УДК 004.93

РОЗПІЗНАВАННЯ КЛАВІАТУРНОГО ПОЧЕРКУ НА ОСНОВІ КЛАСТЕРНО-ОРІЄНТОВАНИХ ЕЛІПСІВ ПРОГНОЗУВАННЯ

Приходько С.Б., д.т.н, проф. [1]; Трухов А.С. PhD [2]
[1,2] Національний університет кораблебудування імені адмірала Макарова
[1,2] Україна, Миколаїв
[1] sergiy.prykhodko@nuos.edu.ua; [2] artem.trukhov@nuos.edu.ua

Анотація. *У роботі запропоновано підхід до розпізнавання клавіатурного почерку, що поєднує кластеризацію даних і побудову еліпсів прогнозування на основі тренувального набору кожного кластера. Отримані результати показують, що кластерно-орієнтований підхід дозволяє підвищити точність розпізнавання за рахунок адаптації моделі до локальних властивостей даних.*
Ключові слова: *розпізнавання клавіатурного почерку; двовимірні дані; нормалізуюче перетворення; кластеризація; еліпс прогнозування; локальне моделювання.*

Вступна частина. Розпізнавання клавіатурного почерку є перспективним напрямом у біометричній ідентифікації користувачів. Точність розпізнавання значною мірою залежить від властивостей даних, їх розподілу та наявності викидів. У традиційних підходах еліпси прогнозування будуються на основі всього набору тренувальних точок, що може знижувати ефективність при неоднорідності даних. Тому актуальним є підхід, який розбиває простір даних шляхом кластеризації та побудови локальних моделей для кожного кластера [1]. Такий підхід дозволяє підвищити точність розпізнавання завдяки поділу характеристик клавіатурного почерку користувача на менші, більш однорідні кластери.

Метою роботи є підвищення точності розпізнавання клавіатурного почерку шляхом побудови кластерно-орієнтованих еліпсів прогнозування.

Основна частина. Для проведення дослідження використано популярний набір даних CMU Keystroke Dynamics Benchmark Dataset, який містить часові характеристики натискань клавіш 51 учасника під час введення пароля «.tie5Roanl». Кожен запис включає такі параметри: H.key – тривалість утримання окремої клавіші (hold time); DD.key1.key2 – інтервал між натисканням двох послідовних клавіш (down–down time); UD.key1.key2 – інтервал між відпусканням однієї клавіші та натисканням наступної (up–down time).

Для спрощення моделі та візуалізації еліпсів прогнозування було обрано лише дві ознаки X={H.o, UD.o.a}, де H.o – тривалість утримання клавіші “o”, а UD.o.a – інтервал між відпусканням клавіші “o” та натисканням клавіші “a”.

Наступним етапом є видалення викидів, що суттєво впливають на статистичні оцінки та значно ускладнюють подальший аналіз. Для цього застосовано підхід на основі квадрата відстані Махаланобіса [2]. Оскільки цей підхід базується на припущенні про багатовимірний нормальний розподіл, попередньо виконано двовимірне перетворення Бокса–Кокса, яке наближує розподіл даних до нормального. Після цього дані очищено від аномальних спостережень ітераційним способом: на кожному кроці було видалено точку з найбільшим значенням квадрата відстані Махаланобіса. У результаті було видалено вісім викидів.

Після видалення викидів виконано кластеризацію нормалізованих даних. Оскільки метод K-means базується на мінімізації відстані Евкліда між точками та центроїдами кластерів, він працює найбільш ефективно тоді, коли дані мають наближено нормальний

розподіл [3]. Саме тому кластеризація застосовувалася до даних, попередньо нормалізованих за допомогою двовимірного перетворення Бокса–Кокса.

Кількість кластерів визначалася за допомогою методу локтя (Elbow Method), який ґрунтується на аналізі зміни значення функції втрат, а саме сумарної квадратичної відстані всередині кластерів при збільшенні кількості кластерів. Точка перегину графіка відповідає оптимальній кількості кластерів, коли подальше збільшення їх числа не дає суттєвого зменшення похибки (рис. 1). Оптимальна кількість кластерів становить сім.

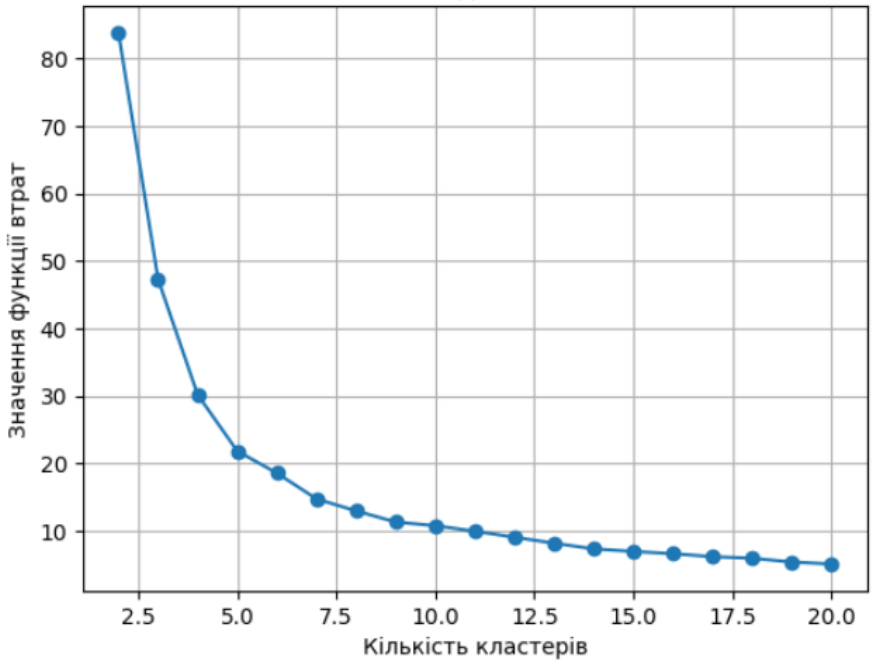

Рисунок 1 – Графік залежності значення функції втрат від кількості кластерів

Схематичне зображення розподілу точок за кластерами та еліпса прогнозування, побудованого на основі всього набору нормалізованих даних, наведено на рис. 2.

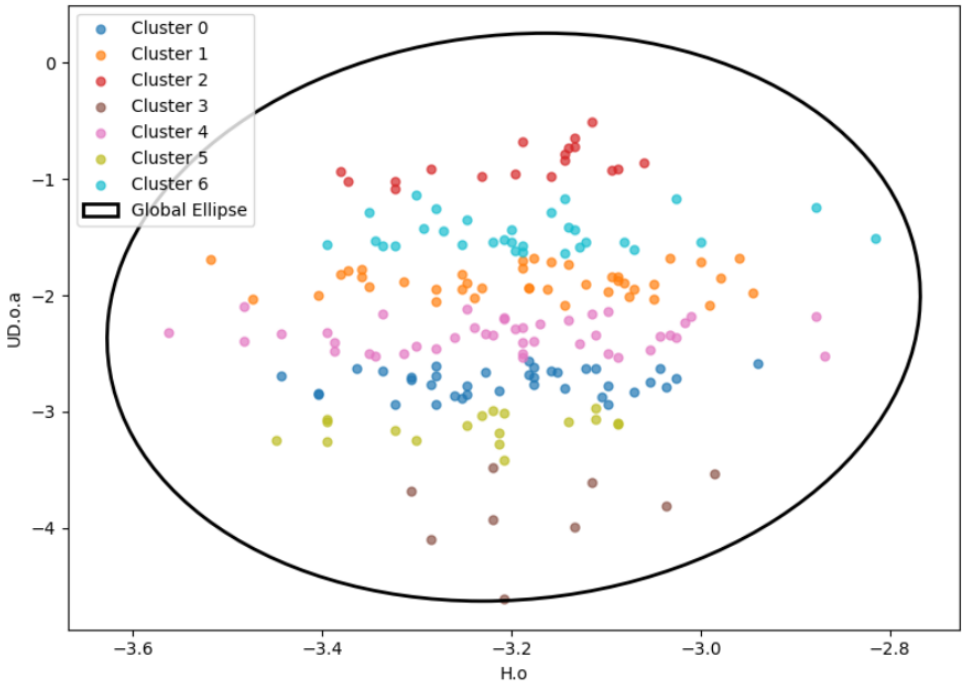

Рисунок 2 – Розподіл точок за кластерами

Для забезпечення однорідності тренувальної та тестової вибірок із кожного кластера було відібрано 50% точок у тренувальний і 50% у тестовий набір. Такий підхід гарантує рівномірне представлення усіх локальних структур даних у процесі побудови та перевірки моделей еліпсів прогнозування. Оскільки деякі кластери містили непарну кількість спостережень, кількість точок у вибірках виявилася нерівною: 194 було включено до тренувального набору, та 198 – до тестового.

Перевірка тренувальних вибірок за тестом Мардіа підтвердила, що дані кожного кластера мають двовимірний нормальний розподіл, що обґрунтовує коректність використання еліпсів прогнозування як статистичних моделей для первинних даних. Для кожного набору початкових тренувальних даних, що належав окремому кластеру, було побудовано еліпс прогнозування (рис. 3).

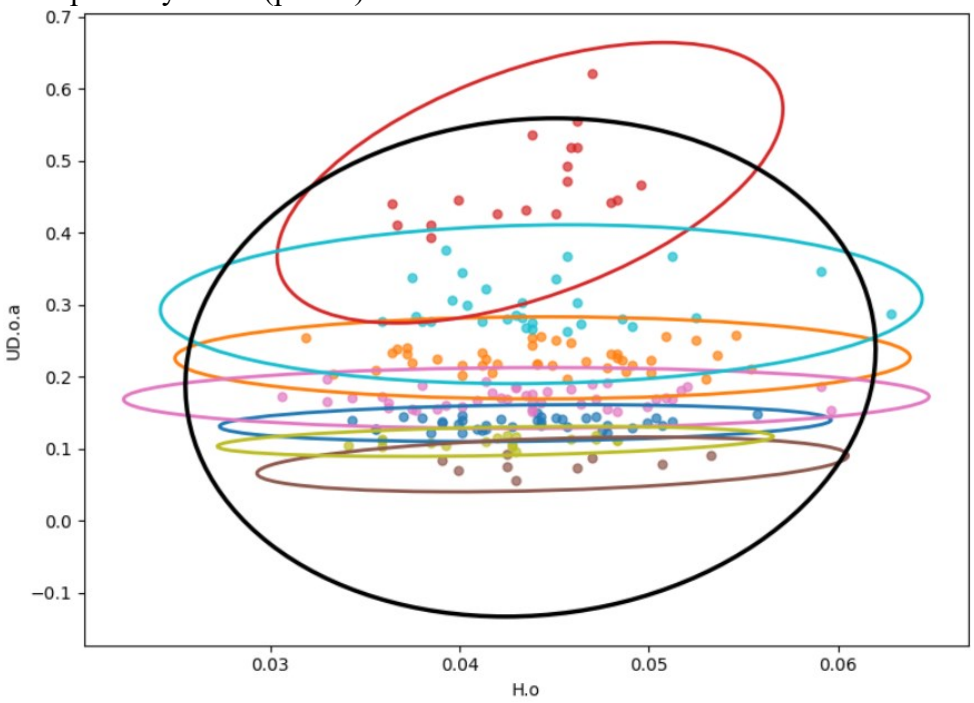

Рисунок 3 – Побудовані кластерно-орієнтовані еліпси прогнозування

За допомогою побудованих кластерно-орієнтованих еліпсів прогнозування (COPE), еліпса прогнозування для негаусівських даних (PENGD) та еліпса прогнозування для нормалізованих даних (PEND) відбувається розпізнавання тестового набору цільового класу та вибірки представника іншого класу (табл. 1). Моделі оцінювалися за трьома метриками: точністю (accuracy), що характеризує загальну правильність розпізнавання; чутливістю (sensitivity), яка визначає здатність моделі правильно ідентифікувати цільові зразки; та специфічністю (specificity), що відображає точність виявлення аномальних точок.

Таблиця 1

Порівняння результатів застосування еліпсів прогнозування

Модель	Accuracy	Sensitivity	Specificity
COPE	0.9883	0.9899	0.9875
PENGD	0.9817	0.9747	0.985
PEND	0.9900	1	0.985

Порівняння результатів показало, що еліпс прогнозування, побудований для негаусівських даних, має найнижчу точність розпізнавання. Кластерно-орієнтовані еліпси,

побудовані окремо для кожного кластера, продемонстрували кращу здатність розпізнавати представників інших класів. Еліпс прогнозування для нормалізованих даних показав найвищу загальну точність розпізнавання.

Висновки. Запропонований кластерно-орієнтований підхід до розпізнавання клавіатурного почерку на основі еліпсів прогнозування є перспективним напрямом підвищення точності біометричної ідентифікації людини. На відміну від загального еліпса, побудованого для негаусівських даних, який охоплює надлишкову область простору ознак, кластерно-орієнтовані еліпси прогнозування локалізуються лише в тих ділянках, де фактично зосереджені спостереження. Подальші дослідження доцільно спрямувати на порівняння результатів із використанням альтернативних алгоритмів кластеризації, зокрема Gaussian Mixture Models, розширення простору ознак, а також аналіз впливу кількості кластерів на точність розпізнавання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Abdallah, Z. S., Gaber, M. M., Srinivasan, B., & Krishnaswamy, S. (2012, December). CBARS: Cluster based classification for activity recognition systems. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 82-91). Berlin, Heidelberg: Springer Berlin Heidelberg.

[2] Prykhodko, S., & Trukhov, A. (2025). Keystroke Dynamics Recognition Using an Eight-Variate Prediction Ellipsoid for Normalized Data with a Reduced Feature Set. In *Proceedings of the 13-th International Conference on Information Control Systems & Technologies (ICST 2025)* (pp. 496-508). Odesa, Ukraine, September 24–26, 2025.

[3] El Khattabi, M. Z., El Jai, M., Lahmadi, Y., Oughdir, L., & Rahhali, M. (2024). Understanding the interplay between metrics, normalization forms, and data distribution in K-means clustering: A comparative simulation study. *Arabian Journal for Science and Engineering*, *49*(3), 2987-3007.

Prykhodko S.B., Trukhov A.S.

Keystroke dynamics recognition based on cluster-oriented prediction ellipses

Abstract. *he paper proposes an approach to keystroke dynamics recognition that combines data clustering and the construction of prediction ellipses based on the training set of each cluster. The obtained results show that the cluster-oriented approach improves recognition accuracy by adapting the model to the local properties of the data.*

Keywords: *keystroke dynamics recognition; two-variate data; normalizing transformation; clustering; prediction ellipsoid; local modeling.*