Biomedical Engineering and Technology
Issue 11(3), 20223 ISSN (online) 2707-8434

УДК 004.81 + 616-006

ОПТИМІЗАЦІЯ РЕЗУЛЬТАТІВ МОДЕЛЮВАННЯ ШЛЯХОМ РОЗБИТТЯ ВИБІРОК ЗА КРИТЕРІЄМ ПОДІБНОСТІ ВІДСТАНІ МАХАЛАНОБІСА

Гупало М. С.
hupalo.mykyta@lll.kpi.ua
Павлов В. А.,
pavlov.volodymyr@lll.kpi.ua
Настенко Є. А.,
nastenko.e@gmail.com
Корнієнко Г. А.
galinakor5555@gmail.com
Кафедра біомедичної кібернетики
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
м. Київ, Україна

Реферат: *Методи створення класифікаційних, апроксимаційних та прогностичних моделей включають процедури розділення робочого набору даних на технологічні частини, які використовуються для розрахунку параметрів, верифікації структури та завершальної оцінки якості моделей. Саме універсальність застосування даних процедур визначає виняткову актуальність вирішення завдання забезпечення ефективності розділення робочого набору даних на технологічні частини з точки зору якості результатів моделювання. Існуючі підходи до розбиття даних не забезпечують стійкої ефективності при побудові моделей (метод простого випадкового відбору, метод проб і помилок, тощо), або є ефективними, однак обмеженими використанням певних типів набору даних (зручний і систематичний відбори, тощо). В роботі для вирішення проблеми пропонується застосувати процедури розподілу даних на основі критерію подібності відстані Махаланобіса, що забезпечує збереження властивості відхилень об'єктів основної робочої вибірки для всіх технологічних підвибірок в умовах нерівних дисперсій змінних та корельованості простору ознак. Оскільки універсальним і найчастіше застосованим методом розбиття даних в даний час є метод випадкового відбору, у роботі саме з ним порівнюється ефективність пропонованого підходу. Аналіз підходів здійснений на даних для прогнозування рівня смертності від раку в округах США, що взяті з ресурсу data.world, та класифікації серцевої недостатності - з ресурсу Kaggle. Порівняння проведено для методів k-найближчих сусідів, логістичної регресії, методу групового урахування аргументів в завданні класифікації та методів k-найближчих сусідів, екстремального градієнтного підсилення (XGB), підвищення градієнта на основі алгоритму дерева рішень (LGBM) в задачі апроксимації. Результати аналізу показали перевагу пропонованого у роботі підходу розбиття даних відповідно до критерію подібності відстані Махаланобіса.*

Ключові слова *– розбиття даних, машинне навчання, відстань Махаланобіса, навчання з учителем, модель класифікації, апроксимаційна модель.*

I. ВСТУП

Всі сучасні технології моделювання даних від нейронних мереж до методів індуктивного моделювання неодмінно включають процедури розбиття робочої вибірки даних на технологічні підвибірки. Властивості одержаних підвибірок відображаються на властивостях одержаних моделей, якості вирішення завдань апроксимації, класифікації. Природною є постановка завдання оптимізації розбиття робочої вибірки на підвибірки з точки зору одержання кращого результату одержаних моделей на «свіжих даних», що можна ототожнювати з екзаменаційною (тестовою) вибіркою.

Існує ряд методів для розбиття даних на підвибірки серед яких найбільш відомим та

універсальним є метод простого випадкового відбору. Випадковий характер розподілу з рівномірною щільністю тим не менш може приводити до випадкових відхиленнях характеристик у вибраних наборах навчання та тестування, отже, немає гарантії, що отримуються репрезентативні набори навчальних даних.

Моделі машинного навчання вирішують багато різнорідних проблем: розпізнавання зображень, мовлення, прогнозування трафіку, рух на автопілоті, виявлення шахрайства в Інтернеті [1, 2], інтелектуальні мережеві програми [3], медична діагностика та інші [4, 5]. Їх використання тільки зростатиме і якість вирішення кожного з цих завдань в певній мірі залежить від ефективності вирішення завдання розподілу даних на підвибірки.

II. ПОСТАНОВКА ЗАДАЧІ

В роботі пропонується метод реалізації розподілу даних робочої вибірки на технологічні підвибірки, що враховує статистичні властивості простору ознак. Метод основано на критерії подібності відстані Махаланобіса задля відтворення схожості властивостей відхилень об'єктів основної вибірки до технологічних підвибірок - тестової, навчальної, валідаційної, тощо.

Метою роботи є розробка алгоритмів та програмного забезпечення, що розподіляє наданий набір даних на технологічні підвибірки на основі критерія подібності відстані Махаланобіса для вирішення задач класифікації та апроксимації.

Досягнення мети передбачає:

1. Обґрунтування та розробку методу розбиття даних відповідно критерію подібності відстані Махаланобіса.

2. Розробка моделей машинного навчання для порівняння точності роботи моделей після навчання внаслідок розбиття набору даних різними методами.

Результати виконаного дослідження дозволять використовувати розроблений метод розбиття даних для широкого класу методів створення моделей машинного навчання, підвищити якість вирішення завдань апроксимації, класифікації.

III. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

Існує ряд методів розбиття даних на підвибірки. Методи відбору умовно можна розділити на наступні категорії за принципами їх дії, мети, алгоритмічної та обчислювальної складності: простий випадковий відбір (Random Simple Sampling - RSS), методи проб і помилок, систематичний відбір, зручний відбір, стратифікована вибірка, самоорганізаційна карта Кохонена (SOM), підхід на основі генетичного алгоритму, алгоритм Кеннарда-Стоуна [6, 7].

Деякі з методів прості та широко використовуються, хоча вони страждають від високої дисперсії продуктивності моделі (RSS, методи проб і помилок). Інші методи є детермінованими та ефективними, однак обмежені певними типами наборів даних (зручний і систематичний відбори). Більш складні методи (наприклад, метод автоматизованого планування експерименту CADEX, удосконалений алгоритм автоматизованого планування експерименту від Рональда Сні DUPLEX) використовують структуру даних для отримання надійних результатів за рахунок більших обчислювальних витрат [8, 9].

Метод простого випадкового відбору - метод імовірнісного відбору, в якому кожний елемент генеральної сукупності має однакові шанси бути обраним до підвибірки. Даний метод є одним із самих популярних і простих в реалізації методів розбиття даних у галузях досліджень імовірностей, математики і статистики. Перевагами методу є відсутність упередженості, оскільки особини вибираються випадково, що у випадках великого набору даних створює більш-менш збалансовану підмножину, та простота використання. Недоліками методу вважають складності отримання максимально репрезентативної популяції, адже точний статистичний показник можна отримати лише з тієї вибірки, де представлений повний перелік досліджуваної сукупності, проте, виконання цієї умови є рідко можливою [10].

Алгоритм Кеннарда-Стоуна - алгоритм, спрямований на вибір репрезентативної підмножини з переліку N вибірок. Робота алгоритму починається з вибору пари точок,

які розташовані найдалі одна від одної. Вони відносяться до калібрувального набору і вилучаються зі списку точок. Потім процедура призначає решту точок до набору калібрування шляхом обчислення відстані між кожною непризначеною точкою та обраними точками, знаходячи точку із найменшою відстанню. Для обчислення використовується Евклідова відстань. Метод добре працює для лінійних моделей, в умовах однакової дисперсії по окремим змінним простору ознак, проте, метод має значну похибку при виборі елементів у просторах вищого порядку [6, 11, 12, 13].

Самоорганізаційні карти Кохонена (SOM) - це клас самоорганізованих штучних нейронних мереж, який представлений у вигляді масиву n-розмірних векторів. SOM виконує перетворення складного високовимірного вхідного простору у простіший низьковимірний дискретний вихідний простір шляхом збереження зв'язків у даних. SOM може генерувати розділення даних шляхом вивчення оптимального розподілу вагових векторів. Даний механізм формує основу стратифікованої вибірки на основі SOM, відбувається розділення даних на страти, з яких потім відбираються дослідження у підвибірки.

Недоліки методу: нейрони, розташовані поблизу на карті, можуть бути далеко в просторі ознак; немає загальноприйнятого емпіричного правила вибору деяких параметрів (розмір карти, сусідня функція), через використання Евклідової відстані погано працює на багатовимірних даних [14, 15].

IV. ОПИС КЛІНІЧНОГО МАТЕРІАЛУ

У ході дослідження пропонується використати 2 набори даних для перевірки роботи методу при вирішенні задач класифікації та апроксимації.

Перший набір даних характеризує пацієнтів з підозрою на серцеву недостатність, джерело даних - інтернет-ресурс Kaggle [16]. Набір даних містить 11 незалежних змінних: вік, стать, тип болю грудної клітини, артеріальний тиск у стані спокою, показник холестерину, рівень цукру в крові натще, результати електрокардіограми в стані спокою, максимальна частота серцевих скорочень, наявність/відсутність стенокардії фізичного навантаження, значення піку ST, нахил сегменту ST при навантаженні і цільова змінна: визначає наявність чи відсутність захворювання. Перед використанням набір даних піддається препроцесингу: категоріальні змінні закодовуються, для пропущенних значень змінної холестерина використовується середнє значення. Завдання бінарної класифікації - спрогнозувати можливу хворобу серця.

Другий набір даних характеризує параметри спільноти різних округів США та призначений для визначення рівня смертності від раку, джерело даних - інтернет-ресурс data.world [17]. Дані зібрані з ряду джерел, включаючи опитування американської спільноти (census.gov), веб-сайти clinicaltrials.gov і cancer.gov та агреговані у єдиний набір даних. Набір містить 3047 записів із 32 незалежними змінними та однією залежною. Серед незалежних змінних: середня кількість зареєстрованих випадків раку, середня кількість зареєстрованих смертей через рак, середній показник захворюваності від раку на душу населення, середній дохід на округ, населення округу, відсоток бідного населення та інші, залежна змінна - середня смертність від раку на 100 000 осіб населення. Завданням моделювання є створення апроксимаційної моделі для прогнозування рівня смертності від раку в округах США.

V. КРИТЕРІЙ ПОДІБНОСТІ

Концепції моделювання передбачають, що робоча вибірка повинна відображати властивості генеральної сукупності даних, які можуть зустрітися моделі в реальних умовах експлуатації. Тому основною проблемою при реалізації розбиття робочої вибірки на технологічні піднабори є збереження для них властивостей з робочої вибірки даних. Розглянуті у критичному аналізі підходи до розбиття вибірок базуються на застосуванні Евклідової відстані, що є природним при однаковій дисперсії змінних та некорельованості простору ознак [18, 19]. Отже, використання

при формуванні розподілу даних міри відстані, що враховує властивості (дисперсії та кореляції) змінних, що формують простір задачі моделювання, - природній резерв для підвищення якості моделей, що забезпечить незміщенність оцінок її параметрів і структури.

Алгоритм розбиття даних за критерієм подібності відстані Махаланобіса для вирішення задач апроксимації та класифікації має забезпечити збереження властивості відхилень об'єктів основної робочої вибірки для всіх технологічних підвибірок в умовах нерівних дисперсій змінних та корельованості простору ознак.

В якості міри подібності пропонується застосувати оцінку схожості гістограм розподілу даних між основною вибіркою та технологічними підвибірками. Для оцінки схожості гістограм застосуємо параметр - косинусну подібність. Гістограми представляються у вигляді векторів, тоді шуканий параметр формується, як косинус кута між двома векторами гістограм: скалярний добуток векторів, поділений на добуток їх довжин. Значення параметру «косинус-подібність» належить інтервалу [−1, 1]. Два пропорційні вектори мають косинусну подібність 1.

Надалі розрахуємо значення введеного вище критерію подібності для двох методів розподілу даних, ефективність яких оцінимо у розділі 7. Для розбиття даних застосуємо підхід, що запропоновано та описано у розділі 6 та найбільш часто застосований на практиці алгоритм простого випадкового відбору RSS. Для розрахунку застосуємо дані задачі прогнозу рівня смертності від раку в округах США, розглянемо розбиття на 2 підвибірки у співвідношенні 25/75 (25% - тестова вибірка, 75% - навчальна). Порівняємо схожість гістограми початкового (робочого) набору з тестовими та навчальними наборами, та тестових і навчальних наборів між собою. Для оцінки методу простого випадкового відбору розбиття набору даних виконується 100 разів, результатом для порівняння є середнє значення косинусної подібності за 100 ітерацій.

Результати оцінки схожості гістограм розподілу даних в результаті розбиття методом RSS та пропонованим детермінованим методом розподілу за критерієм подібності відстані Махаланобіса демонстровано в табл. 1, 2.

Таблиця 1. **Оцінка схожості гістограм RSS**

Порівнювані вибірки	Мін значення схожості	Макс значення схожості	Сер значення схожості
Робоча з тренувальною	0.7592	0.9999	0.9471
Робоча з тестовою	0.5757	0.9996	0.8348
Тестова з тренувальною	0.4079	0.9796	0.7594

Таблиця 2. **Оцінка схожості гістограм при детермінованому розподілу за критерієм подібності відстані Махаланобісу**

Порівнювані вибірки	Значення схожості
Робоча з тренувальною	0.9999
Робоча з тестовою	0.9201
Тестова з тренувальною	0.9203

Схожість гістограм розподілу даних внаслідок розбиття методом розподілу за критерієм подібності Махаланобіса показав суттєво вищий результат схожості отриманих вибірок.

Для порівняння отриманих гістограм наведемо вигляд гістограм розподілу початкової, навчальної та тестової підвибірок при випадковому розбитті та пропонованому у розділі 6 детермінованому методі розподілу за критерієм подібності відстані Махаланобіса (див. рис. 1, 2)

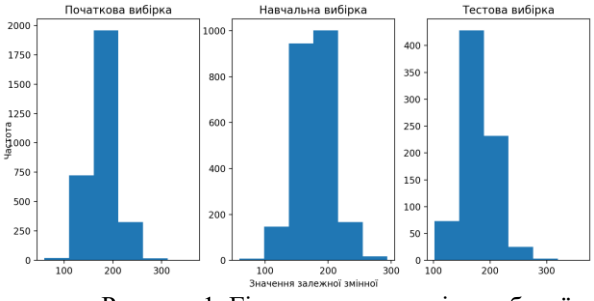

Рисунок 1. Гістограми розподілу робочої вибірки при простому випадковому розподілі

Алгоритм розбиття за критерієм подібності відстані Махаланобіса дозволяє досягти максимальної близькості гістограм вихідної вибірки та підвибірок, отриманих після розбиття.

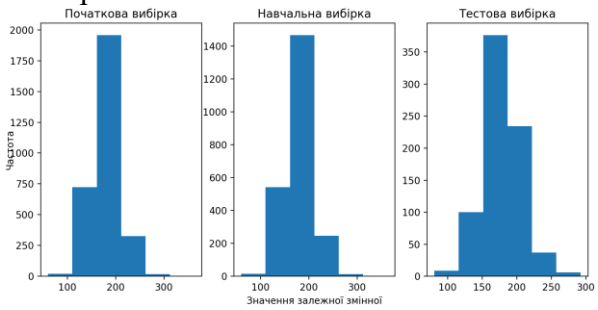

Рисунок 2. Гістограми розподілу робочої вибірки при детермінованому розподілі “за Махаланобісом”

Ефект досягається завдяки тому, що властивість подібності по критерію відстані Махаланобіса дотримується в кожній підвибірці, тому оцінки схожості розподілу даних з основної вибірки з технологічними підвибірками мають найкращі значення. В розділі 8 даний факт обгрунтує результати порівняння якості моделей апроксимації та класифікації одержаних при різних підходах до розбиття даних.

VI. АЛГОРИТМ РОЗБИТТЯ ВИБІРКИ ЗА ВІДСТАННЮ МАХАЛАНОБІСА

В роботі розроблено 2 версії застосування розбиття вибірки за критерієм подібності відстані Махаланобіса, версія для вирішення задачі апроксимації та версія для задачі класифікації. Надалі описана базова процедура розподілу даних на 2 підвибірки.

Кожна з версій містить 2 функціональні алгоритмічні модулі. Перший модуль (надалі - модуль А) представляє алгоритм розрахунку відхилення об’єктів. Алгоритм передбачає створення варіаційного ряду (ВР) набору даних на основі відстані Махаланобіса між кожним об’єктом вибірки та центроїдними значеннями вибірки у порядку збільшення відстані від значень центроїдів. Модуль приймає на вхід матрицю-об’єкт властивості X початкової вибірки із індексами кожного об’єкту у відповідності до початкового набору даних, розраховує значення центроїдів для кожної із незалежних змінних, коваріаційну матрицю (X^TX), для кожного спостереження розраховується відстань Махаланобіса до центроїду, створюється ВР на основі відхилення спостережень від центроїдних значень; створений ВР є виходом роботи модуля А.

Другий модуль (надалі - модуль Б) представляє алгоритм відбору спостережень у підвибірки, який має на увазі розподілення індексів варіаційного ряду до відповідних підвибірок. Алгоритм передбачає розбиття варіаційного ряду на страти і обрання досліджень із страт детермінованим або випадковим методом розподілу до підвибірок. Вхідні параметри алгоритму: вектор індексів варіаційного ряду, розмір меншої підвибірки та метод відбору досліджень зі страти (детермінований або випадковий).

Перший етап алгоритму - розбиття варіаційного ряду на страти складається з двох частин: визначення розміру мінімальної страти та визначення останньої страти. Розмір мінімальної страти визначається, як ціла частина співвідношення 100 відсотків до розміру меншої із двох підвибірок у відсотках (параметр алгоритму). Приклад: якщо модуль отримав параметр розміру меншої підвибірки як 30, тоді розмір мінімальної страти становитиме $\frac{100}{30} = 3{,}33$, тобто 3 елементи складають мінімальну страту. Мінімальна страта утворена фрагментом варіаційного ряду застосовується для реалізації елементарного акту розподілу даних у підвибірки та складається з направлення об’єкту страти до меншої підвибірки та направлення залишку страти (у прикладі - 2 об’єкти) до більшої підвибірки. Формування останньої страти визначається шляхом ділення кількісті елементів варіаційного ряду на розмір мінімальної страти. Якщо залишком є один елемент, він додається до останньої страти, в іншому випадку залишкові елементи формують нову останню страту. Розбиття на страти відбувається з початку варіаційного ряду.

Наступний етап алгоритму реалізує розподіл елементів варіаційного ряду зі страт у підвибірки. При детермінованому методі

відбору зі страт із кожної страти обирається один елемент по центру у меншу підвибірку, всі інші - в другу підвибірку. Якщо кількість елементів у страті є парною, центральним вважається елемент справа від середини. При випадковому методі відбору зі страт обирається один елемент випадковим чином у підвибірку меншості, всі інші елементи страти додаються до підвибірки більшості. На виході метод повертає індекси значень початкового набору даних, які потрапили у відповідні підвибірки.

При завданні розподілу даних більше ніж на дві підвибірки, алгоритм розбиває первинний набір даних на два піднабори: з найменшим розміром та залишковий піднабір. Далі відбувається розбиття залишкового піднабору на 2 піднабори аналогічним чином в залежності від початкових параметрів розмірів підвибірок алгоритму. Отже, при очікуваному розбитті на 4 підвибірки алгоритм буде реалізовувати розподіл 3 рази, результатом першого - отримаємо першу підвибірку, другого - другу підвибірку, третього - третю та четверту підвибірку.

Алгоритм розбиття даних за критерієм подібності відстані Махаланобіса для вирішення задачі класифікації на вхід приймає: матрицю об’єкт-властивості X, вектор значень *Y*, метод обрання елемента зі страти (детермінований, випадковий), параметр розмірів підвибірок (допустиме розбиття до 4-ох підвибірок).

Етапи роботи алгоритму розбиття даних для завдання класифікації: дослідження групуються за унікальними мітками класів вектора *Y* - для кожної із груп з матриці X відбирається відповідний набір об’єктів; індексована матриця X передається на вхід модулю А, де створюється ВР об’єктів відповідного класу на основі відхилення відстані Махаланобіса поточного об’єкту від центроїду класу; отриманий ряд передається на вхід модуля Б, де відбувається розбиття об’єктів на страти та їх обрання у відповідні підвибірки. Дана процедура виконується для кожної групи об’єктів, відповідні групові підвибірки об’єднуються у спільну підвибірку класів. При очікуваному розбитті більше, ніж на 2 підвибірки, виконуються аналогічні кроки, що наведені в алгоритмі апроксимації. На виході алгоритм повертає масиви індексів розрахованих підвибірок у відповідності до індексів значень з початкового набору даних.

VII. РЕЗУЛЬТАТИ

орівняння ефективності методу RSS та розподілу даних на основі критерія подібності відстані Махаланобіса (детермінований та випадковий підхід) для завдань класифікації проведем на наборі даних, що характеризує пацієнтів з підозрою на серцеву недостатність [16]. Будемо розбивати обраний набір даних на дві підвибірки (навчальна, тестова) у співвідношенні 70 до 30, виконувати навчання класифікаційних моделей на навчальній вибірці та порівнювати отримані результати точності класифікації моделі на тестових даних. Для забезпечення стійкості оцінки результату для версій підходів із застосуванням випадкової складової розбиття виконується n=300 разів і порівнюються отримані мінімальні, максимальні та середні значення точності за n ітерацій.. Точність класифікації розраховується як середня точність класифікації у класах на тестовому наборі даних.

Нижче наведено (табл. 3) порівняння результатів класифікації методом k-найближчих сусідів із бібліотеки sklearn, вказано параметр k кількості сусідів - 3.

Таблиця 3. **Точність KNN класифікації**

Метод розбиття	Мін точність	Макс точність	Сер точність
Випадковий	78.26	88.41	82.824
Пропонований випадковий	84.97	85.95	85.36
Пропонований детермінований	-	-	85.95

Далі наведено результати порівняння точності класифікації методом логістичної регресії бібліотеки sklearn (параметр штрафу L2). Умови перевірки аналогічні до попереднього прикладу. Результати класифікації логістичної регресії на наборі даних серцевої недостатності продемонстровано в табл. 4.

Таблиця 4. **Точність класифікації методом логістичної регресії**

Метод розбиття	Мін точність	Макс точність	Сер точність
Випадковий	81.16	90.94	85.673
Пропонований випадковий	85.95	86.93	86.311
Пропонований детерміновани й	-	-	86.27

Третім методом класифікації використано метод групового урахування аргументів МГУА із параметром функції активації - лінійна коваріація, параметром мінімальної кількості шарів - 5. З урахування необхідності розбиття вибірки для цього підходу на 3 технологічні складові виконується розбиття первинного набору даних у співвідношенні 60-20-20 (навчальна, валідаційна, тестова підвибірки).

Результати класифікації серцевої недостатності із використанням алгоритму МГУА продемонстровано в табл. 5.

Таблиця 5. **Точність класифікації МГУА**

Метод розбиття	Мін точність	Макс точність	Сер точність
Випадковий	76.63	90.76	84.7394
Пропонований випадковий	81.52	90.22	86.8103
Пропонований детерміновани й	-	-	86.96

Оцінку ефективності розбиття даних при вирішенні задачі апроксимації реалізуємо на наборі даних хворих на рак[17]. Задача - спрогнозувати рівень смертності населення від раку в округах США.

Первинний датасет розбитий на навчальну та тестову підвибірки у співвідношенні 75- 25 (навчальна, тестова). Для RSS та пропонованої версії розбиття за критерієм подібності відстані Махаланобіса із застосуванням випадкового способу відбору об’єктів з мінімальної страти реалізується 300 ітерацій розбиття даних, навчання та оцінки точності роботи моделі, для оцінки точності застосовано середню абсолютну похибку (Mean Absolute Error) - MAE.

Результати порівняння точності роботи моделей методів «k найближчих сусідів» (KNN), метод екстримального підсилення градієнта (XGB) та метод підсилення градієнта на основі алгоритму дерева рішень (LGBM) внаслідок навчання на піднаборах, отриманих після розбиття продемонстровано в табл. 6-8.

Таблиця 6. **Точність KNN регресії**

Метод розбиття	Мін MAE	Макс MAE	Сер MAE
Випадковий	17.344	19.998	18.555
Пропонований випадковий	17.225	19.677	18.580
Розроблюваний детермінований	-	-	17.840

Таблиця 7. **Точність XGB регресії**

Метод розбиття	Мін MAE	Макс MAE	Сер MAE
Випадковий	14.589	16.846	15.655
Пропонований випадковий	14.544	16.878	15.642
Пропонований детермінований	-	-	14.929

Таблиця 8. **Точність LGBM регресії**

Метод розбиття	Мін MAE	Макс MAE	Сер MAE
Випадковий	13.805	15.978	14.842
Пропонований випадковий	13.814	16.046	14.830
Пропонований детермінований	-	-	14.333

VIII. ОБГОВОРЕННЯ

Порівнюючи результати точності класифікаційних KNN, логістичної регресії, МГУА моделей, отриманих внаслідок навчання на піднаборах даних після розбиття запропонованим методом розподілу даних на основі критерія подібності відстані Махаланобіса та методом RSS, мінімальна та середня точності класифікації об’єктів за Махаланобісом для кожної із моделей машинного навчання виявилися вищими за точність класифікації об’єктів, використовуючи метод RSS (див табл. 3-5).

Результати задачі апроксимації на моделях KNN, XGB, LGBM внаслідок навчання після розбиття запропонованим методом та методом RSS показали, що показник відхилення MAE при використанні детермінованого методу відбору зі страт за Махаланобісом для кожної із моделей машинного навчання надає менше відхилення MAE ніж методи RSS та випадковий відбір зі страт за Махаланобісом, метод розбиття RSS та випадковий відбір зі страт за Махаланобісом показали однакові значення показника MAE (див. табл. 6-8).

Кращі результати, як при вирішенні задачі класифікації, так і задачі апроксимації можна пояснити тим, що алгоритм розбиття даних за критерієм подібності відстані Махаланобіса дозволяє досягти максимальної близькості гістограм відхилень об'єктів вихідної вибірки та підвибірок, отриманих після розбиття, що було продемонстровано у розділі 5. Таким чином, досягається ефект проекції властивості відхилень об'єктів з основної вибірки на технологічні підвибірки і, як результат, модель навчається на більш репрезентативних даних, що забезпечує мінімальні зміщення оцінок параметрів та структури моделей.

IX. ВИСНОВКИ

Критичний огляд робіт-аналогів дозволив запропонувати шляхи до вдосконалення методів розбиття вибірки на технологічні підвибірки. Розроблено алгоритм розбиття даних на основі критерія подібності відстані Махаланобіса для вирішення задач апроксимації та класифікації. Створено програмне забезпечення на мові програмування Python, який реалізовує створений алгоритм і дозволяє застосовувати його на довільних наборах даних.

Пропонований алгоритм було протестовано при вирішенні задачі класифікації на наборі даних пацієнтів з серцевою недостатністю. Одержано класифікаційні моделі k-найближчих сусідів, логістичної регресії, МГУА. Результати точності передбачень внаслідок навчання моделі на даних, отриманих після розбиття запропонованим методом показали кращі результати за навчанням моделей після розбиття методом простого випадкового відбору. Тест розробленого алгоритму (детермінований відбір) при вирішенні задачі прогнозування смертності від раку в округах США (одержано моделі k-найближчих сусідів, XGBoost, LGBM) також показав кращі результати ніж метод RSS.

Фінансування. Дане дослідження не отримувало зовнішнього фінансування.

Конфлікт інтересів. Автори заявляють про відсутність конфлікту інтересів.

ORCID ID та внесок авторів:

0009-0005-4203-0122 (A,B,C,D) Mykyta Hupalo

0000-0003-2104-5745 (B,D) Galina Korniienko

0000-0002-3293-5308 (A,E) Volodymyr Pavlov

0000-0002-1076-9337 (A,F) Ievgen Nastenko

A – Концепція роботи та дизайн, B – аналіз даних, C – Проєктування програмного коду, D – Написання статті, E – Критичний огляд, F – Остаточне схвалення статті.

ПЕРЕЛІК ПОСИЛАНЬ

1. Mohammed M., Khan M. B., Bashier E. B. M. Machine Learning Algorithms and Applications. Boca Raton : CRC Press, 2016. 226 p.
2. J. Sen, Ed. Machine Learning - Algorithms, Models and Applications. IntechOpen, 2021. 154 p.
3. Intelligent Wireless Communications / M. H. Alsharif et al. : IET, 2021. 453 p.
4. Clinical Pharmacology & Therapeutics / S. Badillo et al. : Wiley Periodicals, Inc., 2020. 1033 p.
5. Shen D., Wu G., Suk H. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017. 221-248.
6. Elfil M, Negida A. Sampling methods in Clinical Research. Emerg (Tehran) : Epub. 2017.
7. Mostafa S., Ahmad I. A. Recent Developments in Systematic Sampling. Journal of Statistical Theory and Practice. 2017. Vol. 12.
8. Roshan J.V., Akhil V. Split: An Optimal Method for Data Splitting. Technometrics. 2021. Vol. 64. 1-23 p.
9. Muraina I. Ideal Dataset Splitting Ratios In Machine Laerning Algorithms. Conference: 7th INTERNATIONAL MARDIN ARTUKLU SCIENTIFIC RESEARCHES CONFERENCE. 2022. Turkey, Mardin.
10. Noor S., Tajik O., Golzar, J. Simple Random Sampling. IJELS. 2022. Vol. 1. 78-82 p.

11. Harrington P. B. Multiple Versus Single Set Validation of Multivariate Models to Avoid Mistakes. Critical Reviews in Analytical Chemistry. 2017. Vol. 48.
12. Roshan J.V. Statystical analysis and data mining. 2022. Vol. 15. Iss. 4. 409-538.
13. Saptoro A., Tade M. O. A Modified Kennard-Stone Algorithm for Optimal Division of Data for Developing Artificial Neural Network Models. Chemical Product and Process Modeling. 2012. Vol. 7.
14. Umut A., Secil E. An Introduction to Self-Organizing Maps. ATLANTISCIS. 2012. Vol. 6.
15. May R.J., Maier H.R., Dandy G.G. Data splitting for artificial neural networks using SOM-based stratified sampling. Neural Networks. 2010. Vol. 23. 283-294 p.
16. Data.world. OLS Regression Challenge. URL: https://data.world/nrippner/ols-regression-challenge (дата звернення: 15.05.2023).
17. Kaggle. Heart Failure Prediction Dataset. URL: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction (дата звернення: 15.05.2023).
18. Hyndman R.J: Distance Measures. International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg. 397-398 p.
19. Liberti L., Lavor C., Maculan N. Eclidean Distance Geometry and Applications. 2012. SIAM Review. 56.

REFERENCES

1. M M. Mohammed, Muhammad Badruddin Khan, and M. Bashier, Machine learning: algorithms and applications. Boca Raton: CRC Press, 2017. ISBN: 9781498705387.
2. J. Sen and Sidra Mehtab, Machine learning : algorithms, models and applications. London: Intechopen, 2021. ISBN: 9781839694844.
3. G. Mastorakis, C. X. Mavromoustakis, J. M. Batalla, and E. Pallis, Eds., Intelligent Wireless Communications. Stevenage, England: Institution of Engineering and Technology, 2021. ISBN: 9781839530951.
4. S. Badillo et al., An Introduction to Machine Learning, Clinical Pharmacology & Therapeutics, vol. 107, no. 4, pp. 871–885, Mar. 2020, doi: https://doi.org/10.1002/cpt.1796.
5. S. K. Zhou, H. Greenspan, and D. Shen, Deep learning for medical image analysis. London, United Kingdom: Academic Press is an imprint of Elsevier, 2017, pp. 221-248. ISBN: 9780128104088.
6. M. Elfil and A. Negida, "Sampling Methods in Clinical Research; an Educational Review," Emergency, vol. 5, no. 1, Dec. 2016, doi: 10.22037/emergency.v5i1.15215.
7. S. A. Mostafa and I. A. Ahmad, "Recent developments in systematic sampling: A review," Journal of Statistical Theory and Practice, vol. 12, no. 2. Springer Science and Business Media LLC, pp. 290–310, Aug. 11, 2017. doi: 10.1080/15598608.2017.1353456.
8. V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," arXiv, 2020, doi: 10.48550/ARXIV.2012.10945.
9. V. R. Joseph, "Optimal ratio for data splitting," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 15, no. 4. Wiley, pp. 531–538, Apr. 04, 2022. doi: 10.1002/sam.11583..
10. S. Noor, O. Tajik, and J. Golzar, "Simple Random Sampling," IJELS, vol. 1, no. 2, Dec. 2022, doi: 10.22034/ijels.2022.162982.
11. P. de B. Harrington, "Multiple Versus Single Set Validation of Multivariate Models to Avoid Mistakes," Critical Reviews in Analytical Chemistry, vol. 48, no. 1. Informa UK Limited, pp. 33–46, Oct. 25, 2017. doi: 10.1080/10408347.2017.1361314.
12. A. Vakayil and V. R. Joseph, "Data Twinning," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 15, no. 5. Wiley, pp. 598–610, Feb. 15, 2022. doi: 10.1002/sam.11574.
13. A. Saptoro, M. O. Tadé, and H. Vuthaluru, "A Modified Kennard-Stone Algorithm for Optimal Division of Data for Developing Artificial Neural Network Models," Chemical Product and Process Modeling, vol. 7, no. 1. Walter de Gruyter GmbH, Jul. 31, 2012. doi: 10.1515/1934-2659.1645.
14. U. Asan and S. Ercan, "An Introduction to Self-Organizing Maps," Atlantis Computational Intelligence Systems. Atlantis Press, pp. 295–315, 2012. doi: 10.2991/978-94-91216-77-0_14.
15. R. J. May, H. R. Maier, and G. C. Dandy, "Data splitting for artificial neural networks using SOM-based stratified sampling," Neural Networks, vol. 23, no. 2. Elsevier BV, pp. 283–294, Mar. 2010. doi: 10.1016/j.neunet.2009.11.009.
16. OLS regression challenge (2022) data.world. Available at: https://data.world/nrippner/ols-regression-challenge (Accessed: 15 May 2023).
17. Heart failure prediction dataset (2021) Kaggle. Available at: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction (Accessed: 15 May 2023).
18. B. D. Basic, "Distance Measures," International Encyclopedia of Statistical Science. Springer Berlin Heidelberg, pp. 397–398, 2011. doi: 10.1007/978-3-642-04898-2_626.
19. L. Liberti, C. Lavor, N. Maculan, and A. Mucherino, "Euclidean Distance Geometry and Applications," SIAM Review, vol. 56, no. 1. Society for Industrial & Applied Mathematics (SIAM), pp. 3–69, Jan. 2014. doi: 10.1137/1208759

Biomedical Engineering and Technology
Issue 11(3), 20223 ISSN (online) 2707-8434

UDC 004.81 + 616-006

MODELING RESULTS OPTIMIZATION BASED ON DATA SPLITTING BY MAHALANOBIS DISTANCE SIMILARITY CRITERION

Mykyta Hupalo
hupalo.mykyta@lll.kpi.ua
Volodymyr Pavlov
pavlov.volodymyr@lll.kpi.ua
Ievgen Nastenko
nastenko.e@gmail.com
Galina Korniienko
galinakor5555@gmail.com
Department of Biomedical Cybernetics
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine

Abstract *- The methods of creating classification, approximation, and predictive models include procedures for dividing the initial data set into subsamples, which are used for parameter calculation, structure verification, and the final assessment of model quality. These procedures determine the exceptional urgency of solving the task of ensuring the efficiency of dividing the data set into subsamples that give a high-quality modeling result. Some of the existing data splitting approaches do not provide consistent performance in model building (simple random sampling, trial and error, etc.) or are effective but limited to specific types of data (convenience and systematic sampling, etc.). In order to solve the problem, it is proposed to implement a data distribution procedure based on the Mahalanobis distance similarity criterion, which ensures the preservation of the property of main working sample objects by minimizing deviations between the main working sample and all technological subsamples considering conditions of unequal variables, variances, and feature space correlation. Since the universal and most commonly used data partitioning method at the moment is the random selection method, the effectiveness of the proposed approach is compared with it. The analysis and comparison of the methods were carried out on data sets for predicting cancer mortality rates in US counties taken from the data world resource and a heart failure classification data set taken from the Kaggle resource. Comparisons are made using k-nearest neighbors, logistic regression methods, and group methods of data handling in the classification task and k-nearest neighbor, extreme gradient boosting (XGB), and gradient boosting based on the decision tree algorithm (LGBM) methods in the approximation task. The results showed the advantage of the data division approach according to the Mahalanobis distance similarity criterion.*

Key words – *data partitioning, machine learning, Mahalanobis distance, tutored learning, classification model, approximation model.*