ЗАСТОСУВАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ ДАНИХ

Хиць Руслан Андрійович
Луцький національний технічний університет,
асистент, кафедра комп'ютерних наук, ruslanioc@lntu.edu.ua

APPLICATION OF MACHINE LEARNING FOR DATA CLASSIFICATION

Khyts Ruslan Andriiovych
Lutsk National Technical University,
assistant, Department of Computer Science, ruslanioc@lntu.edu.ua

Machine learning has become a key tool for solving data classification problems across various domains including medicine, finance, cybersecurity and natural language processing. This paper examines the fundamental approaches to classification: decision trees, support vector machines, k-nearest neighbours, naive Bayes classifier, and deep neural networks. Evaluation metrics, preprocessing techniques, and ensemble methods are discussed. Practical recommendations for algorithm selection based on data characteristics are provided, along with an overview of modern software tools for implementation.

Keywords: machine learning, classification, supervised learning, decision tree, neural network, support vector machine, ensemble methods, scikit-learn, data preprocessing.

Вступ. Машинне навчання (МН) є одним із найбільш перспективних напрямів сучасної комп'ютерної науки. За останнє десятиліття інтерес до цієї галузі суттєво зріс завдяки доступності великих обсягів даних, зростанню обчислювальних потужностей та появі ефективних відкритих бібліотек. Задача класифікації – одна з найпоширеніших у практичних застосуваннях: від медичної діагностики та фільтрації спаму до кредитного скорингу, виявлення кібератак і розпізнавання образів. Правильний вибір алгоритму класифікації суттєво впливає на якість, ефективність і масштабованість системи [1].

Постановка задачі класифікації. У задачі класифікації задано навчальну вибірку $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, де x_i – вектор ознак об'єкта, а $y_i \in \{C_1, C_2, \ldots, C_k\}$ – його клас. Метою є побудова функції $f: X \to Y$, що мінімізує помилку класифікації на нових, раніше не бачених даних. Розрізняють бінарну класифікацію ($k = 2$) та багатокласову ($k > 2$). Особливим випадком є багатомічна класифікація, коли один об'єкт може належати одночасно до кількох класів [2].

Основні алгоритми класифікації. Серед методів навчання з учителем (supervised learning) виділяють такі підходи. Дерева рішень (Decision Trees) будують ієрархічну структуру правил розбиття простору ознак. Алгоритм CART використовує індекс Джині як критерій розщеплення, а C4.5 – ентропію інформаційного виграшу. Перевагами є висока інтерпретованість та відсутність потреби у нормалізації даних. Недолік – схильність до перенавчання на малих вибірках [1].

Метод опорних векторів (Support Vector Machine, SVM) шукає гіперплощину максимального зазору між класами у трансформованому просторі ознак. Завдяки ядровим функціям (RBF, поліноміальне, сигмоїдне ядро) SVM ефективно вирішує нелінійно роздільні задачі. Метод демонструє стабільні результати при обмежених обсягах даних і у просторах великої розмірності, що робить його особливо придатним для класифікації текстів та біоінформатики [2].

Алгоритм k найближчих сусідів (k-NN) відносить об'єкт до класу, який найчастіше зустрічається серед k найближчих за евклідовою (або іншою) метрикою сусідів у навчальній вибірці. Метод є непараметричним і легко розширюється на багатокласові задачі, однак повільно працює на великих датасетах через необхідність повного перегляду вибірки під час передбачення [3].

Наївний байєсівський класифікатор (Naive Bayes) базується на теоремі Байєса з припущенням умовної незалежності ознак. Попри спрощеність цього припущення, метод досягає конкурентоспроможних результатів у задачах класифікації текстів, фільтрації спаму та медичної діагностики завдяки обчислювальній ефективності та стійкості до незначних

порушень незалежності ознак [3].

Нейронні мережі та глибоке навчання. Багатошарові перцептрони (MLP) та глибокі нейронні мережі (DNN) здатні автоматично навчатися ієрархічним представленням даних. Згорткові нейронні мережі (CNN) забезпечують точність до 97–99% на задачах класифікації зображень завдяки механізму локальних рецептивних полів. Рекурентні мережі (RNN, LSTM) орієнтовані на послідовні дані – текст і часові ряди. Головний недолік – висока обчислювальна вартість навчання та потреба у великих обсягах розмічених даних [1].

Ансамблеві методи. Random Forest будує множину незалежних дерев рішень на випадкових підвибірках даних і ознак, усереднюючи їхні передбачення. Gradient Boosting (XGBoost, LightGBM, CatBoost) послідовно навчає дерева, кожне з яких виправляє помилки попередніх. Ці методи є де-факто стандартом для табличних даних і регулярно перемагають на змаганнях з машинного навчання (Kaggle) [4].

Метрики оцінювання. Для порівняння алгоритмів використовуються: точність (accuracy) – частка правильних передбачень; прецизійність (precision) і повнота (recall) – особливо важливі при незбалансованих класах; F1-міра – гармонічне середнє precision і recall; площа під ROC-кривою (AUC-ROC) – інтегральна характеристика якості бінарного класифікатора. При незбалансованих класах accuracy є оманливою метрикою, тому рекомендується використовувати F1-macro або AUC-ROC [2].

Попередня обробка даних. Якість класифікації суттєво залежить від підготовки даних. Основні етапи: нормалізація або стандартизація числових ознак (важлива для SVM, k-NN, нейронних мереж); кодування категоріальних змінних (One-Hot Encoding, Label Encoding); заповнення пропущених значень (медіана, середнє, k-NN imputaton); відбір та конструювання ознак (feature selection, PCA). Некоректна обробка може звести нанівець переваги навіть найпотужнішого алгоритму [4].

Програмні засоби реалізації. Бібліотека scikit-learn (Python) надає уніфікований API для більшості класичних алгоритмів

класифікації, включаючи засоби крос-валідації, підбору гіперпараметрів (GridSearchCV, RandomizedSearchCV) та побудови pipeline. TensorFlow і PyTorch є стандартом для проєктування та навчання глибоких нейронних мереж. XGBoost і LightGBM пропонують високооптимізовані реалізації методів градієнтного бустингу з підтримкою GPU [4].

Висновки. Проведений аналіз показав, що не існує єдиного універсального алгоритму класифікації: вибір методу визначається обсягом та природою даних, вимогами до інтерпретованості моделі, допустимим часом навчання та передбачення, а також доступними обчислювальними ресурсами. Для табличних даних оптимальним вибором є методи ансамблювання (Random Forest, XGBoost); для зображень – CNN; для текстів – трансформерні моделі або Naive Bayes при малих вибірках. Подальші дослідження доцільно спрямувати на автоматизований вибір і налаштування моделей (AutoML), методи навчання з малою кількістю прикладів (few-shot learning) та пояснювальний штучний інтелект (XAI).

Список використаних джерел

1. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 3rd ed. Sebastopol: O'Reilly Media, 2022. 861 p.

2. Murphy K. P. Probabilistic Machine Learning: An Introduction. Cambridge: MIT Press, 2022. 864 p.

3. Mitchell T. M. Machine Learning. New York: McGraw-Hill, 1997. 432 p.

4. Breiman L. Random Forests. Machine Learning. 2001. Vol. 45, No. 1. P. 5–32.