УДК 621.369:534

Д.Ю. ГОРЕЛОВ, канд. техн. наук, В.О. АЛЄКСЄЄВ, В.М. БУБЛИК, Д.В. МАСЛІЙ

ДОСЛІДЖЕННЯ СТАТИСТИЧНИХ ВЛАСТИВОСТЕЙ КЛАВІАТУРНОГО ПОЧЕРКУ ДЛЯ ВИРІШЕННЯ ЗАДАЧ АУТЕНТИФІКАЦІЇ КОРИСТУВАЧІВ КОМП'ЮТЕРНИХ МЕРЕЖ

Вступ

Аналіз літератури в області біометричних систем контролю доступу за клавіатурним почерком [1] показує, що найбільш поширеними 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).

Web-доданок для збору та вивчення шаблонів клавіатурного почерку користувачів

Для вирішення завдання збору і аналізу статистичних характеристик клавіатурного почерку користувачів було розроблено клієнт-серверний WEB-доданок. Первинний збір часових параметрів клавіатурного почерку здійснюється безпосередньо в WEB-браузері, після чого дані відправляються на сервер для обчислення і подальшої інтерпретації. Такий підхід не поступається за точністю доданкам, які реалізовані як desktop applications, але разом з тим має ряд істотних переваг, таких як: простота використання для користувача і дослідника, відсутність необхідності інсталяції на кожну робочу станцію, можливість отримання і обробки великого числа статистичних даних в будь-який час з довільної локації.

В рамках даної роботи було зібрано статистичні характеристики клавіатурних профілів 30 користувачів. Всі учасники експерименту – студенти, які є постійними користувачами персональних комп'ютерів з різними навичками швидкості набору клавіатурних символів.

З метою підвищення якості зібраної інформації і більш точного виділення інформативних ознак сформованих профілів в ході досліджень експериментальним шляхом були визначені і реалізовані наступні критерії: часові інтервали диграфів (подвійних подій клавіатури, що можуть бути повністю описані шістьма різними часовими відрізками, рис. 1), що перевищують значення 200 мс, з аналізу виключалися (для захисту від незапланованих довгих натискань), також для більш об'єктивного уявлення характеру розподілу з аналізу виключалися диграфи, число повторень яких у кожному конкретному профілі було менше 10.

На рис. 2 представлено візуальні відмінності декількох сформованих профілів користувачів в однаковому часовому масштабі у вигляді двовимірного лінійного графіка. По осі ординат відкладені значення відповідного часового інтервалу розглянутого диграфу, а по осі абсцис – номер відповідного диграфа в розглянутому профілі.

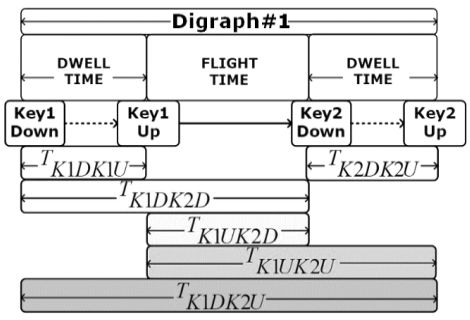

Рис. 1

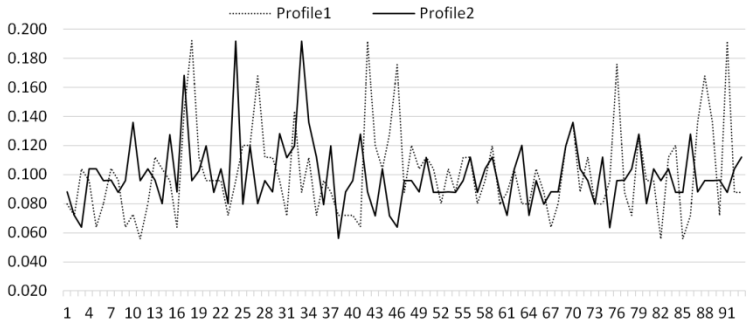

Рис. 2

Дослідження законів розподілів часових інтервалів диграфів

Всі отримані диграфи було розсортовано за частотою повторення. Розглядалися диграфи, що з'явилися максимальну кількість раз. Так, наприклад, для тексту обсягом в 2000 символів (1734 символи без урахування пробілів) найбільша кількість повторень одного і того ж диграфа в профілі дорівнювалась 60 (диграф «СТ»), а найменша – 1 (диграф «ЕА»).

На рис. 3 – 5 наведено гістограми параметрів $T_{K_1DK_1U}$ (тривалості натискання букви «С» диграфу «СТ»), $T_{K_1UK_2D}$ (паузи між натисканнями букв «С» і «Т» диграфу «СТ») та $T_{K_1DK_2U}$ (тривалості диграфу «СТ») для 8 сформованих профілів. З наведених графіків можна зробити висновок про те, що пошук спільного закону розподілу тут абсолютно неможливий, тому є необхідність пошуку інших критеріїв і підходів здатних підвищити точність проходження процедури аутентифікації.

Рис. 3

Рис. 4

Дослідження закону розподілу відношень часових інтервалів диграфів

Часові параметри диграфів клавіатурного почерку в значних межах змінюють свої значення з плином часу. Проте, очевидним є той факт, що кожен користувач має індивідуальний ритм та характер набору тексту. Тому доцільно перейти від аналізу часових параметрів диграфів (див. рис. 1) до їх відношень за аналогією до властивостей подібних фігур з геометрії – де фігура F' є подібною до фігури F, якщо існує відображення фігури F на фігуру F', при якому для будь-яких двох точок X і Y фігури F та їх образів X' і Y' фігури F' відношення відстаней XY і $X'Y'$ є величиною сталою (рис. 6):

$$\frac{AB}{A'B'} = \frac{AC}{A'C'} = \frac{DE}{D'E'} = \text{const.}$$

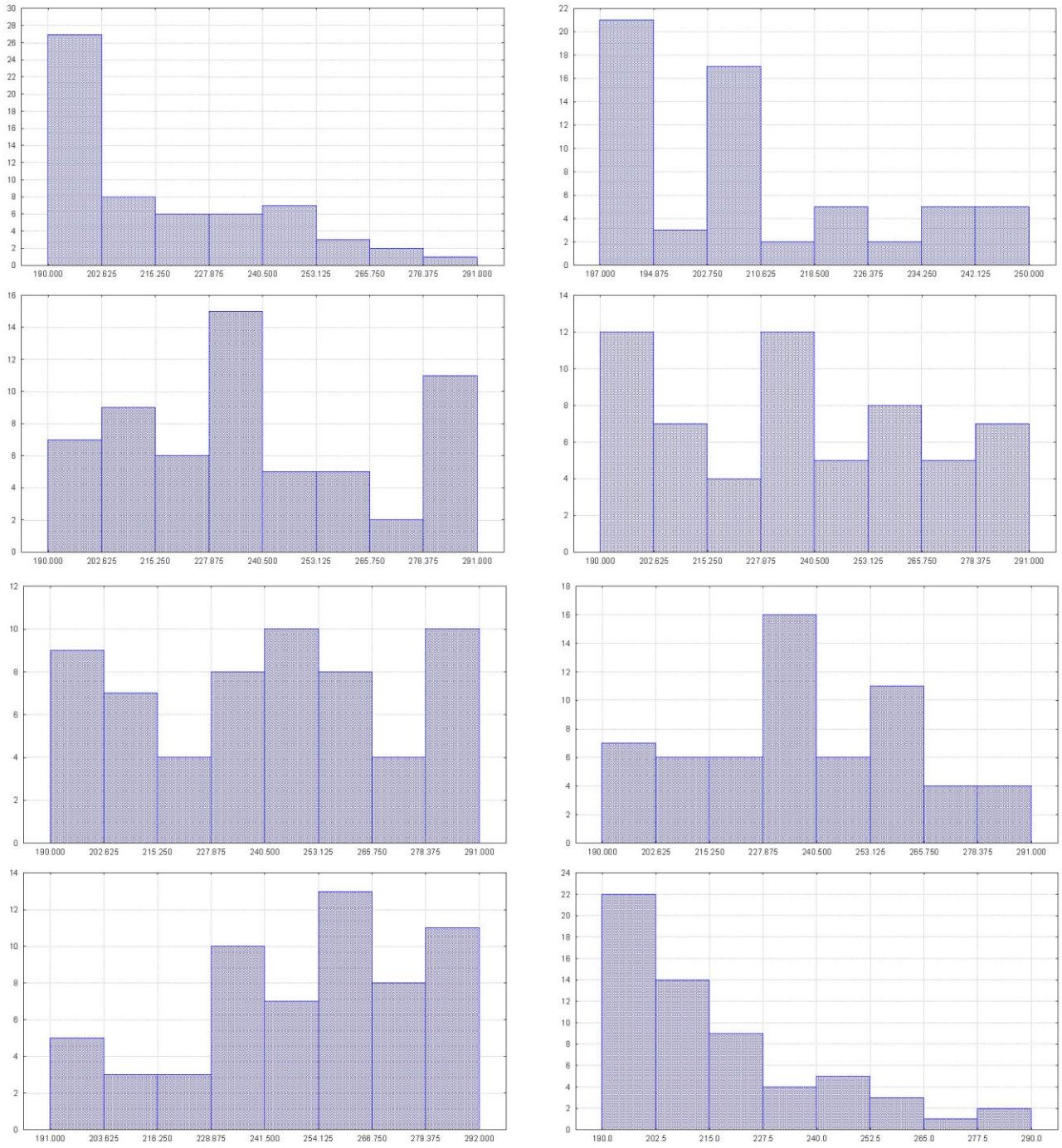

Рис. 5

В ході роботи були розраховані і проаналізовані численні комбінації відношень часових інтервалів диграфів, отриманих при формуванні профілів.

На рис. 7, 8 наведено гістограми відношень $\frac{T_{K_1DK_2D}}{T_{K_1UK_2D}}$ і $\frac{T_{K_1DK_1D}}{T_{K_1UK_2U}}$ (див. рис. 1) диграфа «СТ» для 8 сформованих профілів. На цих же графіках суцільною лінією показана апроксимація отриманих гістограм нормальним законом розподілу, функціонально реалізованим в «Statistica».

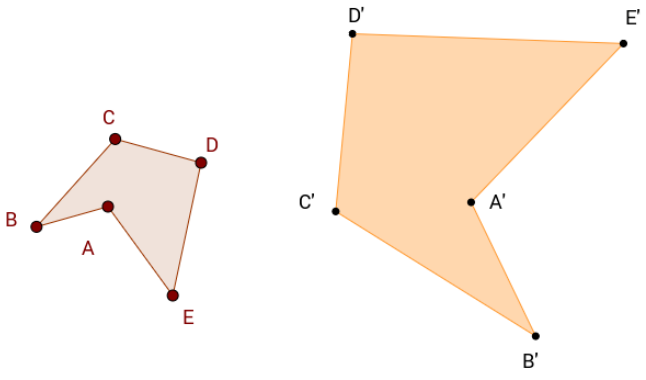

Рис. 6

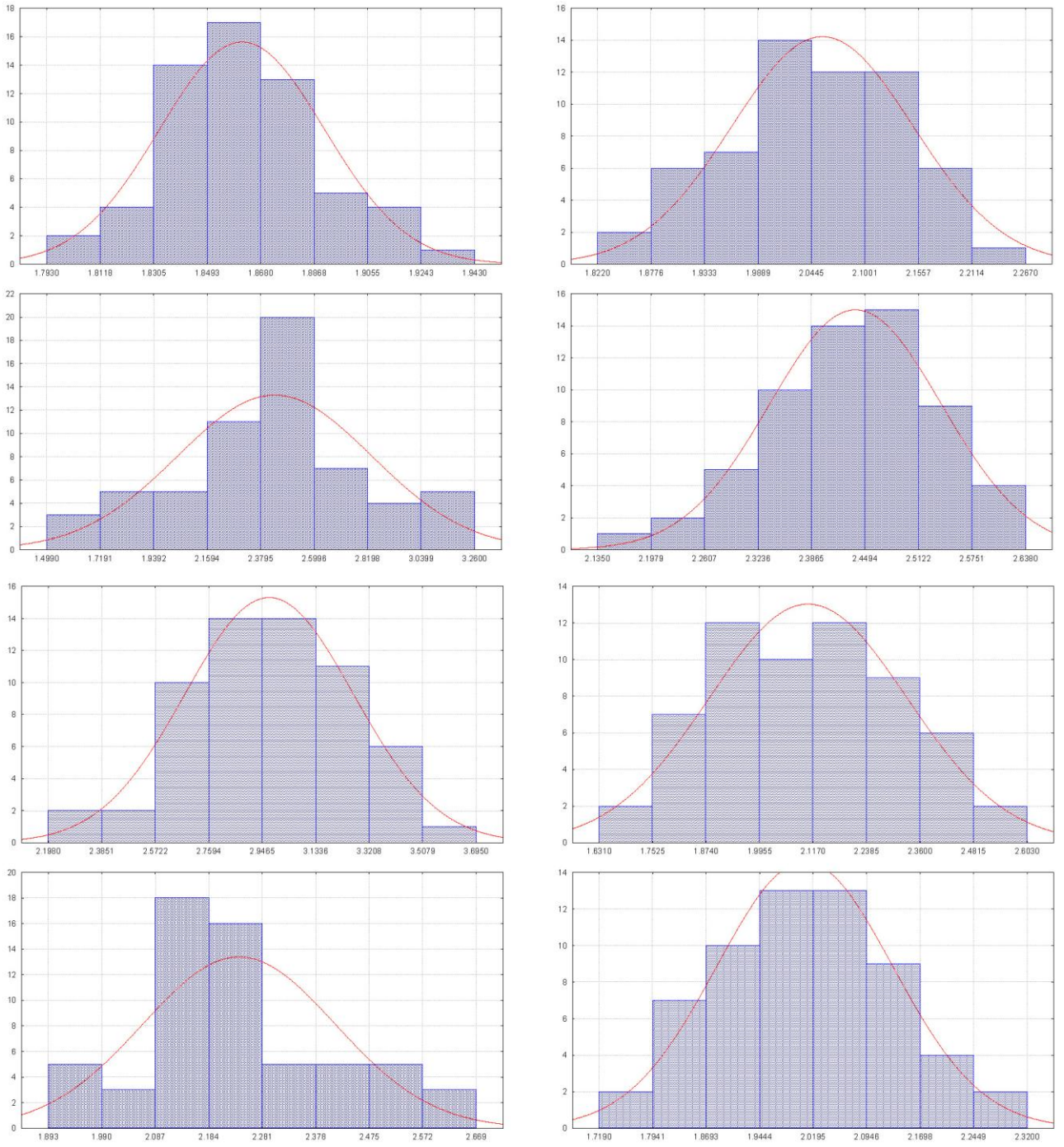

Рис. 7

Як видно з отриманих результатів, закон розподілу відношень часових параметрів дифрафів близький до нормального, тобто «випадковість» зміни параметрів клавіатурного почерку зменшується. Отже, можна говорити про справедливість висунутої гіпотези.

Парольна аутентифікація

Головний недолік відомих способів парольної аутентифікації користувачів на основі порівняння їх клавіатурних почерків – відсутність математично обґрунтованих критеріїв виявлення відмінностей і, отже, можливості заздалегідь встановити прийнятні ймовірності FAR та FRR помилок. В ситуації, коли закон розподілу не є нормальним, найбільш прийнятними для задачі аутентифікації виявляються непараметричні методи, інваріантні до законів розподілу часових параметрів диграфів. У зв'язку з цим, постає проблема вибору з досить великої кількості відповідного рангового критерію порівняння двох або більше наборів вимірювань.

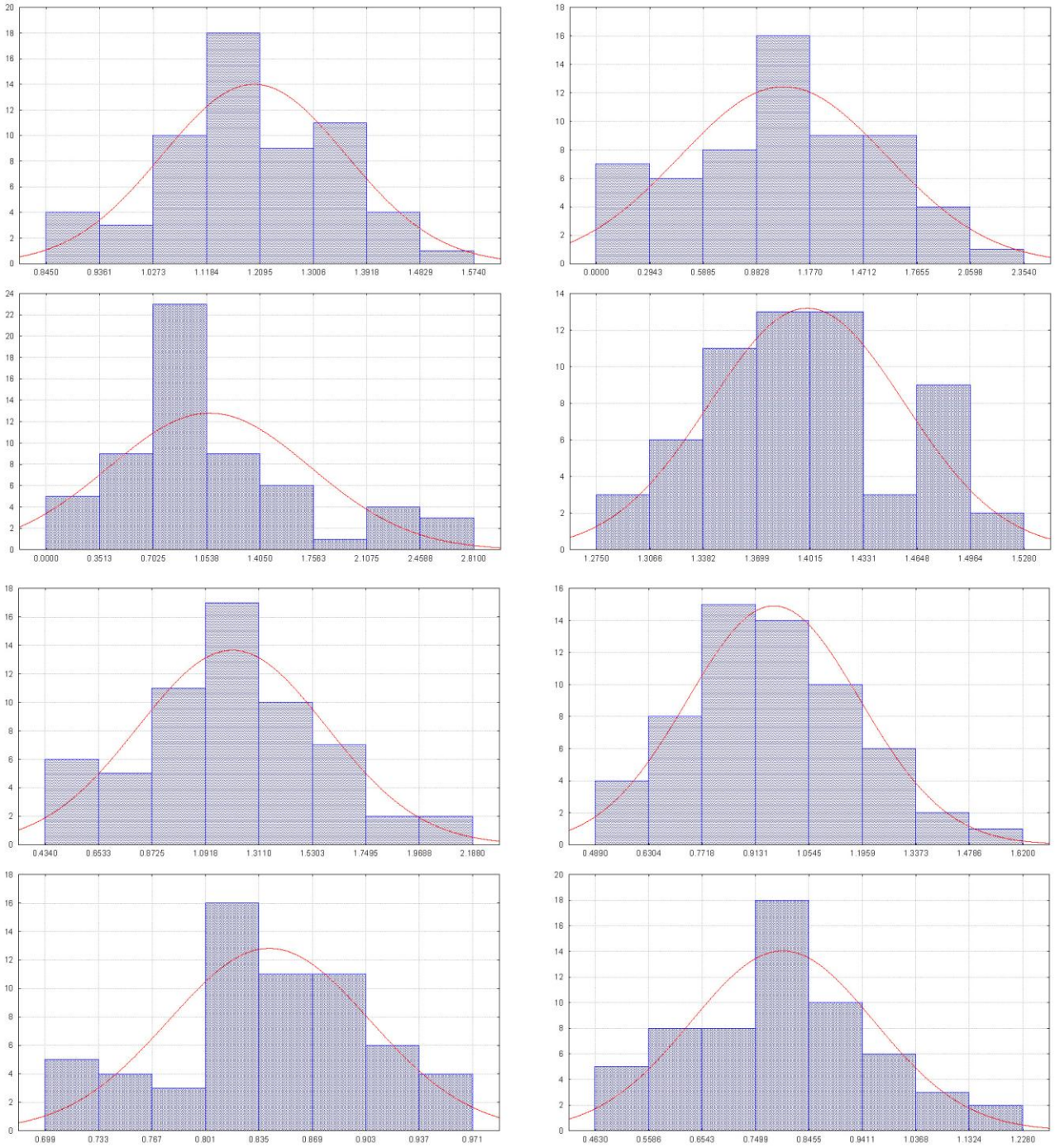

Рис. 8

В даному випадку алгоритм проходження аутентифікації може бути наступним [2].

Перший крок. Кореляційний аналіз подібності послідовностей часових параметрів парольної фрази. Непараметричним аналогом парного коефіцієнта кореляції є ранговий коефіцієнт Спірмена, який використовується для виявлення та оцінки тісноти зв'язку між двома вибірками, що порівнюються.

Другий крок. Висока кореляція свідчить лише про подібність двох вибірок клавіатурного почерку за формою обвідної (подібна ритміка набору). При цьому вони можуть істотно відрізнятися за параметром положення (темп набору) або ступеня розсіювання щодо нього (нестабільність почерку). Очевидна необхідність подальшої перевірки вибірки, що пройшла кореляційний етап, на близькість до еталону за параметрами розподілу. В спеціалізованій літературі рекомендують порівняння параметрів положення і масштабу проводити окремо. З великого арсеналу засобів непараметричної статистики, призначених для порівняння двох

залежних вибірок за параметром положення (тобто що виявляють зсув між ними) доцільно використовувати знаково-ранговий критерій Вілкоксона.

Наступний етап – перевірка відмінностей в параметрі масштабу порівнювальних вибірок. Для цього завдання в ситуації зв'язаних вибірок найчастіше використовують критерій Сендвіка – Олссона.

Іншим варіантом побудови алгоритму парольної аутентифікації є використання DTW-алгоритму [3], ідея якого полягає в наступному: є набір еталонних зразків клавіатурного почерку, які можуть бути закодовані в часовій або в частотної області і які представляють словник для розпізнавання. Саме розпізнавання відбувається шляхом порівняння нових даних з усіма стандартами і визначення найбільш підходящого кандидата відповідно до деякої метрики або за мірою подібності.

Неперервна аутентифікація

Метод прихованого моніторингу клавіатурного почерку повинен застосовуватися спільно з парольною аутентифікацією для зниження ризику, пов'язаного з можливою підміною користувача.

Можна виділити два етапи верифікації користувача: якісний і кількісний. Перший виявляє розбіжності у заздалегідь виявлених індивідуальних особливостях роботи з клавіатурою легального користувача. Це використання альтернативних службових клавіш (наприклад, клавіші Backspace і Delete, CapsLock і правий/лівий Shift); використання клавіш додаткової клавіатури (числа з numpad) тощо. Ці особливості проявляються на підсвідомому рівні, і спроба контролювати їх неминуче відіб'ється на зміні динаміки почерку.

Другий етап передбачає продовження збору і аналізу ключових часових характеристик диграфів після того, як користувач вже увійшов в систему. Таким чином, моніторинг характеру клавіатурної активності виконується протягом всієї робочої сесії під конкретним обліковим записом. По мірі накопичення статистичних даних відбувається уточнення схожості еталона і новосформованої сукупності параметрів. Тут може вирішуватися задача зворотна тій, що була сформульована у випадку парольної аутентифікації, а саме: довести відмінність двох сформованих профілів (іншими словами довести, ту обставину, що вхід в систему зробив не той користувач, за якого він себе видає). В даному випадку утворюються незалежні вибірки, які можуть бути взяті з різних генеральних сукупностей. Для таких випадків можна скористатися непараметричними альтернативами параметричних критеріїв для двох незалежних груп:

– U критерій Манна – Вітні;

– критерій серій Вальда – Вольфовица;

– критерій Колмогорова – Смірнова.

У випадку використання відносних значень часових параметрів диграфів задача аутентифікації зводиться до перевірки гіпотез про подібність таких характеристик часових інтервалів клавіатурного набору як середнє, медіана і стандартне відхилення.

В даному випадку алгоритм проходження аутентифікації може бути наступним [4].

Перший крок. Формуються вектори відносних часових параметрів клавіатурного почерку користувача:

$$DT_1 = \begin{vmatrix} \frac{t_1 X_H Y_H}{t_1 X_B Y_H} \\ \frac{t_2 X_H Y_H}{t_2 X_B Y_H} \\ \vdots \\ \frac{t_M X_H Y_H}{t_M X_B Y_H} \end{vmatrix}, \quad DT_2 = \begin{vmatrix} \frac{t_1 X_H Y_H}{t_1 X_B Y_B} \\ \frac{t_2 X_H Y_H}{t_2 X_B Y_B} \\ \vdots \\ \frac{t_M X_H Y_H}{t_M X_B Y_B} \end{vmatrix}, \tag{1}$$

де M – кількість диграфів «XY».

Другий крок. Перехід до статистичних параметрів диграфів – математичних сподівань m_{DT_1} та m_{DT_2}, СКВ σ_{DT_1} та σ_{DT_2} – і формування еталону користувача:

$$etalon = \begin{vmatrix} m_{1\,DT_1} & \sigma_{1\,DT_1} & m_{1\,DT_2} & \sigma_{1\,DT_2} \\ m_{2\,DT_1} & \sigma_{2\,DT_1} & m_{2\,DT_2} & \sigma_{2\,DT_2} \\ \vdots & \vdots & \vdots & \vdots \\ m_{L\,DT_1} & \sigma_{L\,DT_1} & m_{L\,DT_2} & \sigma_{L\,DT_2} \end{vmatrix}, \quad (2)$$

де L – кількість аналізованих диграфів.

Третій крок. Вхідними значеннями для системи аутентифікації є вектори фактичних значень відносних часових параметрів диграфів (1), що надходять з блоку моніторингу клавіатури, які потім перетворюються в вектори $DTexp_j$ та $DTexp_j$ ($j = 1 \div K$ – кількість диграфів у поточному профілі).

Четвертий крок. За умови нормального закону розподілу середньоквадратичне відхилення – найпоширеніший показник розсіювання значень випадкової величини відносно її математичного сподівання. Тому при порівнянні з еталоном для кожної компоненти векторів $DTexp_j$ та $DTexp_j$ перевіряються умови:

$$m_{1i} - 3\sigma_{1i} \le \frac{t_{jXi_HYi_H}}{t_{jXi_BYi_H}} \le m_{1i} + 3\sigma_{1i}, \quad (3)$$

$$m_{2i} - 3\sigma_{2i} \le \frac{t_{jXi_HYi_H}}{t_{jXi_BYi_B}} \le m_{2i} + 3\sigma_{2i}. \quad (4)$$

В результаті формується вектор узгодженості:

$$V = \begin{vmatrix} v_1 \\ v_2 \\ \vdots \\ v_{2K} \end{vmatrix}, \begin{cases} v_{1\div K} = \begin{cases} 1, & \text{умова (3) виконується;} \\ 0, & \text{умова (3) не виконується;} \end{cases} \\ v_{(K+1)\div 2K} = \begin{cases} 1, & \text{умова (4) виконується;} \\ 0, & \text{умова (4) не виконується.} \end{cases} \end{cases} \quad (5)$$

За нормою вектора узгодженості можна приймати рішення про дійсність суб'єкта:

$$\begin{cases} |V| \le Z_{ВІД} & - \text{ відмова;} \\ |V| \ge Z_{ДОП} & - \text{ допуск;} \\ Z_{ВІД} < |V| < Z_{ДОП} & - \text{ подальший аналіз.} \end{cases} \quad (6)$$

Значення порогів відмови і допуску приймаються рівними $Z_{ВІД} = 0.5L$ і $Z_{ДОП} = 0.6L$ відповідно, де L – кількість одиниць у векторі узгодженості. Також для проходження аутентифікації необхідно, щоб система розпізнала не менше 75 % диграфів.

Висновки

Сформовано підходи до побудови комплексних алгоритмів аутентифікації, що враховують статистичні особливості клавіатурного почерку. У випадку парольної аутентифікації визначальним є обмежений набор даних і, як наслідок, необхідно застосовувати непараметричні критерії виявлення подібностей/відмінностей. Задачу неперервної аутентифікації можна звести до перевірки гіпотез про подібність середнього та стандартного відхилення відносних значень часових інтервалів клавіатурного набору.

Список літератури:

1. Сравнительный анализ перспективных технологий аутентификации пользователей ПК по клавиатурному почерку / В. А. Алексеев, Д. В. Маслий, Д. Ю. Горелов // Радиотехника. 2017. Вып. 189. С. 195-201.

2. Парольная и непрерывная аутентификация по клавиатурному почерку средствами математической статистики / В. Е. Хиценко, Д. С. Крутохвостов // Вопросы кибербезопасности. 2017. № 5 (24). С. 91-98.

3. Alshehri, A., Coenen, F., & Bollegala, D. (2017). Accurate continuous and non-intrusive user authentication with multivariate keystroke streaming. In ICЗK 2017 // Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management Vol. 1 (pp. 61-70).

4. Модифицированный метод диграфов в задаче аутентификации пользователей по клавиатурному почерку / В.А. Алексеев, Ю.А. Синица, Д.Ю. Горелов // Защита информации. Киев, 2017. Вып. 4. С. 252-261.

Харківський національний університет радіоелектроніки

Надійшла до редколегії 17.05.2019