keystroke dynamics, machine intelligence, clustering*

INTRODUCTION

The most widely accepted method for user authentication in a computer system is a password. The method relies on the fact that only the authorised user knows the correct password. There is no security in the use of passwords if an impostor knows the password. Hence, to improve the security of user authentication one option is to replace the passwords with a biometrics identification of the user.

In general, biometric authentication procedures use the features of an individual that are unique to that individual, to identify him or her e.g. fingerprint or iris (eye) recognition. Similarly, typing biometrics based authentication uses an individual's unique typing pattern to separate an authentic user from an impostor. The action of typing the password can be analysed with respect to its physiological characteristics. The latency time between keystrokes, keystroke pressure, key displacement, and key displacement duration are some of the quantifiable components [1,2]

REINFORCED PASSWORDS

Previous research [1,2] has shown that it is possible to identify a user via his or her typing patterns. Our research focuses on using the time period between keystrokes as the measure of the typing pattern, as it can be used as a secondary means of user authentication.

An effective user identification system has to have the following features:

1. Recognise the authentic user and give access
2. Recognise the impostor and disallow access to the system

Another consideration is the fact whether an impostor is able to imitate the typing style of the authorised user if the impostor sees manner in which the authorised user types. Each user's typing pattern can be viewed as a cluster of measurements that can be differentiated from clusters of other users.

METRIC SPACES INVESTIGATED

A traditional metric space [4] is defined by a set and a mapping ρ from pairs of elements in that set to the real numbers such that:

1. $\rho(x,y) = 0 \quad \forall\, x \neq y$
2. $\rho(x,y) = \rho(y,x)$
3. $\rho(x,y) = 0 \;\Rightarrow x = y$
4. $\rho(x,y) + \rho(y,z) \geq \rho(x,z) \quad \forall\, x,y,z$

It is important to remember that the distance measure plays a very significant role in the determination of the clusters. Eight metrics involving Euclidean, Minkowsky, Camberra, Chebyshev, Quadratic, Correlation, City-Block, and Kendall were investigated. This was to explore the effect of different metric spaces for cluster analysis.

GENERAL PARTITIONAL CLUSTER ALGORITHM

The basic idea of partitional clustering is to start with a random initial partition and iteratively assign patterns to clusters so as to reduce the clustering criterion. A general algorithm for iterative partitional clustering [3,5,6,7] may be represented as shown in the following diagram.

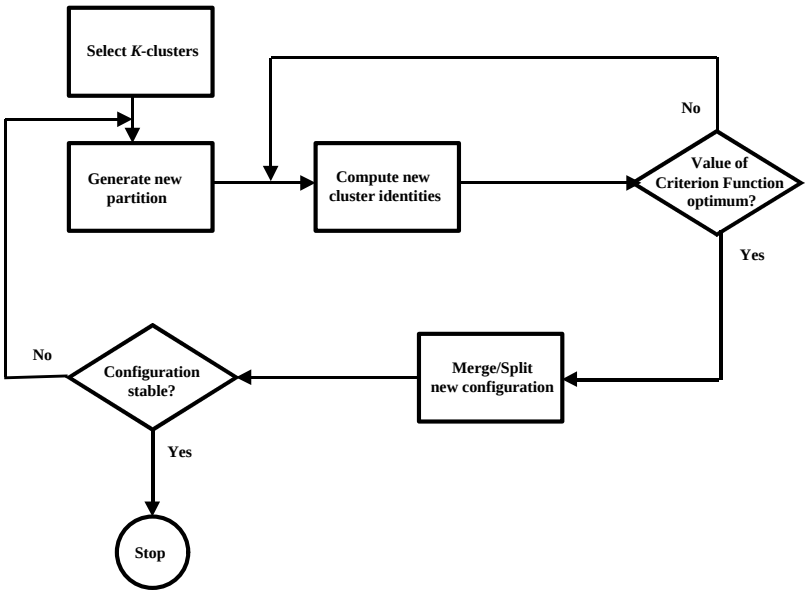


Figure 1: Flow chart for partitional clustering

DESCRIPTION OF CRITERION

As may be seen from the previous section, there is a lot of room for exploration of the partitional clustering algorithm. The key criteria are as follows [5]:

- Initial conditions
- Cluster representation scheme
- Allocation function
- Optimality criterion
- Merge / Split conditions

Since the partitional clustering algorithm is an iterative hill climbing procedure, it does not guarantee the global minimum for a given set of initial conditions. If there is prior knowledge of the domain, then this should be used to determine a more appropriate set of initial partitions. However, for a set of data with unknown properties, the method used is to run a "*large*" number of tests and the cluster configuration that occurs the most often is deemed to be the correct partitioning [3,6].

A cluster representation is a mathematical or geometrical construction that generally characterises objects in the cluster. Among the possible schemes are; the centre of mass of the cluster, the three most extreme objects of the cluster, the line of least inertia, a normal distribution function, a classification tree and conjunctive statements [5]. Given a representation scheme, the representation function determines the best representation for a given cluster. The allocation function is the reverse of the representation function: when given a cluster representation, it determines the objects that belong to each cluster.

The cluster optimality criterion is the measure that determines the correctness of the cluster. It also determines whether to continue with the hill climbing or not. An optimality criterion that is too strict will result in the iterations not converging whereas a loose criterion will produce false clusters.

BASIC K-MEANS CLUSTER ALGORITHM

The K-means cluster algorithm is a special case of the partitional cluster algorithm. The algorithm is the same as the general case. The centre of mass or centroid is chosen as the cluster representation scheme. A point is allocated to a cluster if it is closer to the cluster centre than any other cluster centre. The square error criterion is chosen as the optimality criterion.

Due to the fact that the basic K-means algorithm uses the Euclidean metric as its measure of distance, it intrinsically assumes the Euclidean metric for the representation function and the optimality criterion. Hence to implement the different metrics, a new measure of the cluster centre and the square error criterion was needed. The square error criterion was easily replaced with a similar criterion by observing that the within cluster error is actually the square of the Euclidean distance. Hence the within cluster error, e_K^2 for the general metric space is given by the equation below and the optimality criterion can be obtained.

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

where,

m^K	the mean vector of centroid of the cluster *K*.
$d(X_p, X_q)$	the distance between object represented by attribute vectors X_p and X_q.
$x_i^{(K)}$	vector *i* in cluster *K*.

The centroid or centre of mass of the cluster is more difficult since the equation for the centroid actually assumes the Euclidean metric. However, by looking at the general equation for calculating the centre of mass,

$$m^{(K)} = \frac{\sum_{i=1}^{pK} w_i^{(K)} x_i^{(K)}}{\sum_{i=1}^{pK} w_i^{(K)}} \quad (2)$$

and letting the weights *w* be the reciprocal of the distance measure, the modified centre of mass equation can be easily obtained.

$$m^{(K)} = \frac{\sum_{i=1}^{pK} \frac{x_i^{(K)}}{d(x_i^{(K)}, m_{(K)})}}{\sum_{i=1}^{pK} \frac{1}{d(x_i^{(K)}, m^{(K)})}} \quad (3)$$

With the above modifications, the partitional cluster algorithm can now be used with the different metrics.

EXPERIMENTAL SET-UP

There were three phases to the experiment:

I. Authorised user authentication
II. Impostor rejection
III. Informed impostor rejection

The keystroke latency data consist of the number of clock cycles between the keystrokes. In line with the three phases of the experiment, each user was asked to perform the typing on three different days over a period of two months. On the first day, the user had to type in three sets of 20 passwords. As a result, there were 60 password data samples for each user. During the second session, another set of data was collected. This was in addition to the three other sets of data where each user typed in the passwords of authorised users. However, the subjects have no idea as to the actual typing style of the authorised users prior to typing in the appropriate password. On the third and final session, the subjects were instructed to imitate the passwords of the other users. In this case they had the opportunity to observe the typing style of the appropriate authorised users beforehand.

DATA REPRESENTATION SCHEMES

In addition, several data representation schemes were investigated. There were several reasons for this doing this, viz.

a) to improve the accuracy of the biometric authentication system,
b) to reduce the dependency of the data on the system hardware,
c) to reduce the dependency of the biometric system on the natural variation of the authorised users' keystroke patterns.

Altogether there were 4 different schemes to represent the keystroke latency data collected.

Representation Scheme	Description
Original	*Original representation, without any modifications*
No Last	*Original representation but without the last piece of data in the pattern vector.*
Proportional	*Normalised data from the original scheme*
No Last and Proportional	*Normalised data, but without the last piece of data from the pattern vector.*

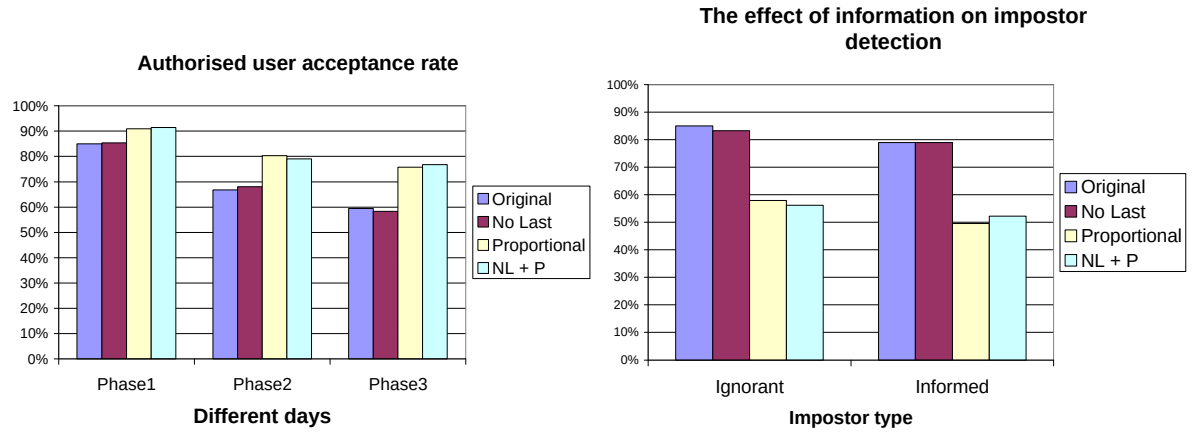
The measurements of the keystroke latency were analysed using the modified K-means clustering algorithm to classify the users. The reference templates for 20 individuals were created from the data of the first session. Each template consisted of a set of 20 measurement vectors, each vector being the latency time between keystrokes of the password.

For each test, a point was chosen from either an authorised user file or an impostor file. This pattern vector was then added to the template file and the modified partitional clustering algorithm was applied to it. If the test point was found to be an outlier, then the point was classified as an impostor, if not, then the sample point was classified as an authorised user. An outlier is defined as a cluster with only one data point in it.

RESULTS

The results of running the cluster algorithm with the authorised user as the test sample are shown in Figure 2(a). These results are for the Camberra metric, which was found to be the most accurate. The results indicated that the user acceptance rate has deteriorated slightly over the period of the experiment. This is probably due to the fact that the password employed in the test was used for the first time in this experiment. Hence, the users were not familiar in typing the password and consequently could not develop a consistent pattern at a later date.

Running the cluster algorithm with the impostor pattern as the test pattern, it was found that there was a slight decrease in the rejection rate of the impostor when the subject observed the authorised user's typing pattern. The results shown in Figure 2(b) are for the Canberra metric.



(a) Authorised users acceptance rate **(b)** Imposter Rejection Rates

Figure 2: Results.

SUMMARY AND CONCLUSIONS

Research to date has shown that typing biometrics, which seeks to analyse the dynamic (behavioural) pattern of a user at a keyboard could be used as a means of personal identification[1]. De Ru and Eloff have applied fuzzy logic to identify the typing patterns from various sources. In this paper, we have reported the results in which clustering techniques have been used instead.

The findings presented herein indicate that cluster analysis can be used as a classification scheme for typing biometrics. The modified cluster algorithm enables the investigation of different metrics to classify a certain domain. All the distance measures investigated here will encapsulate each and every feature of the object to form a useful metric. For the typing biometrics system, it was found that the most effective metric was the Camberra metric. In this investigation, we have basically focussed on three aspects, viz. authorised users authentication, impostor rejection and informed impostor rejection. Partitional clustering is suited to typing biometrics because typing

biometrics is essentially a binary classification problem, and basically the algorithm has to partition the data into two clusters. The modifications made to the basic K-means algorithm allows an extra degree of freedom since we are able to compare the results for different metrics. Therefore, the metric most suitable to the problem domain can be selected. This was found to be the Camberra metric that measures the relative distances, since the numerator contains a difference and the denominator contains a sum. Using the Camberra metric, our average rates of correct classification was 85%. This included detection of authorised users and impostors.

However, there was an observed deterioration in correctly identifying the authorised or correct users after several days. This may be due to a variation in the typing style of the users. To maintain the integrity of the personal passwords that the users use in their daily private transactions, we require the users to make use of a new password in this experiment. Unfortunately, this may mean that the users may not have sufficient time to develop a consistent pattern for this new password. The results obtained have also indicated that the length of the password has a very significant influence on the ability of the classifier to correctly classify the data. This is something that can be addressed in future work.

ACKNOWLEDGEMENTS

The authors wish to thank *Leenesh Kumar Maisuria* for his efforts in obtaining the raw typing biometrics data and all those who participated in this experiment.

REFERENCES

[1] Willen G. De Ru and Jan H.P. Eloff; "*Enhanced Password Authentication through Fuzzy Logic*", IEEE Expert Intelligent Systems, Nov/Dec 1997

[2] E.R. Tee and N. Selvanathan; "*Pin signature verification using wavelet transform*", Malaysian Journal of Computer Science, Vol. 9, No. 2, pp. 71-78, December 1996.

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

[4] Metric spaces, "*Definition of a line*", http://hep.physics.wisc.edu/jnb/line/line.html

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

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

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