DOI 10.34132/mspc2025.01.14.09

Ірина Калініна,
Олександр Гожий

ПОБУДОВА БАГАТОРІВНЕВИХ АНСАМБЛЕЙ В ЗАДАЧАХ МАШИННОГО НАВЧАННЯ

Анотація. *Розглянуто підхід до розв'язання задач машинного навчання, побудований на основі багаторівневих гетерогенних ансамблів. Підхід дозволяє зменшувати помилки прогнозів за рахунок одночасного зменшення зміщення та дисперсії на основі використання багаторівневих гетерогенних ансамблів прогнозних моделей. Розглянуто основні особливості та передумови створення ансамблів. Проведено аналіз загальної помилки алгоритму машинного навчання. Вона складається з трьох компонентів: шум, зміщення та дисперсія. Досліджено та визначено складові цих компонентів. Розглянуто процес створення гетерогенного ансамблю. Проаналізовано найбільш поширені методи агрегування прогнозних значень: bagging, boosting і staking. Подано обґрунтування вибору типу моделей для створення гетерогенної багаторівневої ансамблевої структури. Запропоновано дворівневу архітектуру системи класифікації на основі методів staking та bagging. Для побудови багаторівневого ансамблю моделей з урахуванням різних базових методів розроблено алгоритм. Проаналізовано результати роботи ансамблів. Доведено ефективність багаторівневих гетерогенних ансамблів при вирішенні завдань класифікації.*

Ключові слова*: Машинне навчання, багаторівневі гетерогенні ансамблі, зміщення, дисперсія, bagging, boosting, staking, дворівнева архітектура ансамблю.*

Вступ

Останнім часом спостерігається розвиток різних підходів до вирішення завдань машинного навчання, які призвели до дослідження та створення нових моделей у різних галузях, таких як охорона здоров'я, розпізнавання мовлення, класифікація зображень, прогнозування та багато інших. Один із підходів, який поєднує декілька різних моделей для створення остаточної моделі відомий, як ансамблеве навчання або ансамблева модель. Підхід на основі ансамблів моделей успішно застосовують у широкому спектрі завдань машинного навчання. Ефективність цих моделей пояснюється найкращим поданням ознак за допомогою багатошарових архітектур обробки даних. Ансамблеві моделі в основному використовуються для задач класифікації, регресії та кластеризації. Ансамбль – це набір класифікаторів/регресорів, які вивчають цільову функцію, та їх індивідуальні прогнози поєднуються для класифікації нових прикладів. Ансамбль поєднує в собі кілька моделей, певним чином, наприклад, усередненням або голосуванням, таким чином, щоб отримати нову модель, яка має кращі показники, ніж моделі, що складають ансамбль. Оскільки моделі машинного навчання є обчислювальними та великими за даними, отже, ансамблеві моделі вимагають особливої уваги щодо додаткової інформації для об'єднання кількох алгоритмів у єдиній структурі. Ансамблеві моделі машинного навчання повинні вирішувати кілька завдань, таких як викликати різноманітність серед базових моделей, як зберегти час навчання, а також зниження складності моделей для практичних додатків, як об'єднати передбачення на основі додаткових алгоритмів. Створення ефективних ансамблевих моделей дозволить суттєво підвищити точність розв'язання задач машинного навчання.

Постановка задачі. Особливий інтерес для досліджень представляє створення багаторівневих гетерогенних ансамблів, які здатні істотно підвищити ефективність вирішення завдань машинного навчання. Створення таких ансамблів розглядатиметься у цій роботі.

Методи та моделі

Ефективним підходом до підвищення якості прогнозних значень при розв'язанні завдань класифікації є підхід до прогнозування на основі ансамблевих методів. Він дозволяє зменшувати помилки прогнозів за рахунок компромісу між зміщенням та дисперсією на основі використання багаторівневих *гетерогенних* ансамблів прогнозних моделей.

При реалізації ансамблевого підходу слід враховувати такі особливості:

- Ансамблі, як правило, покращують продуктивність узагальнення набору лише окремих моделей у домені.
- Велика кількість відносно некорельованих моделей, які працюють як домен, перевершують будь-яку з окремих складових моделей.

Для використання ансамблів є ряд передумов: набір класифікаторів зі схожими результатами навчання може мати різні результати узагальнення, об'єднання вихідних даних кількох класифікаторів знижує ризик

вибору погано працюючого (слабкого) класифікатора, якщо обсяг даних для аналізу занадто великий, один класифікатор може не впоратися з ним; необхідно навчати різні класифікатори різних розділах даних. Системи ансамблів також можна використовувати, коли даних замало, за допомогою методів повторної вибірки.

Помилка алгоритму машинного навчання складається з трьох компонентів: *шум, зміщення* та *дисперсія:*

$$Error(x) = Bias(x)^2 + Variance(x) + Noise(x),$$

де *Noise* – це непереборна помилка (випадкові помилки в даних, які неможливо усунути, наприклад, через пошкоджені вхідні дані, помилки введення даних і т. д.); *Bias* – це систематична помилка, яку, як очікується, зробить алгоритм навчання через, наприклад, архітектурний вибір або через недостатні/нерепрезентативні навчальні дані; *Variance* вимірює чутливість алгоритму до конкретного навчального набору та/або гіперпараметрів, що використовуються.

При доборі оптимальної прогнозної моделі необхідно враховувати компроміс між зміщенням і дисперсією. Техніка ансамблювання, тобто агрегування прогнозних значень різних базових моделей для створення однієї «оптимальної» є прийомом, що дозволяє пом'якшити компроміс між зміщенням та дисперсією. На рисунку 1 зображено умовний компроміс між зміщенням та дисперсією. Таким чином, ансамблі можуть допомогти зменшити як зміщення, так і дисперсію (за винятком шуму, який є непереборною помилкою).

На практиці ідея полягає в наступному: якщо використовуються «прості» моделі, навчені на менших вибірках, індивідуальна дисперсія кожної гіпотези може бути вищою, ніж для складнішого «учня», але все ж таки в більшості випадків дисперсія ансамблю нижче, ніж при використанні одного навченого складного.

Створення гетерогенного ансамблю

Застосування ансамблю моделей – це процес, у якому різні та незалежні моделі поєднуються для отримання кращого результату. При побудові ансамблю використовують кілька технік агрегування результатів базових моделей, кожен із яких забезпечує різну точність діагностики. Розглядалися три найбільш поширені методи агрегування прогнозних значень: *bagging, boosting* і *staking.* У таблиці 1 наведено обґрунтування вибору типу моделей для створення гетерогенної багаторівневої ансамблевої структури.

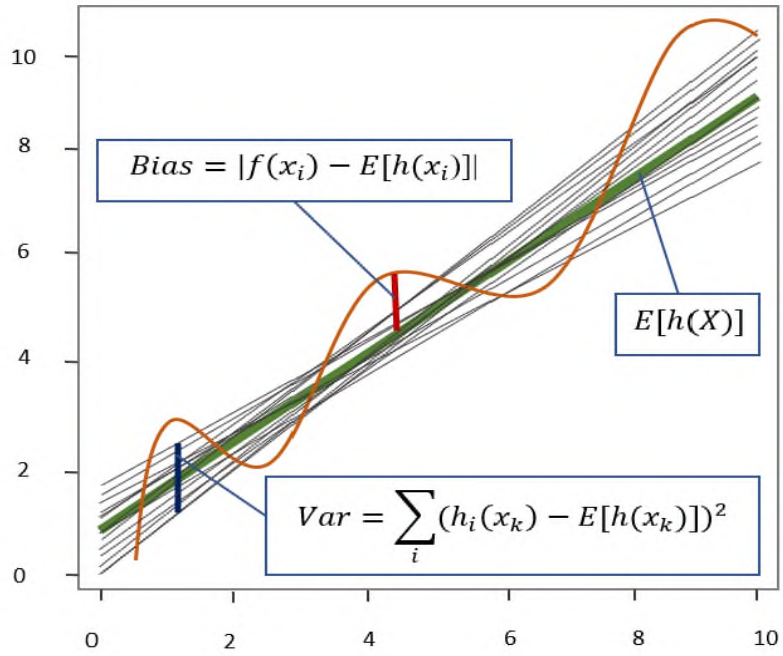

Рис.1 Компроміс між зміщенням та дисперсією.
Джерело: сформовано авторами

Таблиця 1

Обґрунтування вибору типу моделей для створення ансамблю

Тип ансамблю	Тип базових моделей	Характер помилки моделі	Результат агрегації
Bagging (модельне усереднення)	однорідні моделі	моделі з низьким зміщенням та високою дисперсією (перенавчені моделі)	зменшує дисперсію
Boosting (модельне підсилювання)	однорідні моделі	моделі з низькою дисперсією та високим зміщенням (недонавчені моделі)	зменшує зміщення

Stacking (модельне накладення)	різнорідні моделі	моделі з низькою дисперсією та високим зміщенням (недонавчені моделі)	зменшує зміщення, але залежно від вибору метамоделі може зменшувати дисперсію

Після експериментального підбору типу ансамблю розроблено дворівневу архітектуру системи класифікації на основі методів *staking* та *bagging* (рис. 2). Дворівневий класифікатор на обох рівнях агрегації використовує виважену суму. Ваги розраховуються на основі значення *F*-міри кожного з базових методів.

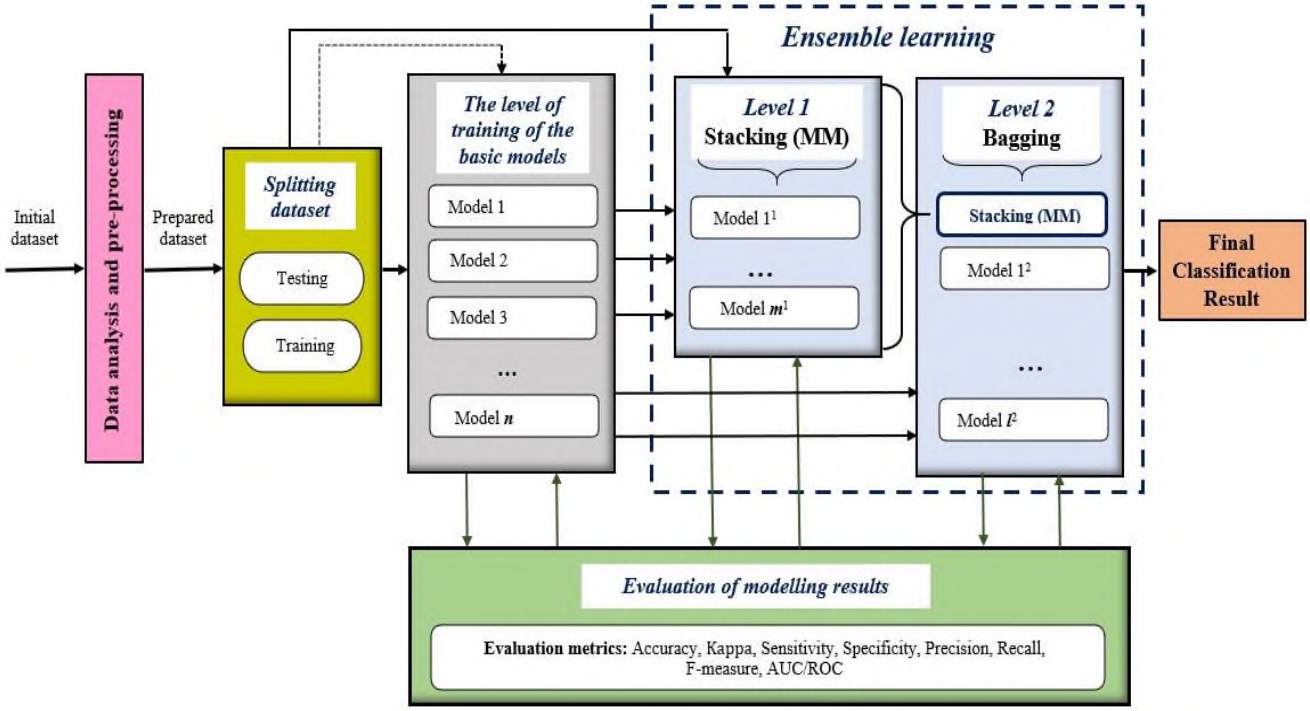

Рис.2. Дворівнева архітектура системи класифікації на основі методів *Staking* та *Bagging*.
Джерело: сформовано авторами

Для побудови багаторівневого ансамблю моделей з урахуванням різних базових методів розроблено алгоритм. Він складається із 7 кроків:

Крок 1. Поділ підготовленого до моделювання набору даних на дві вибірки (навчальну та тестову).

Крок 2. Навчання n різнорідних базових моделей на певній навчальній вибірці. Оцінка якості моделей. Вибір оптимальних параметрів моделей кожного типу.

Крок 3. Розрахунок кожної з n моделей прогнозних значень. Оцінка якості прогнозів за допомогою тестової вибірки.

Крок 4. Обчислення (оцінка) дисперсії та усунення кожної з *n* базових моделей.

Крок 5. Поділ *n* базових моделей на дві групи (з високим усуненням – недоучені та з високою дисперсією – перевчені).

Крок 6. Формування з урахуванням $m(m<n)$ моделей першої групи 1-го рівня ансамблевої структури – *staking*.

Крок 7. Формування з урахуванням $l(l \le n-m)$ моделей другої групи + результат методу *staking* 2-го рівня ансамблевої структури – *bagging*. Результат вважатиме остаточним прогнозним значенням.

Результати роботи класифікаторів

Головне з переваг дворівневого ансамблю полягає у системності використання методів агрегування та підбору базових моделей класифікації для кожного рівня ансамблевого навчання. Більшість моделей мають задовільну якість результату, але результуючі значення потребують покращення. Тому для підвищення якості результату будується дворівнева ансамблева модель. Як вхідні дані вибираються тестові оцінки якості базових моделей і додаються до тестового набору даних.

Розроблена система класифікації, заснована на багаторівневому ансамблевому підході, була перевірена на двох наборах даних. Перший набір даних містить інформацію про 748 донорів крові з Центру служби переливання крові в місті Синьчу, Тайвань. Метою класифікації є визначення кількості добровольців, які, швидше за все, знову здадуть кров. Другий набір даних (Indian Liver Patient Dataset) був підібрано для дослідження пацієнтів із захворюваннями печінки. Метою класифікації є виявлення потенційних донорів печінки, що сприяє більш ефективному та справедливому розподілу донорських органів серед пацієнтів, які очікують трансплантації.

У блоці дворівневого ансамблевого навчання послідовно використовуються *стекінг* та *bagging*. На першому рівні працює метод *staking* на основі моделі логістичної регресії. Модельний *staking* є ефективним методом ансамблювання. Прогнози, сформовані з допомогою базових моделей, використовуються як вхідні дані на навчання першому рівні.

Як базові моделі першого рівня були обрані: дерева рішень (DecisionTree); наївний Байєсов класифікатор (NB); лінійний дискримінантний аналіз (LDA); логістична регресія (LR); метод найближчого сусіда (KNN); штучні нейронні мережі (ANN).

Перший рівень: *Stacking***.** Основні відомості ансамблю моделей першого шару: логістична регресія як метамодель, 10 предикторів, два класи результуючої змінної (*no* та *yes*).

На другому рівні працює метод *bagging* на основі алгоритму *Bagged CART*. Алгоритм створює *N* регресійних дерев, використовуючи *M* початкових навчальних наборів та усереднює отримані прогнози. Кожне дерево має високу дисперсію, але низьку зміщення. Усереднення дерев зменшує дисперсію. Прогнозованими значеннями спостережень є мода (класифікація) чи середнє значення (регресія) дерев. Одним із недоліків *Bagged Trees* є те, що невелика кількість додаткових навчальних спостережень може різко змінити ефективність передбачення навченого дерева. Як базові моделі другого шару вибрано: модель першого рівня (*Stacking LR*); модель випадкового лісу (*RF*); квадратичний дискримінантний аналіз (*QDA*) .

Представлені значення показників ефективності першого та другого наборів даних. Ці дані враховують здатність моделей відрізняти один клас від інших (точність прогнозування, каппа-статистика, чутливість та специфічність, точність та повнота, *F*-міра). Використання дворівневої архітектури гетерогенного ансамблю моделей послідовно підвищує точність розв'язання задачі класифікації в машинному навчанні по всім метрикам якості.

Висновки

Для практичної реалізації багаторівневих гетерогенних ансамблів для вирішення задач класифікації розглянуто два набори даних: Blood Transfusion Service Center та ILPD (*Indian Liver Patient Dataset*). Для оцінки якості класифікаторів було використано систему показників якості. Проаналізовано результати роботи ансамблів. У таблицях представлені значення показників ефективності першого і другого наборів даних. Ці дані враховують здатність моделі відрізняти один клас від інших (точність прогнозування, каппа-статистика, чутливість та специфічність, точність та повнота, *F*-міра). Для першого набору даних на першому етапі в результаті об'єднання деяких базових моделей точність підвищена до 70%, а на другому етапі об'єднання деяких базових моделей та моделі *staking* в ансамбль моделі за допомогою *bagging* підвищило точність моделі до 88%. Для другого набору даних на першому етапі в результаті об'єднання базових моделей точність підвищена до 77%, а на другому етапі об'єднання деяких базових моделей та моделі *staking* в ансамбль моделі за допомогою *bagging* підвищило точність моделі до 82%. Таким чином використання гетерогенного дворівневого ансамблю значно підвищує ефективність класифікаційних моделей.

References

1. P. Bidyuk, I. Kalinina, O. Zhebko, A. Gozhyj and T. Hannichenko, "Classification System Based on Ensemble Methods for Solving Machine Learning Tasks". CEUR- WS. vol. 3426, 2023, pp. 1-11. CEUR-WS.org/Vol-3426/paper5.pdf.

2. V. Pandey, "Ensemble Methods in Practice: Combining the Strengths of Multiple Models and Making Decisions," 2023. [Online]. Available: https://www.linkedin.com/pulse/ensemble-methods-practice-combining-strengths-multiple-pandey (application date: 15.08.2024).

DOI 10.34132/mspc2025.01.14.09

Iryna Kalinina, Oleksandr Gozhyj

CONSTRUCTION OF MULTILEVEL ENSEMBLES IN MACHINE LEARNING PROBLEMS

The approach to solving the problems of machine learning and motivation on the basis of rich heterogeneous ensembles is examined. The approach allows you to change forecast estimates based on one-hour changes in variance and variance based on a variety of diverse heterogeneous ensembles of forecast models. The main features and changes in the creation of ensembles are examined. An analysis was carried out for the purpose of the machine learning algorithm. It consists of three components: noise, displacement and dispersion. The warehouse components have been tracked and identified. The process of creating a heterogeneous ensemble is examined. The most extensive methods of aggregation of forecast values have been analyzed: bagging, boosting and staking. The choice of type of models for the creation of a heterogeneous rich ensemble structure is given. The courtyard architecture of the classification system based on the methods of Staking and Bagging has been proposed. To create a rich

ensemble of models based on various basic methods, an algorithm has been developed. The results of the robotic ensembles were analyzed. The effectiveness of a large number of heterogeneous ensembles with the highest specification of classification has been demonstrated.

***Key words**: Machine learning, rich heterogeneous ensembles, displacement, dispersion, bagging, boosting, staking, courtyard architecture ensemble.*

Відомості про авторів/ Information about the Authors

Ірина Калініна, доктор технічних наук, професор кафедри інтелектуальних інформаційних систем, Чорноморського національного університету імені Петра Могили, м. Миколаїв, Україна. E-mail: irina.kalinina1612@gmail.com, orcid: https://orcid.org/0000-0001-8359-2045

Iryna Kalinina, Doctor of Technical Sciences, Professor of the Department of Intellectual Information Systems,of the Petro Mohyla Black Sea National University, Mykolaiv, Ukraine. E-mail: irina.kalinina1612@gmail.com, orcid: https://orcid.org/0000-0001-8359-2045

Олександр Гожий, доктор технічних наук, професор кафедри інтелектуальних інформаційних систем, Чорноморського національного університету імені Петра Могили, м. Миколаїв, Україна. E-mail: alex.gozhyj@gmail.com, orcid: https://orcid.org/0000-0002-3517-580X.

Oleksandr Gozhyj, Doctor of Technical Sciences, Professor of the Department of Intellectual Information Systems,of the Petro Mohyla Black Sea National University, Mykolaiv, Ukraine. E-mail: alex.gozhyj@gmail.com, orcid: https://orcid.org/0000-0002-3517-580X.

Стаття отримана редакцією 19.05.2025.