DOI 10.31891/2307-5732-2022-315-6-224-233
УДК 004.832.25

СТЕБЕЛЕЦЬКИЙ Мирослав
Хмельницький національний університет
https://orcid.org/0000-0001-7084-8400
e-mail: myroslav.stebeletskyi@gmail.com
МАНЗЮК Едуард
Хмельницький національний університет
https://orcid.org/0000-0002-7310-2126
e-mail: eduard.em.km@gmail.com
СКРИПНИК Тетяна
Хмельницький національний університет
https://orcid.org/0000-0002-8531-5348
e-mail: tkskripnik1970@gmail.com
БАГРІЙ Руслан
Хмельницький національний університет
https://orcid.org/0000-0001-5219-1185
e-mail: bahriiro@khmnu.edu.ua

МЕТОД ПОБУДОВИ АНСАМБЛІВ МОДЕЛЕЙ ДЛЯ КЛАСИФІКАЦІЇ ДАНИХ НА ОСНОВІ КОРЕЛЯЦІЙНИХ ЗВ'ЯЗКІВ РІШЕНЬ

У науковій роботі висвітлюється проблема підвищення точності передбачень бінарної класифікації із використанням алгоритмів машинного навчання. Основою інформаційної системи бінарної класифікації виступає ансамблева модель. Ця модель, в свою чергу, містить набір унікальних комбінацій базових класифікаторів – свого роду алгоритмічні примітиви. Ансамблева модель може розглядатись як деякий мета-алгоритм, який складається із унікальних наборів алгоритмів класифікації машинного навчання (ML). Завданням ансамблевої моделі являється знаходження такої комбінації базових алгоритмів класифікації, яка б давала найвищі показники результативності. Результативність оцінюється згідно з основними метриками ML у завданнях класифікації. Іншим аспектом наукової роботи є створення агрегаційного механізму задля поєднання результатів базових алгоритмів класифікації. Тобто кожна унікальна комбінація у середині ансамблю складається із набору базових моделей (передвісників), результати яких потрібно агрегувати. У даній роботі задля агрегування (усереднення) передбачень базових моделей використовується неієрархічний метод кластеризації. Особливістю цього дослідження є знаходження коефіцієнтів кореляцій базових моделей у кожній комбінації. За допомогою величини кореляцій встановлюється залежність між передбаченням класифікатора (базова модель) та істинним значенням, в результаті чого відкривається простір для подальших досліджень щодо покращення ансамблевої моделі (мета-алгоритму).

Ключові слова: підвищення точності, бінарна класифікація, ансамблева модель, інформаційна система, машинне навчання, коефіцієнт кореляції, унікальна комбінація, модель, алгоритм класифікації.

STEBELETSKYI Myroslav, MANZIUK Eduard, SKRYPNYK Tetyana, BAHRIY Ruslan
Khmelnytskyi National University

METHOD OF BUILDING ENSEMBLES OF MODELS FOR DATA CLASSIFICATION BASED ON DECISION CORRELATIONS

The scientific work highlights the problem of increasing the accuracy of binary classification predictions using machine learning algorithms. Over the past few decades, systems that consist of many machine learning algorithms, also called ensemble models, have received increasing attention in the computational intelligence and machine learning community. This attention is well deserved, as ensemble systems have proven to be very effective and extremely versatile in a wide range of problem domains and real-world applications. One algorithm may not make a perfect prediction for a particular data set. Machine learning algorithms have their limitations, so creating a model with high accuracy is a difficult task. If you create and combine several models by combining and aggregating the results of each model, there is a chance to improve the overall accuracy, this problem is dealt with by ensembling. The basis of the information system of binary classification is the ensemble model. This model, in turn, contains a set of unique combinations of basic classifiers - a kind of algorithmic primitives. An ensemble model can be considered as some kind of meta-algorithm, which consists of unique sets of machine learning (ML) classification algorithms. The task of the ensemble model is to find such a combination of basic classification algorithms that would give the highest performance. The performance is evaluated according to the main ML metrics in classification tasks.

Another aspect of scientific work is the creation of an aggregation mechanism for combining the results of basic classification algorithms. That is, each unique combination within the ensemble consists of a set of basic models (harbingers), the results of which must be aggregated. In this work, a non-hierarchical clustering method is used to aggregate (average) the predictions of the base models. A feature of this study is to find the correlation coefficients of the base models in each combination. With the help of the magnitude of correlations, the relationship between the prediction of the classifier (base model) and the true value is established, as a result of which space is opened for further research on improving the ensemble model (meta-algorithm)

Keywords: accuracy improvement, binary classification, ensemble model, information system, machine learning, correlation coefficient, unique combination, model, classification algorithm.

Мета роботи. Постановка завдання

Метою наукової роботи є розробка метода побудови ансамблів моделей для класифікації даних на основі кореляційних зв'язків рішень. Використовується агрегаційний алгоритм, який усереднює показники передбачень базових моделей у кожній унікальній ансамблевій комбінації. Додатковим завданням являється

вирахування кореляційних показників між результатами алгоритмів класифікації та істинними значеннями, ці показники повинні вираховуватись для кожної унікальної комбінації моделей у ансамблі.

Задля наглядності результативності наукової роботи, слід реалізувати систему візуалізації показників основних метрик алгоритмів. Візуалізація показників – важлива частина роботи, оскільки на базі цього можна наочно оцінити результати моделей, провести аналіз задля подальшого дослідження.

Виходячи із поставленої мети, поставленою задачею буде розробка інформаційної системи бінарної класифікації, яка буде мати наступні особливості:

- система реалізована модульно, кожен етап дослідження (обробка тексту, завантаження базових класифікаторів, складання унікальних комбінацій, усереднення результатів моделей, візуалізація даних) повинен представляти собою окремий програмний модуль;
- можливість заміни базових алгоритмів класифікації;
- можливість заміни функції агрегатора моделей;
- графічна візуалізація результативності базових класифікаторів та ансамблю задля порівняння;
- графічна візуалізація кореляцій між істинними значеннями та передбаченнями ансамблевих комбінацій у вигляді набору гістограм в двовимірному та тривимірному просторах.
- В ході дослідження потрібно виконати наступні завдання:
- провести аналіз сучасних ансамблевих підходів;
- порівняти застосування відомих методів у предметній області на показнику ефективності;
- дослідити ефективність запропонованих рішень методом експериментальних досліджень на відомих корпусах даних.

Особливу увагу слід зазначити на модульності системи класифікації, оскільки така архітектура додатку сприятиме на продуктивність дослідження, адже система дозволятиме виконувати конфігурацію базових моделей (алгоритмів класифікації), можлива заміна функції агрегатора, модулі візуалізації та оцінювання теж підлягають заміні при необхідності.

Складання мета-алгоритму із використанням наборів унікальних комбінацій базових алгоритмів класифікації. Створення ансамблю

Завданням дослідження є покращення результатів мета-алгоритму машинного навчання у задачах класифікації, тому детальне висвітлення інформації стосовно базових алгоритмів класифікації в цій роботі не проводиться, так само як і не розказується про нормалізацію даних (тексту) та навчання і тренування моделей [1, 2]. Також слід зазначити, що набір базових моделей не є константним, його можна замінити при необхідності подальшого дослідження.

В якості базових класифікаторів можна використовувати безліч популярних алгоритмів. В даному випадку, в якості базових моделей використовуються наступні алгоритми:

- метод опорних векторів (Support vector machine) [3];
- наївний Баєсів класифікатор (Naive Bayes classifier) [4];
- метод k-найближчих сусідів (K-neighbors) [5];
- adaptive Boosting (AdaBoost) [6];
- випадковий ліс (Random forest) [7];
- логістична регресія (Logistic regression) [8].

Процес поєднання вище згаданих моделей та усереднення (оптимізація) результатів передбачень цих базових класифікаторів буде називатись створенням ансамблевої моделі.

Суть завдання не полягає в тому, щоб просто створити один ансамбль із набору примітивів, стоїть задача створення набору таких ансамблів, які будуть містити в собі тільки унікальні комбінації базових моделей. Методами комбінаторики потрібно створити набори унікальних ансамблів, щоб в подальшому вибрати найкращий ансамбль із створеного набору.

Функція, яка задовільняє поставлену комбінаторну задачу, наведена у лістингу 1 та реалізована на мові програмування Python [9]. Слід зазначити, що інформаційна система класифікації також реалізована на попередньо зазначеній мові програмування, оскільки дана «платформа» містить в собі багато рішень, пов'язаних із роботою ML (machine learning), що, в свою чергу, полегшує процес проведення дослідження, адже не потрібно відволікатись на реалізацію примітивів при написанні мета-алгоритму.

Лістинг 1 – функція залежності кількості примітивів (x) до кількості унікальних комбінацій (y)

```
def y_function(x):
    y = 0
    for i in range(0, x):
        for j in range(i + 1, x):
            for k in range(j, x):
                y += 1
    return y
```

За допомогою вище зображеного алгоритму створюються унікальні набори комбінацій. Графік залежності кількості базових моделей класифікації до кількості створених унікальних ансамблів зображений на рисунку 1. Даний графік дозволяє орієнтуватись у складності системи, адже дивлячись на нього, можна

обрати оптимальну з точки зору навантаження на систему кількість базових класифікаторів.

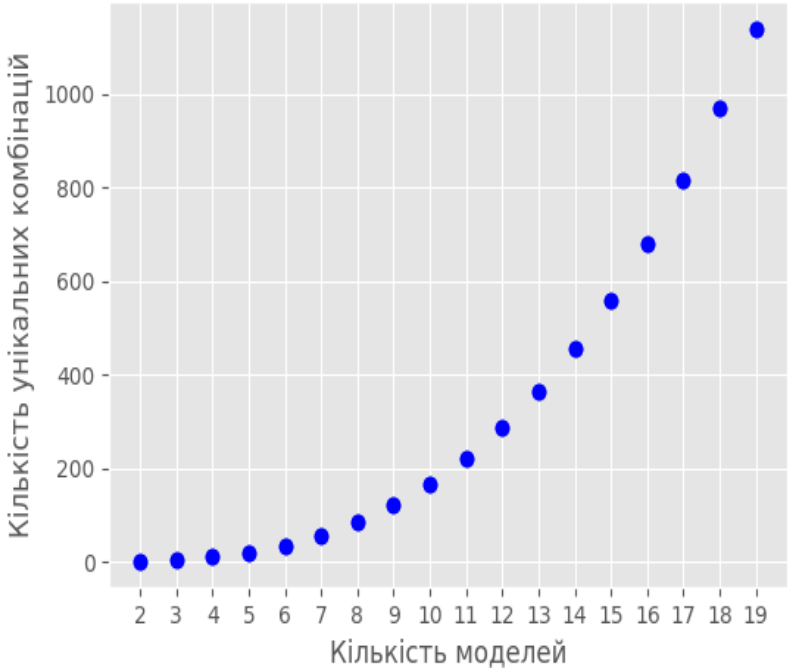

Рис. 1. Залежність кількості базових моделей класифікації до кількості створених унікальних ансамблів

Виходячи із вище зазначеного матеріалу, задача складання мета-алгоритму, який містить в собі набори унікальних ансамблів, вважається виконаною. Наступним кроком буде реалізація алгоритму усереднення (оптимізації) передбачень базових моделей у кожному унікальному ансамблі.

Алгоритм виведення колективного рішення базових моделей у ансамблі на основі неієрархічного методу кластеризації

Після успішного конструювання мета-алгоритму потрібно вирішити наступне завдання, а саме, потрібно узагальнити результати базових класифікаторів у кожному ансамблі, який складається з унікального набору моделей класифікації.

Перед тим як виконувати оптимізацію рішення ансамблю слід зазначити принцип по якому буде відбуватись усереднення передбачень [10]. Особливістю у задачі усереднення передбачень базових класифікаторів є те, що потрібно агрегувати рішення базових класифікаторів у кожному ансамблі тільки на тих порціях даних, де моделі не зійшлись у передбаченнях, тобто, агрегуються результати передбачень кожного алгоритму в ансамблі на даних, передбачення над якими хоча б у однієї моделі з комбінації були відмінні від інших передбачень у даному наборі.

Під час створення колективного рішення для кожного ансамблю моделей використовується кластерний аналіз даних. Кластерний аналіз - це статистичний метод обробки даних, який використовується з матрицями даних, в яких змінні не були попередньо розбиті на підмножини критеріїв та предикторів. Припускається, що в основі даних лежить невпорядкований набір дискретних класів. Всі вони різні, і жоден з них не має більшої ваги, ніж інший. Кластерний аналіз полягає в тому, що елементи об'єднуються в групи, або кластери, на основі того, наскільки тісно вони пов'язані між собою. Його можна окреслити як завдання визначення підгруп даних таким чином, щоб точки даних в одній підгрупі (кластері) були схожі, в той час як точки даних в різних кластерах суттєво відрізнялися.

Процедури кластеризації можна розглядати як "пре-класифікаційні" в тому сенсі, що дослідник не використовував попередні судження для розбиття об'єктів. Однак вважається, що існують неоднорідні групи даних, тобто існують "кластери".

Виконувати агрегацію рішень базових моделей у ансамблі буде алгоритм k-середніх (k-means) [11]. Алгоритм кластеризації k-середніх обчислює центроїди та ітераційно повторюється до тих пір, поки не знайде оптимальний центроїд. Передбачається, що кількість кластерів вже відома. Його також називають алгоритмом плоскої кластеризації [12]. Кількість кластерів, визначених з даних за допомогою алгоритму, позначається «K». У цьому алгоритмі точки даних відносяться до кластеру таким чином, щоб сума квадратів відстаней між точками даних та центроїдом була мінімальною.

Ключовими даними у вище згаданому алгоритмі кластеризації є критерії відстані. Вхідними даними для алгоритму k-means є матриця, що складається з відстаней між кожним об’єктом. Щоб визначити відстані між об’єктами, потрібно мати міру відстаней.

У знаходженні відстані між об’єктами допоможе Евклідова відстань [13]. Це геометрична відстань в багатовимірному просторі. Якщо об’єкти визначаються багатовимірними точками, або мають багато

характеристик (стимулів), $X_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}\}$, де $i = 1 \ldots n$, відстань може бути визначена відстанню між точками $d(X_i, X_j)$, де

$$d\left(X_i, X_j\right) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{in} - x_{jn})^2} \quad . \tag{1}$$

Метод базується на мінімізації суми квадратів відстаней між кожною точкою та центром її кластера, тобто функції:

$$\sum_{i=1}^{N} d\left(x_i, m_j\left(x_i\right)\right)^2 \quad , \tag{2}$$

де d – метрика, x_i – i-й об'єкт даних, а $m_j\left(x_i\right)$ – центр кластера, якому на j-й операції присвоєний елемент x_i.

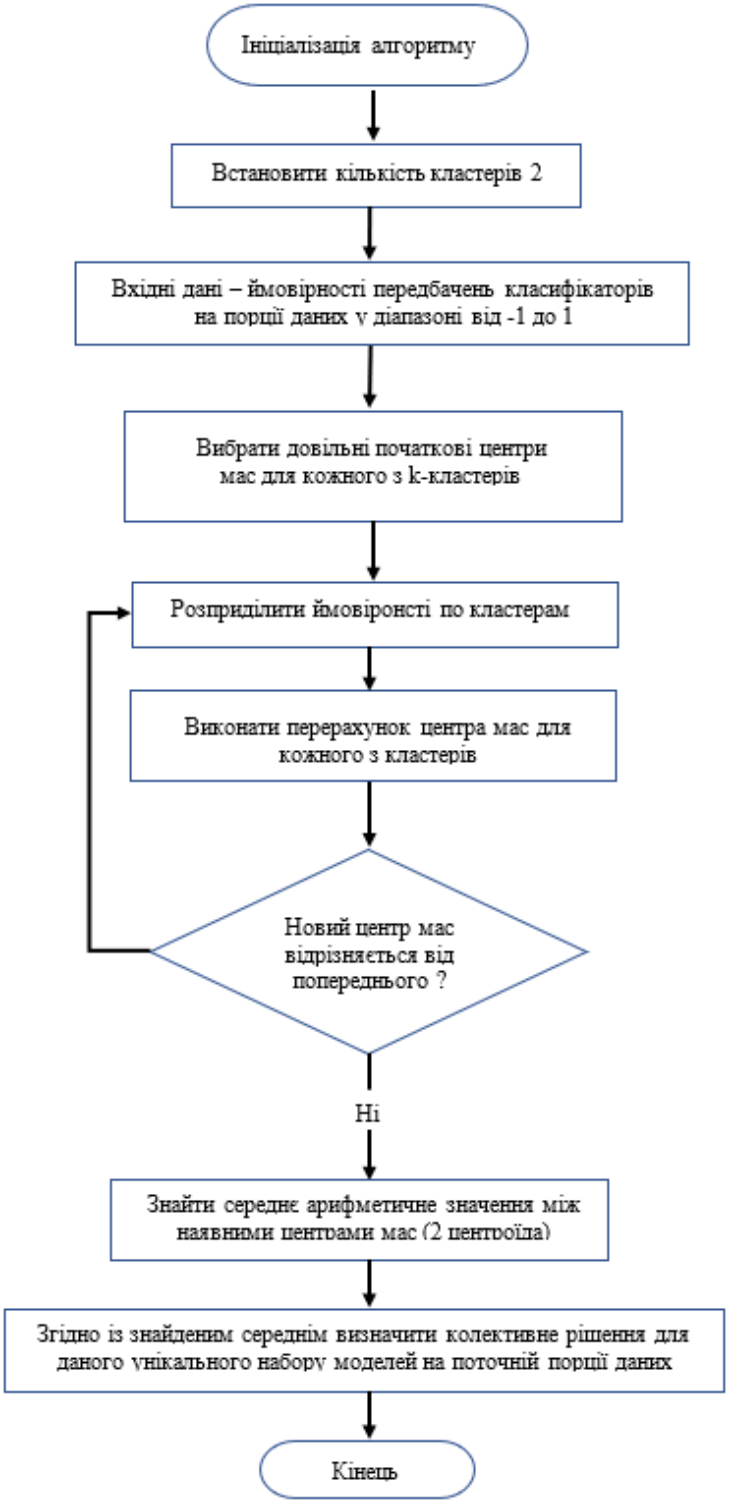

Рис. 2. Блок-схема алгоритму усереднення результатів на основі k-середніх (k-means)

Принцип алгоритму полягає в пошуку таких центрів кластерів та наборів елементів кожного кластера при наявності деякої функції Ф(°), що виражає якість поточного розбиття множини на *k* кластерів, коли сумарне квадратичне відхилення елементів кластерів від центрів цих кластерів буде найменшим:

$$V = arg\min \sum_{i=1}^{k} \sum_{x_j \epsilon S_i} (x_j - \mu_i)^2 \ , \tag{3}$$

де k – число кластерів, S_i – отримані кластери, $i = 1,2,\ldots,k$, μ_i – центри мас векторів $x_j \cdot S_i$.

Процес виконання алгоритму k-means в якості усереднювача (агрегатора) передбачень базових моделей у ансамблі можна розгледіти поетапно на блок-схемі, зображеній на рисунку 2. Зупинка алгоритму проводиться тоді, коли границі кластерів і розташування центроїдів не перестануть змінюватися від ітерації до ітерації, тобто на кожній ітерації в кожному кластері буде залишатися один і той же набір об'єктів. На практиці алгоритм зазвичай знаходить набір стабільних кластерів за кілька десятків ітерацій.

Задля більш наочного представлення роботи даного методу виведення колективного рішення, на рисунку 3 зображений приклад кластеризації даних результатів базових моделей . Вхідні дані на початку алгоритму являють собою ймовірності передбачень кожної моделі на порціях даних, де результати класифікації відрізнялись хоча б у однієї моделі у даному унікальному ансамблі. Порції даних, на яких результат класифікації алгоритмічних примітивів був однаковий, ігноруються.

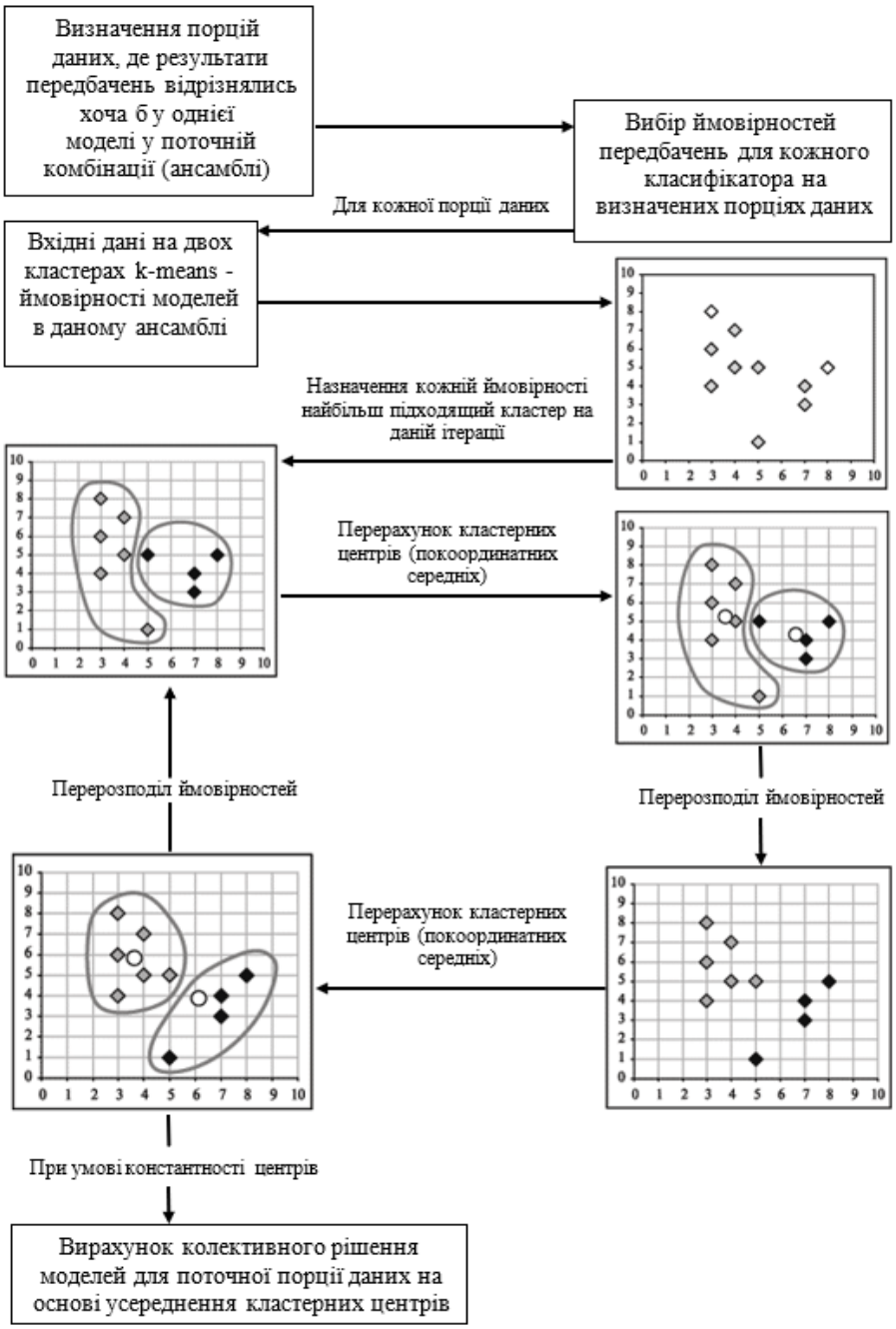

Рис. 3. Приклад роботи алгоритму виведення колективного рішення

В межах цього дослідження вище зазначений алгоритм буде виконувати розбиття вхідних даних на два кластери і працювати наступним чином:

- на вхід подаються ймовірності передбачень базових моделей. Оскільки виконується бінарна класифікація, відповідно, ймовірності можуть бути додатніми та від'ємними;
- вираховуються значення центроїдів кожного кластера, у даному випадку їх два (бінарне розбиття);
- відбувається обчислення середнього значення отриманих на попередньому кроці центроїдів;
- залежно від того, чи середнє значення додатне, чи від'ємне, усереднене рішення моделей на даній порції даних буде наступне: відноситься порція даних до заданої теми, чи ні.

Таким чином, оптимізувавши рішення базових моделей у кожному унікальному ансамблі на тих порціях даних, передбачення над якими було розбіжне хоча б у одного примітивного класифікатора, можна приступити до вирахування коефіцієнтів кореляцій та оцінки продуктивності класифікаційної системи згідно з основними метриками ML.

Результати дослідження. Порівняння продуктивності ансамблевої системи класифікації та базових алгоритмів. Показ кореляційних зв'язків

В якості оцінювальних метрик при порівнянні результативності ансамблевої моделі та базових моделей використовується точність «accuracy» та повнота «recall» [14, 15]. На рисунку 4 зображений графік точності та повноти, побудований на основі оцінювання базових класифікаторів. В даному випадку – це «svm», «naive-bayes», «k-neighbors», «ada-boost», «random-forest», «logistic-regression». На рисунку 5 зображений графік точності та повноти, побудований на основі оцінювання комбінацій (наборів) моделей, або унікальних ансамблів.

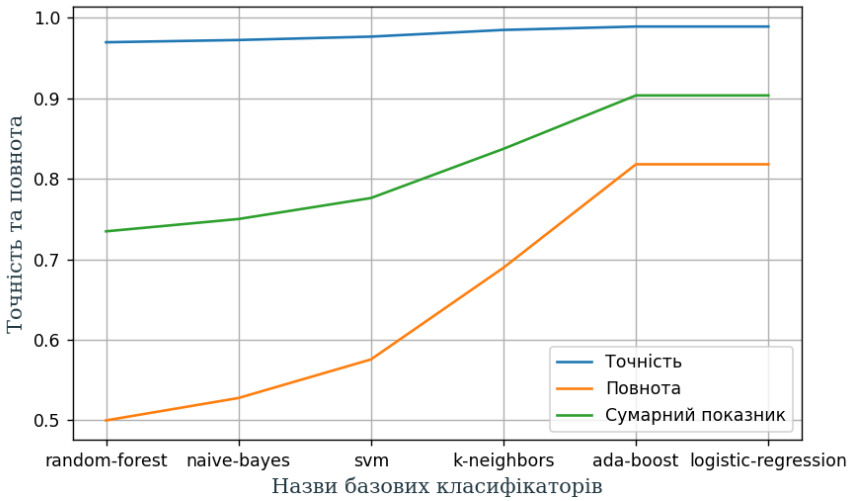

Рис. 4. Оцінка якісних показників класифікації базових класифікаторів: кращий результат модель класифікатора «ada-boost» за точністю 0.9889, повнотою 0.8179

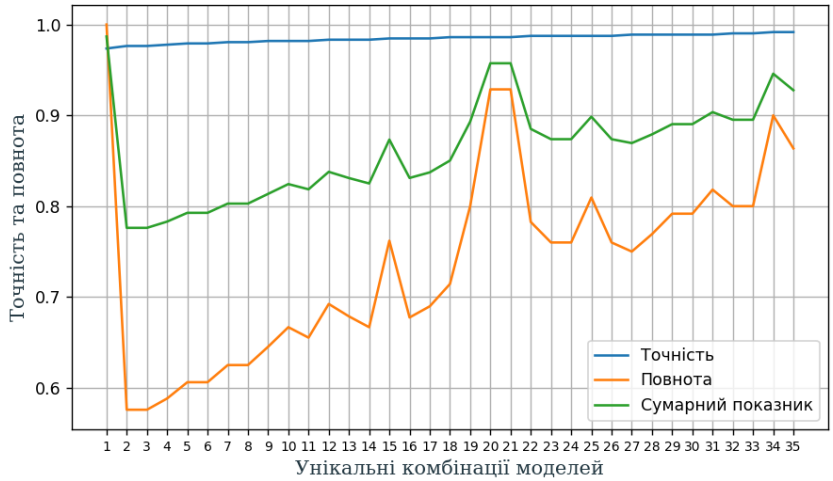

Рис. 5. Оцінка якісних показників класифікації комбінацій (наборів) моделей: кращий результат набір «ada-boost, random-forest» за точністю 0.9736, повнотою 1.0

Виходячи з даних, які зображені вище на рисунках, можна сказати, що одна з цілей дослідження

виконана, адже графіки показують, що продуктивність ансамблевих моделей значно вища за продуктивність базових алгоритмів класифікації.

Ще одним аспектом дослідження є показ коефіцієнтів кореляцій між істинними значеннями та передбаченнями базових класифікаторів у кожному унікальному ансамблі [16].

Слід зазначити, що коефіцієнт кореляції є статистичним показником сили взаємозв'язку між відносними закономірностями двох змінних [17]. Значення коливаються від -1,0 до 1,0. Кореляція -1,0 показує ідеальну негативну кореляцію, тоді як кореляція 1,0 - ідеальну позитивну кореляцію. Кореляція 0,0 показує відсутність лінійної залежності між закономірностями двох змінних.

Негативна кореляція – це зв’язок між двома змінними, в яких одна змінна збільшується в міру зменшення іншої, і навпаки.

Позитивна кореляція – це взаємозв'язок між двома змінними, в яких обидві змінні рухаються в тандемі - тобто в одному напрямку.

Вирахування кореляційних зв’язків робиться для того, щоб дізнатись, наскільки кожен класифікатор у комбінації з іншими класифікаторами в межах ансамблю приносить користь у завданні підвищення точності мета-алгоритму [18].

Кореляційні зв'язки істинних значень та передбачень базових моделей всередині кожного ансамблю у відсортованому вигляді зображені на рисунках 6, 7 в різних перспективах. Варто відзначити, що інформація подається у вигляді наборів даних в тривимірному просторі, це сприяє полегшенню в подальшому дослідженні нових закономірностей між наборами комбінацій, точністю кожної комбінації, величини залежності передбачення кожної моделі від істинного значення в середині кожної комбінації.

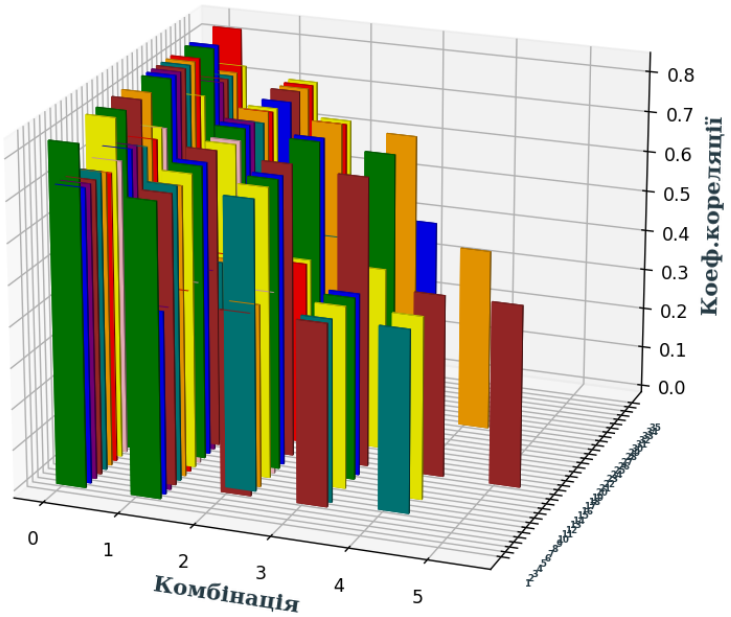

Рис. 6. Графіки гістограм коефіцієнтів кореляцій

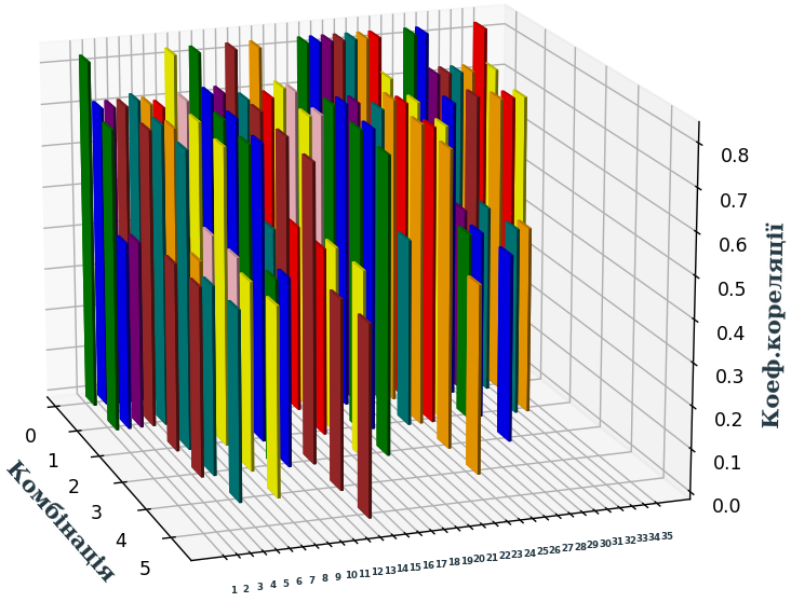

Рис. 7. Графіки гістограм коефіцієнтів кореляцій. Бокова перспектива

На рисунку 7 по осі «y» віднумеровані комбінації базових класифікаторів, список назв алгоритмів, які використовуються у кожному наборі згідно з нумерацією наступний:

1) ada-boost|random-forest;
2) svm|ada-boost;
3) naive-bayes|ada-boost;
4) svm|naive-bayes|k-neighbors|ada-boost;
5) svm|naive-bayes|k-neighbors|ada-boost|logistic-regression;
6) naive-bayes|k-neighbors|ada-boost;
7) svm|k-neighbors;
8) svm|naive-bayes|k-neighbors|ada-boost|random-forest;
9) svm|naive-bayes|ada-boost;
10) naive-bayes|k-neighbors|ada-boost|random-forest;
11) naive-bayes|k-neighbors|ada-boost|logistic-regression;
12) svm|logistic-regression;
13) svm|naive-bayes|k-neighbors|ada-boost|random-forest|logistic-regression;
14) naive-bayes|k-neighbors;
15) svm|random-forest;
16) svm|naive-bayes|k-neighbors;
17) svm|naive-bayes|k-neighbors|logistic-regression;
18) k-neighbors|ada-boost;
19) k-neighbors|ada-boost|random-forest|logistic-regression;
20) ada-boost|random-forest|logistic-regression;
21) random-forest|logistic-regression;
22) naive-bayes|random-forest;
23) naive-bayes|k-neighbors|random-forest;
24) naive-bayes|k-neighbors|ada-boost|random-forest|logistic-regression;
25) k-neighbors|ada-boost|random-forest;
26) k-neighbors|ada-boost|logistic-regression;
27) svm|naive-bayes;
28) svm|naive-bayes|random-forest;
29) svm|naive-bayes|k-neighbors|random-forest;
30) naive-bayes|logistic-regression;
31) ada-boost|logistic-regression;
32) svm|naive-bayes|logistic-regression;
33) naive-bayes|k-neighbors|logistic-regression;
34) k-neighbors|random-forest;

k-neighbors|logistic-regression

Висновок

В ході виконання роботи було досліджене застосування агрегативних підходів для класифікації на базі ансамблевих моделей. Реалізована інформаційна система бінарної класифікації дозволяє проводити дослідження на корпусі даних «Reuters», система розроблена таким чином, щоб уможливити всесторонню конфігурацію компонентів мета-алгоритму (ансамблю), зокрема:

- вибір базових алгоритмів класифікації відбувається у окремому модулі, що сприяє швидкій заміні при необхідності
- алгоритм агрегації рішень наборів базових моделей замінюваний, оскільки реалізований як окремий модуль
- візуалізаційні компоненти реалізовані окремо, потрібно дотримуватись контракту між модулями (інтерфейс взаємодії, зокрема, форма вхідних параметрів)

Під час дослідження застосування ансамблевої моделі на основі наборів унікальних комбінацій було проведено порівняння показників результативності даного ансамблю з результатами базових класифікаторів згідно з основними метриками ML. З'ясовано, що запропонована ансамблева модель являється більш точною та стійкою до похибок.

Важливим аспектом наукової роботи було знаходження коефіцієнтів кореляцій між істинними значеннями та передбаченнями кожного класифікатора у кожній комбінації. З'ясувавши кореляційні показники, було визначено «схильність» конкретної моделі у конкретній комбінації до похибки або правильного рішення. Іншими словами, кореляційна залежність між правильною відповіддю і результатом базової моделі у комбінації з іншими моделями показує наскільки даній моделі «комфортно» у поєднанні з іншими моделями, тобто, дослідивши кореляційні зв'язки, можна винайти такі комбінації базових класифікаторів, в яких кожен примітив (модель) покращує колективне рішення.

Задля подання показників класифікаційної системи використовується графічна система, яка складається із графіків у двох та трьох вимірних просторах. Графіки показують результативність окремих

базових класифікаторів та результативність ансамблевої моделі. У трьох вимірному просторі розміщені гістограми кореляційності кожної моделі у кожній комбінації, що забезпечує хороше проглядання інформації та уможливлює подальше дослідження на основі цих даних.

Одним із шляхів вдосконалення реалізованої інформаційної системи класифікації є добавлення можливості зміни корпусу даних та варіативність у методах нормалізації тексту (винесення в окремий модуль нормалізації даних), що уможливить конфігурування стосовно методу підготовки даних.

Наступним шляхом вдосконалення результативності мета-алгоритму є підбір базових класифікаторів на основі вирахуваного коефіцієнта кореляції, таким чином, система зможе створювати ансамблі з найвищими показниками метрики точності.

Література

1. Conroy B., Eshelman L., Potes C., Xu-Wilson M. A dynamic ensemble approach to robust classification in the presence of missing data. Machine Learning. 2017. P. 443-463.

2. Gomes H. M., Barddal J. P., Enembreck F. and Bifet A. A survey on ensemble learning for data stream classification. ACM Computing Surveys (CSUR). 2017. P. 23.

3. Bogdan Zagajewski, Marcin Kluczek, Edwin Raczko, Ajda Njegovec, Anca Dabija and Marlena Kycko. Comparison of Random Forest Support Vector Machines and Neural Networks for Post-Disaster Forest Species Mapping of the Krkono. Remote Sensing. 2021. Volume 13, № 13. P. 2581.

4. Domingos P., Pazzani M. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. Machine Learning 29. 2017. P. 103–130.

5. Maudes J. J. Rodríguez and C. García-Osorio. Disturbing neighbors diversity for decision forests in Applications of supervised and unsupervised ensemble methods. Heidelberg:Springer. 2018. P. 113-133.

6. Zhang C. X. and Zhang J. S. RotBoost: A technique for combining rotation forest and AdaBoost. Pattern Recognition Letters. 2019. Volume 29, № 10. P. 1524-1536.

7. Kulkarni V. Y. and Sinha P. K. Random forest classifiers: A survey and future research directions. International Journal of Advanced Computer Technology. 2013. Volume 36, № 1. P. 1144-1153.

8. Glenn V. Ostir. Logistic Regression: A Nontechnical Review. American Journal of Physical Medicine & Rehabilitation. 2016. Volume 6. P. 565-572.

9. Python. URL: https://www.python.org/doc/

10. Bifet A., Holmes G., Pfahringer B., Kirkby R. and Gavaldà R. New ensemble methods for evolving data streams/ Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. 2019. P. 139-148.

11. Jigui Sun, Jie Liu and Lianyu Zhao. Clustering algorithms Research, Journal of Software. 2014. Volume 19, № 1. P. 48-61.

11. Hartigan J.A., Wong M.A. Algorithm AS. 136. A k-means clustering algorithm. J. Roy. Stat. Soc. 2021. P. 100-108.

12. Danielsson P. E. Euclidean distance mapping/ Comput. Graphics Image Proc. 2019. Volume 14. P. 227-248.

13. Ling C.X. and Zhang H. Toward Bayesian Classifiers with Accurate Probabilities/ Proc. Sixth Pacific-Asia Conf. Knowledge Discovery and Data Mining. 2012. P. 123-134.

14. Dietterich T. G. Ensemble learning. In The handbook of brain theory and neural networks, Cambridge, MA:MIT Press. 2017. Volume 2. P. 110-125.

15. Cover T. M. The Best Two Independent Measurements Are Not the Two Best/ IEEE Trans. Systems, Man, and Cybernetics. 2010. Volume 4. P. 116-117.

16. Коефіцієнт кореляції. URL: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

17. Bifet A., Frank E., Holmes G. and Pfahringer B. Accurate ensembles for data streams: Combining restricted Hoeffding trees using stacking/ Proceedings of 2nd Asian conference on machine learning. 2018. Volume 13. P. 225-240.

References

1. Conroy B., Eshelman L., Potes C., Xu-Wilson M. A dynamic ensemble approach to robust classification in the presence of missing data. Machine Learning. 2017. P. 443-463.

2. Gomes H. M., Barddal J. P., Enembreck F. and Bifet A. A survey on ensemble learning for data stream classification. ACM Computing Surveys (CSUR). 2017. P. 23.

3. Bogdan Zagajewski, Marcin Kluczek, Edwin Raczko, Ajda Njegovec, Anca Dabija and Marlena Kycko. Comparison of Random Forest Support Vector Machines and Neural Networks for Post-Disaster Forest Species Mapping of the Krkono. Remote Sensing. 2021. Volume 13, № 13. P. 2581.

4. Domingos P., Pazzani M. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. Machine Learning 29. 2017. P. 103–130.

5. Maudes J. J. Rodríguez and C. García-Osorio. Disturbing neighbors diversity for decision forests in Applications of supervised and unsupervised ensemble methods. Heidelberg:Springer. 2018. P. 113-133.

6. Zhang C. X. and Zhang J. S. RotBoost: A technique for combining rotation forest and AdaBoost. Pattern Recognition Letters. 2019. Volume 29, № 10. P. 1524-1536.

7. Kulkarni V. Y. and Sinha P. K. Random forest classifiers: A survey and future research directions. International Journal of Advanced Computer Technology. 2013. Volume 36, № 1. P. 1144-1153.

8. Glenn V. Ostir. Logistic Regression: A Nontechnical Review. American Journal of Physical Medicine & Rehabilitation. 2016. Volume 6. P. 565-572.

9. Python. URL: https://www.python.org/doc/

10. Bifet A., Holmes G., Pfahringer B., Kirkby R. and Gavaldà R. New ensemble methods for evolving data streams/ Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. 2019. P. 139-148. 11. Jigui Sun, Jie Liu and Lianyu Zhao. Clustering algorithms Research, Journal of Software. 2014. Volume 19, № 1. P. 48-61.

11. Hartigan J.A., Wong M.A. Algorithm AS. 136. A k-means clustering algorithm. J. Roy. Stat. Soc. 2021. P. 100-108.

12. Danielsson P. E. Euclidean distance mapping/ Comput. Graphics Image Proc. 2019. Volume 14. P. 227-248.

13. Ling C.X. and Zhang H. Toward Bayesian Classifiers with Accurate Probabilities/ Proc. Sixth Pacific-Asia Conf. Knowledge Discovery and Data Mining. 2012. P. 123-134.

14. Dietterich T. G. Ensemble learning. In The handbook of brain theory and neural networks, Cambridge, MA:MIT Press. 2017. Volume 2. P. 110-125.

15. Cover T. M. The Best Two Independent Measurements Are Not the Two Best/ IEEE Trans. Systems, Man, and Cybernetics. 2010. Volume 4. P. 116-117.

16. Koefitsiient koreliatsii. URL: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

17. Bifet A., Frank E., Holmes G. and Pfahringer B. Accurate ensembles for data streams: Combining restricted Hoeffding trees using stacking/ Proceedings of 2nd Asian conference on machine learning. 2018. Volume 13. P. 225-240.