

РОЗПІЗНАВАННЯ КЛАВІАТУРНОГО ПОЧЕРКУ ЗА ДОПОМОГОЮ ДЕВ'ЯТИВИМІРНИХ ЕЛІПСОЇДІВ ПРОГНОЗУВАННЯ ДЛЯ НОРМАЛІЗОВАНИХ ДАНИХ З РІЗНИМИ КВАНТИЛЯМИ

Приходько С.Б., д.т.н, проф.¹; Трухов А.С.²

^{1,2} Національний університет кораблебудування імені адмірала Макарова

^{1,2} Україна, Миколаїв

¹ sergiy.prykhodko@nuos.edu.ua; ² artem.trukhov@gmail.com

Анотація. Побудовано еліпсоїди прогнозування для нормалізованих за допомогою дев'ятивимірного перетворення Бокса-Кокса даних з квантилями χ^2 -квадрат і F -розподілу для вирішення задачі розпізнавання клавіатурного почерку.

Ключові слова: розпізнавання клавіатурного почерку; χ^2 -квадрат; F -розподіл; еліпсоїд прогнозування; нормалізація даних.

Вступна частина. Технологія розпізнавання клавіатурного почерку знаходить своє застосування в численних галузях, включаючи інформаційну безпеку, автентифікацію користувачів та кібербезпеку. Низька ймовірність автентифікації користувача на основі його клавіатурного почерку може мати серйозні наслідки, що призводить до необхідності подальшого розвитку та вдосконалення методів розпізнавання клавіатурних шаблонів.

Один із поширених підходів у розпізнаванні образів полягає в застосуванні еліпсоїда прогнозування, проте цей метод ґрунтується на припущенні про багатовимірний нормальний розподіл даних [1], але в реальних умовах це не завжди виконується [2], що знижує ймовірність розпізнавання. Тому виникає необхідність застосування нормалізуючих перетворень, які наближають розподіл даних до нормального, що підвищує ефективність застосування методів статистичного аналізу. Одновимірні перетворення, такі як десятковий логарифм або одновимірне перетворення Бокса-Кокса, працюють з кожною характеристикою незалежно, в той час, як багатовимірні, наприклад, багатовимірне перетворення Бокса-Кокса, враховують кореляцію між змінними.

Інший фактор, який впливає на ймовірність розпізнавання - це застосування еліпсоїдів прогнозування з різними квантилями розподілів, таких як χ^2 -квадрат (1) або F -розподіл (2) [3];

$$(X - \bar{X})^T S_X^{-1} (X - \bar{X}) \leq \chi_{k,\alpha}^2, \quad (1)$$

де X - випадкова величина; $\bar{X}$ - вектор вибірових середніх; $\chi_{k,\alpha}^2$ - значення квантиля розподілу χ^2 -квадрат для k ступенів свободи та рівня значущості α ; S_X - вибіркова коваріаційна матриця, що обчислюється $S_X = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(X_i - \bar{X})^T$; N - кількість точок даних;

$$(X - \bar{X})^T S_X^{-1} (X - \bar{X}) = \frac{k(N^2-1)}{N(N-k)} F_{k,N-k,\alpha}, \quad (2)$$

де $F_{k,N-k,\alpha}$ - квантиль F -розподілу з k та $N - k$ ступенями свободи та рівнем значущості α .

Метою роботи є підвищення ймовірності розпізнавання клавіатурного почерку шляхом побудови дев'ятивимірних еліпсоїдів прогнозування для нормалізованих даних з різними квантилями розподілів.

Основна частина. Для роботи було обрано популярний набір даних CMU, що включає в себе метрики натискань клавіш, виміряні в секундах, такі як тривалість утримання та час

між натисканнями 51 учасника при наборі пароля «.tie5Roanl». Колонки з позначкою H.key вказують на тривалість утримання конкретної клавіші, колонки DD.key1.key2 відображають інтервал між натисканням двох послідовних клавіш, UD.key1.key2 позначають інтервал між відпусканням однієї клавіші та натисканням наступної. Для спрощення моделі було вирішено зосередитися на 9 ключових характеристиках, тому вектор ознак має вигляд: $X = \{ H.t, H.i, H.e, H.5, H.R, H.o, H.a, H.n, H.l \}$.

Після виділення векторів ознак наступним важливим етапом є видалення викидів. Популярним методом є використання квадрата відстані Махаланобіса, що базується на припущенні про багатовимірний нормальний розподіл даних. Тому було застосовано багатовимірне перетворення Бокса-Кокса:

$$Z_j = x(\lambda_j) = \begin{cases} (X_j^{\lambda_j} - 1)/\lambda_j, & \lambda_j \neq 0; \\ \ln(X_j), & \lambda_j = 0; \end{cases} \quad (3)$$

де j змінюється від 1 до 9; X_j – випадкова величина, яка нормалізується; Z_j – нормалізована нормально розподілена величина; λ_j – параметр перетворення Бокса-Кокса.

Потім видаляється аномалія, починаючи з найбільших значень квадрата відстані Махаланобіса. Цей процес відбувається ітеративно, всього було видалено 10 точок.

Набір даних був перемішаний та розділений на 2 вибірки по 195 точок: навчальну та тестову. Вектор вибірових середніх навчальної вибірки $\bar{X} = \{0.07635; 0.07052; 0.07875; 0.06254; 0.06955; 0.08806; 0.08752; 0.07447; 0.07495\}$. За результатами тесту Мардіа, багатовимірний розподіл навчальної вибірки демонструє відхилення від багатовимірного нормального, тому було застосовано дев'ятивимірне перетворення Бокса-Кокса. За методом максимальної правдоподібності було отримано наступні оцінки параметрів: $\hat{\lambda}_1 = 1.3676$, $\hat{\lambda}_2 = 1.4807$, $\hat{\lambda}_3 = 1.078$, $\hat{\lambda}_4 = 1.7393$, $\hat{\lambda}_5 = 2.1004$, $\hat{\lambda}_6 = 1.1498$, $\hat{\lambda}_7 = 1.566$, $\hat{\lambda}_8 = 1.1685$, $\hat{\lambda}_9 = 2.1146$.

Унаслідок застосування дев'ятивимірного перетворення Бокса-Кокса з компонентами (3), було отримано вибірку з вектором середніх $\bar{Z} = \{0.70932; 0.66184; 0.86764; -0.57016; -0.47427; 0.81642; 0.62417; -0.81443; -0.47084\}$, яка не відхиляється від багатовимірного нормального розподілу, адже значення тестової статистики багатовимірної асиметрії $B_1 n/6$, яке дорівнює 175.47, не перевищує критичного значення 215.53; тестова статистика багатовимірного ексцесу B_2 , значення якої 99.76, менше за критичне значення 104.19.

На основі (1) рівняння еліпсоїда прогнозування для нормалізованих даних з квантилем розподілу Хі-квадрат набуває вигляду:

$$(Z - \bar{Z})^T S_Z^{-1} (Z - \bar{Z}) \leq \chi_{9, 0.005}^2, \quad (4)$$

де значення квантиля розподілу Хі-квадрат $\chi_{9, 0.005}^2$ для 9 ступенів свободи та рівня значущості 0.005 становить 23.59.

На основі (2) рівняння еліпсоїда прогнозування для нормалізованих даних з квантилем F-розподілу набуває вигляду:

$$(Z - \bar{Z})^T S_Z^{-1} (Z - \bar{Z}) \leq \frac{k(N^2-1)}{N(N-k)} F_{9, 186, 0.005}, \quad (5)$$

де значення квантиля F-розподілу $F_{9, 186, 0.005}$ для 9 та 186 ступенів свободи та рівня значущості 0.005 становить 2.74, а значення коефіцієнта 9.44, відповідно права частина рівняння становить 25.86.

Таблиця 1 містить порівняння результатів використання еліпсоїдів прогнозування для негаусівських даних з квантилем χ^2 -квадрат (1) та квантилем F -розподілу (2), а також еліпсоїдів для нормалізованих даних з квантилем χ^2 -квадрат (3) та квантилем F -розподілу (4). Моделі оцінюються за трьома метриками: точність (accuracy), яка вимірює загальну правильність розпізнавання; чутливість (sensitivity) оцінює здатність моделі виявляти аномалії; специфічність (specificity) оцінює здатність моделі правильно класифікувати цільові екземпляри.

Таблиця 1

Порівняння результатів застосування еліпсоїдів прогнозування

Модель	Accuracy	Sensitivity	Specificity
Негаусівські дані, χ^2 -квадрат	0.9412	0.9225	0.9795
Негаусівські дані, F -розподіл	0.9328	0.9025	0.9949
Нормалізовані дані, χ^2 -квадрат	0.9782	0.9700	0.9949
Нормалізовані дані, F -розподілу	0.9765	0.9650	1

Результати показують, що використання квантилю розподілу χ^2 -квадрат забезпечує вищу загальну точність та краще збалансовані показники чутливості і специфічності порівняно з F -розподілом. У той час, як F -розподіл більше фокусується на правильному розпізнаванні цільових точок, і показує дещо нижчу чутливість до аномалій. Нормалізація даних за допомогою дев'ятивимірного перетворення Бокса-Кокса значно покращила показники всіх метрик для обох еліпсоїдів, що підкреслює її важливість як ключового етапу у розробці ефективних систем розпізнавання.

Висновки. В ході роботи було побудовано еліпсоїди прогнозування для нормалізованих за допомогою багатовимірного перетворення Бокса-Кокса даних із застосуванням квантилів χ^2 -квадрат та F -розподілу. Результати показали, що еліпсоїди прогнозування, побудовані на основі квантиля χ^2 -квадрат, демонструють кращу ймовірність розпізнавання клавіатурного почерку. Важливим є також те, що нормалізація даних значно підвищує ефективність еліпсоїдів прогнозування, не залежно від квантилів розподілу, що підкреслює її важливість у покращенні ймовірності розпізнавання клавіатурного почерку.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Приходько, С., & Трухов, А. (2023). Побудова правил прийняття рішень для розпізнавання облич на основі квадрата відстані Махаланобіса для нормалізованих даних. *Information Technology: Computer Science, Software Engineering and Cyber Security*, (2), 50-58

[2] Oyebola, O. (2023). Examining the distribution of keystroke dynamics features on computer, tablet and mobile phone platforms. *Mobile Computing and Sustainable Informatics: Proceedings of ICMCSI 2023*. pp. 613-620. DOI: 10.1007/978-981-99-0835-6_43

[3] Prykhodko S. B., & Trukhov A. S. (2024). Face recognition using ten-variate prediction ellipsoids for normalized data with different quantiles. *Applied Aspects of Information Technology*. Vol. 7 No. 2: 151–161. DOI: 10.15276/aait.07.2024.11

Prykhodko S.B., Trukhov A.S.
Keystroke dynamics recognition using nine-variate prediction ellipsoids for normalized data with different quantiles
Abstract. Prediction ellipsoids were constructed for normalized data using the nine-dimensional Box-Cox transformation with Chi-square and F-distribution quantiles for solving the problem of keystroke dynamics recognition.
Keywords: keystroke dynamics recognition; Chi-square; F-distribution; prediction ellipsoid; data normalization.