Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

УДК 004.855.5
DOI: https://doi.org/10.30837/ITSSI.2023.26.087

Д. Чернишов, Д. Ситніков

БІНАРНА КЛАСИФІКАЦІЯ НА ОСНОВІ ПОЄДНАННЯ ТЕОРІЇ ПРИБЛИЗНИХ МНОЖИН І ДЕРЕВ РІШЕНЬ

Предмет дослідження – підвищення точності та ефективності алгоритмів класифікації на основі дерев рішень за допомогою інтеграції принципів теорії приблизних множин (*Rough Set*), математичного підходу до апроксимації множин. **Мета дослідження** – розроблення гібридної моделі, яка об'єднує теорію приблизних множин з алгоритмами дерев рішень, тим самим вирішуючи вроджені обмеження цих алгоритмів у роботі з невизначеністю в даних. Ця інтеграція має суттєво покращити точність та ефективність бінарної класифікації на основі дерев рішень, роблячи їх більш стійкими до різних вхідних даних. **Завдання статті** передбачають глибоке вивчення можливих синергій між теорією приблизних множин та алгоритмами дерев рішень. З цією метою комплексно досліджено інтеграцію теорії приблизних множин у межах алгоритмів дерев рішень. Це потребує розроблення моделі, що використовує принципи та алгебраїчні інструменти теорії приблизних множин для більш ефективного відбору ознак у системах, основаних на деревах рішень. Модель застосовує теорію приблизних множин для ефективної роботи з невизначеністю та вагомістю, що дає змогу удосконалювати та розширювати процеси відбору ознак у системах дерев рішень. Проведено серію експериментів на різних наборах даних для демонстрації ефективності та практичності цього підходу. Ці набори даних обрані для подання спектра складностей та невизначеностей із забезпеченням ретельного оцінювання можливостей моделі. **Методологія** використовує передові алгебраїчні інструменти теорії приблизних множин, зокрема формулювання алгебраїчних виразів та розроблення нових правил і технік, для спрощення та підвищення точності процесів класифікації даних за допомогою систем дерев рішень. Знахідки дослідження є важливими, оскільки свідчать про те, що інтеграція теорії приблизних множин у алгоритми дерев рішень може забезпечити більш точні та ефективні результати класифікації. Така гібридна модель демонструє значні переваги в роботі з інформацією із вбудованою невизначеністю, що є загальним викликом у багатьох додаткових сценаріях. Ефективність інтегрованого підходу продемонстровано його успішним застосуванням у сферах кредитного скорингу та кібербезпеки, що вказує на його потенціал як універсального інструмента в галузі видобутку даних і машинного навчання. **Висновки.** З'ясовано, що інтеграція теорії приблизних множин може привести до більш точних та ефективних результатів класифікації. Покращуючи можливість дерев рішень, необхідно зважати на невизначеність і неточність інформації. Дослідження відкриває нові перспективи для надійного й ґрунтовного аналізу та інтерпретації даних у різних галузях – від охорони здоров'я до сфери фінансів тощо. Інтеграція теорії приблизних множин і дерев рішень є важливим кроком у розвитку більш удосконалених, ефективних і точних інструментів класифікації в епоху великих обсягів інформації.

Ключові слова: класифікація дерев рішень; теорія приблизних множин; алгебраїчний підхід; машинне навчання.

Вступ

Класифікація за допомогою дерева рішень набула поширення як основна техніка в галузі видобутку даних і машинного навчання, що визнається своєю простою інтерпретацією та застосуванням у різних галузях, зокрема охорона здоров'я, фінансування, маркетинг і аналіз соціальних медіа. Поява значних обсягів інформації підвищила важливість ефективних методів класифікації, що можуть швидко й точно обробляти й категоризувати великі набори даних. Зокрема, дерева рішень відіграють ключову роль у завданнях, пов'язаних із класифікацією даних на окремі категорії або з виявленням закономірностей і тенденцій у складних наборах даних.

У галузі класифікації за допомогою дерев рішень основна **проблема** полягає в обмеженій здатності алгоритму обробляти невизначеність і неточність інформації. Цей недолік особливо суттєвий у сценаріях, де точна й надійна класифікація важлива, наприклад, у медичній діагностиці, аналізі фінансового ризику та складних процесах прийняття рішень на основі даних. Традиційні моделі дерев рішень, хоча й ефективні в простих сценаріях, часто викликають труднощі в підтриманні високої точності під час зіткнення з неоднозначною або неповною інформацією. Дослідження спрямоване на вирішення цієї проблеми способом інтеграції теорії приблизних множин у фреймворк дерев рішень. Теорія приблизних множин зі своїми міцними математичними структурами для роботи з вагомістю та невизначеністю пропонує перспективне рішення для покращення здатності дерев рішень класифікувати дані точно за таких складних умов.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)

Успішна інтеграція теорії приблизних множин з алгоритмами дерев рішень спрямована на створення більш адаптивної, ефективної та точної моделі дерева рішень, здатної розв'язувати складні питання сучасного аналізу інформації. Для досягнення поставленої мети необхідно виконати завдання – детально дослідити потенційні синергії між теорією приблизних множин та алгоритмами дерев рішень.

Інтеграція теорії приблизних множин у алгоритми дерев рішень спрямована на покращення їх здатності ефективно обробляти й розуміти невизначені дані. Цей підхід особливо актуальний у галузях, де є важливими точність та інтерпретація класифікації, наприклад, у медичній діагностиці або оцінюванні фінансового ризику. Теорія приблизних множин надає міцний математичний фундамент для покращення результативності дерев рішень, особливо в сценаріях із неповною або неоднозначною інформацією.

У класифікації за допомогою дерев рішень застосовуються різні підходи для виявлення важливості ознак, починаючи від евристичних підходів на основі заздалегідь визначених правил і експертних знань, ефективних у простіших сценаріях, до більш класичних підходів, наприклад, алгоритмів дерев рішень, з традиційними критеріями відбору ознак, такими як ентропія. Проте кожен із цих методів має свої обмеження, особливо в умовах невизначеності даних, і прогалини, які має заповнити теорія приблизних множин.

Метою цього дослідження є розроблення гібридної моделі, що об'єднує теорію приблизних множин з алгоритмами дерев рішень, тим самим вирішуючи вроджені обмеження цих алгоритмів у роботі з невизначеністю даних. Щоб досягти окресленої мети, додатково планується провести всебічний порівняльний аналіз ефективності розробленої гібридної моделі та порівняти її з традиційними алгоритмами.

Це порівняльне оцінювання буде зосереджено на ключових метриках, таких як точність класифікації, ефективність оброблення даних і стійкість методологій до різних типів наборів даних. З допомогою цього дослідження сподіваємося визначити більш надійну й універсальну модель класифікації за допомогою дерева рішень, яка може ефективно обробляти складнощі сучасних даних і здійснити вагомий внесок у сфері видобутку інформації та машинного навчання.

Аналіз проблеми й наявних методів

Інтеграція теорії приблизних множин у класифікацію за допомогою дерев рішень є актуальним, проте складним завданням. Дерева рішень – це основний інструмент у галузі добування даних та машинного навчання, відомий своєю інтерпретованістю та застосовністю в різних сферах. Однак дерева рішень часто виявляються неефективними в роботі з нечіткою та невизначеною інформацією, що спричиняє зниження ефективності класифікації. Використання теорії приблизних множин, що ефективно долає нечіткість і невизначеність, має потенціал розв'язати ці проблеми, але така інтеграція не є простою та потребує всебічного аналізу як теоретичних, так і практичних аспектів.

Унаслідок ретельного огляду сучасної літератури можна говорити про значні досягнення як в алгоритмах дерев рішень, так і в теорії приблизних множин. Так, дослідження [1, 2] висвітлюють поліпшення в методологіях дерев рішень, наголошуючи на підвищенні їх адаптивності до різноманітних типів даних. Одночасно теорія приблизних множин усе частіше застосовується в галузях, де є невизначеність і неточність інформації, про що говориться в студіях [3, 4]. Однак злиття цих двох сфер наразі залишається малодослідженим.

У роботі [5] наголошується на важливості невизначеності даних у деревах рішень і прогалині, яку може заповнити теорія приблизних множин. Крім того, у дослідженнях [6, 7] обговорюються алгебраїчні підходи в теорії приблизних множин, що можуть надати систематичний спосіб для їх інтеграції в дерева рішень. Складність такої інтеграції розглядається у статті [8]. ЇЇ автори висвітлюють необхідність нових методологій, які могли б безперешкодно об'єднувати ці дві сфери.

Можливість такої інтеграції підтверджується останніми досягненнями в алгоритмічних підходах до дерев рішень [9, 10]. Ці досягнення створюють основу, до якої може бути долучена теорія приблизних множин. Також вивчення теорії приблизних множин [11, 12] має значний прогрес у алгебраїчних методах, натякаючи на шляхи покращення класифікації дерев рішень.

Практичні наслідки такої інтеграції є вагомими. У таких сферах, як охорона здоров'я та фінансування, де прийняття рішень є критичним і інформація часто містить елемент невизначеності, покращені

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

моделі дерев рішень можуть сприяти більш надійним і точним прогнозам, як це передбачається в дослідженнях, зокрема [13]. Однак ця інтеграція висуває нові виклики, особливо в збереженні простоти та інтерпретованості дерев рішень під час включення складності теорії приблизних множин.

Хоча в літературі йдеться про обнадійливий ефект у покращенні класифікації дерев рішень за допомогою теорії приблизних множин, у ній також наголошується на необхідності інноваційних підходів для подолання викликів, властивих такій інтеграції.

Це дослідження має на меті подолати цей розрив способом розроблення алгебраїчного підходу, що гармонізує надійність теорії приблизних множин із простотою алгоритмів дерев рішень. Постає необхідність створити гібридну модель, що поєднує теорію приблизних множин з алгоритмами дерев рішень, підвищуючи точність класифікації за умови невизначеності даних, а також зберігаючи простоту використання дерев рішень.

Вирішення завдання

У цьому дослідженні, коли розглядаємо граничну ділянку $B(X)$ для множини X, маємо на увазі різницю між верхньою та нижньою апроксимацією X. Математично це може бути виражено так:

$$B(X) = A^*(X) - A_*(X), \quad (1)$$

де $A^*(X)$ – верхня апроксимація X, що є множиною всіх елементів, які можуть бути в X на підставі доступної інформації (2). Вона містить усі елементи, що мають принаймні одну ознаку, спільну з членами X. Ця апроксимація зазвичай більш включаюча й охоплює елементи, які можуть потенційно належати розглянутій множині.

$$A^*(X) = \{x \in U, [x] \cap X \neq \varnothing\}, \quad (2)$$

де U – універсальна множина;

$[x]$ – клас еквівалентності (3).

$$[x] = \{y \in U, xRy\}. \quad (3)$$

З іншого боку, $A_*(X)$ – це нижня апроксимація X, що містить елементи, які точно належать X, на основі наявних даних. Вона має всі елементи, для яких кожна ознака відповідає ознакам в X (4). Ця апроксимація більш виключна і містить лише ті елементи, які обов'язково належать розглянутій множині.

$$A_*(X) = \{x \in U, [x] \subseteq X\}. \quad (4)$$

Після обчислення граничної ділянки наступним кроком є оцінювання важливості кожної ознаки. Це досягається способом систематичного вилучення ознак по одній і спостереження за результатом змін у граничній ділянці. Аналізуючи, як вилучення кожної ознаки впливає на граничну ділянку, можемо визначити відносну важливість кожної ознаки. Цей підхід дає змогу виявити критичні ознаки для визначення ясності або неоднозначності класифікації. Ознаки, що в процесі вилучення значно змінюють граничну ділянку, вважаються більш важливими, оскільки їх присутність істотно сприяє точності та однозначності класифікації. Навпаки, ознаки, вилучення яких призводить до мінімальних змін у граничній ділянці, вважаються менш критичними. Цей метод дає більш детальне розуміння важливості ознак поза традиційними метриками способом прямого зв'язку важливості ознаки з її впливом на динаміку границі класифікації.

$$V(P_i) = \frac{\Delta(BN_i)}{M(X)} \times 100\%, \quad (5)$$

де $M(X)$ – це кількість елементів у множині X.

У запропонованій гібридній моделі важливість ознаки (5) застосовуватиметься як критерій розбиття для побудови дерева рішень.

Матеріали й методи

Отримана гібридна модель була оцінена за допомогою метрики площі під кривою робочої характеристики приймача (*ROC AUC*). Цей метод оцінювання використовується для визначення точності класифікації моделі на наборі даних, який містить як позитивні, так і негативні приклади.

Оцінювання моделі проводитиметься за допомогою випадкової стратифікованої крос-валідації з 10 фолдами. Зазначений метод забезпечує більш точну оцінку ефективності моделі, ніж оцінка на одному наборі даних.

Стратифікована крос-валідація працює шляхом поділу набору даних на 10 рівних частин, або фолдів. Потім модель навчається на 9 фолдах і тестується на 10-му фолді. Цей процес повторюється 10 разів, і результати всіх 10 тестувань використовуються для оцінювання ефективності моделі.

За допомогою цього методу можна оцінити ефективність моделі на трьох різноманітних наборах

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)

даних. Ці набори обиралися для подання різних рівнів складності та невизначеності, забезпечуючи всебічну оцінку можливостей моделі.

Набір даних *Titanic* [14], широко визнаний у сфері машинного навчання й науки про дані, дає ідеальний варіант для оцінювання покращеної моделі дерева рішень. Цей набір даних містить інформацію про пасажирів відомого корабля *Titanic*, зокрема такі відомості, як вік, стать, клас, вартість квитка та статус виживання. Складність і змінність цієї інформації роблять її ефективним вибором для оцінювання можливостей моделі у роботі з реальними, невизначеними й категоріальними даними.

Набір даних *Microsoft Malware Classification Challenge (BIG 2015)* [15] є бенчмарком у сфері кібербезпеки, особливо в класифікації шкідливих програм. Він містить значну кількість об'єктів, кожен з яких є різними типами шкідливих програм і має різні ознаки, такі як байткод, асемблерний код та інші характеристики файлів. Складність цього набору даних і критичне значення його застосування роблять його гарним кандидатом для оцінювання ефективності покращеної моделі дерева рішень в умовах високого ризику та технічної складності.

Набір даних *Home Credit Default Risk* [16], надає вичерпне середовище для оцінювання покращеної моделі дерева рішень у контексті оцінювання кредитного ризику. Цей набір даних особливо актуальний для фінансових установ, мета яких – точно передбачити здатність своїх клієнтів до повернення кредитів. Він має різноманітні ознаки, зокрема кредитна історія, деталі позики та демографічні відомості позичальника, що робить його цінним ресурсом для тестування класифікаційних здібностей моделі у фінансовому середовищі.

У цьому дослідженні всі три набори даних були однаково підготовлені для бінарної класифікації. Підготовка даних передбачала перетворення цільових змінних у бінарний формат. Також були відібрані та бінаризовані відповідні ознаки, здійснено оброблення відсутніх значень за допомогою відповідних технік та нормалізацію або стандартизацію числових ознак. Крім того, кожен набір даних був розподілений на навчальні й тестові набори із заходами, спрямованими на забезпечення балансу даних і подолання проблем нерівномірності класів. Цей стандартизований підхід до підготовки даних забезпечує послідовність між наборами даних і сприяє справедливому оцінюванню продуктивності моделей дерев рішень у різних завданнях бінарної класифікації.

Результати досліджень

У дослідженні проведено порівняльний аналіз продуктивності двох критеріїв для побудови дерев рішень: критерію приросту інформації та критерію зміни граничної ділянки. Для цього було використано три набори даних: *Titanic*, *Malware Classification* та *Credit Default Risk*.

Результати порівняння показали, що критерій зміни граничної ділянки в деяких випадках забезпечує більш високу точність, ніж критерій приросту інформації. Зокрема, для набору даних *Titanic* критерій зміни граничної ділянки забезпечує підвищення точності класифікації на 1,3 %. У табл. 1 подані значення оцінки якості двох критеріїв для кожного набору даних.

Таблиця 1. *Оцінка якості критеріїв*

	Titanic	***Malware Classification***	***Credit Default Risk***
Важливість ознаки відповідно до приросту інформації	0.828	0.5658	0.6155
Важливість ознаки відповідно до змін у граничній ділянці	0.841	0.5658	0.5787

Підвищення точності класифікації на наборі даних *Titanic* можна пояснити тим, що критерій зміни граничної ділянки дозволяє будувати дерева рішень, які краще відображають структуру даних. Зокрема, зазначений критерій бере до уваги не тільки кількість інформації, яка отримується в процесі розбиття на основі ознаки, але також і розмір граничної ділянки, що утворюється в цій ситуації. Це дає змогу обирати ознаки, які забезпечують найбільш чітке розбиття даних на два класи.

Також був проведений візуальний аналіз двох різних графіків дерев рішень, для датасету *Titanic* з параметром максимальної глибини, який був заданий значенням 4. Ці візуалізації є важливими для розуміння впливу інтеграції теорії приблизних множин у моделі дерева рішень. Були візуалізовані графіки як для критерію приросту інформації (рис. 1), так і для критерію різниці граничних ділянок (рис. 2). Така візуалізація дає змогу зрозуміти, як дерево рішень визначає пріоритети різних ознак на основі різних критеріїв.

Було створено три графіки кореляції (рис. 3) для аналізу взаємозв'язку змін у граничній ділянці

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

та ентропії для кожної ознаки в наборах даних. Результати свідчать про слабку кореляцію, яка вказує на те, що зміни в граничній ділянці, хоч і впливають на ентропію, але це не має систематичного характеру. Ця слабка кореляція підкреслює складну природу відношень між коригуваннями граничної ділянки та загальним отриманням інформації в моделях дерева рішень.

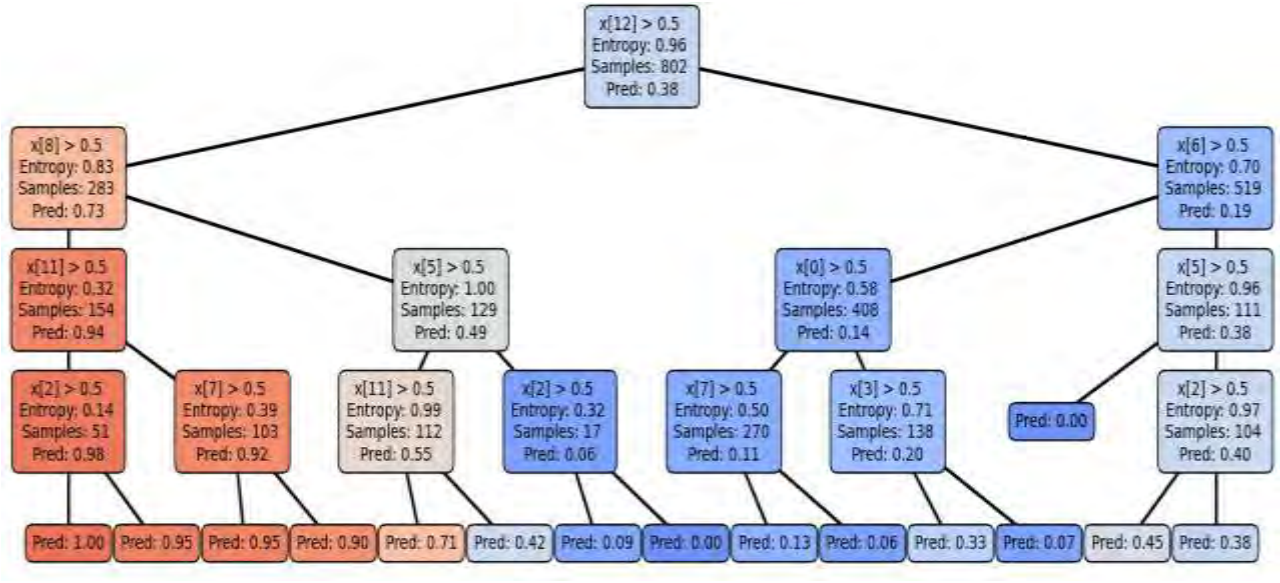

Рис. 1. Дерево рішення на базі критерію приросту інформації

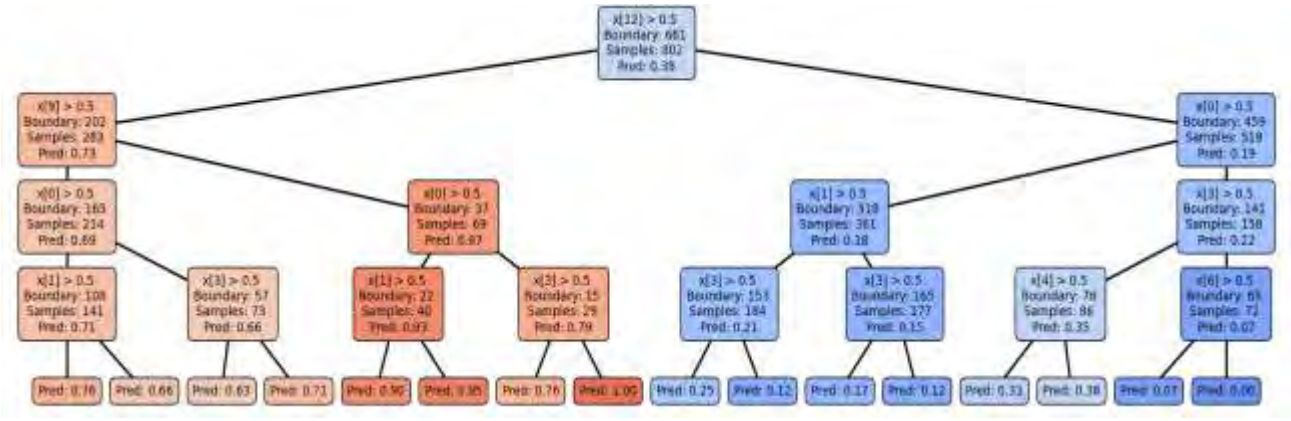

Рис. 2. Дерево на базі критерію різниці граничних ділянок

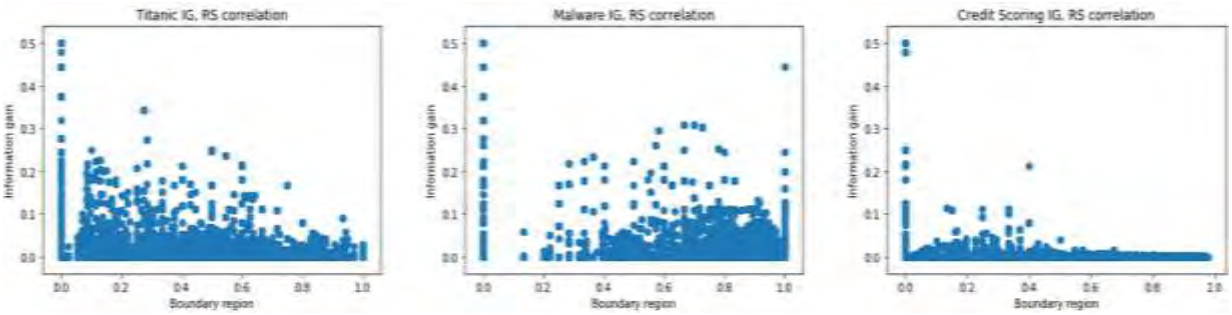

Рис. 3. Кореляція зміни граничної ділянки та приросту інформації

Підсумовуючи, можемо наголосити, що інтеграція теорії приблизних множин у класифікацію за допомогою дерев рішень має перспективні результати в деяких наборах даних, особливо в покращенні

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)

точності класифікації. Візуалізації надають більш глибокі уявлення про те, як ця інтеграція впливає на процес прийняття рішень моделей. Слабка кореляція між змінами в граничній ділянці та ентропією наголошує на недоліках і складнощах цієї інтеграції, указуючи на необхідність подальших досліджень та оптимізації для повного використання потенціалу теорії приблизних множин у нових моделях дерев рішень.

Висновки

На основі результатів цього дослідження можна прийти до висновку, що інтеграція теорії приблизних множин у алгоритми класифікації за допомогою дерев рішень значно підвищує їх продуктивність, особливо в умовах роботи з невизначеністю та вагомістю інформації. Кілька ключових факторів лежать в основі успіху цього гібридного підходу.

1. Об'єднана модель ефективно використовує переваги теорії приблизних множин у роботі з вагомістю ознак та невизначеністю даних. Ця інтеграція може сприяти значному покращенню точності класифікації, особливо в складних наборах даних, де традиційні дерева рішень зазвичай не досягають високої продуктивності.

2. Алгебраїчний підхід для теорії приблизних множин допомагає вдосконалити процес прийняття рішень у завданні класифікації. Це покращення дає змогу моделі виявляти та тлумачити витончені зразки та взаємозв'язки в даних, які можуть бути пропущені традиційними алгоритмами дерев рішень.

3. Інтеграція зберігає фундаментальну простоту та можливість тлумачення дерев рішень, але й додаючи міцності теорії приблизних множин. Ця рівновага є важливою для практичного застосування, де одночасно значущими є як точність, так і зручність використання.

Порівняльний аналіз дослідження, що порівнює покращену модель дерева рішень із традиційними алгоритмами, розкриває переваги та потенціал цього гібридного підходу. Інтеграція не лише покращує точність класифікації, але й розширює застосовність дерев рішень на більш складні та невизначені сценарії оброблення даних.

Саме тому можна зробити висновок, що інтеграція теорії приблизних множин у класифікацію за допомогою дерев рішень є певним проривом. Здатність покращеної моделі вирішувати завдання невизначеності, а також її практичність роблять її дуже ефективним інструментом для завдань класифікації даних у різних галузях. Майбутні дослідження мають зосереджуватися на подальшому вдосконаленні та застосуванні цієї моделі, особливо в галузях, де невизначеність даних є значним викликом. Це дослідження є перспективним для створення більш надійних, ефективних і точних інструментів з метою прийняття рішень в епоху великих обсягів даних та складного оброблення інформації.

Список літератури

1. Costa V. G. and Pedreira C. E. Recent advances in decision trees: an updated survey. *Artificial Intelligence Review, Springer Science and Business Media LLC.* Vol. 56. No. 5. P. 4765–4800. 2022. DOI: 10.1007/s10462-022-10275-5

2. Hafeez M. A., Rashid M., Tariq H., Abideen Z. U., Alotaibi S. S., and Sinky M. H. Performance Improvement of Decision Tree: A Robust Classifier Using Tabu Search Algorithm. *Applied Sciences, MDPI AG.* Vol. 11. No. 15. 6728 p. 2021. DOI: 10.3390/app11156728

3. Wang Z., Zhang X., and Deng J. The uncertainty measures for covering rough set models. *Soft Computing, Springer Science and Business Media LLC.* Vol. 24. No. 16. P. 11909–11929. 2020. DOI: 10.1007/s00500-020-05098-x

4. Geetha M. A., Acharjya D. P., and Iyengar N. Ch. S. N. Algebraic properties and measures of uncertainty in rough set on two universal sets based on multi-granulation. *Proceedings of the 6th ACM India Computing Convention, ACM.* P. 1–8. 2013. DOI: 10.1145/2522548.2523168

5. Qian Y., Xu H., Liang J., Liu B., and Wang J. Fusing Monotonic Decision Trees. *IEEE Transactions on Knowledge and Data Engineering.* Vol. 27. No. 10. P. 2717–2728. 2015. DOI: 10.1109/TKDE.2015.2429133

6. Sitnikov D. and Ryabov O. An Algebraic Approach to Defining Rough Set Approximations and Generating Logic Rules. *Data Mining V, WIT Press.* 10 p. 2004. DOI: 10.2495/data040171

7. Sitnikov D., Titova O., Romanenko O., and Ryabov O. A method for finding minimal sets of features adequately describing discrete information objects. *Data Mining X, WIT Press.* 8 p. 2009. DOI: 10.2495/data090141

8. Wang D., Liu X., Jiang L., Zhang X., and Zhao Y. Rough Set Approach to Multivariate Decision Trees Inducing. *Journal of Computers, International Academy Publishing (IAP).* Vol. 7. No. 4. P. 870–879. 2012. DOI: 10.4304/jcp.7.4.870-879

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

9. Blockeel H., Devos L., Frénay B., Nanfack G., and Nijssen S. Decision trees: from efficient prediction to responsible AI. *Frontiers in Artificial Intelligence, Frontiers Media SA.* Vol. 6. Jul. 26. 2023. DOI: 10.3389/frai.2023.1124553

10. Hu X., Rudin C., and Seltzer M. Optimal Sparse Decision Trees. *arXiv.* 2019. DOI: 10.48550/ARXIV.1904.12847

11. Chiaselotti G., Gentile T., and Infusino F. Decision systems in rough set theory: A set operatorial perspective. *Journal of Algebra and Its Applications, World Scientific Pub Co Pte Lt.* Vol. 18. No. 01. 2019. 1950004 p. DOI: 10.1142/s021949881950004x

12. Xu J., Qu K., Meng X., Sun Y., and Hou Q. Feature selection based on multiview entropy measures in multiperspective rough set. *International Journal of Intelligent Systems, Hindawi Limited.* Vol. 37. No. 10. 2022. P. 7200–7234. DOI: 10.1002/int.22878

13. Duan G., Ding D., Tian Y., and You X. An Improved Medical Decision Model Based on Decision Tree Algorithms. *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), Atlanta, GA, USA IEEE.* 2016. P. 151–156. DOI: 10.1109/BDCloud-SocialCom-SustainCom.2016.33

14. Cukierski W. Titanic – Machine Learning from Disaster. Kaggle. 2012. URL: https://kaggle.com/competitions/titanic

15. Ronen R., Radu M., Feuerstein C., Yom-Tov E., and Ahmadi M. Microsoft Malware Classification Challenge. *arXiv.* 2018. DOI: 10.48550/ARXIV.1802.10135

16. Montoya A., Odintsov K., and Kotek M. Home Credit Default Risk. Kaggle. 2018. URL: https://kaggle.com/competitions/home-credit-default-risk

References

1. Costa, V. G. and Pedreira, C. E. (2022), "Recent advances in decision trees: an updated survey". *Artificial Intelligence Review, Springer Science and Business Media LLC.* Vol. 56, No. 5. P. 4765–4800. DOI: 10.1007/s10462-022-10275-5

2. Hafeez, M. A., Rashid, M., Tariq, H., Abideen, Z. U., Alotaibi, S. S., and Sinky, M. H. (2021), "Performance Improvement of Decision Tree: A Robust Classifier Using Tabu Search Algorithm". *Applied Sciences, MDPI AG.* Vol. 11, No. 15. 6728 p. DOI: 10.3390/app11156728

3. Wang, Z., Zhang, X., and Deng, J. (2020), "The uncertainty measures for covering rough set models". *Soft Computing, Springer Science and Business Media LLC.* Vol. 24, No. 16. P. 11909–11929. DOI: 10.1007/s00500-020-05098-x

4. Geetha, M. A., Acharjya, D. P., and Iyengar, N. Ch. S. N. (2013), "Algebraic properties and measures of uncertainty in rough set on two universal sets based on multi-granulation". *Proceedings of the 6th ACM India Computing Convention, ACM.* P. 1–8. DOI: 10.1145/2522548.2523168

5. Qian, Y., Xu, H., Liang, J., Liu, B., and Wang, J. (2015), "Fusing Monotonic Decision Trees". *IEEE Transactions on Knowledge and Data Engineering.* Vol. 27, No. 10. P. 2717–2728. DOI: 10.1109/TKDE.2015.2429133

6. Sitnikov, D. and Ryabov, O. (2004), "An Algebraic Approach to Defining Rough Set Approximations and Generating Logic Rules". *Data Mining V, WIT Press.* 10 p. DOI: 10.2495/data040171

7. Sitnikov, D., Titova, O., Romanenko, O., and Ryabov, O. (2009), "A method for finding minimal sets of features adequately describing discrete information objects". *Data Mining X, WIT Press.* 8 p. DOI: 10.2495/data090141

8. Wang, D., Liu, X., Jiang, L., Zhang, X., and Zhao, Y. (2012), "Rough Set Approach to Multivariate Decision Trees Inducing". *Journal of Computers, International Academy Publishing (IAP).* Vol. 7, No. 4. P. 870–879. DOI: 10.4304/jcp.7.4.870-879

9. Blockeel, H., Devos, L., Frénay, B., Nanfack, G., and Nijssen, S. (2023), "Decision trees: from efficient prediction to responsible AI". *Frontiers in Artificial Intelligence, Frontiers Media SA.* Vol. 6. Jul. 26. DOI: 10.3389/frai.2023.1124553

10. Hu, X., Rudin, C., and Seltzer, M. (2019), "Optimal Sparse Decision Trees". *arXiv.* DOI: 10.48550/ARXIV.1904.12847

11. Chiaselotti, G., Gentile, T., and Infusino, F. (2019), "Decision systems in rough set theory: A set operatorial perspective". *Journal of Algebra and Its Applications, World Scientific Pub Co Pte Lt.* Vol. 18, No. 01. 1950004 p. DOI: 10.1142/s021949881950004x

12. Xu, J., Qu, K., Meng, X., Sun, Y., and Hou, Q. (2022), "Feature selection based on multiview entropy measures in multiperspective rough set". *International Journal of Intelligent Systems, Hindawi Limited.* Vol. 37, No. 10. P. 7200–7234. DOI: 10.1002/int.22878

13. Duan, G., Ding, D., Tian, Y., and You, X. (2016), "An Improved Medical Decision Model Based on Decision Tree Algorithms". *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), IEEE.* P. 151–156. DOI: 10.1109/BDCloud-SocialCom-SustainCom.2016.33

14. Cukierski, W. (2012), "Titanic – Machine Learning from Disaster" Kaggle. available at: https://kaggle.com/competitions/titanic.

15. Ronen, R., Radu, M., Feuerstein, C., Yom-Tov, E., and Ahmadi, M. (2018), "Microsoft Malware Classification Challenge". *arXiv.* DOI: 10.48550/ARXIV.1802.10135

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)*

16. Montoya, A., Odintsov, K., and Kotek, M. (2018), "Home Credit Default Risk". Kaggle. available at: https://kaggle.com/competitions/home-credit-default-risk

Надійшла 30.11.2023

Відомості про авторів / About the Authors

Чернишов Дмитро Владиславович – Харківський національний університет радіоелектроніки, бакалавр комп'ютерних наук, Харків, Україна; e-mail: dmytro.chernyshov@nure.ua; ORCID ID: https://orcid.org/0009-0003-2773-7467

Ситніков Дмитро Едуардович – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, професор кафедри системотехніки, Харків, Україна; e-mail: dmytro.sytnikov@nure.ua; ORCID ID: https://orcid.org/0000-0003-1240-7900

Chernyshov Dmytro – Kharkiv National University of Radio Electronics, Bachelor of Computer Science, Kharkiv, Ukraine.

Sytnikov Dmytro – PhD (Engineering Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Professor at the Department of System Engineering, Kharkiv, Ukraine.

BINARY CLASSIFICATION BASED ON A COMBINATION OF ROUGH SET THEORY AND DECISION TREES

The subject of the study is to improve the accuracy and efficiency of classification algorithms using decision trees by integrating the principles of Rough Set theory, a mathematical approach to approximating sets. **The aim of the study** is to develop a hybrid model that integrates rough set theory with decision tree algorithms, thereby solving the inherent limitations of these algorithms in dealing with uncertainty in data. This integration should significantly improve the accuracy and efficiency of binary classification based on decision trees, making them more robust to different inputs. Research **objectives** include a deep study of possible synergies between approximate set theory and decision tree algorithms. For this purpose, we are conducting a comprehensive study of the integration of approximate set theory within decision tree algorithms. This includes the development of a model that utilizes the principles and algebraic tools of approximate set theory to more efficiently select features in decision tree-based systems. The model uses the theory of approximate sets to efficiently handle uncertainty and weighting, which allows for improved and extended feature selection processes in decision tree systems. A series of experiments are conducted on different datasets to demonstrate the effectiveness and practicality of this approach. These datasets are chosen to represent a range of complexities and uncertainties, providing a thorough and rigorous evaluation of the model's capabilities. The **methodology** uses advanced algebraic tools of approximate set theory, including the formulation of algebraic expressions and the development of new rules and techniques, to simplify and improve the accuracy of data classification processes using decision tree systems. The findings of the study are important because they show that integrating approximate set theory into decision tree algorithms can indeed provide more accurate and efficient classification results. Such a hybrid model demonstrates significant advantages in dealing with data with embedded uncertainty, which is a common challenge in many complementary scenarios. The versatility and effectiveness of the integrated approach is demonstrated by its successful application in the areas of credit scoring and cybersecurity, which emphasizes its potential as a versatile tool in data mining and machine learning. The **conclusions** show that integrating approximate set theory can lead to more accurate and efficient classification results. By improving the ability of decision trees to account for uncertainty and imprecision in data, the research opens up new possibilities for robust and sophisticated data analysis and interpretation in a variety of industries, from healthcare to finance and beyond. The integration of approximate set theory and decision trees is an important step in the development of more advanced, efficient, and accurate classification tools in the era of big data.

Keywords: decision tree classification; approximate set theory; algebraic approach; machine learning.

Бібліографічні описи / Bibliographic descriptions

Чернишов Д. В., Ситніков Д. Е. Бінарна класифікація на основі поєднання теорії приблизних множин і дерев рішень. *Сучасний стан наукових досліджень та технологій в промисловості*. 2023. № 4 (26). С. 87–94. DOI: https://doi.org/10.30837/ITSSI.2023.26.087

Chernyshov, D., Sytnikov, D. (2023), "Binary classification based on a combination of rough set theory and decision trees", *Innovative Technologies and Scientific Solutions for Industries*, No. 4 (26), P. 87–94. DOI: https://doi.org/10.30837/ITSSI.2023.26.087