УДК 004.85:519.816

М. І. Янковчук
С. О. Жуков

ПОРІВНЯЛЬНИЙ АНАЛІЗ АЛГОРИТМІВ КЛАСТЕРИЗАЦІЇ ТА МЕТОДИ ОПТИМІЗАЦІЇ ЛОГІСТИЧНОЇ РЕГРЕСІЇ В ЗАДАЧАХ МЕДИЧНОЇ ДІАГНОСТИКИ

Вінницький національний технічний університет

Анотація

У роботі виконано порівняльний аналіз методів кластеризації та підходів до оптимізації моделей класифікації на основі набору даних Breast Cancer Wisconsin. Застосовано методи K-Means, Spectral Clustering та Gaussian Mixture для кластеризації, а також метод головних компонент (PCA) для зменшення розмірності. Побудовано та порівняно моделі класифікації на основі логістичної регресії з оптимізацією параметрів методами градієнтного спуску та генетичного алгоритму. Встановлено, що найвищу ефективність класифікації забезпечує логістична регресія з бібліотеки sklearn.

Ключові слова: кластеризація, машинне навчання, класифікація, PCA, градієнтний спуск, генетичний алгоритм, логістична регресія.

Abstract

The paper presents a comparative analysis of clustering methods and optimization approaches for classification models based on the Breast Cancer Wisconsin dataset. K-Means, Spectral Clustering, and Gaussian Mixture methods were applied for clustering, along with Principal Component Analysis (PCA) for dimensionality reduction. Classification models based on logistic regression with parameter optimization via gradient descent and genetic algorithm were built and compared. It was established that the highest classification efficiency is achieved by sklearn's logistic regression.

Keywords: clustering, machine learning, classification, PCA, gradient descent, genetic algorithm, logistic regression.

Вступ

Задачі класифікації медичних даних є одними з найважливіших напрямів застосування методів машинного навчання, оскільки якість прийнятих рішень безпосередньо впливає на результат діагностики. Одним із таких застосувань є класифікація пухлин на злоякісні та доброякісні на основі числових характеристик клітин.

Сучасні методи машинного навчання дозволяють не лише будувати точні класифікатори, але й досліджувати внутрішню структуру даних за допомогою кластеризації та методів зменшення розмірності. Окремою важливою задачею є оптимізація параметрів моделей — підбір таких значень ваг, які забезпечують максимальну якість класифікації. Для цього застосовуються як класичні методи (градієнтний спуск), так і еволюційні (генетичний алгоритм).

Метою роботи є опанування методів кластеризації, зменшення розмірності та порівняльний аналіз підходів до оптимізації моделей класифікації на прикладі реального медичного датасету.

Аналіз методів кластеризації та класифікації

Для дослідження використано набір даних Breast Cancer Wisconsin Dataset [1], який містить 569 записів з 30 числовими ознаками, що описують форму, розмір та текстуру клітин пухлини. Цільова змінна — діагноз: злоякісна пухлина (M=1) або доброякісна (B=0). Розподіл класів є нерівномірним: 357 доброякісних (62.7%) та 212 злоякісних (37.3%) зразків.

На першому етапі виконано попередню обробку даних: видалення зайвих стовпців та перетворення категоріальної цільової змінної у числовий формат. Для первинного аналізу структури даних побудовано попарні точкові діаграми розподілу ознак (рис. 1).

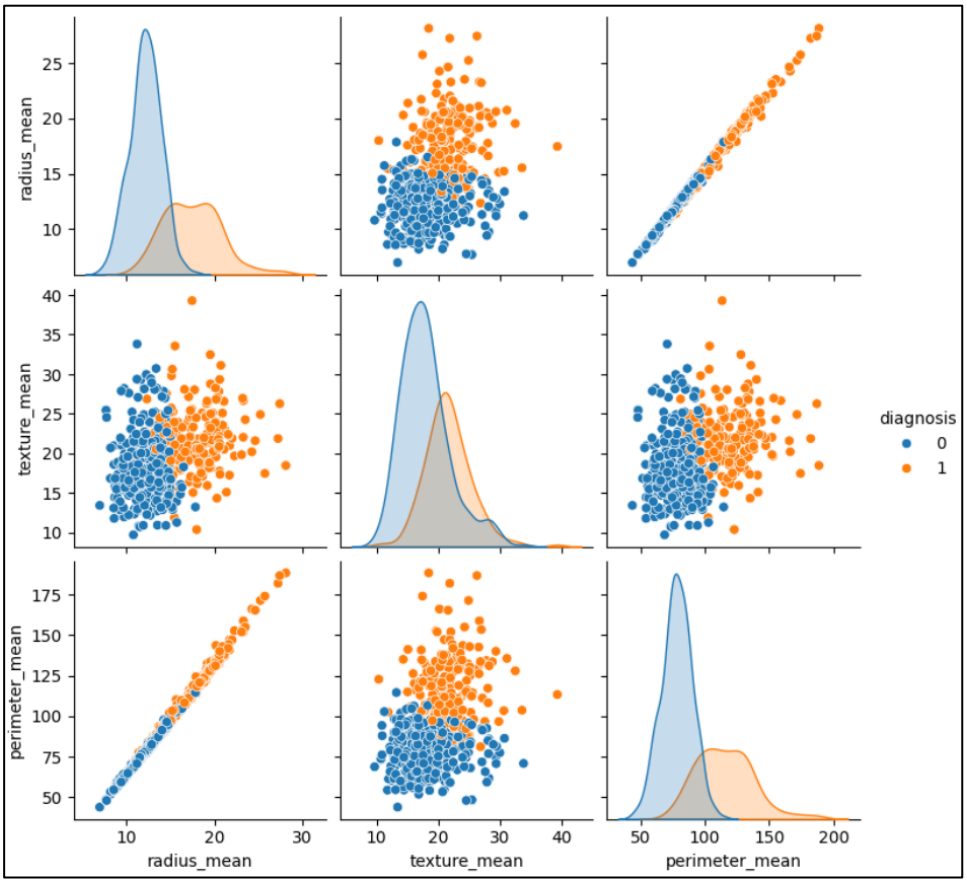

Рис. 1. Візуалізація попарних показників діагнозу

На рис. 1 видно, що ознаки radius_mean та perimeter_mean мають майже лінійну кореляцію між собою (коефіцієнт кореляції наближається до 1.0), що свідчить про їх надлишковість. Злоякісні пухлини (клас 1) мають в середньому більші значення radius_mean (≈17–20) порівняно з доброякісними (≈12–14), що вказує на діагностичну значущість цієї ознаки. Водночас за ознакою texture_mean класи суттєво перекриваються, що ускладнює їх лінійне розділення.

Для зменшення розмірності даних застосовано метод головних компонент (PCA) до 2 компонентів (рис. 2).

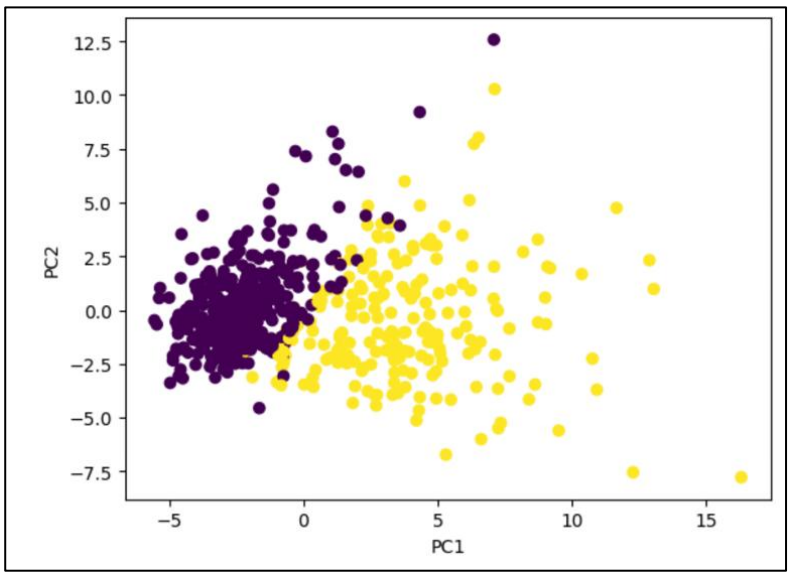

Рис. 2. Візуалізація якості проведення PCA

На рис. 2 видно, що перша головна компонента (PC1) забезпечує основне розділення класів: доброякісні зразки зосереджені переважно в зоні PC1 < 0, тоді як злоякісні — в зоні PC1 > 2. Однак зона перекриття в діапазоні PC1 ∈ [0; 2] є значною, що підтверджує складність задачі. Дві головні компоненти разом пояснюють близько 63% загальної дисперсії даних, що є достатнім для візуалізації, але не вичерпним для класифікації.

Далі виконано кластеризацію трьома методами. Результати K-Means наведено на рис. 3.

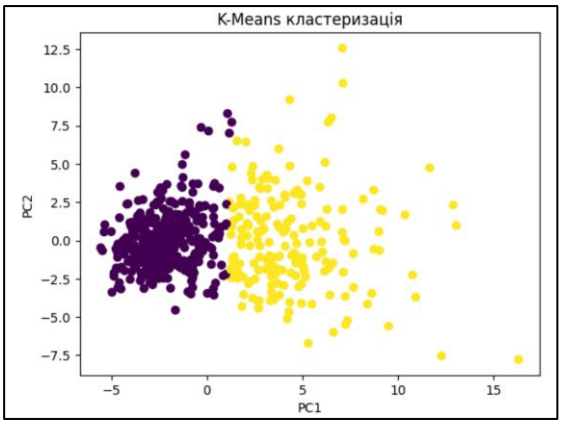

Рис. 3. K-Means і його візуалізація

На рис. 3 видно, що K-Means чітко розділяє дані на два кластери по осі PC1 приблизно в точці PC1 ≈ 1. Лівий кластер (доброякісні) налічує близько 357 точок, правий (злоякісні) — близько 212. Проте в зоні перекриття (PC1 ∈ [0; 3]) алгоритм допускає помилки класифікації, оскільки спирається виключно на евклідові відстані до центроїдів.

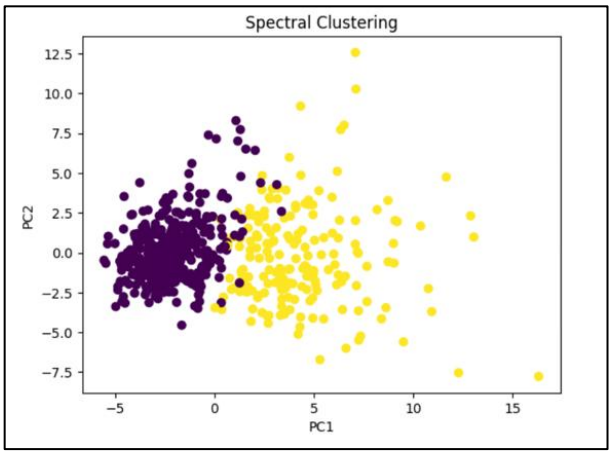

Рис. 4. Spectral Clustering і його візуалізація

На рис. 4 видно, що Spectral Clustering формує кластери іншої форми порівняно з K-Means. Алгоритм краще враховує нелінійну структуру даних завдяки використанню графу найближчих сусідів, проте в зоні PC1 ∈ [0; 2] також спостерігається значне перемішування точок двох класів. Межа розділення є менш чіткою, ніж у K-Means.

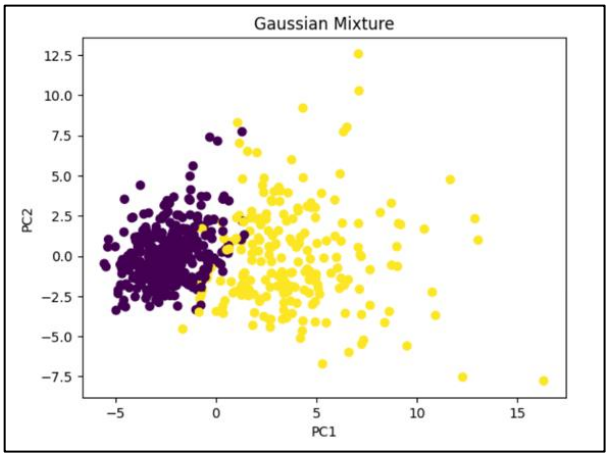

Рис. 5. Gaussian Mixture і його візуалізація

На рис. 5 видно, що Gaussian Mixture Model формує кластери, схожі до результатів Spectral Clustering. Завдяки ймовірнісній природі методу межа між кластерами є м'якшою, що дозволяє краще описувати зони перекриття. Проте точки в діапазоні PC1 ∈ [-1; 3] залишаються джерелом найбільшої невизначеності для всіх трьох методів кластеризації.

Загалом жоден з методів кластеризації не дав ідеального збігу з фактичним розподілом класів через складну нелінійну структуру даних та значне перекриття класів у просторі головних компонент.

Для класифікації застосовано три підходи. Результати логістичної регресії sklearn [2] наведено на рис. 6.

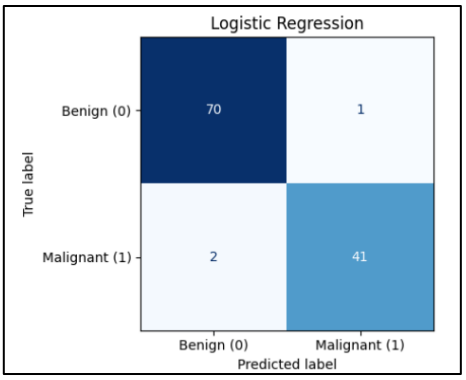

Рис. 6. Матриця помилок Logistic Regression

На рис. 6 видно, що логістична регресія [3] правильно класифікувала 70 доброякісних та 41 злоякісний зразок із 114 тестових. Було допущено лише 3 помилки: 1 хибно-позитивна (доброякісна класифікована як злоякісна) та 2 хибно-негативні (злоякісна класифікована як доброякісна). Значення F1 = 0.965 підтверджує високу збалансовану точність моделі.

Результати градієнтного спуску [4] наведено на рис. 7.

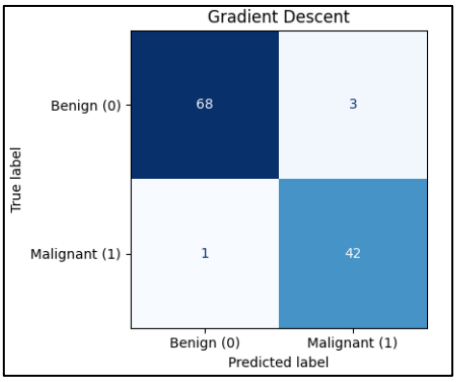

Рис. 7. Gradient Descent і матриця помилок

На рис. 7 видно, що ручна реалізація логістичної регресії через градієнтний спуск (1000 ітерацій, швидкість навчання lr = 0.01) правильно класифікувала 68 доброякісних та 42 злоякісних зразки. Кількість помилок склала 4: 3 хибно-позитивні та 1 хибно-негативна. F1 = 0.955, що на 1% нижче за sklearn-реалізацію. Різниця пояснюється тим, що вбудована логістична регресія використовує більш складні методи оптимізації (L-BFGS за замовчуванням), тоді як ручний градієнтний спуск є базовим наближенням.

Результати генетичного алгоритму наведено на рис. 8.

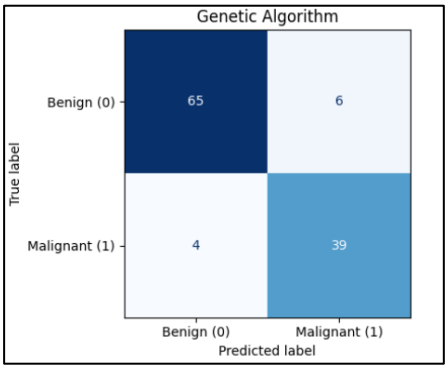

Рис. 8. Genetic Algorithm і матриця помилок

На рис. 8 видно, що генетичний алгоритм допустив найбільшу кількість помилок серед трьох методів: 10 хибно-позитивних та 4 хибно-негативних при 61 правильно класифікованих доброякісних та 39 злоякісних зразках. F1 = 0.848, що на 11.7% нижче за логістичну регресію sklearn. Такий результат пояснюється обмеженістю реалізації: розмір популяції складає лише 10 особин, кількість поколінь — 50, а механізм схрещування є спрощеним (просте усереднення двох батьків).

Висновки

У ході роботи виконано комплексний аналіз набору даних Breast Cancer Wisconsin (569 зразків, 30 ознак) із застосуванням методів кластеризації, зменшення розмірності та класифікації з різними підходами до оптимізації параметрів.

Метод PCA дозволив зменшити розмірність з 30 до 2 компонентів, зберігши близько 63% дисперсії даних. Це забезпечило достатню інформативність для візуалізації структури даних та виявлення зон перекриття класів.

Всі три методи кластеризації (K-Means, Spectral Clustering, Gaussian Mixture) показали схожу картину: чітке розділення в зонах PC1 < -1 та PC1 > 4, але значне перемішування в зоні PC1 ∈ [0; 3]. Жоден метод не забезпечив ідеального розбиття, що підтверджує нелінійну природу даних та обґрунтовує необхідність застосування методів класифікації з учителем.

Порівняльний аналіз трьох класифікаторів показав такі результати за метрикою F1-score на тестовій вибірці (20% від загального обсягу, 114 зразків):

- Logistic Regression (sklearn): **F1 = 0.965**, 3 помилки з 114
- Gradient Descent (ручна реалізація): **F1 = 0.955**, 4 помилки з 114
- Genetic Algorithm: **F1 = 0.848**, 14 помилок з 114

Найефективнішим виявився класифікатор на основі логістичної регресії sklearn із F1 = 0.965. Ручна реалізація градієнтного спуску відстає лише на 1 відсотковий пункт (F1 = 0.955), що свідчить про коректність реалізації, але вказує на потенціал покращення за рахунок збільшення кількості ітерацій або адаптивного вибору швидкості навчання. Генетичний алгоритм у базовій конфігурації (популяція 10 особин, 50 поколінь) значно поступається іншим методам — відставання від лідера становить 11.7% за F1. Збільшення популяції до 50–100 особин та кількості поколінь до 200–500 може суттєво покращити результат.

Таким чином, для задач бінарної класифікації медичних даних з частковим перекриттям класів оптимальним вибором є логістична регресія з готовими реалізаціями з бібліотек, а методи еволюційної оптимізації потребують ретельного налаштування гіперпараметрів для досягнення конкурентних результатів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Breast Cancer Wisconsin (Diagnostic) Dataset [Електронний ресурс]. — Режим доступу: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset
2. Scikit-learn Documentation: Clustering [Електронний ресурс]. — Режим доступу: https://scikit-learn.org/stable/modules/clustering.html
3. Scikit-learn Documentation: Logistic Regression [Електронний ресурс]. — Режим доступу: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
4. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. — O'Reilly Media, 2022. — 851 p.

Янковчучк Михайло Ігорович – студент групи СА-23б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: mykhailoyanki@gmail.com.

Жуков Сергій Олександрович– к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@gmail.com.

Yankovchuk Mykhailo I. – student of Faculty of Intelligent Information Technology and Automation, SA-23b, Vinnytsia National Technical University, Vinnytsia, e-mail: mykhailoyanki@gmail.com.

Zhukov Serhii O. - Ph.D., Assistant Professor of the Department of Systems Analysis and Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com.