https://doi.org/10.31891/2219-9365-2025-81-5
УДК 004.9

ГУЛІЄВ Нурал
Харківський національний університет радіоелектроніки
https://orcid.org/0000-0003-2123-0377
e-mail: nural.huliiev@nure.ua

ДОСЛІДЖЕННЯ МЕТОДІВ ПОБУДОВИ ДЕРЕВ РІШЕНЬ ДЛЯ РЕАЛІЗАЦІЇ АЛГОРИТМУ ВИПАДКОВОГО ЛІСУ В МЕДИЧНІЙ ГАЛУЗІ

В роботі висвітлено алгоритм дерев рішень, їхні переваги та недоліки, приклади їх застосування, проаналізовано різні способи їх побудови.

Ключові слова: випадковий ліс, гіпотиреоз, гіпертиреоз, дерева рішень, психологічні розлади, класифікація, лінійна адитивна згортка, прогнозування.

HULIIEV Nural
Kharkiv National University of Radio Electronics

STUDY OF METHODS FOR CONSTRUCTING DECISION TREES FOR THE IMPLEMENTATION OF THE RANDOM FOREST ALGORITHM IN THE MEDICAL FIELD

Every year, machine learning is increasingly facilitating areas of modern life, ranging from entertainment services to solving difficult tasks related to improving people's work and lives. It is especially important now to apply these analysis methods in the medical field in order to save as many lives as possible, to diagnose diseases as early as possible for easier or timelier treatment. This work is devoted to the same topic, which aims to develop methods to prevent the development of psychological disorders among patients with hypothyroidism and hyperthyroidism. In one of the observations of this topic, it was determined that the best way to predict possible difficulties is the random forest algorithm, which consists in building various decision trees. It is worth noting that it is necessary to choose the right way to develop each tree from such alternatives as ID3, CHAID, C4.5, CART, and XGBoost. All of them were analyzed using linear additive convolution based on such data as the type of tree building algorithm, data distribution criterion, data types, numerical data processing, tree pruning method, tendency to overlearn, algorithm speed, how clear the model interpretation is, and application methods. First, a table was filled out with data from decision tree methods according to the above-mentioned features, then all qualitative indicators were converted into quantitative ones for mathematical calculations when calculating convolution values for each alternative. According to the results of the experiment, greedy CART is the best algorithm for developing a decision tree model that is easy to interpret, fast, and least prone to overtraining, operates with numerical and categorical data, uses the Gini index to divide data into subsets when determining the next attribute from the list of features, and supports pruning of its structure. After conducting the experiment, the advantages and disadvantages of the chosen model of the multicriteria problem of this work are also considered.

Keywords: random forest, hypothyroidism, hyperthyroidism, decision trees, psychological disorders, classification, linear additive convolution, prediction.

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Машинне навчання стає широко розповсюдженим з кожним днем, особливо в галузі фінансів, маркетингу, охорони здоров'я, медицині, а з кожним роком їхня точність та потужність покращуються із розвитком технологій, тому в майбутньому прийняття обґрунтованих рішень стане дедалі простіше та переконливіше.

Дерева рішень є одним із будівельних елементів алгоритмів машинного навчання. Наприклад, у методах прогнозування: випадкові ліси та пакетування – застосовуються ансамблі дерев. Також дерева рішень супроводжуються засобами аналітики та візуальною діагностикою. Навчання на основі дерев використовується в прикладних науках та суміжних дисциплінах. Але варто пам'ятати, що достовірність емпіричних результатів залежить від точності. Дерева рішень теж намагаються слідувати цьому, тому гарантії статистичних даних важко надавати з упевненістю. Проблема полягає в тому, що рекурсивний прохід унеможливлює застосування дерева для аналізу. Незважаючи на вище зазначені недоліки, існують способи адаптування дерев до розрідженості прогнозної моделі [1].

Дерево рішень – метод прийняття рішень, модель якого має форму дерева, який оперує різними умовами: кожен внутрішній вузол відповідає за можливі альтернативи, а гілка – за одну з її значень, а кожен вузол листа містить один із класів.

Метод дерев рішень – сильний інструмент для класифікації, прогнозування, інтерпретації, який використовується в медичних дослідженнях. Вони мають такі переваги:

- легкі для розуміння та інтерпретації;
- надійні до викидів;

– легка обробка відсутніх та спотворених значень без необхідності перетворення даних;
– спрощують складні взаємозв'язки шляхом поділу множин на підмножини на основі атрибутів.

Але, не дивлячись на вище описані плюси моделі, вона також має деякі обмеження: через сильну кореляцію вхідних даних, обирається показник, який розділятиме дані, але не є причинно-наслідковим параметром для остаточного результату, та маленький обсяг даних не допоможе побудувати потужне дерево. Тому слід бути конче уважним під час розробки моделі дерев рішень [2].

Алгоритми на основі дерев рішень завжди містять класифікацію, регресію, вибір ознак, не обмежуючись тільки ними. Ідеєю цього підходу є рекурсивний розділ множин на основі атрибутів на підмножини до моменту досягнення критерію зупинки. В результаті вузли побудованого дерева є рішеннями на основі атрибутів. Кращий та більш точний спосіб розбити дані за певними ознаками досягається такими методами, як індекс Джині, коефіцієнт виграшу, інформаційний приріст.

Розробка моделі прогнозування або класифікації на основі алгоритму дерев рішень проходить через такі кроки, як планування, визначення вибірки, збір даних, розподіл даних на навчальні та тестові, застосування отриманої моделі, обробка результатів.

Одним із найскладніших моментів у будуванні дерев рішень є визначення похибки апроксимації результатів. Дослідження пов'язують значення помилки прогнозування із розміром вузлів та доводять, що вона зменшується або повністю зникає з глибиною дерева, що, у свою чергу, зменшує також помилку апроксимації. Інші спостереження модифікують функцію регресії, не беручи до уваги діаметр вузлів, при чому ефект подібний. Недоліком цих методів є те, що побудова дерев вимагає припущень. Щоб усунути цей мінус, можна використати те, що ймовірно помилка прогнозування обмежена помилкою навчання. Тому можна уникнути значення діаметрів вузлів як показника, від якого залежить значення помилки апроксимації, та пов'язати похибку навчання із даними, наприклад, коефіцієнт кореляції Пірсона, що полегшує процес навчання та дає можливість отримати більш точний результат.

Наведемо приклад задачі, де доцільним буде застосування дерев рішень. Класифікаційна задача: маємо навчальну вибірку з *n* спостережень класу *Y*, яка приймає значення від 1 до *k*, та зі змінними x_1, …, x_p. Метою задачі є проектування моделі, яка передбачатиме значення *Y* на основі значень *x*. Рішення таке: розбиття простору *x* на *k* непересічних множин A_1, …, A_k, так, що передбачуване значення *Y* дорівнює *j*, якщо *x* належить до *Aj*, для $j = 1, \ldots, k$. У випадку, коли значення змінних впорядковані, рішеннями поставленої задачі є застосування дискримінантного аналізу та методу класифікацій найближчого сусіда, які дають множини *Aj* з нелінійними та кусково-лінійними межами, які важко інтерпретувати через велике значення *p*. Методи дерев класифікацій видають прямокутні множини *Aj* за рахунок рекурсивного розбиття набору даних однієї змінної *x*, що полегшує інтерпретацію наборів вхідних показників. Застосування дерев рішень до будь-якої кількості змінних є їхньою перевагою.

Перейдемо до аналізу існуючих спостережень застосування алгоритмів дерев рішень.

АНАЛІЗ ДОСЛІДЖЕНЬ ТА ПУБЛІКАЦІЙ

Машинне навчання відіграє важливу роль у визначенні багатьох відповідей з усіх сфер нашого життя. Методи на основі дерев рішень стали більш поширеними серед алгоритмів машинного навчання через свої зрозумілість та простоту. Такими алгоритмами є ітеративний дихотомізатор 3 (ID3), C4.5, C5.0, CART, CHAID, а також ансамблеві: випадковий та ротаційний ліси. Дані способи аналізу застосовуються в медичній сфері та виявленні шахрайства.

Квінлан розробив метод побудови дерев рішень ID3 у 1986 році, який базувався на алгоритмі Ханта. Новий спосіб полягає у двох кроках: побудові та обрізанні дерева. Перший вимагає сортування у вузлах для вибору найкращої ознаки для поділу даних на підмножини. Якщо вузол нелистяний, то він відповідає за певний атрибут, за яким відбуватиметься черговий поділ на підвибірки, а дуги – його значення. Було проведено дослідження за допомогою даного алгоритму із застосуванням інформаційного приросту та ентропії, яке прогнозувало погодні умови для тенісного матчу на відкритому повітрі. Можливими результатами були сонячно, хмарно та дощ на основі температурних умов, показників вологості, вітру [3].

Метою одного із спостережень було розроблення прикладної моделі, яка в змозі аналізувати дані задля розподілу учнів на класи: тих, хто заслуговує на індонезійську смарт-картку, та тих, хто не має на неї права. Картка дає можливість фінансування шкільної освіти дітям з бідних сімей. Підставою для створення програми розподілу стала похибка, яка спричинила надання картки дітям із багатих сімей, коли учні із сімей без фінансового процвітання не отримали право на програму. Тому необхідно було побудувати правильно модель аналізу задля справедливого розподілу коштів серед дітей, хто справді потребує цього. У дослідженні було застосовано метод дерева рішень алгоритм C4.5.

C4.5 алгоритм – це алгоритм класифікації та прогнозування даних на основі дерев рішень, який найбільш популярний з-поміж інших методів передбачення та класифікації.

Перевагами цього способу є наступні:

– відсутність значення параметру не зупиняє будування дерева рішень;
– змога обробляти дискретні та числові дані;
– простота інтерпретації правил та результатів;
– швидка робота.

Дані збирали у Джатібарангу, Бребес: записи 300 учнів, з яких 240 – навчальна вибірка, а 60 – тестова. Результати дослідження були точними на 97%. Зрештою, бідні діти отримали право на безкоштовне навчання [4].

Як відомо, COVID-19 стала причиною збільшення рівня смертності по всьому світу, застосування моделі класифікації смертності пацієнтів може стати у нагоді задля зменшення цього показника. Проводилося відповідне дослідження в Ірані, де смертність визначалася методом випадкового лісу, а класифікація – за допомогою дерев регресії та класифікації (CART), логістичної моделі, алгоритмів C4.5 і C5.0. Аналіз розроблених дерев показала, що для дерева CART значення показників на основі тестових даних F1-score, Accuracy та Recall 0.8681, 0.7824 і 0.955, а навчальної вибірки – відповідно 0.8667, 0.7834, 0.9385. На всіх етапах лікування визначення відповіді на основі точної моделі є вкрай важливим моментом, в цьому спостереженні саме алгоритм CART якнайкраще надає точний діагностичний результат приналежності пацієнта до певного класу смертності через COVID-19, а також інтерпретує фактори ризику на основі демографічних, клінічних та лабораторних показників [5].

Алгоритм CART – структура, яка застосовується для класифікації у вигляді дерев рішень із множини нерегулярних та невпорядкованих даних. Дивлячись на інші методи класифікації, він краще тим, що він інтуїтивно зрозумілий та простий, швидкий та спроможний видати точний прогноз на основі великого обсягу даних. Але в дослідженні змін цін на житло будування даного дерева ускладнилося за рахунок чутливості та непередбачуваності цін. Спостереження проаналізувало фактори, від яких залежать ціни, відсортувало результати, а розроблена модель дерева застосовувалося для прогнозування вартості за металевий молібден та фактичних цін на житло. Середня абсолютна похибка становила 4.03%, точність тенденції становила приблизно 94.8%, що доводить, що побудоване дерево рішень точне та надійне [6].

Алгоритм CART застосовувався задля прогнозування випуску студентів: його метою було передбачення випуску студентів з університету Пакуана факультету інформатики. Результатом спостереження було побудоване дерево, яке визначало класи за допомогою таких критеріїв, як вчасне та невчасне проходження тестування або атестації. Точність класифікації, яка становила 77.5%, оцінювалася методом розробки матриці помилок техніки перевірки [7].

Також було проведено дослідження того, чи можливе використання способу, заснованого на виявленні знань, прогнозування втрати ваги серед людей із раком голови та шиї, які проходять курс променевої терапії, будуванням дерева класифікації та регресії.

Дані з 2007 по 2015 рік про пацієнтів було зібрано із бази даних Oncospace. Спостереження дійшло до таких висновків:

– аналіз демографічних даних, визначеної дози терапії, цільового об’єму, органів групи ризику показав, що можлива втрата більше 5 кг через 3 місяці променевої терапії;
– цей самий результат отримано за допомогою додаткових даних про якість життя та токсичність лікування може виникнути в кінці лікування.

Для 391 ідентифікованого пацієнта факторами ризику при плануванні променевої терапії були доза на верхні скорочувальні, жувальні та привушні м’язи, діагноз за Міжнародною класифікацією хвороб і вік. А в кінці лікування прогностичними критеріями стали нудота, біль, доза на гортань, привушні м’язи, пероральний прийом, діагноз, відстань між гортанню та цільовим об’ємом. Площею під кривою спостережень під час етапу променевої терапії та після неї в кінці лікування були відповідно 0.773 та 0.821.

Точність передбачення була досягнута за рахунок додаткових даних та може застосовуватись в якості розвитку системи здоров’я [8].

Відомо, що учні після середньої школи вступають в коледжі для продовження набуття необхідних знань задля майбутньої професії, але не всі йдуть таким шляхом – деякі зупиняються. Було проведено дослідження з вибіркою з 25 учнів, метою якого було визначити, чи вступатиме студент у вищий заклад на основі таких факторів, як підтримка батьків, конкуренція, стипендії, самомотивація та можливість працевлаштування. Базуючись на отриманих вхідних даних, було розроблено дерево рішень за допомогою алгоритму C4.5, яке довело, що при наявній стипендії та гарантій знайти роботу за обраною професією, учень може стати студентом коледжу [9].

Алгоритм CHAID відомий тим, що є керованим, тому його можна адаптовувати для вирішення будь-яких поставлених задач. Адже він добре оперує із нелінійними зв’язками та гарантує стабільність моделі, але іноді визначити його точність. Було проведено дослідження, де аналізованими даними були показники задоволеність клієнтів транспортних засобів для порівняння різних алгоритмів будування дерев рішень, а саме вирахування їхніх значень точності. В результаті було встановлено, що CHAID добре аналізує дані, тому впевнено визначає корельовані фактори [10].

Кожен із вище згаданих методів може стати у нагоді в процесі дослідження розробки методів попередження розвитку психологічних розладів серед пацієнтів, хворих на гіпотиреоз та гіпертиреоз.

ФОРМУЛЮВАННЯ ЦІЛЕЙ СТАТТІ

Метою роботи є: дослідження методів дерев рішень задля розв'язку багатокритеріальної задачі, яка полягає у виборі найкращого та найоптимальнішого способу для розробки алгоритму випадкового лісу. Необхідно обрати швидкий та легко інтерпретований, менш схильний до перенавчання алгоритм, яка спроможний оперувати із великим обсягом даних. Кожна із моделей має свої переваги та недоліки, тому слід провести експеримент та визначити, яка з них найбільш підходяща для даної задачі.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Розв'язуючи багатокритеріальні задачі, результатом завжди є кращі альтернативи, які відповідають поставленим вимогам. Найчастіше тут використовуються методи двох видів: перша полягає у виключенні кількості критеріїв оцінки, а друга зменшує кількість варіантів аналізу на його початку. Для нашого дослідження найбільш підходящим є саме метод із першої групи. Такими способами є метод граничних та головного критеріїв, відстані та згорток.

Методи згорток поділяються на лінійні адитивні, мультиплікативні та максимінні. Метою застосування згорток є узагальнення усіх критеріїв аналізу.

Адитивна розраховується за наступною формулою:

$$K(x) = \sum_{j=1}^{n} a_j K_j(x) \quad (1),$$

де K(x) – загальний критерій для альтернативи $x \in X$, ($K_1(x), \dots K_i(x), \dots K_n(x)$) – набір вихідних критеріїв,

n – число вихідних критеріїв,

a_i – нормуючий множник, який вказує на вагу альтернативи.

Найкращий із усіх можливих альтернатив задачі обчислюється за допомогою наступної формули:

$$x^* = arg \max_{x \in X} K(x) \quad (2).$$

Тобто результатом є найбільше значення, отримане методом згортки.

Мультиплікативна згортка розраховується за допомогою такої формули:

$$K(x) = \Pi_{j=1}^{n} K_j^{a_j}(x) \quad (3).$$

Максимінна згортка знаходиться за формулою:

$$K(x) = \max_i \min_j a_{ij} K_j(x) \quad (4).$$

Найкращі результати за мультиплікативною та максимінною згортками обчислюється за формулою (2).

Метод граничних критеріїв застосовується в задачах проектування і планування, в яких порогові значення критеріїв набувають значень $k_j(x) \geq k_{j0};\ j = 1 \dots, n$. Формула обчислення цього способу наступна:

$$K(x) = \min_j \left(\frac{K_j(x)}{K_{j0}(x)} \right) \quad (5).$$

Найкращий результат обирається формулою 2.

Метод відстані використовує відстань, яка є додатковою метрикою. Наприклад, для вибору ідеального рішення цілком достатньо інформації ($K_0, \dots, K_{0n}$). Обчислимо відстань до значення максимуму d(x) для кожної альтернативи. Тоді найкраща альтернатива буде відомою із застосуванням формули:

$$x^* = arg \min_{x \in X} d(x) \quad (6).$$

Застосування методу з першої множини іноді вимагає один із способи із другої, наприклад – принцип Парето: альтернативи, які за всіма критеріями програють іншому або іншим варіантам, видаляються до початку дослідження.

Також бувають випадки, в яких параметри, які неконтрольовані через різні причини, ускладнюють будову моделі для подальшого аналізу. Тут у нагоді може стати метод гарантованого результату, мета якого полягає у визначенні найгіршої реакції та гарантованого значення.

Для спостереження варто використати згортку, адже важко визначити порогові значення критеріїв аналізу, а саме лінійну адитивну, яка є найпоширенішою та найпростішою, та метод із другої множини способів – принцип Парето, якщо одна із альтернатив прозоро гірша за інші.

Спершу необхідно обрати критерії, за якими проводитиметься дослідження: альтернативи порівнюватимуться за допомогою цих ознак.

Значення кожної з них може мати як кількісне, так і якісне походження. Згортка оперує із першими, тому у випадку других, необхідно конвертувати їх у кількісні та побудувати нову таблицю вхідних даних варіантів аналізу.

На третьому кроці виключатимемо альтернативи за допомогою принципу Парето, якщо усі її показники за усіма її ознаками менші з-поміж інших можливих значень інших варіантів експерименту. Варто зазначити, якщо показники альтернатив в різних проміжках або мірах вимірювання, необхідно нормалізувати дані максимізацією або мінімізацією даних, щоб точність та коректність результатів відповідала дійсності.

Четвертим етапом є ранжування показників – обчислення вагомих коефіцієнтів. Існують різні способи, в даному спостереженні можна застосувати один із найбільш популярних методів: для кожного критерію один ділитимемо на суму усіх її значень.

Останнім етапом залишається обчислення значення згортки для кожної із альтернатив: розрахунок суми добутків кожної пари значень вагових коефіцієнтів та критеріїв.

Проведемо експеримент та проаналізуємо способи розробки моделей дерев рішень.

Задача дослідження визначити найкращий та найпідходящий метод будування дерев рішень для алгоритму випадкового лісу, який застосовуватиметься для прогнозування можливого розвитку психологічних розладів у людей з гіпотиреозом та гіпертиреозом.

Заповнимо таблицю (див. табл. 1) з даними про методи побудов дерев рішень: ID3, C4.5, CART, CHAID, XGBoost – за такими критеріями:

- вид алгоритму побудови дерева;
- критерій розподілу даних;
- типи даних;
- обробка числових даних;
- метод обрізки дерева;
- схильність до перенавчання;
- швидкість алгоритму;
- на скільки інтерпретація моделі зрозуміла;
- застосування.

Переведемо якісні показники ознак у кількісні. Вид алгоритму побудов дерев рішень в таблиці (див. табл. 1) набуває трьох значень: жадібний, статистичний метод, на основі χ^2 та градієнтне підсилення – які переводитимуться в 1, 2 та 3 бали відповідно.

Для даних алгоритмів застосовуються такі критерії розподілу даних, як інформаційний приріст, коригований інформаційний приріст, індекс Джині та статистичний тест χ^2, оцінка кожного з яких 1, 2, 3 та 4 відповідно.

Якщо метод дерев рішень не обробляє числові дані, то він матиме найменше значення – 0, якщо ж обробляє, то – 1, а якщо ще й перетворює числові значення на категорії, то це – 2 бали.

Типи даних для всіх способів побудови однакові, тому цю ознаку видаляємо для подальшого спостереження.

Якщо алгоритм у змозі обрізати вузли будь-яким способом, то оцінка – 1, але якщо він не підтримує цю можливість, тоді – 0.

У випадку, коли алгоритм найменш схильний до перенавчання, в такому випадку ця категорія набуває значення 4 для нього, коли менш схильний до перенавчання, то – 3, якщо схильність до перенавчання залежить від регуляризації та підсилення або від налаштувань тесту χ^2, то це 2, а коли все ж схильний до перенавчання – 1.

Дивлячись на значення швидкостей методів розробки алгоритмів побудови моделей дерев рішень, маємо наступне:

- повільний (через різні причини) – 1 бали,
- швидкий – 2 бали,
- швидкий, з великою кількістю оптимізацій – 3 бали.

Таблиця 1

Дані методів будування дерев рішень

Критерій/ Алгоритм	Алгоритм побудови дерева	Критерій розподілу	Типи даних	Обробка числових даних	Метод обрізки дерева	Схильність до перенавчання	Швидкість побудови дерева	Розуміння моделі	Застосування
ID3	Жадібний (greedy)	Інформаційний приріст (Information Gain)	Категоріальні	Не обробляє числові дані	Не підтримує обрізку	Схильний до перенавчання	Швидкий	Легко інтерпретувати	Просте застосування для простих задач
C4.5	Жадібний (greedy)	Інформаційний приріст, коригований (Gain Ratio)	Категоріальні та числові	Перетворює числові значення на категорії	Підтримує обрізку за допомогою перехресної перевірки	Менше схильний до перенавчання	Повільніший, через додаткові етапи	Легко інтерпретувати, але з більшими параметрами	Використовується для складних завдань, де є категорії та числові значення
	Жадібний (greedy)	Індекс Джині (Gini Index)	Категоріальні та числові	Перетворює числові значення на категорії	Підтримує обрізку за допомогою методу мінімізації похибки	Найменше схильний до перенавчання	Швидкий, з великою кількістю оптимізацій	Легко інтерпретувати, але важче налаштувати	Підходить для задач, де важлива точність і контролювання перенавчання
CHAID	Статистичний метод, на основі χ^2	Статистичний тест χ^2	Категоріальні	Перетворює числові значення на категорії	Підтримує обрізку за допомогою статистичних тестів	Залежить від налаштувань тесту χ^2	Повільний, через статистичний аналіз	Може бути складним для інтерпретації через статистику	Найбільше використовується в соціальних науках та маркетингу
XGBoost	Градієнтне підсилення (Gradient Boosting)	Індекс Джині (Gini Index) або ентропія	Категоріальні та числові	Обробляє числові дані	Обрізається через регуляризацію та контроль переобучення	Мінімальна завдяки регуляризації та підсиленню	Повільніший через багато ітерацій градієнтного підсилення	Важче інтерпретувати через багатократне підсилення та регуляризацію	Використовується в задачах з великими даними для класифікації, регресії та прогнозування

Якщо метод важко інтерпретувати, то ця категорія для нього дорівнюватиме – 1, якщо ж легко інтерпретувати, але з деякими умовами, то – 2, а якщо просто легко – 3.

Якщо алгоритм використовується в задачах з великими даними для класифікації, регресії та прогнозування, то становить 4 бали, якщо він використовується для складних завдань, де є категорії та числові значення або підходить для задач, де важлива точність і контролювання перенавчання, то це – 3 бали, коли спосіб побудови дерев рішень найбільш використовується в соціальних науках та маркетингу – 2 бали, а якщо він призначений для простих задач – 1 бал.

Побудуємо таблицю з кількісними показниками (див. табл. 2).

Таблиця 2

Кількісні показники методів дерев рішень

Критерій/ Алгоритм	**Алгоритм побудови дерева**	**Критерій розподілу**	**Типи даних**	**Обробка числових даних**	**Метод обрізки дерева**	**Схильність до перенавчання**	**Швидкість побудови дерева**	**Розуміння моделі**	**Застосування**
ID3	1)	1	0	0	1	2	3	1	**ID3**
C4.5	1	2	2	1	3	1	2	3	**C4.5**
CART	1	3	2	1	4	3	2	3	**CART**
CHAID	2	4	2	1	2	1	1	2	**CHAID**
XGBoost	3	3	1	1	2	1	1	4	XGBoost

На даному кроці можемо застосувати принцип Парето, адже алгоритм C4.5 програє за всіма значеннями критеріїв алгоритму CART, але залишимо його для подальшого дослідження задля отримання більш повної картини.

Розрахуємо значення лінійної адитивної згортки для кожного із альтернатив та матимемо наступну таблицю (див. табл. 3).

Таблиця 3

Розраховані згортки методів дерев рішень

Алгоритм	Значення згортки
ID3	0,94551282
C4.5	1,64255189
CART	2,0528083
CHAID	1,65003053
XGBoost	1,70909646

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

Як бачимо, за результатами лінійної адитивної згортки алгоритм CART – найбільш оптимальний спосіб побудови дерев рішень для моделей випадкового лісу.

Методологія побудови дерев полягає в залежності як від вхідних, так і вихідних даних, що є сприятливим та корисним для алгоритму CART, але може ускладнити математичні розрахунки. Навіть будучи широковживаним засобом аналізу, не багато теоретичних досліджень, але якщо кількість досліджень незначна, теми вичерпні та суттєво важливі для подальших експериментів. Наприклад, для встановлення узгодженості випадкових лісів Бреймана у випадку моделей адитивної регресії досліджують асимптотичні властивості CART в режимі фіксованої розмірності [11].

Зараз точно відомо, що в певних умовах він може ідентифікувати низько розмірну структуру даних, до якої зможе адаптуватись, а саме коли модель є розрідженою: вихід дерева залежить від невеликої кількості вхідних параметрів, що у свою чергу вирішує проблему прокляття розмірності.

Переваги:

– легко інтерпретувати та візуалізувати дерево рішень, що допомагає пояснити прогнози іншим сторонам;
– не працює із припущеннями розподілу даних або взаємозв'язків між значеннями;
– обробляє категоріальні та числові вхідні дані;
– вловлює складно зв'язані показники;
– допомагає визначати важливі ознаки моделі задля того, щоб показати, які саме атрибути найбільш повпливали на результат прогнозу;
– ідеально підходить як для класифікації, так і регресії.

Недоліки:

– є ймовірність перенавчання через занадто складну будову дерева або пристосованість до шуму в даних;
– модель може бути чутливою до даних, що може спричинити різні версії одного і того ж дерева через нестабільні дані;
– якщо кількість даних класу надмірно велика порівняно з іншими у вибірці, то може спостерігатись упередженість до цього домінуючого класу [12-13].

Алгоритм CART застосовує жадібний підхід: він шукає якомога оптимальні розв'язки, що, у свою чергу, може ускладнити розробку найпідходящого дерева рішень.

Отже, найменш схильний до перенавчання жадібний алгоритм CART швидко обробляє категоріальні та числові вхідні дані, розділяючи їх на підмножини за допомогою Індексу Джині, перетворюючи другий тип значень на перший, обрізаючи вузли засобами мінімізації, надає легко інтерпретований результат. Також він підходить для задач, де важлива точність і контролювання перенавчання.

Література

1. Огляд дерев рішень: концепції, алгоритми та застосування [Електронний ресурс] // ResearchGate. – Режим доступу: https://www.researchgate.net/publication/381564302
2. Топуз Е., Караманіс С. Алгоритми дерев рішень у медичній діагностиці [Електронний ресурс] // PMC. – Режим доступу: https://pmc.ncbi.nlm.nih.gov/articles/PMC4466856/
3. Ітеративний дихотомізатор 3 (ID3) дерево рішень: алгоритм машинного навчання для класифікації даних [Електронний ресурс] // ResearchGate. – Режим доступу: https://www.researchgate.net/publication/341884258
4. Дві А.С., Сугито С. Метод дерева рішень із алгоритмом C4.5 для класифікації студентів, які мають право отримувати індонезійську смарт-карту (KIP) [Електронний ресурс] // ResearchGate. – Режим доступу: https://www.researchgate.net/publication/343512262
5. Дерева рішень у сфері охорони здоров'я: огляд [Електронний ресурс] // BMC Medical Informatics and Decision Making. – 2022. – Режим доступу: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-022-01939-x
6. Новий підхід до алгоритмів дерев рішень [Електронний ресурс] // IEEE Xplore. – Режим доступу: https://ieeexplore.ieee.org/document/10145150
7. Дерева рішень у промислових застосуваннях [Електронний ресурс] // IOPscience. – 2020. – Режим доступу: https://iopscience.iop.org/article/10.1088/1757-899X/621/1/012005/pdf
8. Застосування дерев рішень для прогнозування у сфері охорони здоров'я [Електронний ресурс] // PMC. – Режим доступу: https://pmc.ncbi.nlm.nih.gov/articles/PMC6127872/
9. Оцінка продуктивності алгоритму C4.5 [Електронний ресурс] // SHS Web of Conferences. – 2022. – Режим доступу: https://www.shs-conferences.org/articles/shsconf/pdf/2022/19/shsconf_icss2022_01048.pdf

10. Аналіз точності алгоритму CHAID [Електронний ресурс] // ResearchGate. – Режим доступу: https://www.researchgate.net/publication/371410394
11. Скорне Е., Б'ю Г., Вер Ж.-П. Узгодженість випадкових лісів // Annals of Statistics. – 2015. – Т. 43, № 4. – С. 1716–1741.
12. Гунай Д. Пояснення алгоритму CART [Електронний ресурс] // Medium. – Режим доступу: https://medium.com/@denizgunay/cart-4838fec3e405
13. Клусовський Дж. Відновлення розріджених даних за допомогою Lasso: точні нерівності оракула та швидкі темпи збіжності [Електронний ресурс] / Дж. Клусовський. – Принстонський університет. – 2020. – Режим доступу: https://klusowski.princeton.edu/sites/g/files/toruqf5901/files/documents/klusowski2020sparse.pdf

References

1. A Survey of Decision Trees: Concepts, Algorithms, and Applications [Електронний ресурс] // ResearchGate. – Режим доступу: https://www.researchgate.net/publication/381564302
2. Topuz E., Caramanis C. Decision Tree Algorithms in Medical Diagnosis [Електронний ресурс] // PMC. – Режим доступу: https://pmc.ncbi.nlm.nih.gov/articles/PMC4466856/
3. Iterative Dichotomizer 3 (ID3) Decision Tree: A Machine Learning Algorithm for Data Classification [Електронний ресурс] // ResearchGate. – Режим доступу: https://www.researchgate.net/publication/341884258
4. Dwi A.S., Sugito S. Decision Tree Method with C4.5 Algorithm for Students Classification Who is Entitled to Receive Indonesian Smart Card (KIP) [Електронний ресурс] // ResearchGate. – Режим доступу: https://www.researchgate.net/publication/343512262
5. Decision Trees in Healthcare: A Review [Електронний ресурс] // BMC Medical Informatics and Decision Making. – 2022. – Режим доступу: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-022-01939-x
6. A New Approach for Decision Tree Algorithms [Електронний ресурс] // IEEE Xplore. – Режим доступу: https://ieeexplore.ieee.org/document/10145150
7. Decision Trees in Industrial Applications [Електронний ресурс] // IOPscience. – 2020. – Режим доступу: https://iopscience.iop.org/article/10.1088/1757-899X/621/1/012005/pdf
8. Application of Decision Trees in Healthcare Prediction [Електронний ресурс] // PMC. – Режим доступу: https://pmc.ncbi.nlm.nih.gov/articles/PMC6127872/
9. Performance Evaluation of C4.5 Algorithm [Електронний ресурс] // SHS Web of Conferences. – 2022. – Режим доступу: https://www.shs-conferences.org/articles/shsconf/pdf/2022/19/shsconf_icss2022_01048.pdf
10. Performance Analysis of CHAID Algorithm for Accuracy [Електронний ресурс] // ResearchGate. – Режим доступу: https://www.researchgate.net/publication/371410394
11. Scornet E., Biau G., Vert J.-P. Consistency of random forests // Annals of Statistics. – 2015. – Vol. 43, № 4. – P. 1716–1741.
12. Gunay D. CART Algorithm Explained [Електронний ресурс] // Medium. – Режим доступу: https://medium.com/@denizgunay/cart-4838fec3e405
13. Klusowski J. Sparse Recovery via the Lasso: Sharp Oracle Inequalities and Fast Rates [Електронний ресурс] / J. Klusowski. – Princeton University. – 2020. – Режим доступу: https://klusowski.princeton.edu/sites/g/files/toruqf5901/files/documents/klusowski2020sparse.pdf