УДК 004.056.523

ВИКОРИСТАННЯ КРИТЕРІЮ МАННА-ВІТНІ В ЗАДАЧІ ІДЕНТИФІКАЦІЇ КОРИСТУВАЧІВ ЗА КЛАВІАТУРНИМ ПОЧЕРКОМ

Леушина А.А.
Науковий керівник – к.т.н., доц. Горелов Д.Ю.
Харківський національний університет радіоелектроніки,
студентський науковий гурток «Біометричні технології контролю доступу»
каф. КРіСТЗІ, м. Харків, Україна
тел. +38(057) 702-14-30, e-mail: anastasiia.leushyna@nure.ua.

The algorithm of keystroke authentication based on the Mann-Whitney U test is developed. The algorithms of formation of the user profile and its authentication has been designed.

Аналіз літератури в області біометричних систем контролю доступу за клавіатурним почерком показує, що одними з найбільш поширених є методи класифікації клавіатурного почерку на основі параметричних статистичних підходів. Ці методи порівняння параметрів розподілів припускають, що дослідник заздалегідь володіє фундаментальною інформацією – йому відомий вид закону розподілу ймовірностей, найчастіше нормальний закон, що дозволяє звести задачу розпізнавання до перевірки гіпотез про подібність таких характеристик як середнє, медіана і стандартне відхилення. Однак в силу ряду специфічних причин, пов'язаних з нестабільністю клавіатурного почерку, припущення про «нормальність» закону розподілу може привести до спотворення висновків (аж до прийняття рішення, протилежного вірному).

У тих випадках, коли припущення про гіпотетичний закон розподілу ймовірностей не є переконливими, слід застосовувати інші методи, наприклад, непараметричний статистичний U критерій Манна-Вітні.

Алгоритми формування профілю користувача та процесу аутентифікації.

У якості інформативних ознак клавіатурного почерку використовувались наступні часові параметри послідовних подій клавіатури: T_1 – час натискання клавіші; T_2 – час паузи між відпусканням першої клавіші та натисканням другої клавіші; T_3 – час між натисканням першої клавіші та натисканням другої клавіші; T_4 – тривалість слова з двох (трьох) букв.

Параметр T_1 розраховувався для 20 найуживаніших букв англійської мови:

E, A, R, I, O, T, N, S, H, D, L, C, U, M, W, F, G, Y, P, B.

Параметри T_2 та T_3 розраховувались для 20 найуживаніших біграм англійської мови:

IN, TH, TI, ON, AN, HE, AT, ER, RE, ND,
HA, EN, TO, IT, OU, EA, HI, IS, OR, TE.

Параметр T_4 розраховувався для 20 найуживаніших слів англійської мови:

FOR, AND, THE, IS, IT, YOU, HAVE, OF, BE, TO,
THAT, HE, SHE, THIS, THEY, WILL, I, ALL, A, HIM.

Таким чином, для кожного користувача по результатам введеного тексту № 1 формується 20 векторів параметру $\vec{T}_1$ (для кожної з 20 найуживаніших букв англійської мови), 20 векторів параметрів $\vec{T}_2$ та $\vec{T}_3$ (для кожної з 20 найуживаніших біграм англійської мови) та 20 векторів параметру $\vec{T}_4$ (для кожного з 20 найуживаніших слів англійської мови). Цю сукупність з 80-ти векторів можна вважати початковим еталоном користувача:

$$ET = \begin{Bmatrix} \vec{T}_1^E,\ \vec{T}_1^A,\ \vec{T}_1^R, \dots \vec{T}_1^B \\ \vec{T}_2^{IN},\ \vec{T}_2^{TH},\ \vec{T}_2^{TI}, \dots \vec{T}_2^{TE} \\ \vec{T}_3^{IN},\ \vec{T}_3^{TH},\ \vec{T}_3^{TI}, \dots \vec{T}_3^{TE} \\ \vec{T}_4^{FOR},\ \vec{T}_4^{AND},\ \vec{T}_4^{THE}, \dots \vec{T}_4^{HIM} \end{Bmatrix}. \quad (1)$$

За результатами введеного тексту № 2 для кожного користувача за допомогою U критерію для кожного з 20-ти векторів множини $\{\vec{T}_1^E,\ \vec{T}_1^A,\ \vec{T}_1^R, \dots \vec{T}_1^B\}$ перевіряється гіпотеза про відсутність відмінностей між еталонним вектором $\vec{T}_1^X$ та дослідним $\vec{T}_1^{Xдосл}$. Якщо більше 12 і більше (60 % і більше) гіпотез з 20 було прийнято, то вважається, що множина $\{\vec{T}_1^E,\ \vec{T}_1^A,\ \vec{T}_1^R, \dots \vec{T}_1^B\}$ є еталоном. Якщо кількість прийнятих гіпотез менша 12, то користувачеві пропонується знову пройти процес формування еталону. Для формування повного еталону аналогічні розрахунки проводяться для множин $\{\vec{T}_2^{IN},\ \vec{T}_2^{TH},\ \vec{T}_2^{TI}, \dots \vec{T}_2^{TE}\}$, $\{\vec{T}_3^{IN},\ \vec{T}_3^{TH},\ \vec{T}_3^{TI}, \dots \vec{T}_3^{TE}\}$ та $\{\vec{T}_4^{FOR},\ \vec{T}_4^{AND},\ \vec{T}_4^{THE}, \dots \vec{T}_4^{HIM}\}$.

На першому кроці аутентифікації дослідного користувача розраховується 80 векторів:

$$ET = \begin{Bmatrix} \vec{T}_{1exp}^E,\ \vec{T}_{1exp}^A,\ \vec{T}_{1exp}^R, \dots \vec{T}_{1exp}^B \\ \vec{T}_{2exp}^{IN},\ \vec{T}_{2exp}^{TH},\ \vec{T}_{2exp}^{TI}, \dots \vec{T}_{2exp}^{TE} \\ \vec{T}_{3exp}^{IN},\ \vec{T}_{3exp}^{TH},\ \vec{T}_{3exp}^{TI}, \dots \vec{T}_{3exp}^{TE} \\ \vec{T}_{4exp}^{FOR},\ \vec{T}_{4exp}^{AND},\ \vec{T}_{4exp}^{THE}, \dots \vec{T}_{4exp}^{HIM} \end{Bmatrix}. \quad (2)$$

На другому кроці аутентифікації за допомогою U критерію для кожного параметру T_i перевіряються 20 гіпотез про відсутність відмінностей між векторами $\vec{T}_i^{Xexp}$ та еталоном:

$$H_0^{iX}\colon \vec{T}_{iexp}^X = \vec{T}_i^X. \quad (3)$$

Якщо більше 12 і більше з 20 гіпотез було прийнято, то приймається загальне рішення про відсутність відмінностей між вектором аутентифікації та еталоном за параметром T_i.

На третьому кроці аутентифікації приймається загальне рішення про

аутентифікацію: якщо для трьох і більше з чотирьох параметрів T_i прийнято позитивні рішення про відсутність відмінностей, то висувається рішення про позитивну аутентифікацію. В іншому випадку система приймає рішення про негативну аутентифікацію.

Результати проведених експериментів.

Чотирьом користувачам було запропоновано ввести текст, за яким проводилась аутентифікація. Для параметру T_i^X формувались чотири вектори $\left\{\vec{T}_{iexp}^{Xкор1}, \vec{T}_{iexp}^{Xкор2}, \vec{T}_{iexp}^{Xкор3}, \vec{T}_{iexp}^{Xкор4}\right\}$ та вибирались з бази еталонів відповідні вектори $\left\{\vec{T}_i^{Xкор1}, \vec{T}_i^{Xкор2}, \vec{T}_i^{Xкор3}, \vec{T}_i^{Xкор4}\right\}$. Таким чином, можна перевірити 10-ть неперехресних гіпотез про відсутність відмінностей між парами векторів $\left[\vec{T}_{iexp}^{Xкорj}; \vec{T}_i^{Xкорl}\right]$ $(1 \leftrightarrow 2;\ 1 \leftrightarrow 3;\ 1 \leftrightarrow 4;\ 2 \leftrightarrow 3;\ 2 \leftrightarrow 4;\ 3 \leftrightarrow 4;\ 1 \leftrightarrow 1;\ 2 \leftrightarrow 2;\ 3 \leftrightarrow 3;\ 4 \leftrightarrow 4)$ та розрахувати помилки:

$$F_i^X = \frac{кількість\ невірно\ прийнятих\ гіпотез}{10}. \qquad (4)$$

Середні значення помилок аутентифікації склали: $F = 0{,}27$ для параметру T_1; $F = 0{,}21$ для параметру T_2; $F = 0{,}12$ для параметру T_3; $F = 0{,}3$ для параметру T_4.

Таким чином, використання непараметричного критерію Манна-Вітні є досить перспективним для побудови систем контролю доступу на основі клавіатурного почерку; по-друге, найбільш доцільно в якості «аутентифікаційного» параметру за умови вводу тексту англійською мовою слід використовувати тривалість найуживаніших двобуквених сполучень: IN, TH, TI, ON, AN, HE, AT, ER, RE, ND, HA, EN, TO, IT, OU, EA, HI, IS, OR.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ.

1. E. Yu and S. Cho, «Keystroke dynamics identity verification—its problems and practical solutions», Computers and Security, vol.23, no. 5, pp. 428–440, 2004

2. H. Davoudi and E. Kabir, «A new distance measure for free text keystroke authentication», in Proceedings of the 14th International CSI Computer Conference (CSICC ’09), pp. 570–575, October 2009.

3. T. Shimshon, R.Moskovitch, L. Rokach, and Y. Elovici, «Continuous verification using keystroke dynamics», in Proceedings of the International Conference on Computational Intelligence and Security (CIS ’10), pp. 411–415, December 2010.