DOI 10.34185/1562-9945-3-146-2023-07
УДК 004.85 (075.8)

О.П. Гожий, О.О. Жебко, І.О. Калініна, Т.А. Ганніченко

ІНТЕЛЕКТУАЛЬНА СИСТЕМА КЛАСИФІКАЦІЇ НА ОСНОВІ АНСАМБЛЕВИХ МЕТОДІВ

Анотація. В роботі на основі методів машинного навчання досліджено вирішення завдання класифікації за допомогою дворівневої структури ансамблів моделей. Запропоновано архітектуру інтелектуальної системи класифікації. Для покращення результатів прогнозування застосовано ансамблевий підхід: кілька базових моделей навчались для вирішення однієї і тієї ж проблеми, з подальшим агрегуванням і покращенням отриманих результатів. Для пошуку компромісу між зміщенням та дисперсією, властивих моделям машинного навчання, було використано дворівневу структуру ансамблю. На першому рівні реалізовано ансамбль на основі стекінгу. На другому - на основі беггінгу. Проведено дослідження базових моделей класифікації та ансамблевих моделей на основі стекінгу та беггінгу, а також метрик оцінки ефективності використання базових класифікаторів та моделей першого та другого рівня. Визначено наступні параметри для усіх наведених методів у роботі: точність прогнозування та коефіцієнт помилок, Каппа-статистика, чутливість та специфічність, точність та повнота, F-міра та площею під ROC-кривою. Визначено ефективність ансамблю моделей в порівнянні з кожною базовою моделлю.
Ключові слова: Класифікація, ансамблеві моделі, стекінг, бустінг, бегінг, дворівнева архітектура, показники ефективності.

Вступ. Стрімкий розвиток технологій машинного навчання ініціював розвиток нових методів та алгоритмів, які більш ефективно вирішують задачі інтелектуального аналізу даних та прогнозування. Головною особливістю методів машинного навчання є не прямий розв'язок задачі, а навчання на множині прикладів, що дозволяє адаптувати ці методи для вирішення конкретних завдань обробки великих обсягів даних та виявляти в них нові знання. Для реалізації технологій машинного навчання використовуються різноманітні методи, такі як методи математичної статистики, ймовірнісні методи, чисельні методи, методи оптимізації, теорії ймовірностей, теорії графів, різні методи роботи з даними тощо [1-4].

Одними з ефективних методів для вирішення завдань регресії та класифікації є використання ансамбля моделей [5-7]. Застосування ансамблей це про-

цес, у якому різні та незалежні моделі поєднуються для отримання кращого результату. При підборі оптимальної моделі необхідно враховувати компроміс між зміщенням та дісперсією.

Зміщення – це нездатність моделі засвоїти достатньо відомостей про зв'язок між модельними ознаками та мітками, а дисперсія відображає нездатність моделі узагальнити на нових прикладах. Модель з високим зміщенням дуже зпрощує зв'язок і вважається недопідігнаной. Модель з високою дисперсією занадто пристосувалась до тренувальних даних і вважається перенавченой [8]. Будь-яка модель має за мету отримати низькі зміщення і дисперсію, але на практиці досягти і того, і іншого дуже важко. Цей феномен відомий як компроміс між зміщеністю та дисперсією.

Техніка ансамблювання, тобто комбінування різних моделей для створення однієї «оптимальної», є сучасним технічним прийомом, що дозволяє пом'якшити компроміс між зміщенням та дисперсією. Це дає можливість не покладатися на одну єдину модель, яка може бути перенавчена або мати інші недоліки [9].

В данній роботі досліджується процес використання ансамблевих моделей для побудови ефективного класифікатора. Для вирішення задачі класифікації використовуються різноманітні методи і моделі машинного навчання, для реалізації програмних модулів використовується мова програмування *R*, та середовище розробки *RStudio*.

Постановка проблеми. Метою данної роботи є дослідження ансамблевих методів при вирішенні завдань класифікації та розробка інтелектуальної системи класифікації на основі ансамблів моделей.

Аналіз останніх досліджень. Для побудови ансамблю використовується кілька технік агрегування результатів базових моделей, кожен з яких забезпечує різну точність діагностики. Bagging [10,11], boosting [10,12] і stacking [10,13] є найбільш поширеними підходами до побудови ансамблів.

Bagging – це метод паралельного навчання базових моделей, який підходить для різних моделей, що вважаються незалежними одина від одної. Використання методу *bagging* сприяє зменшенню дисперсії в базових моделях. В роботі [14] досліджені різні аспекти процесу *Bagging*. Метод генерує вибіркові дані для навчання з набору даних. Це досягається шляхом випадкової вибірки із заміною вихідного набору даних. Вибірка із заміною може повторювати деякі спостереження в кожному новому навчальному наборі даних. Кожен елемент у *Bagging* з однаковою ймовірністю з’явиться в новому наборі даних. В роботі [11]

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

цей тип ансамблю використовується для паралельного навчання кількох моделей. Розраховується середнє значення всіх прогнозів з різних моделей ансамблю. Під час класифікації враховується більшість голосів, отриманих за допомогою механізму голосування. Різні варіанти побудови ансамблей на основі *Bagging* представлено в роботах [15,16].

В літературних джерелах широко висвітлюється метод створення ансамблей *Boosting* [12,14]. Це послідовний метод агрегування, який ітеративно регулює вагу спостереження відповідно до останньої класифікації. Якщо спостереження неправильно класифіковано, це збільшує вагу цього спостереження. Термін «бустинг» стосується алгоритмів, які перетворюють слабку модель на сильнішу. Це зменшує помилку зміщення та створює надійні прогностичні моделі [17]. Точки даних, неправильно спрогнозовані під час кожної ітерації, виявляються, а їх ваги збільшуються. В роботі [18] алгоритм *Boosting* призначає ваги кожній отриманій моделі під час навчання. Моделі з кращими результатами прогнозування даних навчання присвоєно вищу вагу. Якщо наданий вхід є невідповідним, його вага збільшується. Мета, що стоїть за цим, полягає в тому, щоб майбутня гіпотеза з більшою ймовірністю належним чином класифікувала його шляхом об’єднання всього набору, щоб нарешті перетворити слабкі моделі на моделі з кращою ефективністю [19].

Третім методом створення ансамблей є *Stacking*. Ця методика ансамблю працює шляхом застосування комбінованих прогнозів кількох слабких моделей в рамках *метамоделі*, щоб можна було досягти кращих результатів прогнозування [13]. *Stacking* також відомий як узагальнення з об’єднанням і є розширеною формою методики ансамблю усереднення моделі, в якій усі підмоделі однаково беруть участь відповідно до їх ваг продуктивності та створюють нову модель із кращими прогнозами. Ця нова модель розташована поверх інших, це причина, чому вона віднесена до *Stacking* [15].

В роботах [6,9] архітектури стекінгової моделі розроблені таким чином, що вони складаються з двох або більше базових моделей/моделей та метамоделі, яка поєднує прогнози базових моделей. Ці базові моделі називають моделями рівня 0, а метамодель — моделлю рівня 1. В роботах [18-20] представлені результати досліджень багаторівнених структур ансамблей моделей. В роботі [20] методи сумісного ансамблю включають вхідні (навчальні) дані, моделі первинного рівня, прогноз первинного рівня, модель вторинного рівня та остаточний прогноз. Автори пропонують систему підтримки прийняття рішень, в основі якої - дворівневий класифікатор, що на обох рівнях агрегації використовує

зважену суму. Ваги розраховуються на основі значення *F*-міри кожного з базових алгоритмів.

Опис набору даних. Для створення інтелектуальної системи класифікації на основі ансамблів моделей було використано набір даних E-Commerce *Shipping Data*. Мета моделі – виявити чинники, пов'язані з ризиком невчасної доставки попередніх замовлень клієнтам та інформацію про одержувачів товарів. Дані з такими характеристиками доступні в репозиторії машинного навчання ***Kaggle***. Набір даних містить інформацію про дані доставки, отримані від міжнародної E-Commerce компанії Індії [21].

Набір даних, включає 10999 прикладів доставки товару, а також 12 змінних. Це набір числових і номінальних ознак, що визначають характеристики товару та клієнта. Дані містять таку інформацію, як: ідентифікаційний номер клієнтів; складський блок; спосіб; дзвінки у службу підтримки клієнтів (кількість дзвінків, здійснених клієнтами щодо запиту про відправлення товару); рейтинг клієнтів (від 1 до 5); вартість продукту (вартість продукту в доларах США); попередні покупки (кількість попередніх покупок); важливість продукту (низький, середній, високий); стать покупця (чоловіча та жіноча); знижка; вага в грамах; доставка товару.

Архітектура інтелектуальної системи класифікації. Архітектура розробленої інтелектуальної системи класифікації представлена на рис.1. Система складається з підсистеми збору та зберігання даних, блока попередньої обробки та аналізу даних, блока побудови та навчання базових моделей та блоку побудови і оцінювання ансамблів моделей, який реалізовано на основі дворівненого підходу. Блок попередньої обробки та аналізу даних виконує: аналіз структури набору даних, виявлення нечислових та відсутніх значень, перекодування нечислових ознак, аналіз взаємозв’язку між змінними, відбір ознак, нормалізація числових даних. Далі здійснюється розподіл набору даних на тестову і тренувальну вибірку [22].

Важливим етапом більшості ансамблевих методів є вибір оптимального набору моделей включення до ансамблю. Дві умови, які мають бути виконані для досягнення ансамблю гарної якості, — це точність та різноманітність прогнозів. Тому в інтелектуальній системі в якості базових моделей представлено оптимальну комбінацію класифікаторів, що складається з Decision Tree, NB, LDA, QDA, LR, SVM, RF, KNN, ANN та оцінюється їх якість. В блокі побудови та навчання базових моделей виконуються обчислення на попередньо опрацьованих даних. Кожен окремий класифікатор навчається з використанням навча-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

льних даних. Для підвищення точності в блоці побудови і оцінювання ансамблі моделей формуються на двох рівнях та оцінюється якость остаточного прогнозу [23-25].

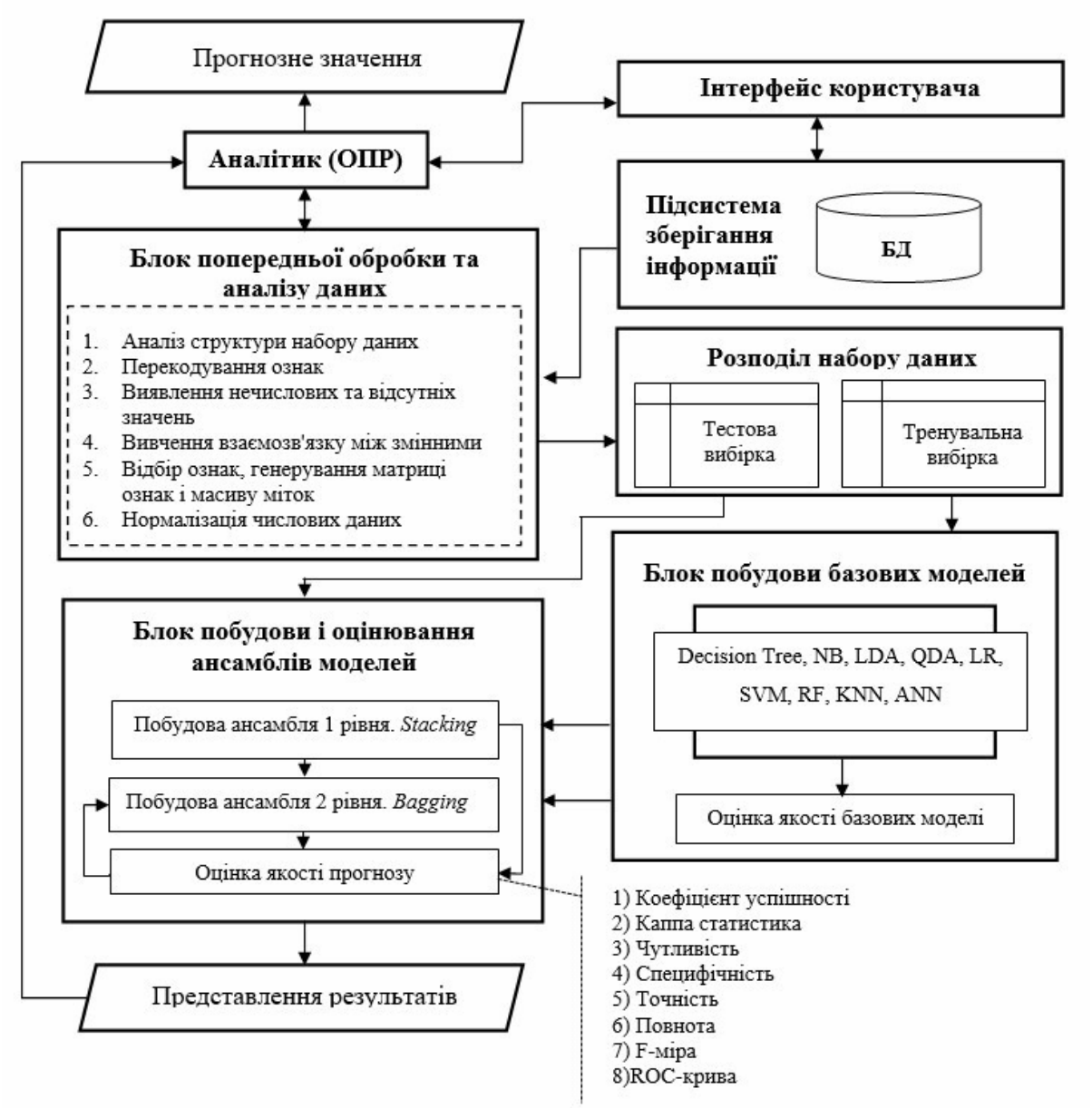

Рисунок 1– Архітектура інтелектуальної системи класифікації

Попередня обробка даних. Підготовка до моделювання. Попередня обробка даних включає в себе перетворення категоріальних даних в числові, нормалізацію даних, перевірку на нечислові та відсутні значення, перевірку на наявність аномальних значень та відбір ознак.

Першим кроком відбувається перетворення факторних змінних набору *Warehouse_block (A→1, B→2, C→3, D→4, E→5), Mode_of_Shipment (Flight→1, Ship→2, Road→3), Product_Importance (high→1, medium→2, low→3)* та *Gender*

(F→1, M→2) на числові, оскільки для більшості моделей машинного навчання на вхід подаються числові значення, а також для стандартизації значень атрибутів з метою отримання кращої точності навчання. Наступним кроком виконується нормалізація даних за допомогою мінімаксного методу. Змінні *Cost_of_the_Product, Discount_offered* та *Weight_in_gms* мають достатньо великі значення та всі дані з набору приводяться до загальної шкали без втрати інформації про відмінність діапазонів.

Нечислові та відсутні значення в наборі даних перевіряються за допомогою функцій *NaValue* та *BlankValue*. При перевірці з'ясовано, що в обраному для досліджень наборі даних немає нечислових чи відсутніх значень в жодному атрибуті. Аномальні значення у наборі даних також відсутні.

Вектор *Reached.on.Time_Y.N* вказує на те, чи змогла компанія вчасно чи з затримкою доставити певний товар покупцю. Загалом приблизно 60 % товару із цього набору даних було доставлено невчасно.

Підготовлений для моделювання набір даних поділено на дві частини: тренувальний набір для побудови базової моделі класифікатора та тестовий набір для оцінки ефективності моделі нових даних. Використано 90% даних для навчання та 10% – для тестування, що дало 1999 записів для моделювання нових претендентів.

Вирішення задачі класифікації за допомогою ансамблю моделей дворівневої класифікації. Для досягнення найкращої якості вирішення завдань класифікації були використанно ансамблеві методи багаторівневої класифікації двох різних типів: *stacking* та *bagging*. Їх ідея полягає у використанні декількох моделей на різних шарах [20]. Одна з переваг даного методу полягає в тому, що забезпечується гнучкість вибору вирішальних функцій, що знаходять «приховану залежність» між об'єктами навчальної вибірки та відповідними значеннями їх класів (через різні комбінації вирішальних функцій базових класифікаторів, що входять до складу ансамблевого). Більшість моделей мають задовільну якість роботи, але результат все ще потребує покращення, оскільки компанія прагне покращити сервіс надавання послуг з продажу та доставки товарів. Тому для покращення результату будується дворівнева ансамблева модель. У якості вхідних даних обираються тестові оцінки якості базових моделей та додаються їх до тестового набору.

На *першому шарі* використано метод стекінгу на основі логістичної регресійної моделі в якості метамоделі. Модельний стекінг є ефективним методом ансамблювання різнорідних за типом моделів. Коли прогнози різних мо-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

делей об'єднуються за допомогою стекінгу, бажано, щоб прогнози, що зроблені базовими моделями, мали низьку кореляцію. Це означає, що моделі є коректними та дозволяє новому класифікатору з'ясувати, як отримати найкраще від кожної моделі для покращеного результату.

Прогнози, які сформовані за допомогою базових моделей, використовуються як вхідні дані для навчання на першому шарі. В якості базових моделей першого шару обрано: дерева рішень (*DecisionTree*), наївний Баєсівський класифікатор (*NB*), квадратичний дискримінантний аналіз (*QDA*), логістична регресія (*LR*), метод опорних векторів (*SVM*), модель випадкового лісу (*RF*).

Основні відомості ансамблю моделей першого шару: логістична регресія в якості метамоделі, 10 предикторів, два класи результуючої змінної ("no" та "yes"). Отримана точність тренувальної моделі 69,3 % та каппа-статистика 0,42. За попереднім оцінюванням отримали покращення точності класифікації в порівнянні з окремими базовими класифікаторами. Остаточне оцінювання відбувалось на тестових даних. Результат показує, що помилка становить 1399 зі 1999, або 30 %. Отже, точність класифікації на тестових даних – 70 %. Невчасно доставлені товари вірно передбачені 607 рази, у той же час невірно класифіковані 25 разів. Товари, які були доставлені вчасно вірно передбачені 792 разів, невірно – 575 рази. У результаті 25 реальних значень "no" були помилково класифіковані як "yes" (хибно-позитивні спрацьовування), а 575 значення "yes" були помилково класифіковані як "no" (хибно-негативні спрацьовування). В результаті об'єднання різних базових моделей в ансамбль моделі за допомогою стекінгу, вдалось підвищити точність моделі на першому рівні до 70 %.

На *другому шарі* було використано метод беггінгу на основі алгоритму ***Bagged CART***. Алгоритм створює *N* регресійних дерев, використовуючи *M* початкових навчальних наборів і усереднює отримані прогнози. Ці дерева вирощуються глибоко та не обрізаються. Кожне окреме дерево має високу дисперсію, але низьке зміщення. Усереднення *N* дерев зменшує дисперсію. Прогнозованими значеннями для спостережень є мода. Одним із недоліків *Bagged Trees* є те, що невелика кількість додаткових навчальних спостережень може різко змінити ефективність передбачення вивченого дерева.

В якості базових моделей другого шару обрано: модель першого рівня (*Stacking LR*), модель штучних нейронних мереж (*ANN*); модель лінійного дискримінантного аналізу (*LDA*) та модель на основі методу найближчого сусіда (*KNN*). Основні відомості ансамблю моделей другого шару: метод *Bagged CART*, 4 предиктора, два класи результуючої змінної: "no" та "yes". Отримана точ-

ність тренувальної моделі 66,3 % та каппа-статистика 0,33. Результат оцінки на тестових даних показує, що помилка становить 232 зі 1999, або 11 %. Отже, точність класифікації на тестових даних – 88 %. Такм чином невчасно доставлені товари вірно передбачені 959 рази, у той же час невірно класифіковані 9 разів. Товари, які були доставлені вчасно вірно передбачені 808 разів, невірно – 223 рази. У результаті 9 реальних значень “no” були помилково класифіковані як “yes” (хибно-позитивні спрацьовування), а 223 значення “yes” були помилково класифіковані як “no” (хибно-негативні спрацьовування).

Слід звернути увагу, що у тестових даних модель вірно спрогнозувала 959 з 1182 реальних невчасних доставок, тобто 81,1%. Отже, на другому етапі в результаті об’єднання деяких базових моделей та моделі стекінгу в ансамбль моделі за допомогою беггінгу, вдалось підвищити точність моделі на другому рівні до 88 %.

Оцінка ефективності розроблених моделей класифікації. Існує багато способів виміряти ефективність класифікатора, проте найкращий показник – той, який визначає, чи є класифікатор успішним, коли застосовується за прямим призначенням. Важливо визначити показники ефективності, які показуватимуть корисність, а не просто точність [26]. Альтернативні показники ефективності отримаємо із матриці невідповідностей.

Матриця невідповідностей – це таблиця, яка класифікує прогнози залежно від цього, чи відповідають вони фактичним значенням. Один із вимірів таблиці – це можливі категорії прогнозованих значень, а друге – ті самі категорії для реальних значень. Якщо прогнозоване значення збігається з реальним, класифікація правильна. Правильні прогнози розміщуються у матриці невідповідностей щодо діагоналі. Інші комірки матриці відповідають випадкам, коли прогнозоване значення відрізняється від реального. Показники ефективності для моделей класифікації засновані на кількості прогнозів, що потрапляють на діагональ та за її межі.

Значення показників ефективності, що враховують здатність моделі відрізняти один клас від інших (*точність прогнозування, каппа-статистика, чутливість та специфічність, точність та повнота, F- міра та площа під ROC – кривою*) наведено в таблиці 1. У якості базових моделей обрано моделі: дерева рішень (Decision Tree), наївний Баєсів класифікатор (NB), лінійний дискримінантний аналіз (LDA), квадратичний дискримінантний аналіз (QDA), логістична регресія (LR), метод опорних векторів (SVM), метод найближчого сусіда (KNN), штучні нейронні мережі (ANN) та модель випадкового лісу (RF).

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Таблиця 1

Показники ефективності моделей

Тип моделі	Коефіцієнт успішності (accuracy)	Каппа-статистика (Kappa)	Чутливість (sensitivity)	Специфічність (specificity)	Точність (precision)	Повнота (recall)	F-міра (F-measure)	ROC-крива (AUC ROC)
Decision Tree	0,7	0,44	0,51	0,98	0,97	0,51	0,67	0,75
NB	0,67	0,36	0,59	0,8	0,81	0,59	0,68	0,69
QDA	0,67	0,38	0,45	0,98	0,97	0,45	0,61	0,72
LR	0,6	0,27	0,56	0,7	0,58	0,6	0,57	0,64
SVM	0,67	0,35	0,6	0,77	0,79	0,67	0,68	0,68
RF	0,67	0,34	0,63	0,72	0,96	0,51	0,69	0,68
Stacking (LR)	**0,7**	**0,44**	**0,51**	**0,96**	**0,96**	**0,51**	**0,67**	**0,74**
LDA	0,64	0,27	0,68	0,59	0,71	0,68	0,69	0,63
KNN	0,6	0,17	0,67	0,5	0,66	0,05	0,67	0,28
ANN	0,33	-0,45	0,05	0,44	0,04	0,63	0,04	0,58
Bagging	**0,88**	**0,77**	**0,81**	**0,98**	**0,98**	**0,81**	**0,89**	**0,9**

Аналіз таблиці показує, що найкращу оцінку специфічності та точності мають дерева рішень та метод квадратичного дискримінантного аналізу, найвищу точність мають дерева рішень. F-міра має найвищій показник для результатів для результатів випадкового лісу та лінійного дискримінантного аналізу, а площа під ROC-кривої найбільша для результатів класифікації з використанням дерев рішень. Загалом, базові класифікатори показали середній результат на даному наборі даних та результати потребують покращення. Найгірші результати дає штучна нейронна мережа.

Використовуючи стекінг для комбінування шести не дуже високих результатів базових класифікаторів, покращено загальний результат. Навчання моделі стекінгу відбувалося шляхом використання логістичної регресії *(glm)*. Використовуючи беггінг для комбінування трьох не дуже високих результатів базових класифікаторів та результату роботи моделі першого рівня – стекінгу, значно покращено загальний результат по всім метрикам оцінювання. Навчання моделі беггінгу відбувалося шляхом використання беггінгу дерев рішень *(treebag)*.

Висновки. Проведено дослідження базових методів прогнозування та ансамблевих моделей на основі стекінгу та беггінгу. Досліджені метрики оцінки ефективності використання базових класифікаторів, моделей першого та дру-

гого рівнів. Визначено наступні параметри для усіх наведених методів у роботі: точність прогнозування та коефіцієнт помилок, Каппа-статистика, чутливість та специфічність, точність та повнота, F-міра та площею під ROC-кривою.

Було визначено та описано ключові характеристики якості моделей, вибір метрики, підбір базових (слабких) моделей, підбір параметрів для базових моделей та ансамблевих методів. Проведено процес збору даних, їх аналіз та інтерпретацію, виконано попередня обробка даних та розділено набір на тренувальну та тестову вибірки й згенеровано вхідні змінні на основі поведінкових даних клієнтів. Наведено результати застосування простих класифікаторів та ансамблю моделі дворівневої класифікації та виконано оцінку ефективності розроблених моделей класифікації.

Метрики оцінки якості роботи базових класифікаторів показали опосередковані результати, що потребують покращення**.** Використовуючи стекінг для комбінування шести не дуже високих результатів базових класифікаторів, було покращено загальний результат. Використовуючи беггінг для комбінування трьох не дуже високих результатів базових класифікаторів та результату роботи моделі першого рівня – стекінгу, значно покращено загальний результат по всім метрикам оцінювання.

ЛІТЕРАТУРА

1. Marsland S. Machine Learning: An Algorithmic Perspective. Palmerston North: Massey University, 2015. 452 p.

2. Wang L., Cheng L., Zhao G. Machine Learning for Human Motion Analysis. Anhui: IGI Global, 2009. 318 p.

3. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. California: Springer–Verlag, 2009. 746 p.

4. Artificial Intelligence: A Modern Approach.
URL: https://towardsdatascience.com/understanding–the–bias–variance–tradeoff (дата звернення: 20.12.2022).

5. Opitz D., Maclin R. Popular ensemble methods: An empirical study: journal of Artificial Intelligence Research No. 11. El Segundo, 1999. P. 169–198.

6. Ensemble Methods to Optimize Machine Learning Models.
URL: https://hub.packtpub.com/ensemble–methods–optimize–machine–learning–models (дата звернення: 20.12.2022).

7. Бармак О. В., Крак Ю. В., Манзюк Е. А. Характеристика для вибору моделей у ансамблі класифікаторів: науковий журнал “Проблеми програмування”. Київ, 2018. С. 171–179.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

8. Understanding the Bias–Variance Tradeoff
URL: http://scott.fortmannroe.com/docs/BiasVariance.html (дата звернення: 20.12.2022).
9. Dietterich T., Ensemble Methods in Machine Learning.
URL: http://web.engr.oregonstate.edu/~tgd/publications/mcs–ensembles.pdf (дата звернення: 25.12.2022).
10. Ensemble methods: bagging, boosting and stacking. URL: https://towardsdatascience.com/ensemble–methods–bagging–boosting–andstacking- c9214a10a205 (дата звернення: 27.12.2022).
11. Moretti F., Pizzuti S., Panzieri S., Annunziato M., Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling, Neurocomputing (2015).
12. Kim M.J., Kang D.K., Kim H.B. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction, Expert Syst. Appl. 42 (3) (2015) 1074–1082.
13. Kang S., Cho S., Kang P. Multi-class classification via heterogeneous ensemble of one-class classifiers, Eng. Appl. Artif. Intell. 43 (2015) 35–43.
14. Бустінг і беггінг як методи формування ансамблей моделей / Вербівський Д. С., Карплюк, С. О., Фонарюк, О. В., Сікора, Я. Б. Житомир: ЖДУ ім. Івана Франка, 2021. С. 163-169.
15. Zhi–Hua Zhou Ensemble Learning. URL: https://cs.nju.edu.cn/zhouzh/zhouzh.files/ publication/springerEBR09.pdf (дата звернення: 20.12.2022).
16. Shah B.R., Lipscombe L.L. Clinical diabetes research using data mining: a Canadian perspective, Can. J. Diabetes 39 (3) (2015) 235–238.
17. Improvements on Cross–Validation: The 632+ Bootstrap Method. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10474007#.U2o7MVdMzTo (дата звернення: 22.12.2022).
18. Ahmad A., Brown G. Random ordinality ensembles: ensemble methods for multi-valued categorical data, Inf. Sci. 296 (2015) 75–94.
19. Sluban B., Lavrac N. Relating ensemble diversity and performance: a study in class noise detection, Neurocomputing 160 (2015) 120–131.
20. Bashir S., Qamar U., Khan F.H. IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework. Journal of Biomedical Informatics. – 2016. – vol.59. – pp.185- 200.
21. E–Commerce Shipping DataSet.

URL: https://www.kaggle.com/datasets/+prachi13/ customer-analytics (дата звернення: 15.11.2022)
22. 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.
23. Maniruzzaman M., Rahman MJ., Ahammed B. Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems, Vol. 8. Texas, 2020. P. 1–14.
24. Бідюк П.І., Кузнєцова Н.В., Терентьєв О.М. Система підтримки прийняття рішень для аналізу даних. Київ: Наукові вісті НТУУ «КПІ», 2011. С. 48–61.
25. Zhan Zh. Introduction to machine learning: k-nearest neighbors. Vienna: Ann Transl Med, 2016. 218 p.
26. Метрики в задачах машинного навчання.
URL: https://habr.com/ru/company/ods/blog/328372 (дата звернення: 20.12.2022).

REFERENCES

1. Marsland S. Machine Learning: An Algorithmic Perspective. Palmerston North: Massey University, 2015. 452 p.
2. Wang L., Cheng L., Zhao G. Machine Learning for Human Motion Analysis. Anhui: IGI Global, 2009. 318 p.
3. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. California: Springer–Verlag, 2009. 746 p.
4. Artificial Intelligence: A Modern Approach.
URL: https://towardsdatascience.com/understanding–the–bias–variance–tradeoff (data zvernennia: 20.12.2022).
5. Opitz D., Maclin R. Popular ensemble methods: An empirical study: journal of Artificial Intelligence Research No. 11. El Segundo, 1999. P. 169–198.
6. Ensemble Methods to Optimize Machine Learning Models.
URL: https://hub.packtpub.com/ensemble–methods–optimize–machine–learning–models (data zvernennia: 20.12.2022).
7. Barmak O. V., Krak Yu. V., Manziuk E. A. Kharakterystyka dlia vyboru modelei u ansambli klasyfikatoriv: naukovyi zhurnal “Problemy prohramuvannia”. Kyiv, 2018. S. 171–179.
8. Understanding the Bias–Variance Tradeoff

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

URL: http://scott.fortmannroe.com/docs/BiasVariance.html (дата звернення: 20.12.2022).
9. Dietterich T., Ensemble Methods in Machine Learning.
URL: http://web.engr.oregonstate.edu/~tgd/publications/mcs–ensembles.pdf (дата звернення: 25.12.2022).
10. Ensemble methods: bagging, boosting and stacking. URL: https://towardsdatascience.com/ensemble–methods–bagging–boosting–andstacking- c9214a10a205 (дата звернення: 27.12.2022).
11. Moretti F., Pizzuti S., Panzieri S., Annunziato M., Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling, Neurocomputing (2015).
12. Kim M.J., Kang D.K., Kim H.B. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction, Expert Syst. Appl. 42 (3) (2015) 1074–1082.
13. Kang S., Cho S., Kang P. Multi-class classification via heterogeneous ensemble of one-class classifiers, Eng. Appl. Artif. Intell. 43 (2015) 35–43.
14. Bustinh i behhinh yak metody formuvannia ansamblei modelei / Verbivskyi D. S., Karpliuk, S. O., Fonariuk, O. V., Sikora, Ya. B. Zhytomyr: ZhDU im. Ivana Fran-ka, 2021. S. 163-169.
15. Zhi–Hua Zhou Ensemble Learning. URL: https://cs.nju.edu.cn/zhouzh/ zhouzh.files/ publication/springerEBR09.pdf (дата звернення: 20.12.2022).
16. Shah B.R., Lipscombe L.L. Clinical diabetes research using data mining: a Canadian perspective, Can. J. Diabetes 39 (3) (2015) 235–238.
17. Improvements on Cross–Validation: The 632+ Bootstrap Method. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10474007#.U2o7MVdMzTo (data zvernennia: 22.12.2022).
18. Ahmad A., Brown G. Random ordinality ensembles: ensemble methods for multi-valued categorical data, Inf. Sci. 296 (2015) 75–94.
19. Sluban B., Lavrac N. Relating ensemble diversity and performance: a study in class noise detection, Neurocomputing 160 (2015) 120–131.
20. Bashir S., Qamar U., Khan F.H. IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework. Journal of Biomedical Informatics. – 2016. – vol.59. – pp.185- 200.
21. E–Commerce Shipping DataSet.
URL: https://www.kaggle.com/datasets/+prachi13/ customer-analytics (дата звернення: 15.11.2022)

22. Kalinina I.O., Hozhyi O.P. Doslidzhennia efektyvnosti metodiv klasyfikatsii pry prohnozuvanni v zadachakh mashynnoho navchannia. Upravlinnia rozvytkom skladnykh system. Kyiv, 2021. № 46. S. 173 – 180, dx.doi.org\10.32347/2412-9933.2021.46.173-180.
23. Maniruzzaman M., Rahman MJ., Ahammed B. Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems, Vol. 8. Texas, 2020. P. 1–14.
24. Bidiuk P.I., Kuznietsova N.V., Terentiev O.M. Systema pidtrymky pryiniattia rishen dlia analizu danykh. Kyiv: Naukovi visti NTUU «KPI», 2011. S. 48–61.
25. Zhan Zh. Introduction to machine learning: k-nearest neighbors. Vienna: Ann Transl Med, 2016. 218 p.
26. Metryky v zadachakh mashynnoho navchannia.
URL: https://habr.com/ru/company/ods/blog/328372 (data zvernennia: 20.12.2022).

Received 05.04.2023.
Accepted 07.04.2023.

Intelligent classification system based on ensemble methods

In the paper, based on machine learning methods, the solution of the classification task was investigated using a two-level structure of ensembles of models. To improve forecasting results, an ensemble approach was used: several basic models were trained to solve the same problem, with subsequent aggregation and improvement of the obtained results. The problem of classification was studied. The architecture of the intelligent classification system is proposed. The system consists of the following components: a subsystem of preprocessing and data analysis, a subsystem of data distribution, a subsystem of building basic models, a subsystem of building and evaluating ensembles of models. A two-level ensemble structure was used to find a compromise between bias and variance inherent in machine learning models. At the first level, an ensemble based on stacking is implemented using a logistic regression model as a metamodel. The predictions that are generated by the underlying models are used as input for training in the first layer. The following basic models of the first layer were chosen: decision trees (DecisionTree), naive Bayesian classifier (NB), quadratic discriminant analysis (QDA), logistic regression (LR), support vector method (SVM), random forest model (RF). The bagging method based on the Bagged CART algorithm was used in the second layer. The algorithm creates N regression trees using M initial training sets and averages the resulting predictions. As the basic models of the second layer, the following were chosen: the first-level model (Stacking LR), the model of artificial neural networks (ANN); the linear discriminant analysis (LDA) model and the nearest neighbor (KNN) model. A study of basic classification mod-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

els and ensemble models based on stacking and bagging, as well as metrics for evaluating the effectiveness of the use of basic classifiers and models of the first and second level, was conducted. The following parameters were determined for all the methods in the work: prediction accuracy and error rate, Kappa statistic, sensitivity and specificity, accuracy and completeness, F-measure and area under the ROC curve. The advantages and effectiveness of the ensemble of models in comparison with each basic model are determined.

Гожий Олександр Петрович – професор кафедри інтелектуальних інформаційних систем Чорноморського національного університету ім. Петра Могили, д.т.н., професор.

Жебко Олександр Олегович – аспірант, Миколаївський Національний аграрний університет.

Калініна Ірина Олександрівна – доцент кафедри інтелектуальних інформаційних систем Чорноморського національного університету ім. Петра Могили, к.т.н., доцент.

Ганніченко Тетяна Анатоліївна – доцент кафедри іноземних мов Миколаївського Національного аграрного університету, к.п.н., доцент.

Gozhyj Oleksandr Petrovych - professor of the Department of Intellectual Information Systems of the Black Sea National University named after Petra Mohyly, Ph.D., professor.

Zhebko Oleksandr Olegovych - PhD student, Mykolaiv National Agrarian University.

Kalinina Iryna Oleksandrivna - Associate Professor of the Department of Intellectual Information Systems of the Black Sea National University named after Petra Mohyly, Ph.D., associate professor.

Hannichenko Tetyana Anatoliivna - associate professor of the Department of Foreign Languages of the Mykolaiv National Agrarian University, PhD, associate professor.