DOI: 10.32347/2412-9933.2021.46.173-180
УДК 004.021

Калініна Ірина Олександрівна
Кандидат технічних наук, доцент кафедри інтелектуальних інформаційних систем, *orcid.org/0000-0001-8359-2045*
Чорноморський національний університет імені Петра Могили, Миколаїв
Гожий Олександр Петрович
Доктор технічних наук, професор кафедри інтелектуальних інформаційних систем, *orcid.org/0000-0002-3517-580X*
Чорноморський національний університет імені Петра Могили, Миколаїв

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МЕТОДІВ КЛАСИФІКАЦІЇ ПРИ ПРОГНОЗУВАННІ В ЗАДАЧАХ МАШИННОГО НАВЧАННЯ

Анотація. *Розглянуто використання методів класифікації для вирішення завдання прогнозування аеродинамічних властивостей матеріалів. Запропоновано і досліджено методологію класифікації методами машинного навчання. Були використані такі методи класифікації: логістична регресія (LR), метод K-найближчих сусідів (KNN), дерева рішень (DT) та випадковий ліс (RF). Методологія складається з таких етапів: збирання даних, розвідувальний аналіз даних, моделювання, оцінювання ефективності моделей та підвищення ефективності моделей. Для реалізації процедури прогнозування проведено попереднє опрацювання даних, яке складається з етапів: збирання даних, розвідувальний аналіз даних. Наступний етап – Моделювання, складається з двох частин: підготовка та вибір моделі. Обрахована точність прогнозів. При аналізі були досліджені результати прогнозування з точки зору точності, як-от: відгук, F-міра, Каппа, значення робочої характеристики (ROC) та частоти помилок, вимірюваних середньою абсолютною помилкою (MAE) і середньоквадратичною помилкою (RMSE). Проведено аналіз точності прогнозування.*

Ключові слова: класифікація; прогнозування; машинне навчання; оцінювання якості прогнозів

Вступ

В галузі інтелектуального аналізу даних класифікація є найбільш відомим методом машинного навчання для вирішення завдань моделювання і прогнозування. Загалом, класифікація визначається як метод навчання, який відображає або класифікує екземпляри даних до відповідних міток класів, які обумовлені в певному наборі даних. Згідно [1], класифікація даних – це двоетапний процес; перший – етап навчання, на якому модель класифікації будується із заданого набору даних; дані, на основі яких вивчається функція або модель класифікації, відомі як навчальний набір, а другий – етап класифікації, на якому модель використовується для тестування або прогнозування міток класів для окремих невидимих даних. Набір даних, який використовується для перевірки класифікаційної здатності моделі або функції, визначається як набір для тестування.

Метою класифікації є точна класифікація міток класів активності елементів вибірки, чиї контекстні особливості або значення атрибутів відомі, але значення класів невідомі [1]. Ефективна класифікація при прогнозуванні параметрів аеродинамічних властивостей матеріалів з використанням методів машинного навчання є складним завданням, оскільки різні класифікатори показують різну якість результатів класифікації в різних контекстах. У такий спосіб необхідно дослідити результативність різних методів класифікації при опрацюванні вхідного набору даних методами машинного навчання для визначення найбільш ефективного. Щоб проаналізувати ефективність моделі класифікації на основі методів машинного навчання, було використано декілька найбільш ефективних методів класифікації [1; 2], а саме: логістична регресія (LR), лінійний дискримінантний аналіз (LDA), квадратичний дискримінантний аналіз (QDA), K-найближчих сусідів (KNN), дерева рішень (DT) та випадковий ліс (RF). Ці методи були розглянуті, оскільки вони є добре відомими класифікаторами в машинному навчанні і часто використовуються при вирішенні завдань порівняльної класифікації.

При побудови моделі на основі класифікатора машинного навчання ефективність кожної моделі вивчається шляхом проведення експериментів на реальному наборі даних. Як набір даних використано набір аеродинамічних властивостей матеріалів [3].

Мета статті

Мета – розробити методологію вирішення завдань класифікації при розв'язанні задач машинного навчання. Дослідити ефективність методів класифікації на прикладі прогнозування заданих аеродинамічних властивостей матеріалів засобами машинного навчання.

Виклад основного матеріалу

Системи класифікації відіграють важливу роль в задачах дослідження складних систем та процесів, класифікуючи доступну інформацію на основі деяких критеріїв. В цьому дослідженні необхідно оцінити відносну ефективність деяких добре відомих методів класифікації на складному наборі даних. Досліджено методи класифікації, засновані на статистичних методах і методах штучного інтелекту.

Для вирішення завдання класифікації була розроблена методологія використання методів машинного навчання, яка представлена у вигляді послідовності етапів на рис. 1.

Вирішення завдання класифікації складається з п’яти етапів.

На першому етапі здійснюється збирання даних, аналіз та їх інтерпретація. Завантажується вхідний набір даних, аналізується структура набору даних, визначаються ознаки та їх типи, а у разі потреби відбувається перекодування цих ознак. В результаті такого попереднього опрацювання первинного набору даних маємо підготовлений набір до наступного етапу – розвідувального аналізу даних.

На другому етапі здійснюються процедури розвідувального аналізу даних. Визначається тип описової статистики, виявляються нечислові та відсутні значення, здійснюється відбір ознак для подальшого моделювання, визначається взаємозв’язок між змінними, генерується матриця ознак і масив міток та відбувається нормалізація числових даних. У результаті цих перетворень отримуємо набір даних, які підготовлені для моделювання.

Третій етап – етап моделювання складається з двох частин: підготовка та вибір моделі. При підготовці відбувається поділення основного набору даних на навчальну (тренувальну) та тестову вибірку, визначається функція втрат і створюються набори для крос-перевірки. При виборі моделі перевіряються алгоритми моделювання на тренувальній вибірці та вибирається найкращий за певними критеріями.

Четвертий етап – визначення параметрів ефективності моделі. Ефективність моделей визначається за допомогою матриці неточності, кривої «точність – повнота», F-міри, Каппа, значення робочої характеристики (ROC) та частоти помилок, вимірюваних середньою абсолютною помилкою (MAE) і середньоквадратичною помилкою (RMSE).

На п’ятому етапі виконуються процедури з метою підвищення ефективності вибраної моделі класифікації. Залежно від вибраного на третьому етапі методу моделювання засобами покращення якості можуть бути: замість нормування числових значень стандартизація по z-оцінках, дослідження декількох варіантів K (при обранні KNN-моделі), додавання адаптивного підсилення, використання матриці штрафів або введення правил класифікації.

Процедури і методи, перелічені в етапах розробленої методології, безпосередньо зв’язані з процесом візуалізації. За допомогою візуалізації на кожному етапі є можливість швидко прийняти рішення з корегування послідовності дій та повернення на попередні етапи.

Важливо, що кожний з наведених етапів має певні особливості, які враховуються залежно від структури даних початкового набору, особливостей предметної галузі, для якої вирішуються завдання класифікації та засобу його реалізації.

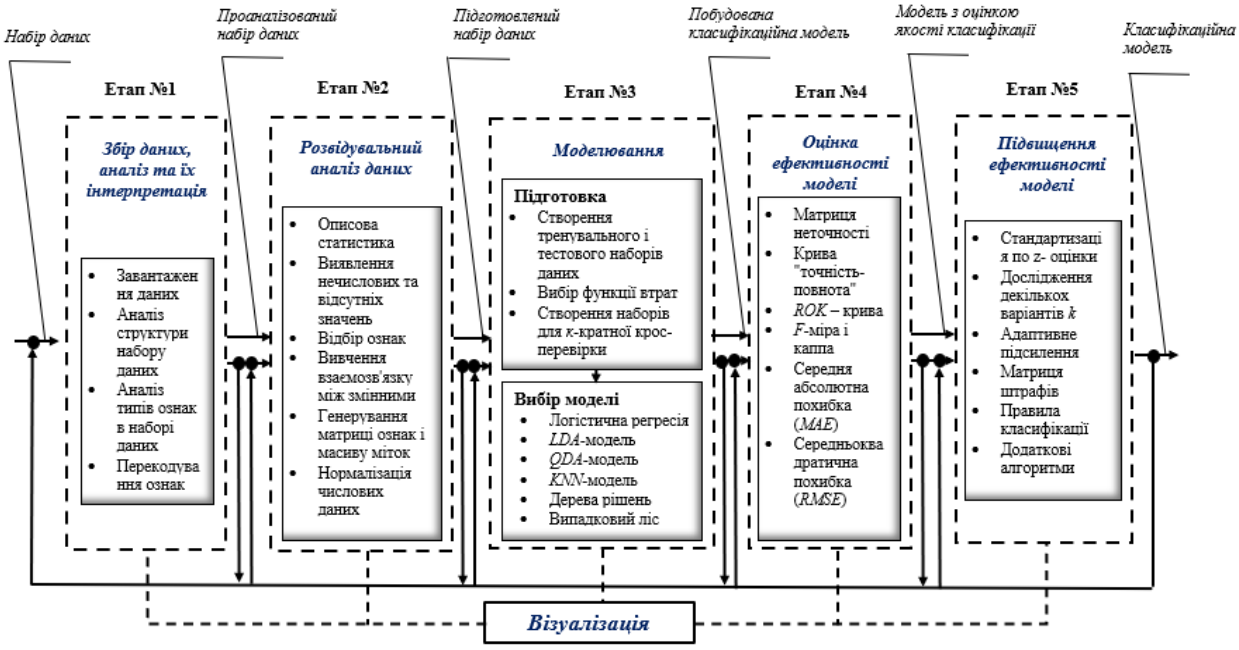

Рисунок 1 – Послідовність етапів вирішення завдань класифікації

У літературі з інтелектуального аналізу даних запропоновано багато алгоритмів класифікації, що дають змогу робити прогнози. Для вирішення завдання класифікації визначимо низку відомих методів класифікації, за допомогою яких ефективно вирішуються завдання класифікації.

K-найближчих сусідів (KNN)

Алгоритм класифікації, K-найближчих сусідів (KNN) [4] – один з найвідоміших методів класифікації в машинному навчанні. За цим методом класифікації враховується локальна апроксимація, тож і всі обчислення відкладаються до класифікації. Метод зберігає всі доступні спостереження в даному наборі даних і класифікує нові спостереження на основі функцій відстані, таких як евклідова відстань, міри схожості та інших. Після цього *K* найбільш порівнянних випадків, що називаються сусідами, контролюються шляхом пошуку по всьому набору підготовки для іншої точки тестової інформації. Отже, очікуваний результат виходить шляхом скорочення змінної прибутковості для тих *K* випадків, які залежать від більшої частини, що віддала більшість голосів сусідів.

Логістична регресія (LR)

Логістична регресія (LR) [5] – ще одна популярна імовірнісна статистична модель, використовувана для вирішення завдань класифікації. Зазвичай логістична регресія оцінює ймовірності за допомогою логістичної функції, яку також називають сигмоїдною функцією. Гіпотеза логістичної регресії має тенденцію обмежувати функцію між 0 і 1. Цей класифікатор вимірює взаємозв'язок між категоріальною залежною змінною і однією або декількома незалежними змінними для цього набору даних. Залежна змінна – це цільовий клас, який необхідно передбачити. Однак незалежні змінні – це атрибути або контекстні функції, які використовуються для прогнозування цільового класу.

Дерево рішень (DT)

Дерево рішень – це відомий і часто обговорюваний метод класифікації та подальшого використання для прогнозування [6]. Основний алгоритм побудови дерев рішень ID3, запропонований в роботі [7]. Алгоритм ID3 будує дерево рішень, використовуючи спадний підхід, за якого жадібний пошук по заданому набору навчальних даних використовується для перевірки кожного атрибута або контексту на кожному з вузлів. Він обчислює ентропію й інформаційний приріст, який є статистичною властивістю, що використовується для вибору, який атрибут перевіряти в кожному вузлі дерева [7]. На основі алгоритму ID3 Куінланом запропоновано розширення, а саме алгоритм C4.5 [8]. C4.5 будує дерева рішень з набору навчальних даних за тією ж процедурою, що і ID3. C5.0 – ще один модифікований алгоритм дерева рішень, використовуваний для прогнозного моделювання [9]. Дерево рішень C5.0 значно швидше і ефективніше з точки зору пам'яті за C4.5.

Випадковий ліс (RF)

Класифікатор випадкових лісів, запропонований в роботі [10] – це метод ансамблевого машинного навчання, який враховує декілька алгоритмів навчання разом при генерації результату прогнозування. Випадковий ліс поєднує в собі завантажувальну агрегацію (упаковку) [11] і випадковий вибір функцій [12], щоб побудувати набір дерев рішень, які демонструють керовану варіацію. Отже, випадковий ліс генерує кілька дерев рішень, а не одне дерево рішень.

Застосування розробленої методології класифікації

Для вирішення завдання аналізу ефективності класифікації було використано набір даних *Airfoil Self-Noise Data Set*, за допомогою якого досліджувались аеродинамічні властивості матеріалів [3]. Набір даних NASA включає аеродинамічні поверхні NACA 0012 різного розміру при різних швидкостях і кутах атаки в аеродинамічній трубі. Розмах профілю та положення спостерігача в усіх експериментах були однаковими. Набір складається з 1506 спостережень та 6 таких атрибутів:

1. Частота в герцах.
2. Кут атаки в градусах.
3. Довжина хорди в метрах.
4. Швидкість потоку, що набігає, в метрах в секунду.
5. Товщина витіснення на стороні всмоктування в метрах.

Результуюче значення:

6. Масштабований рівень звукового тиску в децибелах.

Для вирішення завдання бінарної класифікації створена додаткова бінарна змінна на додаток до залежної змінної. Вона приймає значення, що дорівнює 1, якщо залежна змінна перевищує власну медіану, і 0, якщо вона менше власної медіани.

Створена таблиця даних була досліджена графічно для встановлення зв'язку між бінарною міткою і предикторами. Для аналізу використані діаграми розсіювання і діаграми розмахів (рис. 3, 4). Всі обчислення та візуалізація проводилися в середовищі R [13]. На рис. 2 представлено структуру набора даних після додавання бінарної мітки.

На рис. 3 представлено матрицю розсіювання для числових ознак з набору даних, який досліджується.

```
 num [1:1503] 1 0 1 1 1 0 0 0 0 0 ...
'data.frame':   1503 obs. of  7 variables:
 $ V1        : int  800 1000 1250 1600 2000 2500 3150 4000 5000 6300 ...
 $ V2        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ V3        : num  0.305 0.305 0.305 0.305 0.305 ...
 $ V4        : num  71.3 71.3 71.3 71.3 71.3 71.3 71.3 71.3 71.3 71.3
 ...
 $ V5        : num  0.00266 0.00266 0.00266 0.00266 0.00266 ...
 $ V6        : num  126 125 126 128 127 ...
 $ binary_01: num  1 0 1 1 1 0 0 0 0 0 ...
> |
```

Рисунок 2 – Структура даних

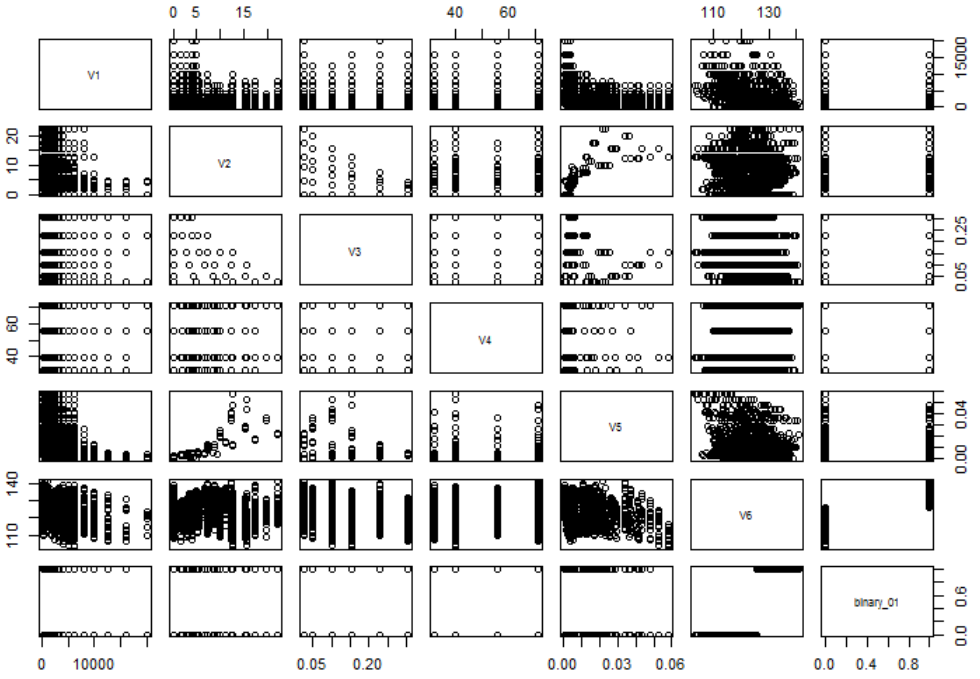

Рисунок 3 – Матриця розсіювання для числових ознак з набору даних

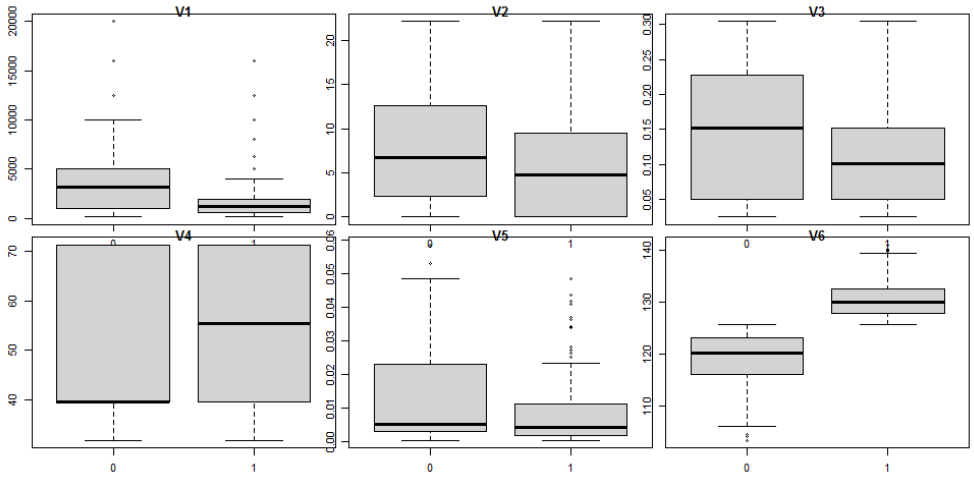

Рисунок 4 – Діаграми розмаху для ознак з набору даних

За допомогою діаграми розмаху досліджується вплив на результуюче значення (рис. 4). Найбільший вплив на результуюче значення має змінна V4.

На етапі моделювання спочатку відбулась підготовка до навчання моделі. Набір даних поділено на дві вибірки: для навчання моделі і для тестування. Для розподілу використана пропорція 70/30, спостереження відбиралися за допомогою генератора випадкових чисел.

На другому етапі моделювання обраховано оптимальні параметри для вибраної моделі класифікації. Як перша модель для вирішення

завдання класифікації вибрана логістична регресія. Розрахована матриця неточності і загальна частина правильно класифікованих випадків.

```
glm.prediction   0   1
             0 455 115
             1 142 490
[1] 0.7861897
```

Рисунок 5 – Матриця неточності по моделі логістичної регресії

Результат показує, що похибка (або рівень помилкової класифікації) (рис. 5) дорівнює 21,4%.

Наступною побудована LDA-модель на вибірці навчальних даних для передбачення бінарної мітки на основі змінних, які виглядали найбільш тісно пов'язаними з нею. Результат показує, що похибка дорівнює 21,4%. Перевірка на контрольній вибірці дала 24.3% похибок.

Експерименти з QDA-моделлю дали кращі результати класифікації порівняно з LDA-моделлю. Отже, оцінка якості показує, що у 79% випадків модель дає правильний результат (тобто 21% похибки). Результати на контрольній вибірці – 23.7% похибки.

KNN-модель потребує вибору значень *K* – кількості найближчих сусідів. Підгонка KNN-моделі виконана для різних значень *K* з метою обрання кращого значення параметра. В табл. 1 представлені результати дослідження якості класифікації KNN-моделі при різних значеннях *K*.

Таблиця 1 – ***Порівняльна таблиця якості класифікації для моделі типу KNN***

Тип моделі	Частота похибок	
	На навчальній вибірці, %	На контроль-ній вибірці, %
KNN-модель, K = K_1	40,2	40,0
KNN-модель, K = K_3	38,2	35,7
KNN-модель, K = K_5	32,9	32,9
KNN-модель, K = K_8	30,9	32,1
KNN-модель, K = K_{10}	29,9	30,6
KNN-модель, K = K_{25}	21,9	22,7
KNN-модель, K = K_{50}	18,5	20,6
KNN-модель, K = K_{100}	26,9	30,8
KNN-модель, K = K_{200}	30,0	33,2

Найкращий результат отримано для *K* = 50, похибка класифікації становить 18,5% та 20,6% відповідно на навчальній та контрольній вибірках.

Як ще одну альтернативну модель використано модель дерева рішень. Навчальні алгоритми на базі дерев рішень – це потужні класифікатори, в яких використовується деревоподібна структура для моделювання відносин між ознаками і можливими результатами. Алгоритм використовує структуру розгалужених рішень, по якій приклади перенаправляються до остаточного спрогнозованим значенням класу. Результат моделювання та візуалізація моделі представлені на рис. 6, 7 відповідно.

```
        Conditional inference tree with 19 terminal nodes

Response:  class
Inputs:  V1, V2, V3, V4
Number of observations:  1052

1) V1 <= 2500; criterion = 1, statistic = 105.542
  2) V2 <= 11.2; criterion = 1, statistic = 51.821
    3) V4 <= 39.6; criterion = 1, statistic = 29.681
      4) V3 <= 0.1524; criterion = 1, statistic = 15.937
        5)*  weights = 155
      4) V3 > 0.1524
        6) V1 <= 1250; criterion = 1, statistic = 16.02
          7)*  weights = 93
        6) V1 > 1250
          8)*  weights = 27
    3) V4 > 39.6
      9) V4 <= 55.5; criterion = 0.973, statistic = 7.299
        10)*  weights = 103
      9) V4 > 55.5
        11) V1 <= 1600; criterion = 0.967, statistic = 6.943
          12)*  weights = 95
        11) V1 > 1600
          13) V3 <= 0.1524; criterion = 0.998, statistic = 12.524
            14)*  weights = 26
          13) V3 > 0.1524
            15)*  weights = 7
  2) V2 > 11.2
    16) V4 <= 55.5; criterion = 0.981, statistic = 7.946
      17)*  weights = 112
    16) V4 > 55.5
      18)*  weights = 63
1) V1 > 2500
  19) V3 <= 0.0508; criterion = 1, statistic = 97.964
    20) V2 <= 12.7; criterion = 1, statistic = 40.072
      21) V1 <= 8000; criterion = 1, statistic = 26.379
        22) V2 <= 4.8; criterion = 1, statistic = 18.369
          23)*  weights = 53
        22) V2 > 4.8
          24) V1 <= 4000; criterion = 0.992, statistic = 9.453
            25)*  weights = 18
          24) V1 > 4000
            26)*  weights = 12
      21) V1 > 8000
        27)*  weights = 27
    20) V2 > 12.7
      28)*  weights = 26
  19) V3 > 0.0508
    29) V3 <= 0.1016; criterion = 0.999, statistic = 14.354
```

Рисунок 6 – Звіт побудови дерева рішень

Оцінка рівня помилкової класифікації на навчальній і тестовій вибірках дала такий результат. Отже, величина похибки на навчальній вибірці – 14,86%, а на тестовій – 20,5%, що значно краще результатів, отриманих на попередніх моделях.

Останньою була побудована модель випадкового лісу. Для побудови алгоритму випадкового лісу використано ансамблевий метод. Рівень помилок моделі при значенні параметра n_{tree} = 500 (кількість дерев за замовчуванням) дорівнює 14,6%. Алгоритм випадкового лісу допомагає побудувати графік помилок для дерев, який наведено на рис. 8. На графіку по осі *X* показано кількість дерев, а на осі *Y* – рівень помилок. Три різні криві показують помилку кожного класу і загальну частоту помилок за порядком спадання. Визначено, що при значенні n_{tree} = 200 величина похибки є оптимальною.

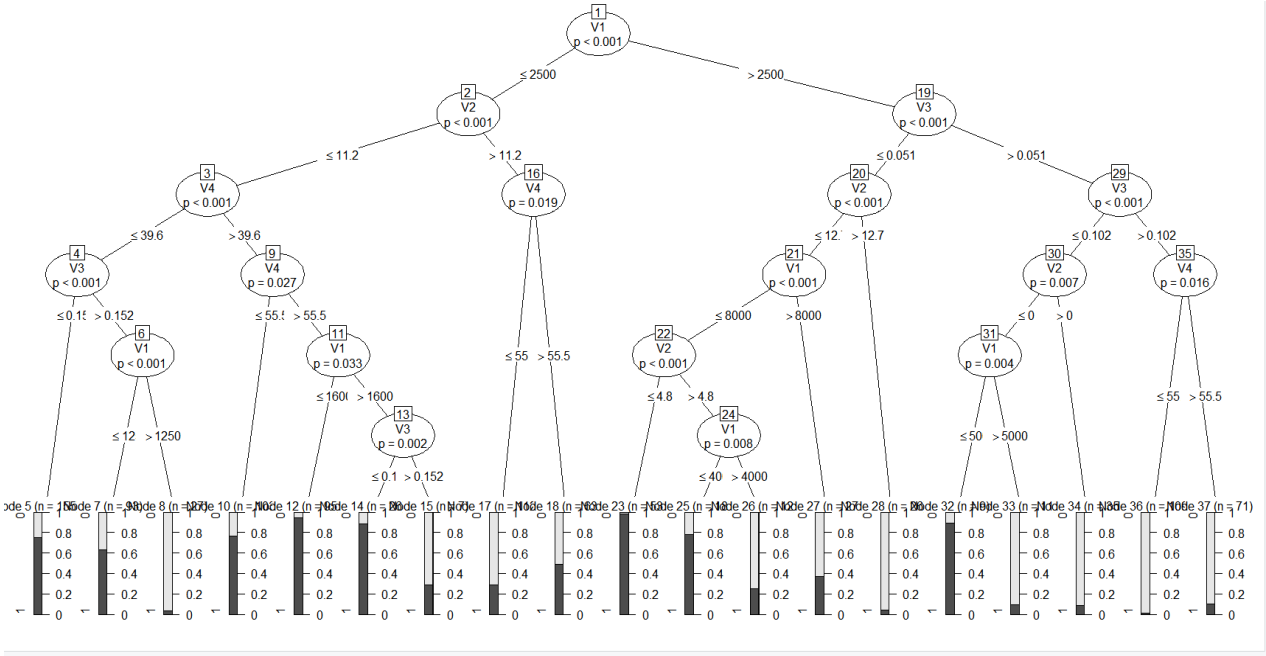

Рисунок 7 – Візуалізація моделі дерева рішень для класифікації

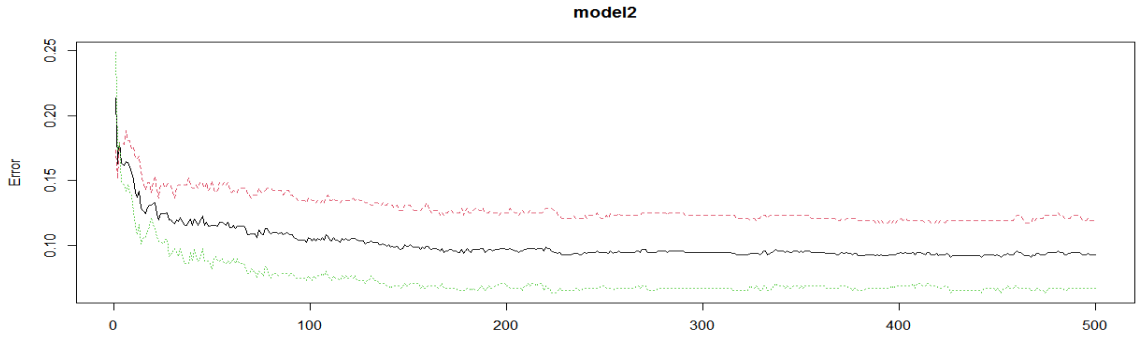

Рисунок 8 – Графік помилок для моделі випадкового лісу

Оцінювання точності моделей

Результати прогнозування були досліджені з точки зору точності, як-от: відгук, *F*-міра, Каппа, значення робочої характеристики (ROC) і частоти помилок, вимірюваних середньою абсолютною помилкою (MAE) і середньоквадратичною помилкою (RMSE).

Точність – це міра, яка являє собою співвідношення між кількістю правильно прогнозованих значень і загальною кількістю прогнозованих значень (як правильно, так і неправильно). Відгук – це міра повноти, яка являє собою співвідношення між кількістю правильно прогнозованих значень і загальною кількістю релевантних значень. Вони розраховуються з використанням значень істинно позитивної швидкості, непозитивної швидкості і помилково негативної швидкості в результатах прогнозування.

Робоча характеристика приймача (ROC) [14] може використовуватися як ще одна метрика, яка також включає частоту справжніх позитивних і помилкових спрацьовувань для оцінки якості вихідних даних класифікатора. Якщо *TP*, *FP* і *FN* позначають справжні спрацьовування, помилкові спрацьовування і помилкові заперечення, то формальне визначення точності і відкликання буде [14]:

$$\tau = \frac{TP}{TP+FP}, \qquad \upsilon = \frac{TP}{TP+FN},$$

де τ – точність; υ – відгук.

F-міра – це показник, який об'єднує точність і чутливість в єдиній оцінці, яка являє собою гармонічне середнє значення точності і чутливості. Каппа – це показник, який порівнює спостережувану точність з очікуваною точністю (випадковий шанс). Формальне визначення F-заходи і Каппи [14]:

$$K_{measure} = 2 * \frac{\tau * \upsilon}{\tau + \upsilon}, Kappa = \frac{(OA-EA)}{(1-EA)},$$

де *OA* (*observed accuracy*) – спостережувана точність; *EA* (*expected accuracy*) – очікувана точність.

Середня абсолютна помилка (MAE) і середньоквадратична помилка (RMSE) використовуються для розрахунку частоти помилок кожної моделі на основі класифікатора, яка являє точність прогнозування в завданнях машинного навчання [15; 16]. Якщо прогнозовані значення в тестових примірниках дорівнюють $p_1, p_2, \dots, p_n$, а фактичні значення $a_1, a_2, \dots, a_n$, для n точок даних MAE і RMSE формально визначені, як показано нижче

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - a_i|,$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - a_i)^2}.$$

Для кожної моделі використано один і той самий набір даних, щоб точно порівняти методи. Для оцінювання розробленої методології було використано десятикратну перехресну перевірку для вихідного набору даних. Метод 5-кратної перехресної перевірки розділяє підготовлений набір на 5 наборів розміром N/5. Після цього кожну модель навчають на чотирьох наборах і тестують на останньому наборі. Згідно з процедурою перехресної перевірки це повторюється п’ять разів. Як результат беруться середні результати прогнозу для кожної моделі.

У табл. 2 наведено результати прогнозування кожної моделі на основі класичної оцінки ефективності класифікатора з точки зору швидкості CCI (правильно класифіковані екземпляри), швидкості ICI (неправильно класифіковані екземпляри), швидкості середньої абсолютної помилки (MAE), середньоквадратичної помилки (RMSE) та значення робочої характеристики приймача (ROC).

Результати обчислювального експерименту в порівнянні методів прогнозування, наведені в табл. 2, свідчать про те, що правильно класифіковані екземпляри, отримані на основі моделі випадкового лісу (RF) і дерева рішень (DT), становлять 85,37% і 82,66% відповідно, що вище, ніж у інших моделей. Значення ROC, які представляють справжню позитивну частоту порівняно з помилковою позитивною частотою цих моделей класифікації на основі дерева, також дають результати, які кращі за інші моделі на основі класифікатора, наведені в табл. 2. На додаток до цих вимірів, деревоподібні моделі також дають нижчу частоту помилок з точки зору значень ICI, MAE і RMSE в задачах прогнозування аеродинамічних характеристик матеріалів.

Таблиця 2 – ***Підсумкова таблиця оцінок ефективності моделей на основі експериментів з набором даних***

Тип моделі	Результати				
	CCI (%)	ICI (%)	MAE	RMSE	ROC
Логістична регресія (LR)	79,77	20,22	0,11	0,25	0,93
K-найближчих сусідів (KNN)	79,43	20,56	0,14	0,27	0,91
Дерева рішень (DT)	82,66	17,34	1,10	0,23	0,94
Випадковий ліс (RF)	85,37	14,62	0,08	0,21	0,96

Висновки

Проведено аналіз методів класифікації на основі розробленої і дослідженої методології класифікації для завдань машинного навчання. Була розглянута задача прогнозування аеродинамічних властивостей матеріалів. Як методи класифікації в методології були використані такі: логістична регресія (LR), метод K-найближчих сусідів (KNN), дерева рішень (DT) та випадковий ліс (RF). Методологія складається з таких етапів: збирання даних, розвідувальний аналіз даних, моделювання, оцінювання ефективності моделей та підвищення ефективності моделей. Для реалізації процедури прогнозування проведене попереднє опрацювання даних. Обчислено показники якості прогнозів. Проведено порівняльний аналіз методів визначення якості прогнозів. Експерименти показали, що характеристики даних значно впливають на ефективність класифікації методів. Результати дослідження можуть допомогти в розробленні систем класифікації, в яких можна використовувати кілька методів класифікації для підвищення надійності і узгодженості класифікації.

Список літератури

1. Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. New York: Elsevier.
2. Witten, I. H., & Frank, E. (2005). Data mining: practical machine learning tools and techniques.
3. Machine Learning Repository. URL: http://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise.
4. Aha, D. W., Kibler, D. & Albert, M. K. (1991). Instance-based learning algorithms. *Mach Lear.,* 6(1), 37–66.

5. Le Cessie, S. & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *J R Stat Soc*., 41(1), 191–201.
6. Quinlan, J. R. (1993). C4.5: Programs for machine learning. *Machine Learning*.
7. Kuhn, M. & Johnson, K. (2013). Applied predictive modeling. New York: Springer; 26.
8. Breiman, L. (2001). Random forests. *Mach Learn*., 45(1), 5–32.
9. Breiman, L. (1996). Bagging predictors. *Mach Learn*., 24(2), 123–40.
10. Amit, Y. & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Comput*., 9 (7), 1545–1588.
11. Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G. & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with java implementations.
12. Lanz, B. (2020). Machine Learning in R: Expert Techniques for Predictive Analysis. SPb.: Peter, 464. ISBN: 978-5-4461-1512-9.
13. James, G., Whitton, D., Hasti, T., Tibshirani, R. (2017). Introduction to statistical learning with examples in R. DMK Press, Moscow, 456. ISBN: 978-5-97060-495-3.
14. Bidyuk, P., Gozhyj, A., Kalinina, I. & Vysotska, V. (2020). Methods for forecasting nonlinear non-stationary processes in machine learning. In: Data Stream Mining and Processing. DSMP 2020. Communications in Computer and Information Science. vol. 1158, pp. 470–485. Springer, Cham, (2020). https://doi.org/10.1007/978-3-030-61656-4 32.
15. Bidyuk, P., Kalinina, I. & Gozhyj, A. (2021). An Approach to Identifying and Filling Data Gaps in Machine Learning Procedures. International Scientific Conference "Intellectual Systems of Decision Making and Problem of Computational Intelligence" ISDMCI 2021: Lecture Notes in Computational Intelligence and Decision Making, pp. 164–176.

Received 24.05.2021

Kalinina Iryna
PhD (Eng.), PhD, Associate Professor of the Department of Intelligent Information Systems, *orcid.org/0000-0001-8359-2045*
Petro Mohyla Black Sea National University, Mykolaiv
Gozhyj Alexander
DSc (Eng.), prof., Professor of the Department of Intelligent Information Systems, *orcid.org/0000−0002−3517−580X*
Petro Mohyla Black Sea National University, Mykolaiv

STUDY OF THE EFFICIENCY OF CLASSIFICATION METHODS IN FORECASTING IN MACHINE LEARNING TASKS

Abstract. *The article considers the use of classification methods to solve the problem of predicting the aerodynamic properties of materials. The methodology of classification by methods of machine learning is offered and investigated. The following logistic regression (LR), K-nearest neighbors (KNN) method, decision trees (DT) and random forest (RF) were used as classification methods. The methodology consists of the following stages: data collection, exploratory data analysis, modeling, evaluation of model efficiency, and improving model efficiency. To implement the forecasting procedure, preliminary data processing was performed, which consists of stages: Data collection and Intelligence data analysis. The next stage – Modeling, consists of two parts: Preparation and Selection of the model. The accuracy of forecasts is calculated. The analysis examined the prediction results in terms of accuracy, such as response, F-measure, Kappa, performance value (ROC) and error rate measured by the mean absolute error (MAE) and the root mean square error (RMSE). The analysis of forecasting accuracy is carried out.*

Keywords: classification, forecasting, machine learning, quality assessment of forecasts

---

Link to the article

APA Kalinina, Iryna & Gozhyj, Alexander. (2021). Study of the efficiency of classification methods in forecasting in machine learning tasks. *Management of Development of Complex Systems*, 46, 173–180, dx.doi.org\10.32347/2412-9933.2021.46.173-180.

ДСТУ Калініна І. О., Гожий О. П. Дослідження ефективності методів класифікації при прогнозуванні в задачах машинного навчання. *Управління розвитком складних систем.* Київ, 2021. № 46. С. 173 – 180, dx.doi.org\10.32347/2412-9933.2021.46.173-180.