Вісник Національного технічного університету "ХПІ", 2021, № 1 (5)
ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

УДК 004.8: 004.89: 519.7 DOI: 10.20998/2411-0558.2021.01.02

І. Ф. ПОВХАН, д-р тех. наук, доц., ДВНЗ "Ужгородський національний університет", Ужгород

МЕТОД АЛГОРИТМІЧНИХ ДЕРЕВ КЛАСИФІКАЦІЇ НА ОСНОВІ ОБМЕЖЕНЬ

Розглянута загальна задача побудови алгоритмічних дерев розпізнавання (класифікації) на основі обмеженого методу в теорії штучного інтелекту. Об‘єктом дослідження є концепція алгоритмічного дерева класифікації на базі обмеженого методу. Предметом дослідження є актуальні методи, алгоритми та схеми (обмежені методи) побудови алгоритмічних дерев класифікації. Пропонується обмежений метод побудови алгоритмічних дерев класифікації, який для заданої навчальної вибірки довільного розміру будує деревоподібну структуру (модель дерева алгоритмів), яка складається з набору автономних алгоритмів класифікації та розпізнавання, оцінених на кожному кроці побудови дерев алгоритмів за даною початковою вибіркою. Тобто пропонується обмежений метод побудови алгоритмічного дерева класифікації, основна ідея якого полягає в по кроковій апроксимації начальної вибірки довільного об‘єму та структури набором незалежних алгоритмів класифікації та розпізнавання. Метод при формуванні поточної вершини алгоритмічного дерева забезпечує виділення найбільш ефективних автономних алгоритмів класифікації з початкового набору та добудову лише тих шляхів в структурі дерева, де відбувається найбільша кількість помилок класифікації. Такий підхід при побудові результуючого дерева класифікації дозволяє значно скоротити розмір та складність дерева, підвищити якість його наступного аналізу (інтерпретабельність), можливість декомпозиції, та будувати структури дерев класифікації в умовах обмежених апаратних ресурсів. Обмежений метод побудови алгоритмічного дерева класифікації дозволяє будувати різнотипні деревоподібні моделі розпізнавання з наперед заданою точністю для широкого класу задач теорії штучного інтелекту. Розроблений та представлений в роботі обмежений метод алгоритмічного дерева класифікації отримав програмну реалізацію та був досліджений і порівняний з методами логічних дерев класифікації, методами алгоритмічного дерева класифікації (першого та другого типу) при розв‘язку задачі розпізнавання реальних даних геологічного типу. Іл.: 2. Табл.: 1. Бібліогр.: 34 назв.

Ключові слова: алгоритмічне дерево класифікації; розпізнавання реальних даних; класифікація; алгоритм класифікації; обмежений метод.

Вступ. Задачі, які об’єднуються тематикою розпізнавання образів, дуже різноманітні та виникають у сучасному світі в усіх сферах економіки та соціального контенту діяльності людини, що приводить до необхідності побудови та дослідження математичних моделей відповідних систем. Станом на зараз не існує універсального підходу до їх розв‘язання, запропоновано декілька досить загальних теорій та підходів, що дозволяють вирішувати багато типів (класів) задач, але їх прикладні застосування відрізняються досить великою чутливістю до

специфіки самої задачі або предметної області застосування. Багато теоретичних результатів отримано для спеціальних випадків та підзадач, причому слід відмітити, що вузьким місцем вдалих реальних систем розпізнавання залишається необхідність виконання величезного об‘єму обчислень та орієнтація на потужний апаратний інструментарій. Проте велика кількість прикладних задач, в різних областях природознавства, наприклад, в геології, геофізиці, геохімії, медицині, соціології, археології, біології та інше, де вирішуються задачі класифікації з використанням програмних та апаратних систем, визначає інтенсивність та актуальність такого напрямку досліджень. Так подолання даної проблеми для задач теорії штучного інтелекту при автоматизації алгоритмічного та програмного конструювання конкретних систем розпізнавання (СР) у вигляді моделей дерев класифікації є запорукою їх високої ефективності для кожної реальної задачі, а отже забезпечить швидкий розвиток різних галузей науки й техніки [1 – 3]. Причому на сьогоднішній день відомо більше чотирьох тисяч алгоритмів розпізнавання (заснованих на різноманітних підходах та концепціях), які мають певні обмеження при їх використанні (точність, швидкодія, пам'ять, універсальність, надійність, тощо), крім того кожний з алгоритмів обмежений певною специфікою задач застосування, а це безумовно є найслабкішим місцем не тільки даних алгоритмів, але й систем розпізнавання, які базуються на відповідних концепціях [4]. Так об‘єктом даного дослідження є логічні дерева класифікації (дерева рішень). Відомо, що представлення навчальних вибірок (дискретної інформації) великого об‘єму у вигляді структур логічних дерев має свої суттєві переваги в плані економічного опису даних та ефективних механізмів роботи з ними [5 – 7].

Відомо, що представлення масивів дискретної інформації (навчальних вибірок) великого об‘єму у вигляді структур логічних або алгоритмічних дерев має свої суттєві переваги в плані економічного опису та зручного аналізу даних, ефективних механізмів (процедур) роботи з ними [8]. Так ефективне покриття навчальної вибірки набором елементарних ознак у випадку структури логічного дерева класифікації (ЛДК), або покриття навчальної вибірки фіксованим набором автономних алгоритмів розпізнавання та класифікації у випадку конструкції алгоритмічного дерева класифікації (АДК), породжує фіксовану деревоподібну структуру даних (модель дерева), причому яка забезпечує стиск та перетворення початкових даних навчальних вибірок. Підкреслимо, що такий підхід дозволяє суттєву оптимізацію та економію апаратних ресурсів інформаційної системи [9]. Галузь застосування концепції дерев рішень (структур ЛДК/АДК) в даний час надзвичайно

об‘ємна, а абсолютна більшість сучасних схем методів побудови дерев класифікації відома з літератури під назвою – розділення та захоплення [10 – 13]. Відомо, що структура дерева класифікації (ЛДК/АДК) представлена у вигляді послідовності гілок та вузлів, причому на гілках дерева розташовуються деякі мітки від яких залежить цільова функція (у випадку АДК – незалежні алгоритми класифікації, набори узагальнених ознак (УО)), а в вузлах знаходяться значення функції розпізнавання (ФР) – значення класів належності або розширені атрибути переходів. Принциповими питаннями концепції дерев класифікації залишаються питання вибору критерія розгалуження (побудови або відбору вершин), критерію зупинки розгалуження (закінчення побудови структури дерева класифікації) та критерію відкидання гілок (структурних блоків) дерева класифікації. Принциповою особливістю більшості існуючих методів обробки навчаючих вибірок (масивів дискретної інформації) в задачах розпізнавання при побудові правил та схем класифікації є те, що вони не дозволяють регулювати їх складність, точність, інформаційну ємність (параметричну складність УО) у процесі конструювання моделі [4]. Принциповою особливістю методу алгоритмічного дерева – є можливість комплексного використання для розв‘язання кожної конкретної задачі побудови схеми розпізнавання багатьох відомих алгоритмів (методів) розпізнавання. В основі концепції АДК лежить єдина методології – оптимальної апроксимації навчаючої вибірки набором узагальнених ознак (автономних алгоритмів), які входять в деяку схему (оператор), побудовану в процесі навчання [4, 5]. Обмежений метод побудови моделей алгоритмічних дерев класифікації генерує деревоподібні схеми, які складаються з незалежних та автономних алгоритмів класифікації та являють собою в певній мірі новий алгоритм розпізнавання (зрозуміло, що синтезований з відомих алгоритмів та методів).

Постановка задачі. Нехай на деякій множині G об'єктів x задане розбиття R на скінчене число k підмножин (класів, образів)

$$H_i\ (i=1,..,k)\ ,\ G=\bigcup_{i=1}^{k} H_i\ .$$

Відповідні множини $H_1,...,H_k$ будемо називати образами, а елементи множини G – зображеннями або представниками образів $H_1,...,H_k$. Об'єкти (зображення) x задаються наборами значень деяких ознак $x_j\,(j=1,...,n)$. Якщо $x\in H_i$ то будемо рахувати, що даний об‘єкт належить образу H_i. В загальному випадку образи $H_1,...,H_k$ можуть бути задані імовірнісними розподілами $p(H_1/x),...,p(H_k/x)$, де $p(H_i/x)$ – імовірність (або в неперервному випадку щільність

імовірності) належності $(x \in G)$ образу H_i. Нехай умовою задачі задана деяка початкова навчальна вибірка (НВ) у вигляді послідовності навчальних пар наступного вигляду:

$$(x_1, f_R(x_1)),...,(x_m, f_R(x_m)). \quad (1)$$

Причому крім початкової НВ задана також тестова вибірка (ТВ) – набір об‘єктів відомої класової належності, як деяка частина початкової НВ. Отже за початковою умовою НВ – це сукупність (фіксована послідовність) деяких наборів (дискретних об‘єктів), причому кожний набір це сукупність значень деяких ознак (атрибутів) та значень деяких функцій (ФР) на цьому наборі. Тоді сукупність значень ознак – це деяке зображення (дискретний об‘єкт), а значення функції (ФР) відносить це зображення до відповідного образу [5].

Отже в роботі буде стояти задача побудови моделі АДК в умовах обмежених апаратних ресурсів з параметрами *p*, структура *L* якої була би оптимальною $F(L(p, x_i), f_R(x_i)) \to opt)$ по відношенню до початкових даних НВ.

Аналіз літературних джерел. Дослідження продовжує цикл робіт, які присвячені проблематиці деревоподібних схем розпізнавання (моделей класифікації ЛДК/АДК) дискретних об‘єктів [7 – 9, 14]. В них піднімаються принципові питання побудови, використання, та оптимізації структур дерев класифікації. Так з роботи [7] відомо, що результуюче правило класифікації (схема), яке побудоване довільним методом або алгоритмом РВО, має деревоподібну логічну структуру, причому логічне дерево складається з вершин (ознак, атрибутів), які групуються по ярусам (рівням) і які отримані на певному кроці (етапі) побудови дерева розпізнавання [15]. Так для методів АДК важливою проблемою, яка виникає є задача синтезу дерев розпізнавання, які будуть представлятися фактично деревом (графом) алгоритмів (методи АДК) або деревом моделей класифікації. На відміну від існуючих методів, головною особливістю деревоподібних систем розпізнавання є те, що важливість окремих ознак (групи ознак чи алгоритмів) визначається відносно функції, яка задає розбиття об‘єктів на класи [16]. Так в роботі [15] піднімаються принципові питання стосовно генерації дерев рішень для випадку малоінформативних ознак, питання оцінки якості побудованих моделей, причому здатність структур дерев класифікації виконувати одномірне розгалуження (вибір ознак, атрибутів) для аналізу впливу (важливості, якості) окремих змінних (вершин) дає можливість працювати зі змінними різних типів у вигляді предикатів, узагальнених ознак, для випадку АДК – відповідними автономними алгоритмами

класифікації та розпізнавання. Загальна концепція дерев класифікації активно використовується в інтелектуальному аналізі даних (LGMB, XGBoost), де кінцева мета полягає в синтезі моделі (фіксованої схеми), яка прогнозує значення цільової змінної на основі набору початкових даних (масивів даних НВ) на вході системи [17].

Домінуючими підходами на сьогодні в концепції дерев рішень є системи на основі методів CART (спрямованих для розв‘язку задач класифікації та регресивного аналізу), а також системи на основі схеми C4.5 та її сучасних модифікації (для розв‘язку задач розпізнавання та класифікації) та ID3. Схема ID3 базується на використанні обмеженого ентропійного критерію – структура ЛДК будується до тих пір, поки для кожної результуючої вершини (листа дерева) не залишаться лише об‘єкти одного фіксованого класу, або доки сама процедура розгалуження в дереві, що будується дає зменшення початкового ентропійного критерію. Схема C4.5/C5.0 базується на відомому критерії Gain-Ratio (нормативний ентропійний критерій), причому в якості критерію зупинки процедури розгалуження (побудови дерева) використовується обмеження на кількість об‘єктів для результуючої вершини (листа структури ЛДК). Відмітимо, що процедура відсіканні в структурі ЛДК проводиться за схемою Error-Based Pruning, яка базується на загальній оцінці здатності узагальнення для прийняття рішення щодо видалення гілок та вершин конструкції дерева класифікації. Схема CART в своїй роботі використовує критерії Джині, причому процедура відсіканні в структурі ЛДК проводиться за схемою *Cost-Complexity Pruning*, а для випадку наявних пропусків атрибутів використовується базова схема сурогатних предикатів.

Відмітимо, що базову ідею методів розгалуженого вибору ознак (вершин алгоритмів) в структурі АДК можна визначити як оптимальну апроксимацію деякої початкової НВ набором ранжованих алгоритмів класифікації (ознак, атрибутів об‘єкту у випадку ЛДК), то на перший план виходить центральне питання – задача вибору ефективного критерію розгалуження (відбору вершин, атрибутів, ознак дискретних об‘єктів для ЛДК та алгоритмів для АДК). Ці принципові задачі розглядаються в роботах, де піднімаються питання якісної оцінки окремих дискретних ознак, їх наборів та фіксованих сполучень, що дозволяє запровадити ефективний механізм реалізації розгалуження.

Відомо, що структури моделей дерев класифікації (ЛДК/АДК) характеризується компактністю з одного боку та нерівномірністю заповнення (розрядженістю) ярусів з іншого боку в порівнянні з конструкціями регулярних дерев [18 – 24], причому важливими питаннями залишаються питання збіжності процесу побудови дерев

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

класифікації за методами розгалуженого вибору ознак та питання вибору критерію зупинки процесу синтезу логічного дерева [25 – 34]. Відмітимо, що концепції дерев класифікації не протирічить можливість в якості ознак (вершин структури) дерева класифікації використовувати не тільки окремі атрибути (ознаки) об‘єктів їх сполучення (ідея узагальненої ознаки, розглядалась в роботі [9]) та набори, але якщо піти далі та не розглядати в якості розгалужень атрибути об‘єктів (ознаки) – а відбирати окремі незалежні алгоритми розпізнавання, то на виході буде отримане нова структура – АДК [20]. Саме структурам АДК в розрізі обмеженого методу і буде присвячена дана робота.

Мета роботи та задачі дослідження. Отже, зважаючи на все вище сказане, метою даної роботи є створення простого та ефективного обмеженого методу побудови деревоподібних моделей розпізнавання та класифікації на основі алгоритмічних дерев класифікації для навчальних вибірок дискретної інформації великого об‘єму – який характеризується структурою отриманих дерев класифікації з незалежних алгоритмів розпізнавання оцінених на основі функціоналу розрахунку їх загальної ефективності для широкого класу прикладних задач.

Саме для досягнення даної мети – був поставлений набір таких завдань:

1) Аналіз концепції синтезу структур дерев класифікації.

2) Побудова загального методу синтезу обмежених структур АДК на основі модифікованої процедури препрунингу.

Обмежені методи дерев класифікації. На даному етапі роботи запропонуємо обмежені методи для структур ЛДК та АДК (дерев класифікації), які дозволяють подолати певні негативні концептуальні моменти та обмеження, які притаманні даним схемам з точки зору ресурсних потреб та результуючої складності побудованих моделей дерев класифікації [25].

Так спочатку звернемо увагу, що в загальній схемі методу побудови моделі ЛДК (на основі покрокової селекції елементарних ознак), яка була описана в роботах [21] раніше – є принциповий недолік, який пов‘язаний з тим що зі зростанням кількості вершин (ярусів структури дерева класифікації) в конструкції ЛДК кількість елементарних ознак φ_i^j (тут i – номер елементарної ознаки в наборі, j номер ярусу розташування відповідної ознаки) в дереві значно збільшується. Звичайно таке ускладнення результуючої моделі ЛДК (конструкційної складності) негативно впливає на апаратні можливості системи класифікації (пам‘ять, процесорний час) та загальну можливість сприйняття і аналізу побудованої моделі без зовнішнього виділення правил класифікації в

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

структурі дерева. Для того, щоби подолати ці принципово негативні моменти методу дерева класифікації запропонуємо наступну модифікацію методу ЛДК.

Обмежений метод побудови ЛДК. На початковому етапі зафіксуємо деяке додатне число Z. Нехай маємо побудоване ЛДК (дерево класифікації після визначеної кількості кроків побудови моделі ЛДК) наступної загальної структури – (Рис. 1), яке відображає деякий предикат (побудовану узагальнену ознаку) $p_1(x)$.

Звернемо увагу, що в роботі [9] при представленні методу побудови ЛДК (на основі селекції елементарних ознак) на етапі тесту обчислювали деяке число S, яке фігурує у співвідношенні:

$$\frac{S}{m} \geq \delta . \tag{2}$$

Тепер, крім числа S, для кожного незакінченого шляху $r_1 r_2 r_3$ в структурі логічного дерева (рис. 1) розраховуємо ще число $S_{r_1 r_2 r_3}$, де $S_{r_1 r_2 r_3}$ – кількість всіх пар $(x_i, f_R(x_i))$ з НВ, які фактично належать шляху $r_1 r_2 r_3$ і для яких виконується співвідношення:

$$f_R(x_i) \neq l(r_1 r_2 r_3) . \tag{3}$$

Таким чином, $S_{r_1 r_2 r_3}$ – це число всіх тих помилок, які здійснюється деяким предикатом $p_1(x)$ (узагальненою ознакою), яке представляється даним ЛДК загальної структури (Рис. 1) на фіксованому шляху $r_1 r_2 r_3$ в конструкції даного дерева.

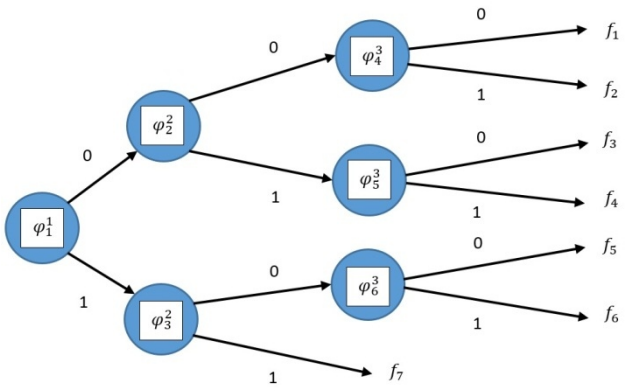

Рис. 1. Структура ЛДК, побудованого за даними початкової НВ на основі селекції елементарних ознак

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

На наступному етапі вибираємо число Z таких шляхів $(r_1r_2r_3)_1,...,(r_1r_2r_3)_Z$, для яких число $S_{r_1r_2r_3}$ буде, по можливості, найбільшим.

Приклад. Нехай задано $Z=3$, та має місце наступне базове співвідношення:

$$S_{000} \geq S_{100} \geq S_{101} \geq S_{001} \geq S_{010} \geq S_{011}. \quad (4)$$

Тоді вибираються тільки шляхи 000, 100, 101. Наступна добудова відбір вершин (елементарних ознак) $\varphi_{r_1r_2r_3}$ здійснюється тільки для цих шляхів.

Зафіксуємо – що дану модифіковану схему побудови дерева класифікації (ЛДК на основі селекції елементарних ознак) будемо називати обмеженим методом побудови ЛДК.

Відмітимо, що за даною схемою, в процесі побудови дерева класифікації продовжуються тільки ті шляхи (загальної структури ЛДК) за якими відбувається найбільша кількість помилок класифікації.

В кінці слід зауважити, що при застосуванні тільки що вказаного процесу в кінці шляхів $r_1r_2r_3$, які не входять у відібрані Z шляхів, значення $l(r_1r_2r_3)$ зберігаються, причому процес модифікованого методу побудови ЛДК можна застосовувати в тому випадку, коли початкова НВ не є фіксованою – тобто коли на кожному кроці процедури побудови дерева класифікації подається своя вибірка (частина НВ).

Отже, запропонована вище схема побудови дерева класифікації дозволяє фактично запровадити механізм регулювання точності моделі дерева, яка будується враховуючи загальну кількість помилок класифікації на тому чи іншому шляху (етапі побудови) загальної структури логічного дерева.

Зрозуміло, що дана ідея може працювати і на рівні структури АДК, причому враховуючи певні її особливості. Отже зважаючи на все вище сказане запропонуємо наступну модифікацію методу побудови структури АДК першого типу, який був представлений в роботах [20].

Обмежений метод побудови АДК. Нехай, на початку задана деяка НВ загального типу (1) – у вигляді послідовності навчальних пар $(x_i, f_R(x_i))$, потужністю – m, розмірністю ознакового простору – n та фіксований набір різнотипних алгоритмів класифікації $(a_1,...,a_M)$. Відмітимо, що робота побудованих моделей дерев класифікації перевіряється на масиві даних ТВ, потужністю – T (класова належність яких також відома).

Зауважимо, що тут дані початкової НВ задають деяке розбиття R на класи $(H_1,...,H_k)$, а відповідні алгоритми a_i можуть бути не зв‘язані єдиною концепцією розпізнавання, а реалізовувати різноманітні методи та алгоритми класифікації (для прикладу це можуть бути звичайні геометричні алгоритми – принцип роботи яких полягає в апроксимації навчальної вибірки відповідними геометричними об‘єктами, алгоритми обчислення оцінок, потенціальних функції, тощо). Відмітимо, що результатом роботи кожного з зафіксованих (відібраних з бібліотеки алгоритмів деякої інформаційної системи) автономних алгоритмів класифікації та розпізнавання a_i, на відповідному кроці генерації АДК, є одна або декілька узагальнених ознак – f_j (певних правил класифікації НВ), які і описують (апроксимують) визначену частину початкової навчальної вибірки. Так для випадку відомих геометричних алгоритмів розпізнавання – відповідними результуючими узагальненими ознаками будуть геометричні об‘єкти, які покривають НВ в ознаковому просторі задачі розмірності – n.

Зрозуміло, що в реальних прикладних задачах можливі випадки, коли відповідний алгоритм класифікації a_i не може побудувати узагальнену ознаку f_j – в зв‘язку зі складним розташуванням класів H_k в ознаковому просторі задачі, або певними концептуальними та реалізаційними обмеженнями самого алгоритму класифікації. Тоді, по аналогії з ЛДК можливий випадок, коли побудовані алгоритмом класифікації a_i (побудовані узагальнені ознаки f_j) неповністю апроксимують початкову НВ, або така ситуація передбачена самою схемою алгоритму генерації АДК (як приклад, наявність початкового обмеження в схемі алгоритму дерева класифікації – про генерацію не більше однієї узагальнені ознаки f_j на кожному етапі побудови моделі АДК).

Зауважимо, об‘єкти початкової НВ, які не підпадають під побудовану схему апроксимації вибірки послідовністю узагальнених ознак f_j (на останньому етапі процедури синтезу АДК) відносяться до відмов (помилок) класифікації першого типу – En_{tr} і аналогічно для даних ТВ неправильно класифіковані дискретні об‘єкти – також відносять до помилок першого типу – Et_{tr}.

Звернемо увагу, що АДК першого типу складається з ярусів, кожний з яких фактично відповідає певному кроку (етапу) побудови (апроксимації даних початкової НВ) дерева класифікації. Причому для кожного алгоритму класифікації на тому чи іншому кроці апроксимації

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

можна розрахувати його ефективність відносно робочих даних – $(S / P_{pt}(TS^{-}))$, причому ця величина має бути більше або дорівнювати ніж задане на початку обмеження δ (зрозуміло, що в деяких реалізаціях схеми АДК величина δ може бути використана в якості критерію зупинки K_{Stop} процедури розгалуження в структурі дерева класифікації) .

Зауважимо, що тут $S-$ загальна кількість помилок класифікації для фіксованого алгоритму на певному кроці (етапі генерації АДК), а $P_{pt}(TS^{-})$ потужність (об‘єм) підмножини початкової НВ яка подається на вхід даного алгоритму на відповідному ярусі (рівні або етапі) дерева класифікації, що будується.

Тоді доцільно по аналогії з обмеженим методом ЛДК для кожного ярусу структури АДК (етапу побудови дерева класифікації) розраховувати величини $S_{a_1,\ldots,a_i}$ які характеризують кількість всіх пар $(x_i, f_R(x_i))$ з масиву початкової НВ, які не можуть бути апроксимовані послідовністю фіксованих алгоритмів класифікації $a_1,\ldots,a_i$. Таким чином, $S_{a_1,\ldots,a_i}$ це число всіх тих помилок класифікації, які здійснюється деякою послідовністю УО (побудованих на відповідних рівнях структури АДК), яке представляється даним АДК загальної структури (рис. 2) для фіксованого набору алгоритмів розпізнавання та класифікації $a_1,\ldots,a_i$.

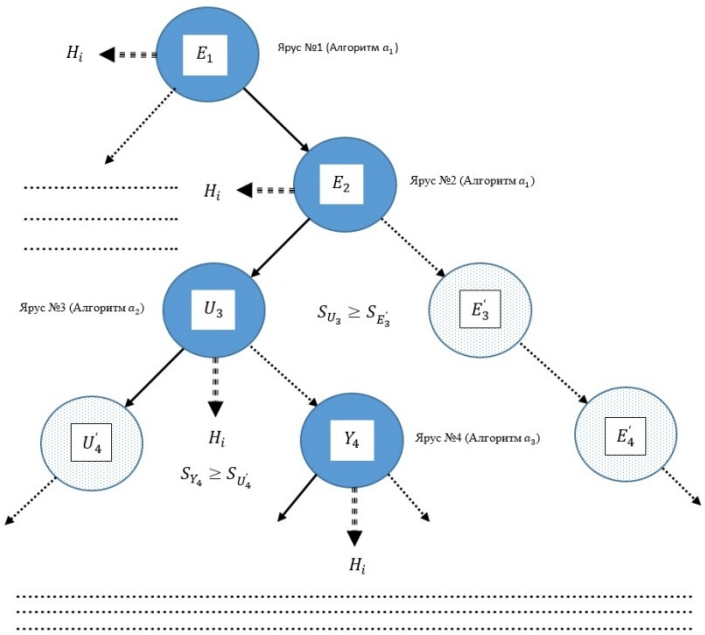

Рис. 2. Фрагмент структури дерева класифікації, побудоване обмеженим методом АДК

Тоді за даною схемою побудови структури АДК вибирається лише той шлях або певні шляхи в конструкції дерева (в залежності від типів дерев класифікації, які можуть бути різні), для який величина $S_{a_1,...,a_i}$ буде по можливості максимальною – тобто добудовується шлях в структурі АДК з найбільшою кількістю помилок класифікації. Причому наступна добудова та відбір вершин (алгоритмів класифікації – параметрів УО) $a_1,...,a_i$ здійснюється тільки для цих шляхів в структурі АДК.

Зауважимо при такому підході в побудові обмежених методів ЛДК та АДК після побудови моделі дерева класифікації можлива його фактична добудова (донавчання), що дає пряму можливість впливу на точність побудованої моделі системи класифікації.

Відмітимо, що за обмеженою схемою АДК, в процесі побудови дерева класифікації продовжуються (вибираються для добудови УО) тільки ті шляхи (загальної структури АДК) за якими відбувається найбільша кількість помилок класифікації. Зрозуміло, що при застосуванні тільки що вказаного процесу побудови структури АДК в будь який момент є можливість повернення до відкинутих шляхів для добудови структур (шляхів) дерева класифікації.

Підкреслимо, що запропонована вище схема побудови дерева класифікації (обмежений метод АДК) дозволяє регулювання точності (ефективності) моделі дерева, яка будується враховуючи загальну кількість помилок (всіх типів) класифікації на тому чи іншому шляху (етапі побудови) загальної структури алгоритмічного дерева. Важливою є принципова можливість побудови моделі АДК з наперед заданою точністю відносно масиву даних початкової НВ. Така можливість досягається шляхом обмеження кількості кроків процедури генерації АДК, системою обмежень щодо інформаційної ємності, кількості та параметрів узагальнення (області НВ, що апроксимується) набору узагальнених ознак, які будуються на відповідних етапах конструювання результуючого дерева класифікації.

Експериментальна частина. Запропонована в даному дослідженні схема побудови обмеженої структури АДК дозволяє досить гнучко регулювати складність моделі дерева класифікації, що будується, або будувати модель розпізнавання з наперед заданою точністю відповідно до умов поточної задачі, причому задача відбору моделі дерева класифікації (серед фіксованого набору побудованих структур ЛДК/АДК) для конкретної задачі визначається множиною параметрів, які мають визначальну важливість відносно поточної прикладної задачі (наборів

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

даних НВ/ТВ). Зрозуміло, що для порівняння та відбору конкретної моделі дерева класифікації з фіксованого набору необхідно виділити найбільш важливі їх характеристики (розмірність ознакового простору, кількість вершин, кількість переходів конструкції дерева тощо) та визначити їх похибку відносно масиву вхідних даних.

Принципово важливим моментом на даному етапі дослідження є аналіз критеріїв якості отриманих інформаційних моделей, які залежать від похибки моделі, потужності початкового масиву даних НВ та ТВ (кількість навчальних пар та розмірність ознакового простору задачі), кількості структурних параметрів моделі і так далі. Зрозуміло, що критично важливими параметрами побудованої моделі АДК, які необхідно мінімізувати є помилки моделі відповідно на масивах даних НВ, ТВ та для кожного з класів (частин, підмножин початкової НВ), які задані початковою умовою поточної прикладної задачі.

Відмітимо, що одним з найважливіших показників, який характеризує базові властивості отриманих моделей АДК є базовий показник узагальнення даних початкової НВ деревом класифікації (моделлю класифікації) який розраховується наступним чином:

$$I_{Main} = \frac{m \cdot O_{Uz}}{V_{All} + N_{All} + 2P_{All}}. \qquad (5)$$

Даний показник узагальнення моделі дерева класифікації (структури АДК) відображає його базові параметри (характеристики) дерев класифікації та може бути застосован в якості критерію оптимальності в процедурі оцінки довільної деревоподібної схеми розпізнавання, наприклад у випадку методів побудови та відбору випадкових дерев класифікації (з врахуванням їх відповідних структурних параметрів). Для довільної прикладної задачі важливо максимізувати параметр I_{Main} (показник узагальнення моделі АДК), що дозволяє добитися оптимальної структури дерева класифікації та забезпечує фактично максимальний стиск даних початкової НВ (представити масив початкових даних мінімальним за структурною складністю деревом). Також слід підкреслити, що принциповим моментом при побудові структур ЛДК/АДК (моделей класифікації) залишається питання зменшення складності структури дерева (мається на увазі кількість вузлів, вершин, ознак, алгоритмів в структурі дерева класифікації, загальна кількість переходів в структурі моделі), параметри загальних витрат пам‘яті та процесорного часу інформаційної системи. Так важливим показником якості побудованої моделі у вигляді дерева класифікації з врахуванням параметрів структури моделі АДК – є

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

загальний інтегральний показник якості представлений в наступній формі:

$$Q_{Main} = \frac{Fr_{All}}{O_{Uz} \cdot \sum_i p_i} \cdot e^{-\frac{Er_{All}}{M_{All}}}. \quad (6)$$

Відмітимо, що тут параметр Er_{All} – загальна кількість помилок моделі АДК на масивах даних початкових тестової та навчальної вибірки – $Er_{All} = En_{tr} + Et_{tr}$, відповідно M_{All} – загальна потужність цих двох масивів даних – $M_{All} = m + T$. Параметр Fr_{All} – характеризує загальну кількість вершин отриманої моделі АДК з результуючими значеннями f_R (ФР, тобто листів дерева класифікації), а параметр O_{Uz} – представляє загальну кількість всіх узагальнених ознак в структурі моделі АДК. Набір параметрів p_i представляє собою найбільш важливі характеристики дерева класифікації (відповідно до структур ЛДК/АДК), що оцінюється (наприклад кількість елементарних ознак або узагальнених ознак, що використовуються в моделі дерева класифікації, кількість переходів між вершинами, ярусами дерева класифікації, тощо).

Відмітимо, що даний інтегральний показник якості моделі АДК буде приймати значення в межах нуля та одиниці. Чим менший він буде тим гірша буде якість побудованого дерева класифікації, а чим більший буде показник тим краще буде отримана модель. Так запропонована інтегральна оцінка якості моделі дерева класифікації (структури АДК) відображає його базові параметри (характеристики) дерев класифікації, також може бути застосована в якості критерію оптимальності в процедурі оцінки довільної деревоподібної схеми розпізнавання (відповідно до своїх параметрів моделі).

Так в Ужгородському національному університеті був розроблений програмний комплекс "Оріон ІІІ" для генерації автономних систем розпізнавання та класифікації де алгоритмічна бібліотека системи нараховує 13 алгоритмів (схем розпізнавання) серед яких запропонована вище алгоритмічна реалізації побудови АДК обмеженим методом. Причому базова задача на якій перевірялася робота обмеженого методу побудови АДК була задача класифікації масиву геологічних даних (задача про розділення нафтоносних пластів). Так для розпізнавання об‘єктів використовувався набір з 22 елементарних ознак (атрибутів), в НВ представлена інформації про об‘єкти двох класів, а на етапі екзамену побудована система класифікації (модель АДК) має забезпечити ефективне розпізнавання об‘єктів невідомої класифікації відносно цих двох класів. На початковому етапі роботи програмної системи навчальна

вибірка була автоматично перевірена на коректність (пошук та видалення однакових об‘єктів різної належності – помили першого роду).

Так в початковій НВ представлена інформації про розбиття R на два класи. На етапі екзамену (на основі ТВ) побудована система класифікації має забезпечити ефективне розпізнавання (класифікацію) об‘єктів невідомої класифікації відносно цих двох класів. Зауважимо, що в масиві навчаючої інформації переважали навчальні пари класу H_1 (нафтоносні пласти) в пропорції $\approx(1.5/1)$, а масив НВ складався з 1250 об‘єктів (наборів відомої класифікації), причому ефективність сконструйованої системи розпізнавання оцінювалася на тестовій виборці об‘єму 240 об‘єктів. Відмітимо, що масив ТВ представляв собою відокремлену частину початкової НВ (складався з дискретних об‘єктів відомої класифікації). Зауважимо, що звичайно такий об‘єм тестової вибірки не достатній для всебічного аналізу якості побудованих моделей дерев класифікації, але в зв‘язку з обмеженим характером самої НВ навіть така ТВ дає змогу оцінити та проаналізувати основні параметри синтезованих структур ЛДК/АДК. Дані масивів навчаючих та тестових вибірок отримані на основі геологічної розвідки на території Закарпатської області в період з 2001 року по 2011 рік. Так фрагмент основних результатів, приведених вище експериментів, порівняльних тестів методів побудови моделей АДК (структур дерев класифікації) на масиві даних даної прикладної задачі, представлений в (табл. 1).

Підкреслимо, побудовані моделі (структури) АДК забезпечили необхідний рівень точності, ефективності заданих умовою прикладної задачі, необхідну швидкодію та витрати робочої пам‘яті системи, але показували різну структурну складність побудовах дерев класифікації (параметрів складності конструкції ЛДК) та набору узагальнених ознак (в порівняльних випадках моделей алгоритмічного дерева класифікації – структур АДК).

Відмітимо, що представлена в дослідженні загальна оцінки якості моделі АДК (зрозуміло, що вона може бути адаптована і для випадку структур ЛДК) фіксує найважливіші характеристики (параметри) побудованих дерев класифікації та може бути застосований в якості критерію оптимальності в процедурі побудови АДК та фінального відбору (ранжуванні моделей) з множини моделей АДК.

Звернемо увагу, що структура АДК як і обмежений метод алгоритмічного дерева оперує лише вже готовими (побудованими) наборами, узагальненими ознаками (алгоритмами класифікації), та його може зовсім не цікавити, яким алгоритмом чи способом (схемою, правилом, методом) вони отримані, причому кожна із сконструйованих схем за обмеженим методом алгоритмічного дерева буде являти собою

загальну систему розпізнавання (модель АДК), яку можна застосовувати для практичної роботи (обробки великих масивів експериментальних даних у вигляді дискретних наборів).

Таблиця 1

Порівняльна таблиця моделей / методів дерев класифікації (ЛДК/АДК)

No	Метод (схема) синтезу структури (моделі) дерева класифікації (ЛДК/АДК)	Інтегральний показник якості моделі дерева класифікації Q_{Main}	Загальна кількість помилок моделі на НВ та ТВ Er_{All}
1	Метод повного ЛДК на основі селекції елементарних ознак (наборів ознак)	0,004789	2
2	Модель ЛДК з одноразовою оцінкою важливості ознак (наборів ознак)	0,002263	3
3	Обмежений метод побудови структури ЛДК (Z=5)	0,003168	4
4	Обмежений метод побудови структури ЛДК (Z=9)	0,003029	3
5	Метод алгоритмічного дерева (типу I)	0,005234	0
6	Метод алгоритмічного дерева (типу II)	0,002941	0
7	Метод АДК (типу I) на основі алгоритму гіперсфер в якості УО	0,005445	0
8	Метод АДК (типу I) на основі алгоритму гіперкуба в якості УО	0,005139	1
9	Обмежений метод побудови моделі АДК (Z=6)	0,003018	2
10	Обмежений метод побудови моделі АДК (Z=10)	0,003176	1

Важливим моментом є те, що отримана схема класифікації буде представляти собою в певній мірі новий алгоритм розпізнавання (зрозуміло, що синтезований з відомих алгоритмів та методів), причому отримана структура АДК (нова схема класифікації) характеризується високою універсальністю відносно прикладного застосування та відносно компактною структурою самої моделі, однак вимагає відносно великих апаратних витрат для зберігання узагальнених ознак (або їх наборів) та початкової оцінки якості алгоритмів класифікації за даними НВ. Причому моделі АДК в порівняні з структурами ЛДК мають високу швидкодію правил класифікації, порівняні апаратні витрати для зберігання та роботи самої структури дерева та високу якість класифікації.

Висновки. В роботі вирішена задача побудови обмеженого методу синтезу моделей алгоритмічних дерев класифікації на основі апроксимації НВ набором незалежних алгоритмів класифікації в умовах обмежень на шляхи добудови структури АДК.

Наукова новизна отриманих результатів базується в тому, що вперше запропонований обмежений метод побудови структур АДК на основі оцінки та ранжування набору автономних алгоритмів розпізнавання та класифікації для генерації структури дерева класифікації (моделі АДК) з наборами обмежень на напрямки добудови конструкції АДК (моделі дерева класифікації). Причому на кожному кроці розгалуження дерева класифікації апроксимується певна частина НВ (або її підмножина). Відмітимо критерій розгалуження для структур АДК (в обмеженому методі побудови АДК) можна використовувати не тільки для оцінки якості окремих алгоритмів класифікації, але й для розрахунку ефективності зв‘язаних наборів алгоритмів, що в перспективі дозволяє досягти більш оптимальної структури синтезованого АДК за початковими даними НВ. В роботі запропонований набір загальних показників (параметрів), який дозволяє ефективно представити загальні характеристики моделі АДК, можливе його використання для відбору найбільш оптимального АДК з набору побудованих на основі методів випадкових дерев класифікації.

Практична цінність отриманих результатів полягає в тому що запропонований обмежений метод побудови моделей АДК (структур ЛДК/АДК) був реалізований в бібліотеці алгоритмів універсальної програмної системі “ОРІОН ІІІ” для розв‘язку різноманітних практичних задач класифікації (розпізнавання) різнотипних масивів дискретних об‘єктів.

Відмітимо, що проведені практичні випробовування підтвердили працездатність математичного забезпечення та запропонованих моделей АДК, розробленого програмного забезпечення, що дозволяє зробити рекомендацію щодо використання даного підходу (обмеженого методу моделей ЛДК/АДК) та його програмної реалізації для широкого спектру прикладних задач класифікації та розпізнавання дискретних об‘єктів.

Перспективи подальших досліджень можуть бути спрямовані в бік розвитку методів алгоритмічних дерев класифікації (методів бустингу структур АДК) оптимізації програмних реалізацій запропонованого обмеженого методу побудови АДК, а також його практичної апробації на множині реальних задач класифікації та розпізнавання.

Список літератури:

1. *Srikant R.* Mining generalized association rules / *R. Srikant, R. Agrawal* // Future Generation Computer Systems. – 1997. – Vol. 13. – № 2. – P. 161-180.

2. *Hastie T.* The Elements of Statistical Learning / *T. Hastie, R. Tibshirani, J. Friedman*. – Stanford, 2008. – 768 p.
3. *Quinlan J.R.* Induction of Decision Trees / *J.R. Quinlan* // Machine Learning. – 1986. − № 1. – P. 81-106.
4. *Vasilenko Y.A.* Construction and optimization of recongnizing systems / *Y.A. Vasilenko, E.Y. Vasilenko, A.I. Kuhayivsky, I.O. Papp* // Scientific and technical journal "Information technologies and systems". – 1999. – № 1. – P. 122-125.
5. *Povhan I.* Designing of recognition system of discrete objects / *I. Povhan* // 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2016, Ukraine. – Lviv, 2016. – P. 226-231.
6. *Mitchell T.* Machine learning / *T. Mitchell*. – New York: McGrawHill, 1997. – 432 p.
7. *Povhan I.* General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects / *I. Povhan* // Collection of proceedings "Electronics and information technology". – 2019. – Vol. 11. – P. 73-80.
8. *Breiman L.L.* Classification and regression trees / *L.L. Breiman, J.H. Friedman, R.A. Olshen et al.*]. – Boca Raton: Chapman and Hall/CRC, 1984. – 368 p.
9. *Vasilenko Y.A.* Automating the construction of classification systems based on agent - schemes / *Y.A. Vasilenko, F.G. Vashuk, I.F. Povkhan* // Mathematical modeling, optimization and information technologies: International Joint Conference MDIF-2012, Kisheneu, Moldova, 2012. – Kisheneu, 2012. – P. 444-446.
10. *Vtogoff P.E.* Incremental Induction of Decision Trees / *P.E. Vtogoff* // Machine Learning. – 1989. – № 4. – P. 161-186.
11. *Amit Y.* Joint induction of shape features and tree classifiers / *Y. Amit, D. Geman, K. Wilder* // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1997. – Vol. 19. – № 11. – P. 1300-1305.
12. *Dietterich T.G.* Machine learning bias, statistical bias and statistical variance of decision tree algorithms [Electronic resource] / *T.G. Dietterich, E.B. Kong*. – Corvallis: Oregon State University, 1995. – 14 p. – Access mode : http://www.cems.uwe.ac.uk/~irjohnso/coursenotes/uqc832/trbias.pdf
13. *Mingers J.* An empirical comparison of pruning methods for decision tree induction / *J. Mingers* // Machine learning. – 1989. – Vol. 4. – № 2. – P. 227-243.
14. *Povhan I.* Question of the optimality criterion of a regular logical tree based on the concept of similarity / *I. Povhan* // Collection of proceedings "Electronics and information technology". – 2020. – Vol. 13. – P. 12-16.
15. *Subbotin S.A.* Construction of decision trees for the case of low-information features / *S.A. Subbotin* // Radio Electronics, Computer Science, Control. – 2019. – № 1. – P. 121-130.
16. *Lupei M.* Identification of authorship of Ukrainian-language texts of journalistic style using neural networks / *M. Lupei, A. Mitsa, V. Repariuk, V. Sharkan* // Eastern-European Journal of Enterprise Technologies. – 2020. – Vol. 1 (2 (103)). – P. 30-36. DOI: https://doi.org/10.15587/1729-4061.2020.195041
17. *Bodyanskiy Y.* Hybrid neuro-neo-fuzzy system and its adaptive learning algorithm / *Y. Bodyanskiy, O. Vynokurova, G. Setlak and I. Pliss* // Computer Sciences and Information Technologies (CSIT): Xth Scien. and Tech. Conf., Lviv, 2015. – Lviv. – 2015. – P. 111-114.
18. *Karimi K.* Generatio n and Interpretation of Temporal Decision Rules / *K. Karimi, H.J. Hamilton* // International Journal of Computer Information Systems and Industrial Management Applications. – 2011. – Vol. 3. – P. 314-323.
19. *Kotsiantis S.B.* Supervised Machine Learning: A Review of Classification Techniques / *S.B. Kotsiantis* // Informatica. – 2007. – № 31. – P. 249-268.
20. *Povkhan I.F.* Features of synthesis of generalized features in the construction of recognition systems using the logical tree method / *I.F. Povkhan* // Information technologies

and computer modeling ITKM-2019: materials of the international scientific and practical conference, Ivano- Frankivsk, May 20–25, 2019. – Ivano-Frankivsk, 2019. – P. 169-174.
21. *Vasilenko Y.A.* The importance of discrete signs / *Y.A. Vasilenko, F.G. Vashuk, I.F. Povkhan* // XX International Conference Promising ways and directions of improving the educational system, Uzhgorod, November 16–19, 2010. – Uzhgorod, 2010. – Vol. 21. – № 1. – P. 217-222.
22. *Deng H.* Bias of importance measures for multi-valued attributes and solutions / *H. Deng, G. Runger, E. Tuv* // Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN), Espoo, Finland, Jun 14–Jun 17, 2011. – Espoo, 2011. – P. 293-300.
23. *Kamiński B.* A framework for sensitivity analysis of decision trees / *B. Kamiński, M. Jakubczyk, P. Szufel* // Central European Journal of Operations Research. – 2017. – Vol. 26 (1). – P. 135-159.
24. *Dietterich T.G.* An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization / *T.G. Dietterich* // Machine learning. – 2000. – Vol. 40. – № 2. – P. 139-157.
25. *Povhan I.* Generation of elementary signs in the general scheme of the recognition system based on the logical tree / *I. Povhan* // Collection of proceedings "Electronics and information technology". – 2019. – Vol. 12. – P. 20-29.
26. *Subbotin S.* The dimensionality reduction methods based on computational intelligence in problems of object classification and diagnosis / *S. Subbotin, A. Oliinyk* // Recent Advances in Systems, Control and Information Technology / eds.: R. Szewczyk, M. Kaliczyńska. – Cham : Springer, 2017. – P. 11-19. – (Advances in Intelligent Systems and Computing, vol. 543).
27. *Subbotin S.A.* Methods and characteristics of locality-preserving transformations in the problems of computational intelligence / *S.A. Subbotin* // Radio Electronics, Computer Science, Control. – 2014. – № 1. – P. 120-128.
28. *Koskimaki H.* Two-level clustering approach to training data instance selection: a case study for the steel industry / *H. Koskimaki, I. Juutilainen, P. Laurinen, J. Roning* // Neural Networks: International Joint Conference (IJCNN-2008), Hong Kong, 1–8 June 2008: proceedings. – Los Alamitos: IEEE, 2008. – P. 3044–3049. DOI: 10.1109/ijcnn.2008.4634228
29. *Subbotin S.* The neuro-fuzzy network synthesis and simplification on precedents in problems of diagnosis and pattern recognition / *S. Subbotin* // Optical Memory and Neural Networks (Information Optics). – 2013. – Vol. 22. – № 2. – P. 97-103. DOI: 10.3103/s1060992x13020082
30. *Subbotin S.A.* Methods of sampling based on exhaustive and evolutionary search / *S.A. Subbotin* // Automatic Control and Computer Sciences. – 2013. – Vol. 47. – № 3. – P. 113-121. DOI: 10.3103/s0146411613030073
31. *De Mántaras R.L.* A distance-based attribute selection measure for decision tree induction / *De Mántaras R.L.* // Machine learning. – 1991. – Vol. 6. – № 1. – P. 81-92.
32. *Alpaydin E.* Introduction to Machine Learning / *E. Alpaydin*. – London: The MIT Press, 2010. – 400 p.
33. *Painsky A.* Cross-validated variable selection in tree-based methods improves predictive performance / *A. Painsky, S. Rosset* // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. – Vol. 39. – № 11. – P. 2142-2153. DOI:10.1109/tpami.2016.2636831
34. *Miyakawa M.* Criteria for selecting a variable in the construction of efficient decision trees / *M. Miyakawa* // IEEE Transactions on Computers. – 1989. – Vol. 38. – № 1. – P. 130-141.

References:

1. Srikant, R. (1997), "Mining generalized association rules", *Future Generation Computer Systems*,Vol. 13, No. 2, pp. 161-180.
2. Hastie, T. (2008), *The Elements of Statistical Learning*, Stanford, 768 p.

3. Quinlan, J.R. (1986), "Induction of Decision Trees", *Machine Learning*, No. 1, pp. 81-106.
4. Vasilenko, Y.A., Vasilenko, E.Y., Kuhayivsky, A.I., and Papp, I.O. (1999), "Construction and optimization of recongnizing systems", *Scientific and technical journal "Information technologies and systems",* No.1, pp. 122-125.
5. Povhan, I. (2016), "Designing of recognition system of discrete objects", *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, Ukraine, Lviv, 2016, pp. 226-231.
6. Mitchell, T. (1997), *Machine learning*, New York: McGraw-Hill, 432 p.
7. Povhan, I. (2019), "General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects", *Collection of proceedings "Electronics and information technology"*, Vol. 11, pp. 73-80.
8. Breiman, L.L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984), *Classification and regression trees*. Boca Raton: Chapman and Hall/CRC, 368 p.
9. Vasilenko, Y.A. Vashuk, F.G., and Povkhan, I.F. (2012), "Automating the construction of classification systems based on agent – schemes", *Mathematical modeling, optimization and information technologies: International Joint Conference MDIF-2012, Kisheneu, Moldova*, 2012, Kisheneu, pp. 444-446.
10. Vtogoff, P.E. (1989), "Incremental Induction of Decision Trees", *Machine Learning*, No. 4, pp. 161-186.
11. Amit Y., Geman, D., and Wilder, K. (1997), "Joint induction of shape features and tree classifiers", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 19, No. 11, pp. 1300-1305.
12. Dietterich, T.G., and Kong, E.B. (1995), "Machine learning bias, statistical bias, and statistical variance of decision tree algorithms [Electronic resource]", Corvallis: Oregon State University,. – 14 p. – Access mode: http://www.cems.uwe.ac.uk/~irjohnso/coursenotes/uqc832/trbias.pdf
13. Mingers, J. (1989), "An empirical comparison of pruning methods for decision tree induction", *Machine learning*, Vol. 4, No. 2, pp. 227-243.
14. Povhan, I. (2020), "Question of the optimality criterion of a regular logical tree based on the concept of similarity", *Collection of proceedings "Electronics and information technology"*, Vol. 13, pp. 12-16.
15. Subbotin, S.A. (2019), "Construction of decision trees for the case of low-information features*", Radio Electronics, Computer Science, Control*, № 1, pp. 121-130.
16. Lupei, M., Mitsa, A., Repariuk, V., and Sharkan, V. (2020), "Identification of authorship of Ukrainian-language texts of journalistic style using neural networks", *Eastern-European Journal of Enterprise Technologies,* Vol. 1 (2 (103)), pp. 30-36. DOI: https://doi.org/10.15587/1729-4061.2020.195041
17. Bodyanskiy Y., Vynokurova O., Setlak G. and Pliss I. (2015), "Hybrid neuro-neo-fuzzy system and its adaptive learning algorithm", *Computer Sciences and Information Technologies (CSIT): Xth Scien. and Tech. Conf.*, Lviv, 2015, Lviv, pp. 111-114.
18. Karimi, K. and Hamilton, H.J. (2011), "Generation and Interpretation of Temporal Decision Rules", *International Journal of Computer Information Systems and Industrial Management Applications*, Vol. 3, pp. 314-323.
19. Kotsiantis, S.B. (2007), "Supervised Machine Learning: A Review of Classification Techniques", *Informatica*, No. 31, pp. 249-268.
20. Povkhan, I.F. (2019), "Features of synthesis of generalized features in the construction of recognition systems using the logical tree method*", Information technologies and computer modeling ITKM-2019: materials of the international scientific and practical conference, Ivano- Frankivsk*, May 20–25, 2019, Ivano-Frankivsk, pp. 169-174.
21. Vasilenko, Y.A. Vashuk, F.G., and Povkhan, I.F. (2010), "The importance of discrete

signs", *XX International Conference Promising ways and directions of improving the educational system*, Uzhgorod, November 16–19, Uzhgorod, 2010, Vol. 21, No. 1, pp. 217-222.

22. Deng, H., Runger, G., and Tuv, E. (2011), "Bias of importance measures for multi-valued attributes and solutions", *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN),* Espoo, Finland, Jun 14–Jun 17, Espoo, 2011, pp. 293-300.

23. Kamiński, B., Jakubczyk M., and Szufel, P. (2017), "A framework for sensitivity analysis of decision trees", *Central European Journal of Operations Research*, Vol. 26 (1), pp. 135-159.

24. Dietterich, T.G. (2000), "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization", *Machine learning*, Vol. 40, No. 2, pp. 139-157.

25. Povhan, I. (2019), "Generation of elementary signs in the general scheme of the recognition system based on the logical tree", Collection of proceedings "Electronics and information technology", Vol. 12, pp. 20-29.

26. Subbotin, S., and Oliinyk, A. (2017), "The dimensionality reduction methods based on computational intelligence in problems of object classification and diagnosis", *Recent Advances in Systems, Control and Information Technology* / eds.: R. Szewczyk, M. Kaliczyńska, Cham: Springer, pp. 11-19, (Advances in Intelligent Systems and Computing, vol. 543).

27. Subbotin, S.A. (2014), "Methods and characteristics of localitypreserving transformations in the problems of computational intelligence", *Radio Electronics, Computer Science, Control*. No. 1, pp. 120-128.

28. Koskimaki, H., Juutilainen, I., Laurinen, P., and Roning, J. (2008), "Two-level clustering approach to training data instance selection: a case study for the steel industry", *Neural Networks: International Joint Conference (IJCNN-2008), Hong Kong,* 1–8 June 2008: proceedings, Los Alamitos: IEEE, pp. 3044-3049. DOI: 10.1109/ijcnn.2008.4634228

29. Subbotin, S. (2013), "The neuro-fuzzy network synthesis and simplification on precedents in problems of diagnosis and pattern recognition", *Optical Memory and Neural Networks (Information Optics),* Vol. 22, No. 2, pp. 97-103. DOI: 10.3103/s1060992x13020082

30. Subbotin, S.A. (2013), "Methods of sampling based on exhaustive and evolutionary search", *Automatic Control and Computer Sciences*, Vol. 47, No. 3, pp. 113-121. DOI: 10.3103/s0146411613030073

31. De Mántaras, R.L. (1991), "A distance-based attribute selection measure for decision tree induction", *Machine learning*, Vol. 6, No. 1, pp. 81-92.

32. Alpaydin, E. (2010), *Introduction to Machine Learning,* London: The MIT Press., 400 p.

33. Painsky, A., and Rosset, S. (2017), "Cross-validated variable selection in tree-based methods improves predictive performance", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 39, No. 11, pp. 2142-2153. DOI:10.1109/tpami.2016.2636831.

34. Miyakawa, M. (1989), "Criteria for selecting a variable in the construction of efficient decision trees", *IEEE Transactions on Computers*, Vol. 38, No. 1, pp. 130-141.

Статтю представив доктор техн. наук, проф. Дмитрієнко В.Д.

Надійшла/Received: 23.05.2021 р.

Povhan Igor, Doctor of Technical Sciences, assistant professor,
DHSU "Uzhhorod National University"
Uzhgorod national university, Ukraine, Narodna Square 3, Uzhgorod, Ukraine, 88000
e-mail: Igor.povkhan@uzhnu.edu.ua
ORCID ID: 0000-0002-7034-8702

УДК 004.8: 004.89: 519.7

Повхан І.Ф. Метод алгоритмічних дерев класифікації на основі обмежень / Повхан І.Ф. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2021. – № 1 (5). – С. 17 – 38.

Розглянуто спільне завдання побудови алгоритмічних дерев розпізнавання (класифікації) на основі обмеженого методу в теорії штучного інтелекту. Об'єктом дослідження є концепція алгоритмічного дерева класифікації на базі обмеженого методу. Предметом дослідження є актуальні методи, алгоритми і схеми (обмежені методи) побудови алгоритмічних дерев класифікації. Пропонується обмежений метод побудови алгоритмічних дерев класифікації, який для заданої навчальної вибірки довільного розміру будує деревоподібну структуру (модель дерева алгоритмів), яка складається з набору автономних алгоритмів класифікації і розпізнавання, оцінених на кожному кроці побудови дерева по вихідній вибірці. Пропонується обмежений метод побудови алгоритмічного дерева класифікації, основна ідея якого полягає в по крокової апроксимації початкової вибірки довільного обсягу і структури набором незалежних алгоритмів класифікації і розпізнавання. Даний метод при формуванні поточної вершини алгоритмічного дерева (вузла, узагальненої ознаки дерева алгоритмів) забезпечує виділення найбільш ефективних автономних алгоритмів класифікації з початкового набору і добудову тільки тих шляхів в структурі дерева де відбувається найбільша кількість помилок класифікації. Обмежений метод побудови алгоритмічного дерева класифікації дозволяє будувати різнотипні деревовидні моделі розпізнавання з наперед заданою точністю для широкого класу задач теорії штучного інтелекту. Розроблений і представлений в роботі обмежений метод алгоритмічного дерева класифікації отримав програмну реалізацію і був досліджений і зрівняний з методами логічних дерев класифікації, методами алгоритмічного дерева класифікації (першого і другого типу) при вирішенні задачі розпізнавання реальних даних геологічного типу. Іл.: 2. Табл.: 1. Бібліогр.: 34 назв.

Ключові слова: алгоритмічне дерево класифікації; розпізнавання реальних даних; класифікація; алгоритм класифікації; критерій розгалуження; обмежений метод.

УДК 004.8: 004.89: 519.7

Повхан И.Ф. Метод алгоритмических деревьев классификации на основе ограничений / Повхан И.Ф. // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2021. – № 1 (5). – С. 17 – 38.

Рассмотрена общая задача построения алгоритмических деревьев распознавания (классификации) на основе ограниченного метода в теории искусственного интеллекта. Объектом исследования является концепция алгоритмического дерева классификации на базе ограниченного метода. Предметом исследования являются актуальные методы, алгоритмы и схемы (ограниченные методы) построения алгоритмических деревьев классификации. Предлагается ограниченный метод построения алгоритмических деревьев классификации, который для заданной обучающей выборки произвольного размера строит древовидную структуру (модель дерева алгоритмов), которая состоит из набора автономных алгоритмов классификации и распознавания, оцененных на каждом шаге построения дерева по исходной выборке. Предлагается ограниченный метод построения алгоритмического дерева классификации, основная идея которого заключается в по шаговой аппроксимации начальной выборки произвольного объема и структуры набором независимых алгоритмов классификации и распознавания. Данный метод при формировании текущей вершины алгоритмического дерева (узла, обобщенной признаки дерева алгоритмов) обеспечивает выделение наиболее эффективных автономных алгоритмов классификации с начального набора и достройку

только тех путей в структуре дерева где происходит наибольшее количество ошибок классификации. Ограниченный метод построения алгоритмического дерева классификации позволяет строить разнотипные древовидные модели распознавания с наперед заданной точностью для широкого класса задач теории искусственного интеллекта. Разработанный и представленный в работе ограничен метод алгоритмического дерева классификации получил программную реализацию и был исследован и сравнен с методами логических деревьев классификации, методами алгоритмического деревья классификации (первого и второго типа) при решении задачи распознавания реальных данных геологического типа. Ил.: 2. Табл.: 1. Библиогр.: 34 назв.

Ключевые слова: алгоритмическое дерево классификации; распознавание реальных данных; классификация; алгоритм классификации; критерий ветвления; ограниченный метод.

UDC 004.8: 004.89: 519.7

Povhan I.F. Method of algorithmic classification trees based on constraints / Povkhan I.F. // Herald of the National Technical University "KhPI". Series of "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2021. – № 1 (5). – P. 17 – 38.

The general problem of constructing algorithmic recognition (classification) trees based on a limited method in the theory of artificial intelligence is considered. The object of this research is the concept of an algorithmic classification tree based on a bounded method. The subject of the research is actual methods, algorithms and schemes (limited method) for constructing algorithmic classification trees. A limited method for constructing algorithmic classification trees is proposed, which for a given initial training sample of any size builds a tree structure (algorithm tree model), which consists of a set of autonomous classification algorithms and recognition evaluated at each stage of constructing algorithm trees based on this initial sample. That is, a limited method for constructing an algorithmic classification tree is proposed, the main idea of which is a step-by-step approximation of the initial sample of an arbitrary volume and structure by a set of independent classification and recognition algorithms. This method, when forming the current vertex of the algorithmic tree, ensures that the most efficient (high-quality) autonomous classification algorithms are selected from the initial set and only those paths in the tree structure where the greatest number of classification errors occur are completed. The limited method of constructing an algorithmic classification tree makes it possible to build different types of tree-like recognition models with predefined accuracy for a wide class of problems in the theory of artificial intelligence. The limited method of the algorithmic classification tree developed and presented in this paper received a software implementation and was investigated and compared with the methods of logical classification trees, methods of the algorithmic classification tree (first and second types) when solving the problem of recognizing real geological data. Figs.: 2. Tabl.: 1. Refs.: 34 titles.

Keywords: algorithmic classification tree; pattern recognition; classification; classification algorithm; branching criterion; restricted method.