УДК 004. 732.056 **doi: 10.26906/SUNZ.2022.3.079**

О. А. Горносталь, С. Ю. Гавриленко

Національний технічний університет “Харківський політехнічний інститут”, Харків, Україна

МЕТОД ІДЕНТИФІКАЦІЇ СТАНУ КОМП'ЮТЕРНОЇ СИСТЕМИ НА ОСНОВІ АНСАМБЛЕВИХ КЛАСИФІКАТОРІВ З ПОКРАЩЕНОЮ ПРОЦЕДУРОЮ ГОЛОСУВАННЯ

Анотація. Об'єктом дослідження є процес ідентифікації стану комп'ютерної системи. **Предметом дослідження** є методи ідентифікації стану КС. **Метою дослідження** є підвищення якості та швидкодії ансамблевих класифікаторів шляхом оптимізації процедури голосування. **Методи, що використовуються**: методи машинного навчання, ансамблеві класифікатори, метод обрізки ансамблів, процедура зваженого адаптивного голосування. **Отримані результати**: розроблено ансамблевий метод ідентифікації комп'ютерних систем на основі мета-алгоритму бегінг зі спеціальною процедурою зменшення кількості базових класифікаторів та їх ранжування. Досліджено ефективність різних підходів обрізки базових класифікаторів на основі дерев рішень для підвищення якості мета-алгоритму. Розглянуто різні види методів розрахунку вагових коефіцієнтів для реалізації зваженого голосування з використанням різних метрик якості. Експериментальні дослідження дозволили оцінити розглянуті підходи окремо, а також підтвердили ефективність їх комплексного використання. **Висновки**. За результатами дослідження запропоновано удосконалений ансамблевий класифікатор ідентифікації стану комп'ютерної системи на основі мета-алгоритму бегінг, який відрізняється від відомих комплексним використанню методів обрізки базових класифікаторів ансамблів та використанням процедури адаптного зваженим голосуванням За рахунок удосконалення класифікатору вдалося підвищити його точність до 2.5%. Перспективами подальших досліджень може бути підбір та налаштування базових класифікаторів з використанням різних методів машинного навчання.

Ключові слова: класифікація, машинне навчання, ансамблі, бегінг, зважене голосування, обрізка ансамблів, точність, швидкодія.

Вступ

У сучасному інформаційному суспільстві комп'ютерні системи та комп'ютерні мережі стали невід'ємною частиною повсякденної діяльності людей та організацій. Вони відіграють ключову роль у бізнесі, науці, освіті та інших галузях людського життя. При цьому зростання залученості інформаційних технологій у ці сфери призводить до постійного зростання складності та різноманітності мережевих технологій, а також до збільшення обсягу інформації, що передається.

Такі явища, як Інтернет речі (IoT), хмарні обчислення та розподілені мережі, стали звичайними у нашому повсякденному житті, і це призвело до нових викликів у галузі ідентифікації та моніторингу стану комп'ютерних ресурсів. Сучасні системи та мережі стикаються з ризиками кібератак, збоїв, неполадок та перевантажень, що може призвести до серйозних наслідків, наприклад, до простою бізнес-процесів та витоку конфіденційної інформації [1]. Це робить питання ідентифікації стану комп'ютерних систем та мереж критично важливими. Існує гостра необхідність у розробці нових та вдосконаленні існуючих методів ідентифікації та контролю стану комп'ютерних систем та мереж, які дозволять оперативно виявляти та усувати проблеми, що виникають.

Об'єктом дослідження є процес ідентифікації стану комп'ютерної системи.

Предметом дослідження є методи та засоби ідентифікації стану КС.

Огляд пов'язаних наукових публікацій. Процес ідентифікації стану комп'ютерної системи представляє собою процедуру збору інформації про поточний стан та параметри системи та її аналізу з метою забезпечення оптимальної продуктивності та безперебійної роботи. Основні характеристики, які слід враховувати в процесі ідентифікації, включають апаратну конфігурацію, обсяг оперативної пам'яті, використання процесорного ресурсу, наявність актуальних оновлень програмного забезпечення та налаштування мережевих параметрів, інтенсивність використання ресурсів мережі, особливості використання системних функцій та інші [2]. Актуальність цього процесу полягає в забезпеченні ефективності та безпеки роботи системи, вчасному виявленні можливих проблем або вразливостей, а також вдосконаленні ресурсного використання. На процес ідентифікації можуть вплинути зміни в апаратному забезпеченні, програмних оновленнях, налаштуваннях безпеки, зміна завдань, які виконує комп'ютерна система, а також несанкціоноване втручання в роботу комп’ютерної системи. В будь-кому випадку, ці події відображаються в основних характеристиках її роботи.

Для ідентифікації стану комп'ютерної системи використовуються різні методи класифікації, проте попередні дослідження показали, що найбільш якісними є ансамблеві класифікатори. Актуальність використання ансамблевих класифікаторів полягає в їх здатності покращити точність та стійкість моделей шляхом поєднання декількох базових класифікаторів, що допомагає знизити вплив випадкових помилок та підвищити загальну точність прогнозів [3].

Одним з популярних методів ансамблювання є бегінг (bootstrap aggregating method), який використовує вибірки даних, згенеровані методом бутстрепу (вибірка з повторенням), для навчання декількох базових класифікаторів. Цей підхід сприяє зниженню впливу випадкових помилок та підвищенню загальної точності прогнозів. Результати роботи цих класифікаторів потім агрегуються за допомогою голосування або обчислення середньої оцінки, для отримання кінцевого прогнозу [4].

Вибір ансамблевих класифікаторів, таких як бегінг, у порівнянні з іншими методами машинного навчання обумовлений низкою переваг, які роблять його привабливими для багатьох класифікаційних завдань [5]:

1. Стійкість до перенавчання: Однією з основних проблем у машинному навчанні є перенавчання моделей на тренувальних даних, що призводить до поганої узагальнюючої здатності нових даних. Ансамблеві методи, включаючи бегінг, працюють на основі композиції декількох моделей, що зменшує ймовірність перенавчання та підвищує стійкість.

2. Поліпшення стабільності та узагальнюючої здатності: Ансамблеві методи поєднують прогнози кількох базових моделей, що дозволяє знизити вплив аномалій та шуму в даних. Це призводить до більш стабільних і надійних передбачень, які краще узагальнюються на нові дані.

3. Підвищення точності: Комбінування прогнозів кількох моделей може значно покращити точність класифікації. Ансамблеві методи, такі як бегінг, дозволяють знизити дисперсію помилки та підвищити загальну точність моделі.

4. Робота з різними типами базових моделей: Ансамблеві методи можуть застосовуватися з різними базовими моделями, такими як вирішальні дерева, лінійні моделі, метод опорних векторів та інші. Це дозволяє вибирати найкращі базові моделі для конкретного завдання.

5. Масштабованість: Ансамблеві методи легко масштабуються, моделей, що робить їх придатними для аналізу великих обсягів даних та складних завдань.

6. Простота реалізації: Більшість ансамблевих методів, включаючи бегінг, відносно прості в реалізації та потребують незначного налаштування Це спрощує їх використання у практичних завданнях.

Разом взяті, ці переваги роблять ансамблеві методи, такі як бегінг, потужним та універсальним інструментом для вирішення різних завдань класифікації. Однак, як і з будь-яким іншим методом машинного навчання, важливо проводити експерименти з налаштування параметрів для оптимізації продуктивності моделі для конкретнію задачі та конкретних вихідних даних.

Постановка проблеми. Основною метою поточного дослідження є підвищення точності та швидкодії бегінг класифікаторів застосовно до завдання ідентифікації стану комп'ютерних систем. Попередні наші дослідження були пов’язані з попередньою обробкою даних, яка суттєво впливає на точність класифікації стану комп'ютерної системи. Зараз, метою роботи є дослідження ефективності використання різних технік процедури голосування ансамблевих беггінг класифікаторів:

1. Обрізка ансамблів. Цей метод включає визначення оптимального числа базових класифікаторів, які повинні бути включені в ансамбль. Для цього проводиться оцінка якості прогнозів класифікатору на валідаційній вибірці для різного числа класифікаторів та вибирається оптимальна кількість базових моделей ансамблю. Обрізка ансамблів допомагає уникнути перенавчання та підвищує його швидкодію.

2. Зважене голосування з адаптивними вагами. У цьому методі присвоюються різні ваги кожному з базових класифікаторів в ансамблі, ґрунтуючись на їхній продуктивності на навчальній та валідаційній вибірках. Адаптивне присвоєння ваги дозволяє дати більшу вагу більш точним класифікаторам, що покращує загальну точність ансамблю.

3. Голосування з вагами, заснованими на вихідних даних. У цьому методі ваги кожного класифікатора визначаються з урахуванням характеристик вихідних даних, як-от їх складність чи подібність. Це дозволяє гнучкіше налаштовувати внесок кожного класифікатора у підсумкове голосування.

4. Калібрування впевненості. Цей метод спрямований на корекцію впевненості, з якою ансамбль робить прогнози. Це особливо корисно у випадках, коли ансамбль акцентує увагу на неправильних прогнозах. Калібрування впевненості дозволяє покращити якість імовірнісних прогнозів.

5. Метанавчання з метаознаками. У цьому методі використовується другий рівень ансамблю, який навчається на основі передбачень першого рівня. Це навчання дозволяє використовувати інформацію про те, як кожен класифікатор працює на різних прикладах, щоб поліпшити підсумкові передбачення.

За попередньою оцінкою, найкращі результати досягаються при застосуванні зваженого голосування [6] з вагами, що базуються на продуктивності кожного з класифікаторів та техніки обрізки ансамблів з використанням різних оцінок якості класифікаторів [7]. Основне припущення дослідження полягає у комплексному використанні цих методів відносно створення більш збалансованого і точного ансамблю класифікаторів для ідентифікації стану комп'ютерної системи. Комбінуючи їх, прагнемо покращити точність процедури голосування та забезпечити більш надійні результати під час ідентифікації стану комп'ютерної інфраструктури, а також зменшити час роботи отриманої системи ідентифікації за рахунок зменшення кількості класифікаторів.

Огляд підходів та методів

Стандартне просте голосування в бегінг ансамблях передбачає рівний внесок кожного з базових незалежних класифікаторів. Проте визначення вагових коефіцієнтів кожного класифікатору, який залежать від різних оцінок його якості, можуть підвищити точність ансамблю [8]. Рівняння для розрахунку вагового коефіцієнта є таким:

$$w_{di} = \frac{Perf_i}{\sum_{k=1}^{n} Perf_k}, \quad (1)$$

де w_{di} – динамічна (адаптивна) вага і-го класифікатора; $Perf_i$ – оцінка якості і-го класифікатору; $\sum_{k=1}^{n} Perf_k$ – сума оцінок продуктивності усіх класифікаторів.

У якості критерія оцінки якості класифікації можуть використовуватися наступні показники: точність (1), F1 Score (2), функція втрат 0-1 Zero-One Loss function (6), площа під ROC-кривою (ROC AUC Score), функція логарифмічних втрат Log Loss function (7), збалансована точність Balanced Accuracy Score (8).

$$f_{01}(\hat{y}_i, y_1) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\hat{y}_i \neq y_1}, \quad (2)$$

де $f_{01}(\hat{y}_i, y_1)$ – функція втрат 0-1 Zero-One Loss function; n – кількість спостережень; $\hat{y}_i$ – прогнозований вихід класифікатору; y_1 – істинне значення; $\delta_{\hat{y}_i \neq y_1}$ – бінарна функція, яка приймає значення 1, коли класифікатор робить вірну класифікацію.

$$f_{\log loss} = \\ = -\frac{1}{n}\sum_{i=1}^{n} (y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)), \quad (3)$$

де $f_{\log loss}$ – функція логарифмічних втрат; n – кількість спостережень; $\hat{y}_i$ – прогнозований вихід класифікатору; y_i – істинне значення.

$$f_{bas} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right), \quad (4)$$

де f_{bas} – функція збалансованої точності; TP – кількість правильно передбачених позитивних подій; TN – кількість правильно передбачених негативних подій; FP – кількість неправильно передбачених позитивних подій; FN – кількість неправильно передбачених негативних подій.

Основна ідея обрізки ансамблевих класифікаторів полягає в тому, щоб побудувати ансамбль незалежних класифікаторів з використанням мета-алгоритму бегінга, а потім оцінити внесок або продуктивність кожного з класифікаторів у загальний результат і вилучити частину найменш продуктивних із них за певним правилом [9. 10], що зменшить час навчання та ілентифікації КС. Для оцінки продуктивності можна використовувати абсолютну точність та оцінку F1 Score:

$$Acuracy = \frac{TP + TN}{TP + TN + FP + FN}; \quad (5)$$

$$f_1 score = \\ = \frac{2}{\frac{1}{Pr\,e\,cision} + \frac{1}{Re\,c\,all}} = \frac{TP}{TP + 0.5(FP + FN)}. \quad (6)$$

Позначимо обидві характеристики продуктивності як Perf. Тоді функція вибору базового індивідуального класифікатора буде виглядати так: для обрізки ансамблю за певним значенням продуктивності (вираз 3) і для обрізки ансамблю за віддаленістю продуктивності від середнього значення (4).

$$f = Perf(n) > Perf_T, \quad (7)$$

де f – функція, що визначає необхідність використання n-й класифікатору у фінальному голосуванні; Perf (n) – оцінка продуктивності n-го класифікатору; $Perf_T$ – заздалегідь обране порогове значення.

$$f = \left|Perf(n) - Perf_{Avg}\right| \leq Perf_T, \quad (8)$$

де f – функція, що визначає, необхідність використання n-й класифікатору у фінальному голосуванні; Perf (n) – оцінка продуктивності n-го класифікатору; $Perf_{Avg}$ – середнє значення оцінки продуктивності усіх класифікаторів; $Perf_T$ – заздалегідь обране порогове значення. Задавши порогові значення та вибравши за їх допомогою якісні базові моделі, необхідно поєднати їх для фінального голосування.

Формулювання теоретичних очікувань. Основною ідеєю бегінг класифікаторів є побудова великої кількості незалежних базових класифікаторів, які навчаються на різних наборах даних. Саме тому, класифікатори можуть робити різні прогнози. Крім того, на точність класифікації ансамблю може впливати незбалансованість класів. Саме у таких випадках зважене голосування може підвищити точності за рахунок ранжування складників ансамблю за якістю їх роботи. Разом із тим, зважене голосування потребує збільшення часу навчання та, що найголовніше, збільшення часу ідентифікації КС, адже під час голосування необхідно враховувати вагові коефіцієнти кожної моделі.

Вирішити цю проблему можливо за рахунок обрізки ансамблю, що призведе до зменшення кількості базових класифікаторів [11, 12]. Комплексне використання наведених підходів підвищити точність роботи ансамблю при незначному збільшенні часу класифікації.

Експериментальні дослідження та оцінка ефективності

Перша частина експерименту полягає у використанні техніки зваженого голосування для підвищення точності класифікації ансамблевого бегінг класифікатору. У якості вихідних даних, використано штучно згенеровані значення з великим вмістом шумів, що імітують показники функціонування КС. Розроблено програмне забезпечення, яке спочатку навчає стандартний бегінг класифікатор. Надалі виконується аналіз якості кожної із базових моделей та розраховуються для них вагові коефіцієнти на основі різних показників якості їх роботи: точність (accuracy), F1 Score, функція втрат 0-1 (Zero-One Loss function), площа під ROC-кривою (ROC AUC Score), функція логарифмічних втрат (Log Loss function), збалансована точність (Balanced Accuracy Score). Результати дослідження представлені на рис 1 – 3.

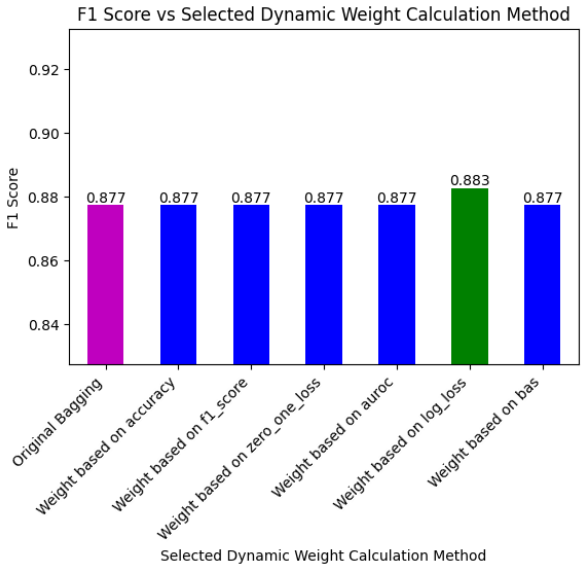

Рис. 1. Залежність метрики F1 Score ансамблю від обраного методу розрахунку вагових коефіцієнтів для голосування

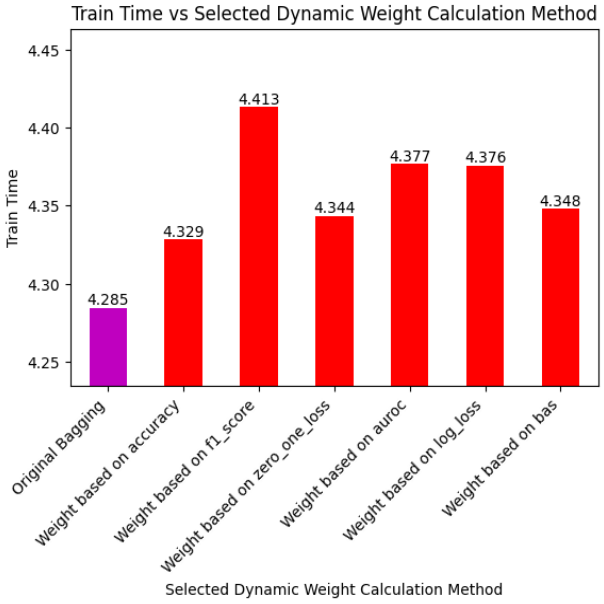

Рис. 2. Залежність часу тренування ансамблю від обраного методу розрахунку вагових коефіцієнтів для голосування

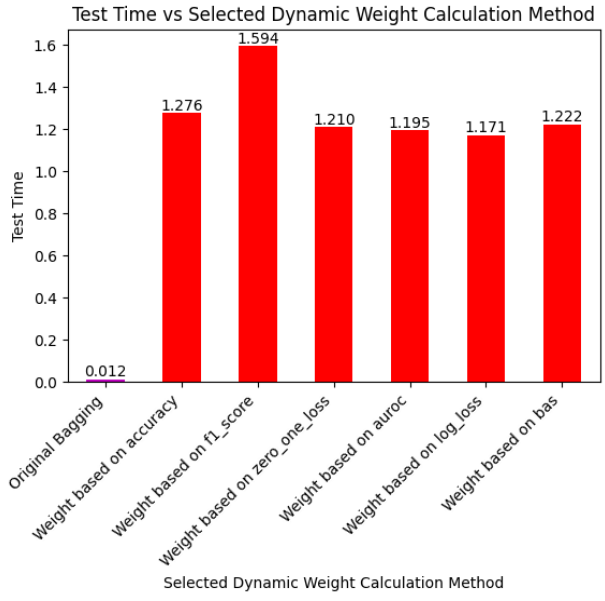

Рис. 3. Залежність часу ідентифікації ансамблю від обраного методу розрахунку вагових коефіцієнтів для голосування

За результатом проведених експериментів отримано, що використання процедури зваженого голосування дозволяє підвищити точність роботи ансамбля на 0.5%, значення F1 Score на 0.6%, значення Presicion на 0.1% та значення Recall на 1%. При цьому, найкращі результати отримано при розрахунку вагових коефіцієнтів на основі методу логарифмічних втрат.

У зв'язку з необхідністю розрахунку вагових коефіцієнтів, збільшуються час тренування та тестування ансамблю (рис.5-6). Це пов'язано з необхідністю врахування вагових коефіцієнтів базових класифікаторів у процесі фінального голосування. Збільшення часу тестування є несуттєвим і складає одну секунду.

Друга частина експерименту полягає у дослідженні впливу різних технік обрізки ансамблю з ранжуванням порогових значень критеріїв прийняття рішень на ефективність роботи ансамблевого класифікатору. Для цього було розглянуто чотири техніки обрізки, де у якості критерія прийняття рішень використовувались такі показники якості класифікації: Accuracy, Presicion, Recall та F_1 score. Результати дослідження представлені на рис. 4 – 6.

За результатом проведених експериментів можна стверджувати, що використання процедури обрізки ансамблю з ранжуванням порогового значення дозволяє підвищити якість класифікації. Вибір оптимального порогу призводить до підвищення точність роботи ансамбля на 2%, значення F1 Score на 2.1%, значення Presicion на 1.9% та значення Recall на 3.8%. При цьому відмічається незначне збільшення часу тренування ансамбля через необхідність видалення певних класифікаторів. З іншого боку – значно зменшується час тестування, що пов'язано зі зменшенням кількості базових класифікаторів. Найкращі результати вдається отримати при використанні обрізки за точністю з пороговими значеннями 0.9062 та 0.9162, а також при використанні обрізки за значенням метрики якості F1 Score з пороговими значеннями 0.8992 та 0.9092.

Третя частина експерименту полягає у дослідженні впливу комплексного використання обрізки ансамблю з ранжуванням критеріїв прийняття рішень і їх порогових значень та технік зваженого голосування. Результати дослідження представлені на рис. 7 – 9.

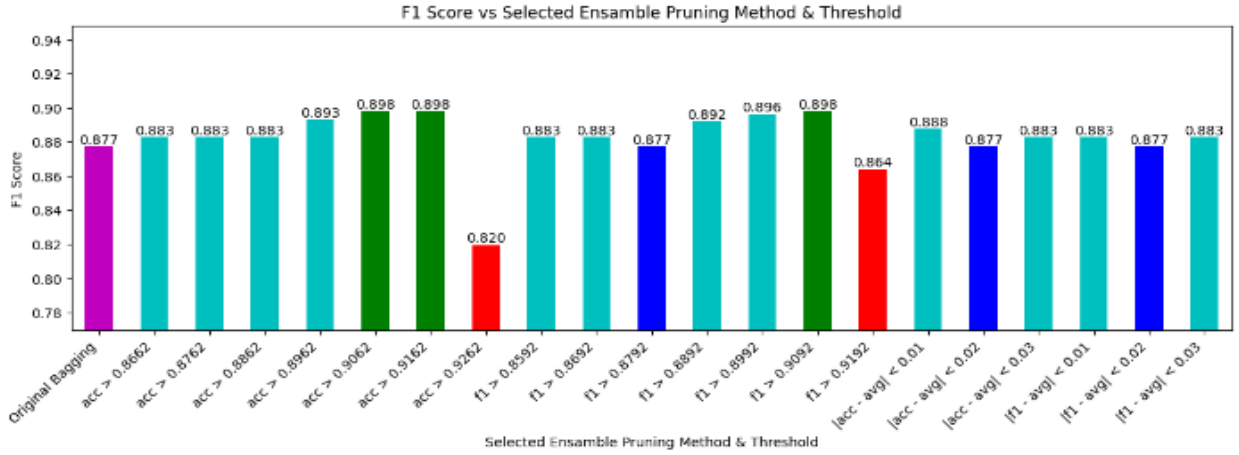

Рис. 4. Залежність метрики F1 Score ансамблю від обраної техніки обрізки та від обраного порогового значення

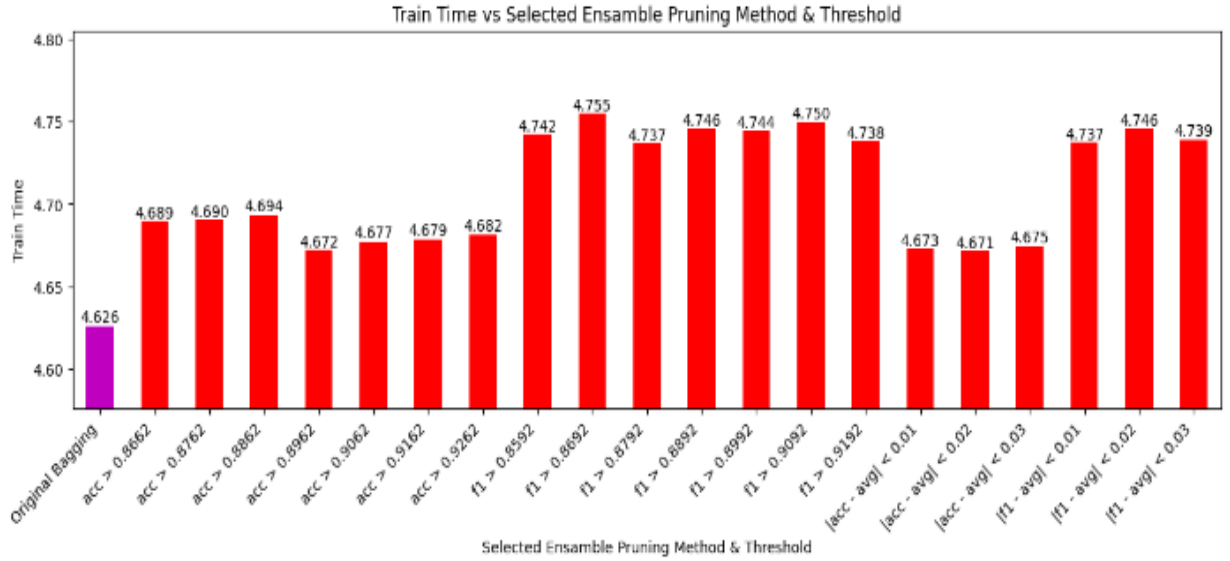

Рис. 5. Залежність часу навчання ансамблю
від обраної техніки обрізки та від обраного порогового значення

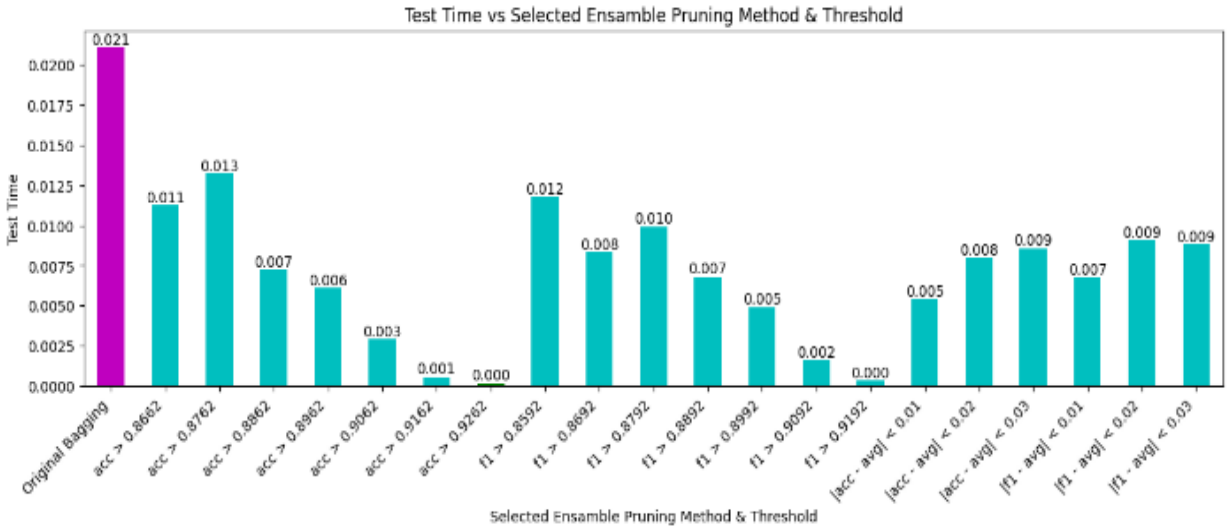

Рис. 6. Залежність часу ідентифікації ансамблю
від обраної техніки обрізки та від обраного порогового значення

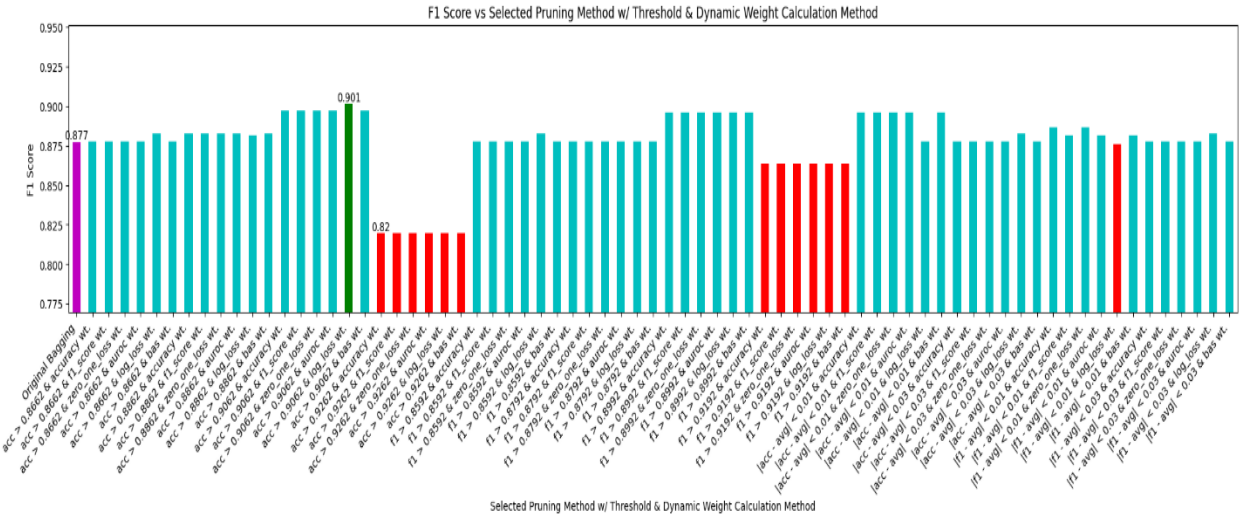

Рис. 7. Залежність значення метрики F1 Score ансамблю від обраної техніки обрізки,
обраного порогового значення та методу розрахунку вагових коефіцієнтів для голосування

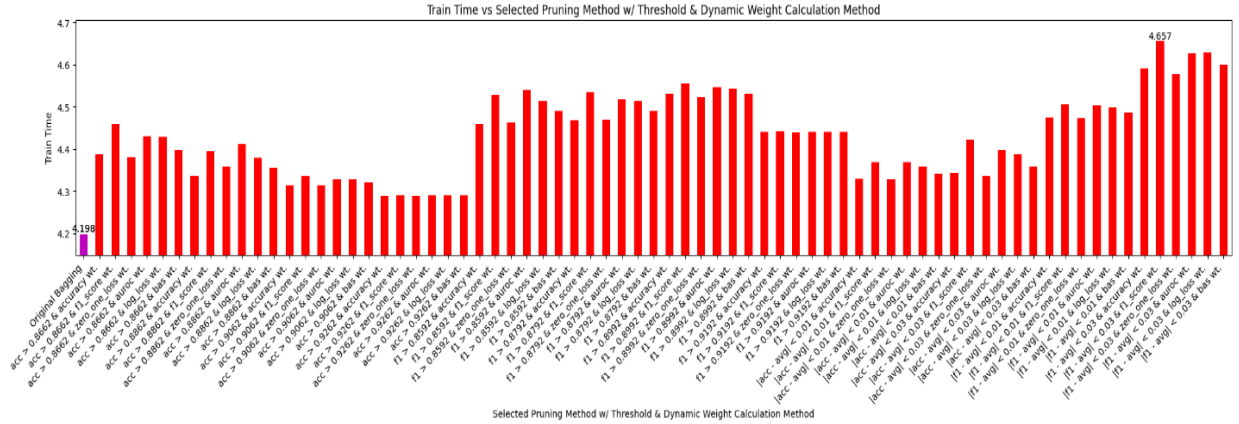

Рис. 8 Залежність часу тренування ансамблю від обраної техніки обрізки, обраного порогового значення та методу розрахунку вагових коефіцієнтів для голосування

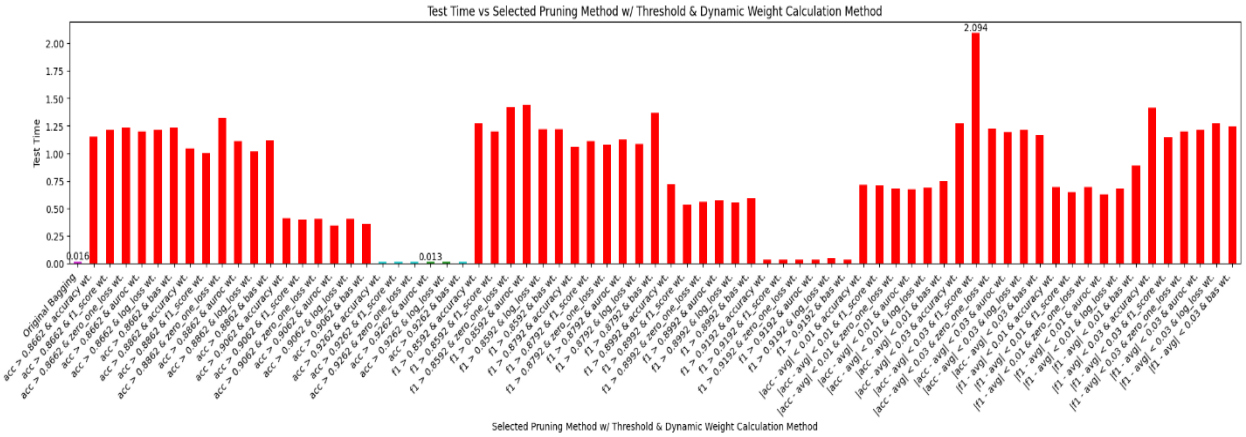

Рис. 9. Залежність часу ідентифікації від обраної техніки обрізки, обраного порогового значення та методу розрахунку вагових коефіцієнтів для голосування

За результатом проведених експериментів отримано, що комплексне використання процедури обрізки ансамблю та зваженого голосуванням дозволяє, в цілому, підвищити показники якості класифікації запропонованого ансамблю: Accuracy – на 2.5%, F1 Score – на 2.4%, Presicion – на 2% та Recall – на 2.8%. При цьому відмічається збільшення часу тренування ансамбля до 1 секунди через необхідність видалення певних класифікаторів та розрахунку вагових коефіцієнтів. З іншого боку, збільшення часу тестування до 1 секунди є не значним та прийнятним для ідентифікації стану комп'ютерних систем з щосекундним збором статистичних даних.

Результати експерименту підтвердили підвищення якості класифікації за рахунок використання процедури зваженого голосування на основі функції логарифмічних втрат та обрізки ансамблю, засновану на використанні точності у якості критерію прийняття рішення з пороговим значенням 0.9062.

Таким чином, процедури обрізки ансамблів з різними техніками та пороговими значеннями, а також методика зваженого голосування дають можливість підвищити ефективність роботи ансамблю. При цьому їх комплексне використання дозволяє отримати максимальне значення метрик якості при допустимому підвищенні часу тренування та часу тестування.

Висновки

У рамках дослідження було проведено аналіз різних підходів до підвищення ефективності бегінг-класифікаторів для ідентифікації стану комп'ютерної системи.

У якості вихідних даних, використано штучно згенеровані значення з підвищеним вмістом шумів, що імітують показники функціонування КС. Було досліджено техніку зваженого голосування з використанням різних метрик якості класифікаторів для оцінки їх ваг та процедура зменшення кількості базових класифікаторів.

Проведені дослідження підтвердили припущення, що зважене динамічне голосування та обрізка ансамблів сприяють підвищенню якості класифікації. Ці методи покликані забезпечити більш надійну та точну ідентифікацію стану комп'ютерної системи, що важливо для своєчасного виявлення аномалій у їх роботі.

Обрізка ансамблів, крім підвищення точності, має ще одну важливу перевагу - скорочення часу розпізнавання на тестовій вибірці. Це особливо актуально у ситуаціях, де швидкодія є критичним чинником успішної ідентифікації стану системи.

З іншого боку, зважене динамічне голосування надає можливість точніше врахувати продуктивність

кожного класифікатора в ансамблі. Особливо важливою є гнучкість у виборі функції оцінки якості, яка може бути адаптована до конкретного завдання.

Комплексне використання обох методів має найбільший позитивний ефект. Груба обрізка ансамблів скорочує розмір ансамблю та прискорює процес класифікації, а зважене динамічне голосування збільшує якість моделі. Такий комплексний підхід забезпечує найкращу ефективність ідентифікації стану комп'ютерної системи.

За результатами дослідження отримано, що комплексне використання методів може бути корисними інструментами для підвищення точності та надійності ідентифікації стану комп'ютерних систем, особливо в умовах наявності даних, що знаходяться на межі розмежування класів. Отримані результати можуть бути важливим керівництвом для практиків та дослідників, які прагнуть підвищити ефективність роботи бегінг-класифікаторів в аналогічних прикладних задачах.

Список літератури

1. Sathiya Devi, S., Rajakumar, R. (2021). Network Intrusion Detection Using Cross-Bagging-Based Stacking Model. In: Pandian, A., Fernando, X., Islam, S.M.S. (eds) Computer Networks, Big Data and IoT. Lecture Notes on Data Engineering and Communications Technologies, vol 66. Springer, Singapore. https://doi.org/10.1007/978-981-16-0965-7_58
2. Челак В. В. Розробка методу ідентифікації стану комп'ютерної системи на основі нечітких дерев рішень / С.Ю. Гавриленко та В.В. Челак // Системи управління, навігації та зв'язку Випуск 1 (71).– Полтава, Україна, 2023.– с. 78-83. https://doi.org/10.26906/SUNZ.2023.1
3. Andrea Campagner, Davide Ciucci, Federico Cabitza, Aggregation models in ensemble learning: A large-scale comparison, Information Fusion, Volume 90, 2023, pp. 241-252. https://doi.org/10.1016/j.inffus.2022.09.015
4. O. Hornostal and S. Gavrylenko, "Development of a method for identification of the state of computer systems based on bagging classifiers", A.I.S., vol. 5, no. 4, pp. 5–9, Dec. 2021. https://doi.org/10.20998/2522-9052.2021.4.01
5. O. Hornostal, S. Gavrylenko and V. Chelak "Ensemble Approach Based on Bagging and Boosting for Identification the Computer System State," 2021 XXXI International Scientific Symposium Metrology and Metrology Assurance (MMA), Sozopol, Bulgaria, 2021, pp. 1-7. https://doi.org/10.1109/MMA52675.2021.9610949
6. J. A. Sáez and J. L. Romero-Béjar, "On the Suitability of Bagging-Based Ensembles with Borderline Label Noise," Mathematics, vol. 10, no. 11, p. 1892, Jun. 2022. https://doi.org/10.3390/math10111892
7. A. Dogan and D. Birant, "A Weighted Majority Voting Ensemble Approach for Classification," 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 2019, pp. 1-6. https://doi.org/10.1109/UBMK.2019.8907028
8. Pinto, F., Soares, C., Mendes-Moreira, J. (2015). Pruning Bagging Ensembles with Metalearning. In: Schwenker, F., Roli, F., Kittler, J. (eds) Multiple Classifier Systems. MCS 2015. Lecture Notes in Computer Science, vol 9132. Springer, Cham. https://doi.org/10.1007/978-3-319-20248-8_6
9. Sannasi Chakravarthy, S.R., Rajaguru, H. (2022). Ensemble-Based Weighted Voting Approach for the Early Diagnosis of Diabetes Mellitus. In: Karrupusamy, P., Balas, V.E., Shi, Y. (eds) Sustainable Comm. Networks and Application. Lecture Notes on Data Engineering and Comm. Technologies, vol 93. Springer, Singapore. https://doi.org/10.1007/978-981-16-6605-6_33
10. Wenyu Zhang, Dongqi Yang, Shuai Zhang, A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring, Expert Systems with Applications, Volume 174, 2021. https://doi.org/10.1016/j.eswa.2021.114744
11. Li, D., Zhang, Z. & Wen, G. Classifier subset selection based on classifier representation and clustering ensemble. *Appl Intell* (2023). https://doi.org/10.1007/s10489-023-04572-x
12. S. Kwak and H. Kim, "Comparison of ensemble pruning methods using Lasso-bagging and WAVE-bagging," Journal of the Korean Data and Information Science Society, vol. 25, no. 6. Korean Data and Information Science Society, pp. 1371–1383, 2014. https://doi.org/10.7465/jkdi.2014.25.6.1371
13. George D.C. Cavalcanti, Luiz S. Oliveira, Thiago J.M. Moura, Guilherme V. Carvalho, Combining diversity measures for ensemble pruning, Pattern Recognition Letters, Volume 74, pp. 38-45, 2016. https://doi.org/10.1016/j.patrec.2016.01.029

Received (Надійшла) 12.06.2023
Accepted for publication (Прийнята до друку) 23.08.2023

Method of identifying the state of a computer system based on ensemble classifiers with an improved voting procedure

Oleksii Hornostal, Svitlana Gavrylenko

Abstract. **The object of research** is the process of identifying the state of the computer system. **The subject of research** is the methods of identifying the state of CS. **The purpose of research** is to improve the quality and performance of ensemble classifiers by optimizing the voting procedure. **Methods used**: machine learning methods, ensemble classifiers, ensemble pruning method, weighted adaptive voting procedure. **The results were obtained**: an ensemble method of identification of computer systems based on the begging meta-algorithm with a special procedure for reducing the number of basic classifiers and their ranking was developed. The effectiveness of various approaches to pruning basic classifiers based on decision trees to improve the quality of the meta-algorithm was investigated. Different types of methods for calculating weighting coefficients for the implementation of weighted voting using various quality metrics are considered. Experimental studies allowed to evaluate the considered approaches separately, and also confirmed the effectiveness of their integrated use. **Conclusions**. Based on the results of the research, an improved ensemble classifier for identifying the state of the computer system based on the begging meta-algorithm is proposed, which differs from the known ones in the complex use of pruning methods of basic ensemble classifiers and the use of the adaptive weighted voting procedure. Due to the improvement of the classifier, it was possible to increase its accuracy to 2.5%. Prospects for further research may be the selection and adjustment of basic classifiers using various machine learning methods.

Keywords: classification, machine learning, ensembles, begging, weighted voting, ensemble pruning, accuracy, performance.