Data classification and analyses for performing biometric identification based on keystroke patterns

Paweł Kasprowski[1], **Piotr Kuźniacki**[1], **Adrian Kapczyński**[2]

Streszczenie: The paper presents the way that data classification methods may be used to identify a person who types a given word on a keyboard. Two parameters: latency of keystroke and delays between them where taken into consideration. The data was converted with Linear Discriminant Analysis method and than classified with C4.5 classification methods. The paper presents the results of this experiment.

Słowa kluczowe: keystroke recognition, biometric identification, classification

1. Introduction

Security issues seem to be one of the most important problems of contemporary computer science. One of the most important branches of security is identification of users. Identification may be required for access control to buildings, rooms, devices or information. In case of computer systems we say about access to software and data. The basic aim of identification is to make it impossible for unauthorized persons to access to the specified resources. There are generally three solutions for performing secure identification:

- Token methods (something you have),
- Memory methods (something you know).
- Biometric methods (somebody you are).

The token method has two significant drawbacks. Firstly, the token may be lost or stolen. A person who finds or steals a token may have an access to all the resources that the proper owner of the token was able to access, and there is no possibility to find out if they are the person they claim to be. Secondly, the token may be copied. The easiness of making a copy is of course different for different kinds of tokens, but it is always technically possible.

Memory based methods identify people by checking their knowledge. The most popular memory methods are of course different kinds of passwords. The main drawback of this kind of methods is the unconscious selectivity of human memory. People may do their best to remember a password but they cannot guarantee

[1] Institute of Informatics, Silesian University of Technology ul. Akademicka 16, 44-100 Gliwice
e-mail: {pawel.kasprowski,piotr.kuzniacki}@polsl.pl

[2] Department of Computer Science and Econometrics, Silesian University of Technology ul. Roosevelta 26-28, 41-800 Zabrze
e-mail: {adrian.kapczynski}@polsl.pl

that the information will not be forgotten. Similarly to the token method when a malicious user knows a password it is impossible to check if they are the person they claim to be.

The problems with token and memory-based methods are the main cause of increasing interest in methods of identification based on biometric information of a person.

The terms Biometrics and Biometry have been used since the early 20th century to refer to the field of development of statistical and mathematical methods applicable to data analysis problems in the biological sciences (Tadeusiewicz, 1993). Biometric techniques are frequently used in medicine, agriculture or biology. Biometric identification uses a fact that measurements of biological properties often gives different results for different people. As some measurements are very similar to whole or most of the population - for example body temperature or pulse frequency - biometric identification methods seek measurements, which are characteristic of a single human being only and therefore unique.

Nearly every part of human body has been already used to identification. There are well-established methods for measuring fingerprints, iris, eye retina, face, palm, teeth, ears and even smell (Kasprowski, 2004). There are methods that measure human's behavior patterns such as way of walking (gait), shape of signature or mouse signature.

Unfortunately there isn't one ideal method. Some of biometric methods are not very distinctive - different people may have the same results. Other are very expensive as they need specialized equipment to be measured. Acceptability is also a very important factor. This is for instance problem for fingerprint recognition which is commonly joined with criminal investigations and therefore not acceptable for many people (Maltoni, 2003).

2. Evaluation of biometric methods

The main problem of every biometric method is uncertainity of measurement. Every measurement of the same property usually gives different results. The main goal of identification method is to establish algorithms that are able to extract properties of the measurements that are as constant as possible for subsequent trials. These methods focus on creating general models that describe specific properties of human being which may be found in every trial (Zhang, 2000).

Of course creating such an exact model is difficult and, in most cases, even impossible. That is why biometric identification methods use statistical algorithms to answer the question what is the probability that user is who he claims to be.

There are generally two techniques of biometric identification (Wayman, 2000):

- Identification.
- Authorization.

During the identification process, system collects a sample and than tries to match it with one of the stored templates. Commonly it counts for each template a probability that the sample was collected from the user and chooses one with the highest probability.

Another kind of test is an authorization test. In such test users are first explicitly asked for their names or logins and then system measures a sample of their biometric attributes. After that the system evaluates similarity of the sample to the template of the specified person and accepts or rejects authorization. It is obvious that authorization is much more reliable than identification. Furthermore it is easier to provide and generally faster to perform.

The main issue of biometric methods is how to measure the reliability of the given method. In case of authorization it may be done with two kinds of tests:

There are two kinds of tests when considering authorization (two class) system:

- Genuine test - when a sample is given with correct identification information (login). In another words 'the identified person is telling the truth'. In such case the rate of improper rejections may be measured. This measure is often called a False Rejection Rate (FRR) or False Non-Match Rate.

- Impostor test - when a sample is given with incorrect login. In another words 'the identified person is lying'. Now a rate of improper acceptances may be measured. This measure is called a False Acceptance Rate (FAR) or False Match Rate.

3. Object of experiment

The experiment described in this paper examines possibilities of human identification basing on the way the person types a given word.

Keystroke dynamics, also known as typing rhythms, has a very long tradition in biometrics and is one of the most eagerly developed of all biometric technologies (Joyce, 1990). It was observed in the end of the 19th century that telegraph operators could identify each other only by listening to the rhythm of their Morse code keying patterns. But the first intentional use of keystroke dynamics for person identification was in 1975 (Spillane, 1975). Since then it is generally approved that keystroke biometrics measure typing characteristics are unique to individual person and thus difficult to duplicate (Monrose, 2000).

First stage in each biometric process is collecting a set of 'samples' from every user who should be identified by the system. A sample is a set of biometric data measured for a person in a single measurement. The biometric data may be a different kind of psycho-physiological measurements. Next stage in most methods is creating a 'template' for each user based on previously collected samples. A template is a kind of mean from all samples collected for this user. The process of creating a template is called an 'enrolment' of the user.

To make the process of enrollment easy and convenient for users there was an ActiveX application prepared which was connected to a web page. Examined persons used this web page to enter their login and then they where asked to type a keyword 'POLITECHNIKA'. During the typing our application stored information about:

- dwell time - the time one keeps a key pressed (11 values in ms),
- flight time - the time it takes a person to jump from one key to another (12 values in ms).

The 23 attributes collected during every measurement were then sent to the server together with information about the person identity and stored in the database.

There were overall 1883 probes taken from 47 different persons. The collected probes where then used in experiment examining if it may be used for person's identification.

4. Data coversions

As it was mentioned above, the database consisted of 1883 probes signed by 47 participants of the experiment. Every probe consisted of 23 integer numbers and could be later treated as a vector in 23 dimensional space. The next task was to find relevant elements of this vector. Assuming that we have a set of vectors of attributes, we can check if attributes are relevant to the classification. Each vector X consists on n attributes X1…Xn and is accompanied by a label (or class) Y=y. One of the most obvious definitions of the relevancy may be that attribute Xi is relevant if knowing its value can change estimates for Y, or, in other words, if Y is conditionally dependent of Xi. However the important problem that must be solved is correlation among attributes.

The problem of correlations between attributes may be solved by using algorithms calculating linear conversion of dataset of vectors. That linear conversion may be defined as:

$$Z_{mk} = A_{mn} X_{nk}, \tag{1}$$

In the equation above Xnk represents an input dataset consisting of k probes with n attributes each. Matrix Amn is the matrix converting the data set into another one Zmk consisting of k probes with m attributes each. The value of m is often (but not always) less than n. What is important, after calculating the Amn matrix for the Xnk dataset, it is possible to use it for recalculating every new n-attribute sample to the new m-dimensional one. There are several different linear conversion methods. the most popular method is PCA.

The goal of Principal Component Analysis (called often also Karhunnen-Loeve transform or Hotelling transform) is to explain as much variance as possible with the smallest number of variables (Calvo, 1998). The assumption is made that attributes have a normal distribution, so all information about correlations between attributes is contained in the covariance matrix. The method creates a new dataset that should maintain as much of the original data structure as possible. The classic algorithm calculates eigenvectors and eigenvalues of the covariance matrix. These eigenvectors correspond to the directions of the principal components of the original data, their statistical significance is given by their corresponding eigenvalues.

The problem with PCA is that it does not take into account any information about vectors classifications. Such methods are called unsupervised methods. In contrast to that, Linear Discriminant Analysis (LDA) is an example of a supervised method that takes into account the vectors classification and tries to find a linear conversion of vectors that maximizes class separability. The most common approach is to maximize the ratio of between-class variance to the within-class variance (Balakrishnama, 1998). This approach was used in described experiment. Similarly to PCA, LDA algorithm is used to create conversion matrix. This matrix was then used to convert input matrix to a new one which consists of - hopefully - more relevant attributes.

5. Classification of attributes

The data created by the algorithms described in the previous section may be used to establish a technique of sample's classification.

The general algorithm that has been used may be divided into two phases: learning and testing. The first phase uses a dataset of samples (vectors) with known class assignment (so called training-set) to learn classification rules and to create a classification model. The model is then used to classification of unknown samples (so called test-set) in the testing phase.

Of course the quality of the created model is dependent mostly on the quality of the training sample (representativness for the whole problem) but the choice of proper algorithm for model creation is an important factor as well.

There are plenty of techniques that use training data to create a classification model. The most popular include: k nearest neighbors (kNN), Bayes methods, decision trees, support vector machines (SVM) or artificial neural networks (ANN) (Witten, 1999). It was the most popular version of decision tree - C4.5 algorithm used in the experiment described in this paper (Quinlan, 1993). The algorithm is one of the family of divide and conquer algorithms. It starts with the whole dataset and tries to find the best split of this data set. Each split is based on the value of one attribute. For each possible split of each attribute a value of gain on this split is computed as the difference between entropies before and after the split.

6. Experiment

The first step of the experiment was choosing 13 persons for whom there was more than 40 probes taken. Then the classification experiment was provided for every such person independently.

Firstly the whole dataset was divided into positive and negative probes. There was at least 40 positive probes for every person. The next step was random division of the whole dataset into 10 folds. Every fold consisted of both positive and negative probes with the same distribution as in the whole dataset. Every set of nine folds were used to create a classification model which was then tested on the 10th fold. The result of each single classification task was the number of errors (both false acceptances and false rejections) found for each nine-fold training set and one-fold testing set.

The classification process was performed using previously mentioned C4.5 decision tree algorithm. There was the decision tree for every training set build and then used for classification of probes from the testing set. The random fold division and classification process was repeated 10 times for every person giving overall average FAR and FRR results.

As it was mentioned the experiment was conducted for every person independently. As could be suspected final average false acceptance and false rejection rates were different for every person. The average value of false acceptance rate (FAR) calculated for all 13 classified persons was 4,23% and the average false rejection rate (FRR) was 18,36%. The results are obviously far from perfect and our system cannot yet be used in practice. It is worth mentioning however, that relatively high false rejection may be easily reduced by allowing user to perform several attempts. But of course it also results in increasing a false acceptance rate because malicious user also has more possibilities to mislead the system.

7. Future work

The experiment described in this paper is only the first step in our researches. It shows that even with classic classification techniques without any feature specific enhancements it is possible to obtain results that are showing that there is something personal in typing patterns.

The results would be probably better if the longer text sample was studied. The other problem is accuracy of measurement. There were Microsoft Windows measuring mechanisms used which are of course not very reliable. Furthermore one should take into consideration the influence of the keyboard on the way people type. People were in most cases using their own keyboards in our experiment and it should be examined what was the 'added value' of the way the keyboard was working.

To have more accurate measurements the specialized device called KeyScanner was build. The device is connected between keyboard and computer and is able to measure keystroke delays very reliable. The results of that analyses will be hopefully published in future articles.

Literatura

Balakrishnama, S. and Ganapathiraju, A. (1998) Linear Discriminant Analysis - A Brief Tutorial. Institute for Signal and Information Processing, Mississippi State University, MS State, MS, USA.

Calvo, R. A., Partridge, M. and Jabri, M. A. (1998) A Comparative Study of Principal Component Analysis Techniques. In Proc. Ninth Australian Conf. on Neural Networks, Brisbane, QLD.

Joyce, R., Gupta, G. (1990) Identity Authentication Based on Keystroke Latencies. Comm. ACM, vol. 33, no. 2, 1990, pp. 168-176

Kasprowski, P. and Ober, J. (2004) Eye Movement in Biometrics. Proceedings of Biometric Authentication Workshop, European Conference on Computer Vision in Prague 2004, LNCS 3087, Springer-Verlag, Berlin.

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

Monrose, F., Rubin, A. D. (2000) Keystroke Dynamics as a Biometric for Authentication. Future Generations Computing Systems, vol. 16, no. 4, 2000, pp. 351-359

Quinlan, J. R. (1993) C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann.

Spillane, R. (1975) Keyboard Apparatus for Personal Identification. IBM Technical Disclosure Bulletin vol. 17, no. 3346.

Tadeusiewicz, R. (1993) Biometria. Wydawnictwa Akademii Górniczo Hutniczej, Kraków.

Wayman, J. L. (2000) Fundamentals of Biometric Authentication Technologies. National Biometric Test Center Collected Works 1997-2000, San Jose University Press.

Witten, I. H. and Frank, E. (1999) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.

Zhang, D. D. (2000) Automated Biometrics Technolgies and Systems. Kluwer Academic Publishers.

Using classification methods and tools for data analysis to perform biometric identification based on keyboard typing patterns

The paper shows an example how methods for data analyses and classification algorithms may be used to perform identification of people basing on the way they are typing a specific word. The algorithm records intervals between key-presses and tries to find author of the key pattern. The method uses Linear Discriminant Analysis to preprocess data used then as the training-set for creation of C4.5 decision tree. The tree is able to classify any given probe (result of key-presses recording) as positive - belonging to the specified user - or negative - not belonging to the user (possibly a result of the fraud).