Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

USE OF INFORMATION CRITERION FOR CLASSIFICATION OF MEASUREMENT DATA

T. I. Lapina*, D. V. Lapin, E. A. Petric

Southwest State University 305040, Kursk, st.50 years of October, 94

Published online: 08 August 2017

ABSTRACT

This paper presents the method to analyze and classify measurement data for the purpose of identification and authentication of users during online network activity. The proposed method increases the accuracy of classification of signals in authorization systems.

Keywords: analysis and classification of signals, identification and authentications of user, access control system

INTRODUCTION

In modern world, every business activity makes intensive use of computerized information systems, which make the utilization of information resources much less labor consuming. Computerized information systems are essential for communication of geographically distributed offices, branches and clients of a company, because they provide interaction based on telecommunication information networks. Among a wide range of business activities demanding new information technologies we want to specifically point out the remote management of company's activity by means of telecommunication networks. The importance of this activity increases due to the growing range of online payments such as utility bills, mobile and credit card payments, and using credit cards to pay for online purchases. Development of network communications demands constant interaction between equipment, services, and the software, and in all cases the acknowledgement of the user's personality engaging in an information exchange must be guaranteed.

Author Correspondence, e-mail: author@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v9i2s.84

Journal of Fundamental and Applied Sciences is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. Libraries Resource Directory. We are listed under Research Associations category.

This guarantee is provided by registration of a user in a system, and also by identification, authentication and authorization procedures.

The software introduced in this paper is the prototype of computerized system for authentication of a user by dynamic signature on the screen of a computer or a mobile device. The proposed system increases the robustness of conventional password-based protection. It is based on multifactor authentification through the analysis of dynamic signature carried out by means of mathematical statistics and artificial neural networks.

Authentification of personality by handwriting and dynamically written control phrases (signature) relies on uniqueness and identity of a person's handwriting, which are measured, converted to a digital form and are subjected to computer processing. Thus, the authentification is made not for the result of writing, but the process of writing itself. Biometric authentification by signature passes the following stages:

− submission of a biometric image by a user, i.e. writing of a password (signature) on a graphic tablet;

− digitization of input electrical signals, measurement of the given biometric parameters in the submitted image;

− normalization of input signals which results in a certain reference value;

− saving of the biometrical reference sample of the identified personin in a system database, creation of template, i.e. profile of a user;

− system training;

− comparison of submitted user's profile with the saved one;

− prediction of a type I and II error level for the obtained biometric profile and making a decision.

Authentification of user based on control of falling in the area of distributed of reference samples

Assume at a stage of registration (training) an authorized user has submitted L signatures, this corresponds L to mappings of a vector of biometric parameters $\mathbf{V} = \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_L\}$.

For the tasks of dynamic biometrics, in most cases, we may assume that distribution of a vector $\mathbf{V}$ in N-dimensional space is close to normal, hence, vectors $\mathbf{V}_i$, $i = \overline{1, L}$ lie in N-dimensional area which for $L \to \infty$ in an orthogonal coordinate system is described by a dispersion hyperellipsoid [3]. Generally, the components of biometric vectors $\mathbf{V}_i$, $i = \overline{1, L}$ are intercorrelated, i.e. major axes of a hyperellipsoid of dispersion are not parallel to coordinate axes. Hence, having received the formula of this hyperellipsoid, we can authenticate a user by

controlling the fall of a vector of user's biometric parameters **V** inside the *N*-dimensional area described by a hyperellipsoid of dispersion [1]. For the normal law of distribution of *N*-dimensional random correlated values the frequency distribution function F-distribution is given as

$$f(v_1, v_2, ..., v_N) = \frac{1}{\sqrt{(2\pi)^N \det_j(\lambda_{jk})}} \exp\left[-\frac{1}{2} \sum_{j=1}^{N} \sum_{k=1}^{N} \Lambda_{jk} (v_j - \xi_j)(v_k - \xi_k) \right], \quad (1)$$

where

$$\lambda_{jk} = \lambda_{kj} = M(v_j - \xi_j)(v_k - \xi_k) = \begin{cases} D_{v_j} = \sigma_j^2 \text{ при } j = k, \\ \operatorname{cov}\{v_j, v_k\} \text{ при } j \neq k, \end{cases}$$

$$j, k = 1, 2, ..., N.$$

Coefficients λ_{jk} form a correlation matrix $[\lambda]$, and coefficients Λ_{jk} make a matrix $[\Lambda]$, which is inverse to the correlation matrix. To calculate the coefficients of the matrix $[\Lambda]$ we use the formula

$$\Lambda_{jk} = (-1)^{j+k} \frac{M_{jk}}{|\lambda|}, \quad (2)$$

where $|\lambda|$ is the determinant of the correlation matrix, and M_{jk} is the minor of this determinant obtained from it by elimination of j th row and k th column. Note that $|\Lambda| = \frac{1}{|\lambda|}$. The hyperellipsoid of dispersion has equal density of distribution of N-dimensional random variables, therefore its expression can be derived from the condition

$$f(v_1, v_2, \ldots, v_N) = \text{const}. \quad (3)$$

It follows from the expression (2) that the condition (3) is met if

$$\frac{1}{2} \sum_{j=1}^{N} \sum_{k=1}^{N} \Lambda_{jk} (v_j - \xi_j)(v_k - \xi_k) = \text{const}. \quad (4)$$

From all possible solutions of the equation (4) for different constants in the right part we are to select the only one which corresponds to a so-called individual hyperellipsoid which major semiaxes correspond to the mean square deviations $\sigma_1, \sigma_2, \ldots, \sigma_N$:

$$\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\Lambda_{jk}(v_j-\xi_j)(v_k-\xi_k)=1 \quad (5)$$

Due to the restrained statistics of the biometric samples submitted at a stage of registration by the authorized user there is always a probability that the sample submitted by the same user during authentification, will fall outside the limits set by the reference sample. To reduce this probability value we additionally set the tolerance level between areas for authorized and unauthorized users in the form of the Student's coefficient *With* [L, ($_{1-P1}$)], given the type I error (probability $_{P1}$of false rejection of an authorized user) and the number of L submitted on samples at thestage of registration. Having added this tolerance level to the equation (5) we rewrite it as

$$\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\Lambda_{jk}(v_j-\xi_j)(v_k-\xi_k)=C[L,(1-P_1)]^2. \quad (6)$$

Authentification procedure reduces to checking whether the **V** vector of biometric parameters submitted by the user falls within the area described by expression (6) throughinequality

$$\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\Lambda_{jk}(v_j-\xi_j)(v_k-\xi_k)\le C[L,(1-P_1)]^2. \quad (7)$$

If the inequality holds for the vector of biometric parameters V submitted by the user, this vector is decided to belong to an authorized user and access is granted, otherwise the vector is decided to belong to an unauthorized user and access is denied.

Approach to classification of digital signals based normalization

Classification of signals in real time requires processing of incoming data by a limited number of counts, which in its turn, requires the analysis of sample distribution instead of common data-flow algorithms. This paper introduces a novel approach to classification of signals and building of the set of information bearing traits significant to the dynamics of the process on the basis of normalization of sampled distributions. The proposed approach reduces the computing complexity of classification, which is essential for the design of access control systems.

Given the vector which consists of input variables, let's interpret it as a multivariate time series. Then the task of identification comes to estimation of statistical characteristics of a time series. Here the task of differentiating the signals comes to changing the structure of random process or the times series.

In this paper, the given task is solved using a method of normalization of data and conceptual model of representation of distributions pairs [2].

Let the restrained space contains the distributions

$$F_{\hat{x}_2}(x) \text{ and } F_{\hat{x}_1}(x) = \frac{(x_r - x^-)}{(x^+ - x^-)} = r_x, \quad (8)$$

where $F_{\hat{x}_2}(x)$, $x \in [x^-, x^+]$ israndom distribution; and

$F_{\hat{x}_1}(x)$, $x \in [x^-, x^+]$ is uniform distribution.

Let's solve $F_{\hat{x}_1}(x)$ with x_r : $x_r = r_x(x^+ - x^-) + x^-$ and substitute x_r into the function $F_{\hat{x}_2}(x)$

$$F_{\hat{x}_2}(x_r) = F_{\hat{x}_2}\ (r_x(x^+ - x^-) + x^-) = F_{\hat{r}}(r_x),\ r_x \in [0,1], \quad (9)$$

where the system of functions $F_{\hat{r}_1}(r_x) = r_x$ and $F_{\hat{r}_2}(r_x)$ is a conceptual model of representation of pairs of distribution $F_{\hat{x}_1}(x)$ и $F_{\hat{x}_2}(x)$.

In fact, we obtain $F_{\hat{r}_2}(r_x)$ by shifting and scaling operations r = ax + b, где $a = \frac{1}{x^+ - x^-}$,

$b = -\frac{x^-}{x^+ - x^-}$, which are derived from the set of equations $\begin{cases} ax^- + b = 0, \\ ax^+ + b = 1. \end{cases}$

Applying the shifting and scaling operations we bring the analyzed sampled data to the [0,1] interval without structural changes. This allows solving different tasks of statistical analysis including classification of signals in a completely new way.

For $F_{x_1}(x)$ and $F_{x_2}(x)$ we introduce the concept of ordering $x \in [x_1^-, x_2^+]$ in the form of conversion tx=t (x) $x \in [x^-, z^+]$ $t_x \in [t_x^- = x^-, t_x^+ = x^+]$. Then by using the common method for determination of law of distribution of functions from random variables t_x given the laws

of distribution of initial random variables $\hat{x}$ we have $f_{t_x}(t_x) = f_x(x)\left|\frac{\partial t_x}{\partial x}\right| \Rightarrow f_x(t_x^{-1})\left|\frac{\partial t_x}{\partial x}\right|$, from which we have

$$f_x(x) = f_{t_x}(t_x)\left|\frac{\partial x}{\partial t_x}\right|. \qquad (10)$$

The property of the conceptual model given (3) gives us

$$\frac{\partial F_{r_2}(r)}{\partial r} = \frac{\partial F_{r_2}(r)}{\partial F_{r_1}(r)} = \frac{f_{x_2}(x_2)}{f_{x_1}(x_1)}. \qquad (11)$$

To illustrate the conceptof ordering let's assume that a certain set ofa set of random variables $\{x_j\}_{\langle N\rangle}$ has two different probability distributions

$$\{P_1(x_j)\}_{\langle N\rangle} \text{ и } \{P_2(x_j)\}_{\langle N\rangle}, \quad \sum_{J=1}^{N} P_i(x_j), \ j=1,2.$$

According to the probability distribution $\{P_1(x_j)\}_{\langle N\rangle}$ and $\{P_2(\hat{x}_j)\}_{\langle N\rangle}$ for the elements $\{x_j\}_{\langle N\rangle}$ allocated on the axis T in the selected order t, we form the probability distributions $F_{r_{t_1}}(r_t)$ and r_t in the coordinate system, which represents a unit square (see Fig. 1).

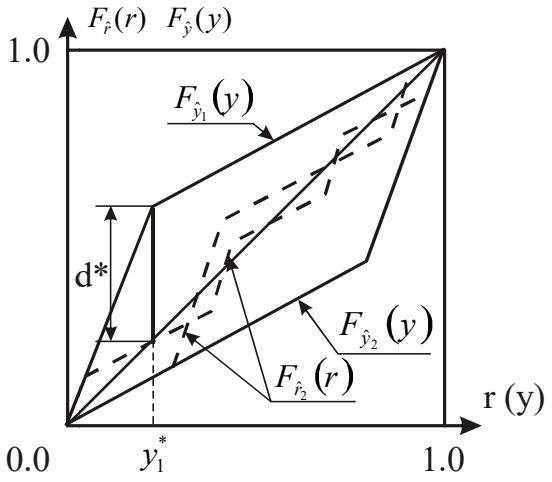

Fig.1. Differentiation measures for functions $\{f_{\hat{r}_2}(r)\}$, $F_{\hat{y}_1}(y)$, $F_{\hat{y}_2}(y)$ for signal estimates

As a differentiation measure for compared distributions, we use an average amount of Kullback information [1]:

$$I = \int_{x \in X} f_{\hat{x}_2}(x) \cdot \ln\left(\frac{f_{\hat{x}_2}(x)}{f_{\hat{x}_1}(x)}\right) dx . \quad (12)$$

Applying a conceptual model for representation of pairs of distributions (13), we transform the formula (5) as follows

$$I = \int_{x \in X} \frac{f_{\hat{x}_2}(x)}{f_{\hat{x}_1}(x)} \cdot \ln\left(\frac{f_{\hat{x}_2}(x)}{f_{\hat{x}_1}(x)}\right) \cdot f_{\hat{x}_1}(x) dx = \int_0^1 f_{\hat{r}_2}(r) \cdot \ln(f_{\hat{r}_2}(r)) dr \ , \ r \in [0,1],$$

$$\text{or} \quad I = \int_0^1 \frac{f_{\hat{r}_2}(r)}{f_{\hat{r}_1}(r)} \cdot \ln\left(\frac{f_{\hat{r}_2}(r)}{f_{\hat{r}_1}(r)}\right) \cdot f_{\hat{r}_1}(r) dr, \qquad r \in [0,1] \quad (13)$$

where $f_{\hat{r}_1}(r) = 1$ is a uniform distribution in the [0,1] interval.

The latter expression is used as the information criteria for classification. It characterizes the average amount of information in the distribution $f_{\hat{r}_2}(r)$ given in the interval [0,1] in relation to $f_{\hat{r}_1}(r)$, i.e. in relation to the uniform distribution.If we consider $f_{\hat{r}_2}(r)$ as a certain distribution brought to the interval [0,1], then for $f_{\hat{r}_1}(r) = 1$ theequality

$$I = \int_0^1 f_{\hat{r}_1}(r) \cdot \ln(f_{\hat{r}_1}(r)) dr = -\int_0^1 1 \cdot \ln 1 dr = 0$$

holds, and the expression

$$I = \int_0^1 f_{\hat{r}_2}(r) \cdot \ln(f_{\hat{r}_2}(r)) dr \quad (14)$$

is considered the differentiation measure for sampled data $\{f_{\hat{r}_2}(r)\}$ in $r \in [0,1]$ interval with equal values of I in relation to $f_{\hat{r}_1}(r) = 1$.

For the function $y = \varphi(r), \ r \in [0,1], \ y \in [0,1]$, which represents the ordering of r in relation to increasing/decreasing $f_{\hat{r}_2}(r)$, i.e. $y_1 \in Y_1, \ y_2 \in Y_2, \ y_1 < y_2, \ Y_1 = \varphi(R_1), \ Y_2 = \varphi(R_2)$ the following equality

$$d = \max_{y \in [0,1]} \left| F_{\hat{y}}(y) - y \right| . \quad (15)$$

holds.

CONCLUSION

Comparison of two proposed methods of classification proves that the usage of representation of input measurement data in the form of discrete signals allows to use the differentiation measure for sampled data I and the distance as a criterion for classification of signals significantly reduces computational complexity of classification. This benefit contributes to the design of systems based on analysis and classification of measurement data, for example, systems that control authorization of users and access to information resources in telecommunication systems.

Representation of biometric data as the multi-component signal allows estimating each fragment of the signal decomposition individually. This simplifies the process of segmentation and analysis of the initial signal image.

We introduce the technique for building the biometric image based on extraction of dynamic features of signals from the multi-element movement sensor using the Haar transform. Implementation of Haar functions allows considering local features of signals in the form of short-time spikes. Besides, this technique is computationally cost-effective, hence it guarantees extraction of highly informative dynamic features.

Use of a measure of distinction of selective data of I and distance as criterion of classification of signals allow to simplify significantly procedure of classification at creation of monitoring systems of access.

To determine identification characteristics and layout of AIC applicability ofthe proposedbiometricimage studies were performed sequence variability of the method based on the evaluation of errors of the first and of the second the first kind for the null hypothesis – "man - a known user of the system".

According to theresearch conducted with the false acceptance rate FAR = 0,0006 (is the measure of the likelihood that the biometric security system will incorrectly accept an access attempt by an unauthorized user), the false rejection rate FRR= 0,42 (is the measure of the likelihood that the biometric security system will incorrectly reject an access attempt by an authorized user).

REFERENCES

1.V.G.Grigorovich Information methods of management of quality[Text] / Grigorovich V.G., Udin S.V., Shildin V.V. -M.:RIA «Standards and quality», 2001. – 107p.

2. T.I. Lapina Information approach to creation of models of objects in monitoring systems [Text] / Lapina T.I.// Information measuring and managing directors of system .№7, т. 8, 2010. p. 39-42.

3. I.G. Urazbahtin Creation of measures of a structural variety of laws of distribution of probabilities on the basis of application of streamlining of random variables [Text] / Urazbahtin I.G., Lapina T.I.//News KGTU, №4, 2000.

How to cite this article:
Lapina T I, Lapin D V, Petric E A. Use of information criterion for classification of measurement data. J. Fundam. Appl. Sci., 2017, *9(2S), 1099-1107.*