УДК 004.8: 004.89: 519.7
DOI: 10.25140/2411-5363-2020-2(20)-126-138

Ігор Повхан

МЕТОД ПОБУДОВИ АЛГОРИТМІЧНОГО ДЕРЕВА ДРУГОГО ТИПУ НА ОСНОВІ АПРОКСИМАЦІЇ НАВЧАЛЬНОЇ ВИБІРКИ НАБОРОМ АЛГОРИТМІВ КЛАСИФІКАЦІЇ

Актуальність теми дослідження. *Сучасні інформаційні технології, засновані на математичних моделях розпізнавання образів у вигляді ЛДК (логічних дерев класифікації), широко використовуються в соціально-економічних, екологічних та інших системах первинного аналізу та обробки великих масивів інформації, зрозуміло, що це пояснюється тим фактом, що такий підхід дозволяє усунути набір існуючих недоліки добре відомих класичних методів та досягти принципово новий результат. Робота присвячена тематиці моделей ЛДК, пропонує ефективний метод побудови моделей алгоритмічних дерев класифікації (АДК), які складаються з незалежних та автономних алгоритмів класифікації і будуть являти собою в певній мірі новий алгоритм розпізнавання (зрозуміло, що синтезований з відомих алгоритмів та методів).*

Постановка проблеми. *На сьогоднішній час відомі різні підходи та методи побудови моделей ЛДК (відомо більше ніж 3600 алгоритмів розпізнавання заснованих на різноманітних концепціях, які мають певні обмеження при їх використанні – точність, швидкодія, пам'ять, універсальність, надійність, тощо), проте всі вони, як правило, зводяться до побудови одного дерева класифікації за даними початкової навчальної вибірки. Зрозуміло, що доцільно не розробляти новий алгоритм, а запропонувати деяку концепцію раціонального використання вже накопиченого потенціалу алгоритмів та методів класифікації у вигляді моделей АДК, і саме тому дана робота має намір хоча би частково подолати ці обмеження та присвячена розробці методу побудови моделей алгоритмічних дерев класифікації.*

Аналіз останніх літературних даних. *Були розглянуті останні публікації у відкритому доступі, які присвячені проблематиці підходів, методів та алгоритмів логічних дерев класифікації (концепції дерев рішень) у задачах розпізнавання образів.*

Виділення недосліджених частин загальної проблеми. *Можливість ефективної та економної роботи запропонованого методу побудови алгоритмічного дерева класифікації на основі масивів навчальних вибірок великого об'єму.*

Постановка завдання. *Розробка простого та якісного методу побудови моделей АДК для великих масивів початкових вибірок шляхом синтезу мінімальних форм дерев класифікації та розпізнавання, які забезпечують ефективну апроксимацію навчальної інформації набором автономних та незалежних алгоритмів класифікації.*

Виклад основного матеріалу. *Виявлення простого та ефективного механізму за допомогою якого можна було би будувати алгоритмічне дерево класифікації (модель АДК) за фіксованою початковою інформацією у вигляді початкової навчальної вибірки (НВ). Дане алгоритмічне дерево класифікації буде безпомилково розпізнавати всю навчальної вибірки за якою побудоване дерево класифікації мати мінімальну структуру (структурну складність) та складатися з автономних алгоритмів класифікації в якості вершин конструкції (атрибутів дерева).*

Висновки відповідно до статті*. Запропонований метод побудови моделей АДК другого типу дозволяє працювати з навчальними вибірками великого об'єму та забезпечує високу швидкість та економність апаратних ресурсів в процесі генерації кінцевої схеми класифікації, будувати дерева класифікації з наперед заданою точністю.*

Ключові слова: *задачі розпізнавання, дерева класифікації, алгоритмічне дерево, схема розпізнавання, дискретний об'єкт, узагальнена ознак.*

Рис.: 2. Табл.: 2. Бібл.: 16.

Актуальність теми дослідження. Інформаційні технології, які засновані на математичних моделях розпізнавання образів у вигляді ЛДК (моделей логічних дерев класифікації), широко використовуються в соціально-економічних, екологічних та інших системах обробки інформації. Це пояснюється тим фактом, що такий підхід дозволяє усунути набір недоліки класичних методів та досягти принципово новий результат, ефективно та раціонально використовуючи потужності обчислювальних систем [1; 2]. Причому на сьогоднішній день відомо більше трьох 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 [3]. 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 [4]. Тобто покриття навчальної вибірки набором елементарних ознак у випадку ЛДК, або покриття навчальної вибірки фіксованим набором автономних алгоритмів розпізнавання та класифікації у випадку АДК (алгоритмічних дерев класифікації), породжує фіксовану деревоподібну структуру

даних, яка в якісь мірі забезпечує навіть стиск та перетворення початкових даних НВ (навчальної вибірки) – а отже дозволяє суттєву оптимізацію та економію апаратних ресурсів інформаційної системи [5]. Відмітимо, що галузь застосування концепції ЛДК в даний час надзвичайно об‘ємна, а множина задач та проблем, які розв’язуються даним апаратом може бути зведена до таких трьох базових сегментів – задачі опису структур даних, задачі розпізнавання та класифікації, задачі регресії [6].

Так, здатність ЛДК виконувати одномірне розгалуження для аналізу впливу (важливості, якості) окремих змінних дає можливість працювати зі змінними різних типів у вигляді предикатів (у випадку АДК – відповідними автономними алгоритмами класифікації та розпізнавання) [7]. В даному випадку структура логічного дерева представлена у вигляді гілок та вузлів, причому на гілках дерева розташовуються деякі мітки (атрибути, значення) від яких залежить цільова функція (у випадку ЛДК – функція розпізнавання), а в вузлах (вершинах) знаходяться значення функції розпізнавання (ФР) або розширені атрибути переходів. Відмітимо, що при побудові ЛДК центральними питаннями залишаються питання вибору критерія атрибуту (вершини ЛДК), за якою відбудеться розбиття початкової НВ, критерію зупинки навчання (побудови структури ЛДК) та критерію відкидання гілок логічного дерева (піддерев ЛДК). Саме на цьому етапі виникає принципове питання теорії ЛДК – питання можливої побудови всіх варіантів логічних дерев, які відповідають початковій НВ та відбору мінімального за глибиною (кількістю ярусів) логічного дерева. Тут слід відмітити, що дана задача є *NP* – повною (це було зафіксовано ще Л. Хайфілем та Р. Рівестом), а отже не має простих та ефективних методів розв’язку.

Постановка проблеми. Нехай на початку задачі задана деяка НВ стандартного вигляду:

$$\left(x_1, f_R(x_1)\right), \dots, \left(x_M, f_R(x_M)\right). \quad (1)$$

Зауважимо, що тут $x_i \in G$ $(G -$ деяка множина початкових сигналів), а відповідно $f_R(x_i) \in \{0,1,2, \dots, k-1\}, (i = 1,2, \dots, M)$, причому $M -$ загальна кількість навчальних пар (об‘єктів відомої класифікації) початкової НВ в задачі.

Відповідно $f_R(x_i) = l, (0 \le l \le k-1)$ означає, що об‘єкт $x_i \in H_l, H_l \subset G$. Причому тут $f_R -$ деяка скінчено значна функція (функція розпізнавання – ФР), яка задає початкове розбиття R множини G, яке складається з підмножин (образів, класів) $H_0, H_1, H_2, \dots H_{k-1}$ (заданої НВ).

Отже, можна зафіксувати, що початкова НВ – це сукупність (точніше послідовність) деяких наборів (об’єктів відомої класифікації), причому кожний набір – це сукупність значень деяких ознак та значення деякої функції (ФР) на цьому наборі. Іншими словами, можна підсувати, що сукупність значень ознак – це деяке зображення (дискретний об’єкт), а значення функції (ФР) відносить це зображення до відповідного образу [8-10].

Отже, зважаючи на вищезазначене, на цьому етапі дослідження буде стояти задача побудови конструкції L деякого ЛДК, структурні параметри p якого були б оптимальними щодо початкових даних НВ.

Відмітимо, що головна ідея методу поетапної селекції елементарних ознак (або набору алгоритмів) полягає в тому, щоби максимізувати величину якості ознаки (алгоритму) $W_M(f)$ [11; 12]. Останнє означає, що в алгоритмах логічного дерева має бути знайдена для навчальної вибірки типу (1) така узагальнена ознака f, для якої величина $W_M(f)$ є, по можливості, найбільшою [13].

Аналіз останніх досліджень і публікацій. Це дослідження продовжує цикл робіт, які присвячені проблемі деревоподібних схем розпізнавання (класифікації) дискретних об‘єктів [2-7]. У них порушуються питання побудови, використання та оптимізації логічних дерев. Так, з [2] відомо, що результуюче правило класифікації (схема), яке побудоване довільним методом або алгоритмом розгалуженого вибору ознак, має деревоподібну логічну структуру. Логічне дерево складається з вершин (ознак), які групуються за ярусами і які отримані на певному кроці (етапі) побудови дерева розпізнавання [14]. Ва-

жливою задачею, яка виникає з [15], є синтез дерев розпізнавання, які будуть представлятися фактично деревом (графом) алгоритмів. На відміну від наявних 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 [5]. Так, у роботі [13] порушуються принципові питання стосовно генерації дерев рішень для випадку малоінформативних ознак. Здатність ЛДК виконувати одномірне розгалуження для аналізу впливу (важливості, якості) окремих змінних дає можливість працювати зі змінними різних типів у вигляді предикатів (у випадку АДК – відповідними автономними алгоритмами класифікації та розпізнавання). Така концепція логічних дерев активно використовується в інтелектуальному аналізі даних, де кінцева мета полягає в синтезі моделі, яка прогнозує значення цільової змінної на основі набору початкових даних на вході системи [10].

Виділення недосліджених частин загальної проблеми. Пошук можливостей ефективної та економної роботи запропонованого методу побудови алгоритмічного дерева класифі-кації на основі масивів навчальних вибірок великого об'єму.

Мета роботи. Метою цієї роботи є вивчення особливостей генерації правил (схем, моделей) класифікації в задачах розпізнавання на основі алгоритмічних дерев (моделей АДК другого типу), розробка загальної схеми методу побудови АДК для навчальних вибірок великого об'єму. Результатом роботи є простий метод синтезу моделей АДК для задач класифікації дискретних об'єктів.

Виклад основного матеріалу. Припустимо, на початку задана деяка НВ загального типу (1) – у вигляді послідовності навчальних пар $\left(x_i, f_R(x_i)\right)$, потужністю – M, розмірністю ознакового простору – n та фіксований набір різнотипних алгоритмів класифікації $(\alpha_1, \alpha_2, \dots, \alpha_m)$. Зауважимо, 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 $(H_1, H_2, \dots, H_k)$, а відповідні алгоритми α_i можуть бути не пов'язані єдиною концепцією розпізнавання, а реалізовувати різноманітні методи та алгоритми класифікації (наприклад, це можуть бути звичайні геометричні алгоритми – принцип роботи яких полягає в апроксимації навчальної вибірки відповідними геометричними об'єктами, алгоритми обчислення оцінок, потенціальних функції тощо).

Треба зазначити, що результатом роботи кожного із зафіксованих (відібраних із бібліотеки алгоритмів деякої інформаційної системи) автономних алгоритмів класифікації та розпізнавання a_i, на відповідному кроці генерації АДК, є одна або декілька узагальнених ознак (УО) – f_j (певних правил класифікації), які й 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 прикладах можливі випадки, коли відповідний алгоритм класифікації a_i не може побудувати узагальнену ознаку f_j у зв'язку зі складним розташуванням класів H_k в ознаковому просторі задачі, або певними концептуальними та реалізаційними обмеженнями самого алгоритму класифікації. Тоді, за аналогією, можливий випадок, коли побудовані алгоритмом класифікації a_i узагальнені ознаки f_j неповністю апроксимують початкову НВ або така ситуація передбачена самою схемою алгоритму генерації АДК (наприклад, наявність початкового обмеження у схемі алгоритму дерева – про генерацію не більше ніж одна узагальнена ознака 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, 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}.

Отже, зважаючи на все вищевказане, можна зробити припущення, що структура АДК (типу II) буде мати загальну конструкцію вигляду, як на рис. 1, де кожний ярус такого логічного дерева визначає етап побудови АДК у вигляді апроксимації поточним алгоритмом класифікації a_i певної частини НВ та завдяки такому підходу дозволяє регулювати фінальну складність (точність) отриманої моделі дерева класифікації.

Також зазначимо, що на кожному кроці генерації моделі АДК (рис. 1) подається свій алгоритм α_i класифікації та своя відповідна НВ (або підмножина початкової НВ), причому початкова НВ в повному складі подається лише на першому кроці, далі з наступними етапами побудови дерева класифікації потужність масиву даних НВ буде падати за рахунок набору побудованих УО f_j, які будуть відрізати (описувати) певну частину даних початкової НВ. Також важливо зауважити, що в залежності від структури схеми побудови АДК та особливостей поточного алгоритму α_i на кожному кроці можливо генерувати більше однієї УО f_j.

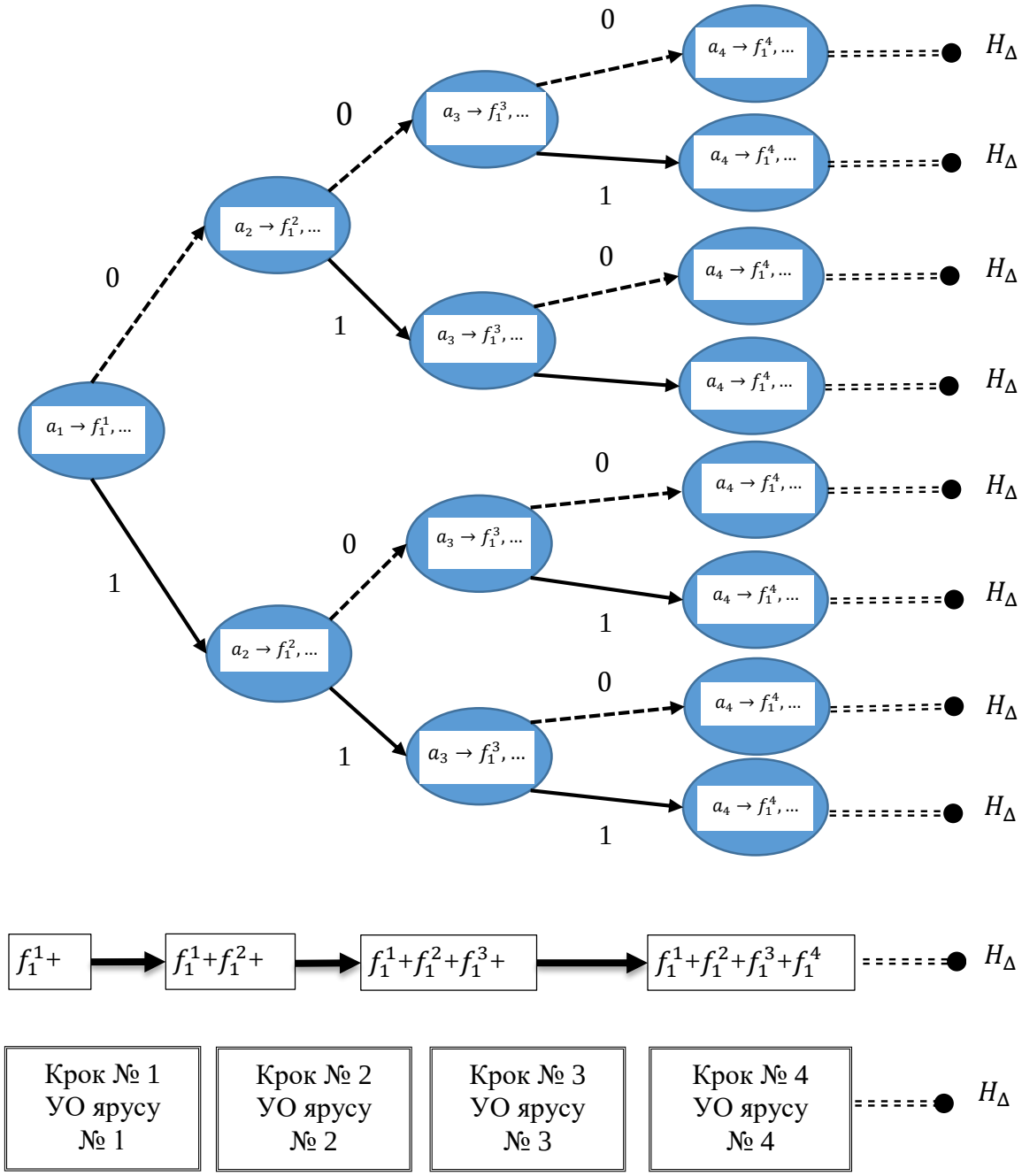

Рис. 1. Загальна схема структури АДК типу II

На наступному етапі дослідження для методу АДК введемо два базові критерії побудови моделі дерева класифікації – критерій зупинки процедури розгалуження K_{stop} (він фактично регулює складність та точність отриманої моделі АДК) та критерії відбору розгалуження $W(a)$ (вибору алгоритму класифікації на поточному кроці) для дерева класифікації що будується.

Отже, зважаючи на вищезазначене, доцільно ввести критерій зупинки K_{stop} процесу розгалуження типу $(boolean)$ процедури побудови АДК, який полягає в перевірці потужності $P_{pt}(HB)$ навчальної вибірки такого вигляду:

$$K_{stop} = \begin{cases} 0, if\ P_{pt}(HB) = 0 \\ 1, if\ P_{pt}(HB) > 0 \end{cases}. \quad (2)$$

Зауважимо, що процедура побудови дерева класифікації продовжується доти, поки $K_{stop} = 1$, а протилежна ситуації – коли $K_{stop} = 0$ сигналізує про завершення етапу синтезу моделі АДК та необхідності переходу до кроку тестової перевірки за даними ТВ та оцінки якості отриманої моделі дерева класифікації.

Зауважимо, що в методі АДК стає принциповим питання вибору критерію розгалуження (вибору поточного алгоритму класифікації a_i) у структурі моделі дерева класифікації, що будується. Зрозуміло, що за аналогією з методом апроксимації НВ набором ранжованих елементарних ознак як критерію розгалуження, можна запропонувати початкову оцінку ефективності набору алгоритмів $(\alpha_1, \alpha_2, \dots, \alpha_5)$ в наступному вигляді:

$$W(a_i) = \frac{\frac{1}{P_{pt}(HB)} * \sum_{j=1}^{k} (T_{\text{УЗ}} + S_{\text{УЗ}} + \frac{E_{\text{УЗ}}}{S_{\text{УЗ}}})}{k}. \quad (3)$$

У запропонованому функціоналі (3) введені величини мають таку інтерпретацію:

1) k – загальна кількість класів поточної задачі, які задані розбиттям R даних початкової НВ.

2) $T_{\text{УЗ}}$ – характеризує загальний час (апаратний час), який витрачається на побудову поточної УО f_j;

3) $E_{\text{УЗ}}$ – інформаційна ємність (структурна складність) побудованої УО f_j на поточному кроці генерації моделі АДК;

4) $S_{\text{УЗ}}$ – являє собою загальну кількість дискретних об'єктів x_i НВ, які узагальнює (описує) ця УО f_j;

5) $P_{pt}(HB)$ – потужність (об'єм) початкової НВ (або її фіксованої частини для поточного кроку алгоритму побудови АДК).

Зауважимо тут, що в формулі (3) сумування ведеться по всіх класах, які задані масивом даних початкової НВ (хоча можуть бути й обмеження по додаванню, які обумовлені структурою (параметрами) самого алгоритму побудови дерева класифікації).

Важливим моментом у схемі побудови моделі АДК (рис. 1) є те, що на кожному кроці алгоритму дерева фактично будується своя фіксована (одна або декілька – залежно від структури самого алгоритму АДК) УО f_j, причому їх загальна кількість збільшується з кожним кроком алгоритму дерева класифікації, а саме АДК із набором алгоритмів класифікації $(\alpha_1, \alpha_2, \dots, \alpha_m)$ породжує (реплікує) деревоподібну конструкцію – дерево узагальнених ознак (ДУО) з відповідними набором УО $(f_1, f_2, \dots, f_z)$.

Таким чином, зважаючи на все вищевказане, можна запропонувати одну з можливих алгоритмічних схем побудови АДК (типу II).

Схема побудови АДК (типу II).

Етап 1. Отже, на першому етапі побудови АДК відбирається на основі відповідного критерію ефективності (або в інтерактивному режимі, випадковим чином) набір автономних алгоритмів класифікації та розпізнавання $(\alpha_1, \alpha_2, \dots, \alpha_m)$, ранжованих відповідним чином. Тут також відбираються як самі алгоритми класифікації, так і їх

загальна кількість у наборі (величина m) залежно від умов та аспектів прикладної задачі. Саме цей етап має принципову важливість, тому що безпосередньо впливає на фінальну складність отриманої моделі АДК.

Етап 2. На другому етапі побудови АДК будується повне регулярне логічне дерево, де на кожному з ярусів цієї структури розташовується відповідний алгоритм класифікації ранжованої послідовності $(\alpha_1, \alpha_2, \dots, \alpha_m)$. У цьому логічному дереві кожна вершина має по два переходи на наступний ярус (два нащадки), які позначаються значенням з бінарної множини $\{0,1\}$. Оскільки маємо справу з регулярним логічним деревом, то на кожному з ярусів цієї структури розташовані мітки (змінні) одного типу (порядку), це стосується лише самих алгоритмів класифікації a_i, а не УО f_j, які вони генерують.

Так, на другому етапі генерації АДК (тип II) послідовно на вхід алгоритмам класифікації a_i подається масив даних НВ (відповідно до структури побудованого дерева класифікації (рис. 1) з метою отримати на виході набір відповідних УО f_j, причому їх загальна кількість у конструкції дерева та кількість для кожного алгоритму класифікації (кроку у схемі дерева, ярусу логічного дерева) залежить від початкових параметрів ініціалізації алгоритму побудови АДК (задаються в інтерактивному режимі або автоматично) та особливостей прикладної задачі, для якої будується модель АДК.

Після побудови набору всіх УО f_j для цієї прикладної задачі, вони розташовуються у відповідних вершинах отриманого дерева класифікації з метою завершення процедури його побудови. Принциповим моментом цього етапу є те, що набір побудованих УО має перекривати весь масив даних НВ для забезпечення стовідсоткового розпізнавання начальних даних. Причому тут можуть бути певні відхилення, якщо будується модель АДК з наперед заданою точністю та складністю (це обмеження умов задачі може бути реалізовано за рахунок зміни кількості та потужності УО f_j, які будуються на другому етапі). Зауважимо, що ця умова може бути також реалізована за рахунок обмеження кроків (кількості ярусів конструкції) в процедурі побудови моделі АДК, додатковими обмеженнями кількості алгоритмів класифікації, які використовуються в структурі дерева класифікації.

Етап 3. На третьому етапі схеми побудови АДК після побудови основної конструкції дерева класифікації можна переходити безпосередньо до режиму тестування отриманої моделі АДК. Причому для кожного тестового об'єкта, який подається на вхід дерева класифікації, обчислюються відповідні значення $\varphi(\alpha_j)$ (за допомогою набору побудованих раніше УО – для кожної вершини відповідного ярусу дерева) які забезпечують (визначають) відповідний маршрут у структурі побудованого АДК другого типу. Так УО кожної з вершин АДК – у випадку можливої апроксимації об'єкта невідомої класифікації забезпечують інкрементують відповідний лічильник класу належності та залишають його без змін у випадку відмови (неможливості) класифікації. На виході структури АДК об'єкт невідомої класифікації належить до того класу, лічильник належності якого буде максимальним, у випадку їх нульової рівності маємо справу з відмовою класифікації.

Зауваження. Зі схеми побудови АДК другого типу, яка була представлена вище, можна бачити, що кількість УО (параметрична складність, потужність) які генеруються тим самим відібраним алгоритмом класифікації α_j на деякому ярусі дерева класифікації для кожного зі шляхів структури АДК може бути різною, причому слідуючи в цьому напрямку прийдемо до того, що конструкція моделі АДК не обов'язково має належати до класу регулярних структур (логічних дерев), тобто в кожному ярусі конструкції АДК що будується, разом з різною кількістю та типом (загальними параметрами) УО допускається наявність різник алгоритмів класифікації та розпізнавання α_j.

Етап експериментальної перевірки. Треба зазначити, що запропоновані схеми побудови АДК дозволяє регулювати складність моделі дерева класифікації, що будується, будувати моделі з наперед заданою точністю, а сама структура дерева класифікації складається з різнотипних автономних алгоритмів класифікації як будівельних модулів (компонентів). Причому задача відбору моделі дерева класифікації серед набору побудованих АДК для конкретної задачі визначається набором параметрів, які мають визначальну важливість відносно поточної прикладної задачі (набору даних НВ).

Зрозуміло, що для порівняння та відбору конкретної моделі АДК з фіксованого набору необхідно виділити найбільш важливі їхні параметри (розмірність ознакового простору, кількість вершин, переходів, алгоритмів тощо) та визначити їх похибку відносно масиву вхідних даних.

Принципово на цьому етапі дослідження розглянути критерії якості отриманих інформаційних моделей, які залежать від похибки моделі, потужності початкового масиву даних НВ, об'єму ТВ (кількість навчальних пар та розмірність ознакового простору задачі), кількості параметрів моделі тощо.

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, які задані початковою умовою поточної прикладної задачі.

Зауважимо, що принциповим моментом залишається питання зменшення складності структури АДК (мається на увазі кількість ознак, алгоритмів у структурі АДК, загальна кількість вершин моделі АДК та загальна кількість переходів у структурі АДК), 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 ЛДК є загальний інтегральний показник якості в такій формі:

$$Q_{Main} = \frac{Fr_{All}}{V_{All} \cdot \sum_i p_i} \cdot e^{-\frac{Er_{All}}{M_{All}}}. \quad (4)$$

Відмітимо, у в формулі (4) набір параметрів p_i являє собою найбільш важливі характеристики побудованого дерева класифікації, що оцінюється:

1) Er_{All} – загальна кількість помилок моделі АДК на масивах даних початкових тестової та навчальної вибірки;

2) M_{All} – загальна потужність (об'єм) масивів даних навчальної та тестової вибірки;

3) Fr_{All} – характеризує кількість вершин отриманої моделі АДК з результуючими значеннями f_R (ФР, тобто листів дерева класифікації);

4) V_{All} – представляє загальну кількість всіх типів вершин в структурі моделі АДК;

5) O_{Uz} – загальна кількість узагальнених ознак, що використовуються в моделі дерева класифікації;

6) P_{All} – загальна кількість переходів між вершинами в структурі побудованої моделі дерева класифікації;

7) N_{Alg} – загальна кількість різних автономних алгоритмів класифікації, що використовуються в моделі дерева класифікації.

Цей 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.

Далі розглянемо наступний приклад побудови моделі АДК з відповідними початковими параметрами:

1) Фіксований набір різнотипних алгоритмів класифікації та розпізнавання $(\alpha_1, \alpha_2, \dots, \alpha_5)$, $(m = 5)$.

2) У масиві початкової НВ задана інформації про розбиття R на класи, що не перетинаються – $(H_1, H_2, \dots, H_4)$, $(k = 4)$.

3) Початкова НВ вигляду (1) має потужність в 2000 начальних пар (об'єктів відомої класифікації), $(M = 2000)$.

4) Кожний із дискретних об'єктів НВ x_i характеризуються набором ознак, атрибутів – $(x_i^1, x_i^2, \dots, x_i^{20})$, $(n = 20)$.

5) Для перевірки отриманої моделі АДК задана ТВ потужністю 500 елементів, $(T = 500)$.

Представлена в цьому прикладі початкова НВ містить дані компонентного хімічного аналізу вмісту дизельного (вуглеводного) пального (задача оцінки якості пального) у спрощеному варіанті (кількість класів НВ задачі зменшена до чотирьох, розмірність ознакового простору з 38 до 20, а кількість алгоритмів класифікації також обмежена на початковому етапі шляхом відбору лише геометричних класифікаторів) заради демонстрації самої концепції алгоритмічного дерева.

На першому етапі процедури побудови дерева класифікації, оцінимо ефективність кожного з відібраних алгоритмів класифікації, на основі яких і буде побудована загальна схема класифікації (модель АДК), стосовно даних початкової навчальної вибірки (за кількістю узагальнених ознак, що генеруються поточним алгоритмом та відмовам класифікації).

Таблиця 1

Оцінка ефективності фіксованих алгоритмів класифікації дискретних об'єктів відносно початкової начвальної вибірки

(Номер класу/ Тип алгоритму)	***Алгоритм a_1***	***Алгоритм a_2***	***Алгоритм a_3***	***Алгоритм a_4***	***Алгоритм a_5***
Клас H_1	0/32	0/12	0/11	0/9	18/10
Клас H_2	0/16	12/17	1/16	14/6	12/8
Клас H_3	0/8	0/10	0/17	0/12	0/10
Клас H_4	0/11	15/3	9/16	16/6	14/11

У комірках наведеної табл. 1 представлена ефективність кожного з відібраних для задач алгоритмів класифікації відносно класів початкової навчальної вибірки, причому перше число відповідає за кількість об'єктів, яким відмовлено в класифікації відповідним алгоритмом (помилкам, відмовам класифікації), а друге – за кількість узагальнених ознак (для цього типу алгоритмів – геометричних об'єктів), якими апроксимований відповідний клас початкової вибірки. Залежно від початкового вибору алгоритму як вершина дерева класифікації (моделі АДК), процедура побудови результуючої схеми класифікації може завершитися з різною кількістю кроків. Одна з можливих побудованих схеми АДК представлена на (рис. 2).

Так, з табл. 1 можна бачити, що ефективність усіх алгоритмів, за винятком алгоритму a_5 (геометричному алгоритму гіперплощин) відносно класу H_1 становить 100 %, тому для його апроксимації можна застосувати довільний алгоритм (зрозуміло, за винятком a_5. На всіх наступних етапах побудови схеми розпізнавання (ярусах структури АДК) доцільно знову зафіксувати алгоритм a_1 (геометричний алгоритм гіперсфер), який виявився найбільш ефективним та економним проти всіх інших класів даних початкової вибірки. Зокрема, його особливістю є велика універсальність щодо можливості побудови узагальненої ознаки, навіть у тих випадках, коли інші геометричні алгоритми цього зробити не можуть і дають великий відсоток відмов (помилок) класифікації для об'єктів начальної вибірки (випадок складного, заплутаного розташування класів в ознаковому просторі задачі). Важливим моментом також є той факт, що кожна з узагальнених ознак, згенерованих алгоритмом класифікації a_1, являє собою набір координат центру гіперсфери (в ознаковому просторі задачі) і її радіус та потребує мінімальний об'єм пам'яті інформаційної системи для свого зберігання та прості механізми роботи з набором таких УО.

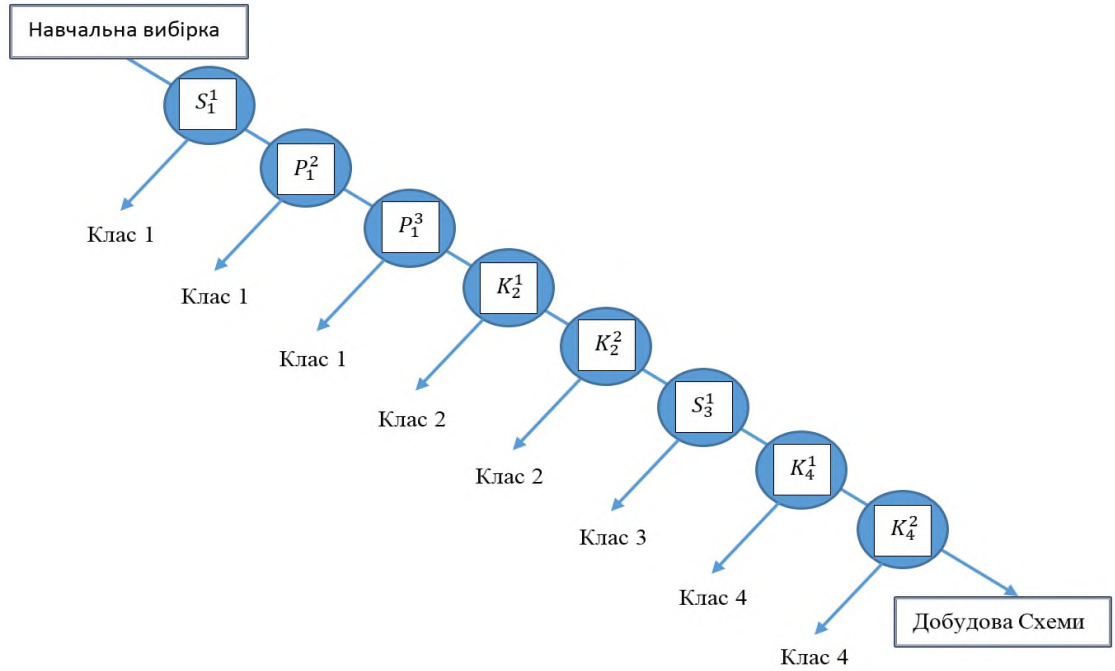

Рис. 2. Приклад сконструйованої моделі АДК

Треба зазначити, що модель дерева класифікації (рис. 2) побудована алгоритмами, ефективність яких оцінювалася щодо кількості узагальнених ознак (якими апроксимується початкова навчальна вибірка), причому для повної апроксимації масиву НВ даної задачі було достатньо лише трьох алгоритмів класифікації. Так, для апроксимації класу H_1 на (з першого по третій крок) було застосовано два алгоритми – спочатку алгоритм a_5 (геометричний алгоритм гіперплощин) побудував узагальнену ознаку S_1^1, яка лише частково його описувала. На другому етапі застосований алгоритм a_4 (геометричний алгоритм гіперпаралелепіпедів) – ознаки P_1^2 та P_1^3, які остаточно й завершили розпізнавання (апроксимацію) цього класу H_1.

На наступних етапах побудови моделі АДК знову же застосований алгоритм класифікації a_1 (алгоритм гіперсфер) (ознаки $K_1^2, K_2^2, K_4^1, K_4^2$) та алгоритм a_5 (узагальнена ознака S_3^1).

Тут треба звернути увагу, що для побудови цієї схеми класифікації (рис. 2) застосовані три різні алгоритми розпізнавання (з п'ятьох початково відібраних), які безпосередньо не впливають на роботу один одного, тобто на їхньому місці могли бути зовсім різні за принципом та ідеологією алгоритми класифікації, з яких можна сконструювати схему розпізнавання (модель АДК) довільної складності та ефективності. Важливим є лише ефективність кожного з них на фіксованій виборці та інформаційна ємність узагальнених ознак, генерованих ними.

Підкреслимо, що метод алгоритмічного дерева оперує лише вже готовими (побудованими) узагальненими ознаками, і його може зовсім не цікавити, яким алгоритмом чи способом (правилом, методом) вони отримані. Причому у схемі АДК (рис. 2) показана покрокова схема генерації не одиничних, а цілих наборів УО K_i^j, S_i^j, P_i^j, де $j-$ номер УО для відповідного класу апроксимації (номер етапу генерації АДК, відносно класу), а $i-$ номер кроку процедури побудови дерева (класу який апроксимується).

Зрозуміло, що ця схема розпізнавання (модель АДК), сконструйована на основі методу дерева класифікації (ЛДК/АДК), та може бути представлена як певна алгоритмічна схема (оператор), яка побудована за деяким алгоритмом мінімізації або максимізації відповідного

функціонала, на основі якого оцінюється важливість ознаки, групи ознак або ефективність автономного алгоритму класифікації, однозначно пов'язаного з помилками класифікації (перетинається з методами логічних дерев класифікації).

Зауважимо, що метод алгоритмічного дерева на основі вхідних даних (даних навчаючої вибірки) та асортименту (набору) алгоритмів формування узагальнених ознак, які зберігаються в його бібліотеці, конструює (генерує) оптимальну за витратами пам'яті (складності) та ефективності розпізнавання (систему) певну схему (дерево класифікації). Під схемою в цьому випадку будемо розуміти набір числових параметрів для набору УО K_i^j, S_i^j, P_i^j, які найкращим чином апроксимують масив початкових даних. Зокрема, у нашому випадку аргументи сконструйованої схеми розпізнавання – ознаки класів (гіперсфери, гіперпаралелепіпеди та інші) або міжкласові ознаки (гіперплощини). Параметри вказаних ознак та загальна структура АДК (схеми класифікації) зберігаються в пам'яті комп'ютера (інформаційної системи).

Кожна зі сконструйованих схем за 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). Зауважимо також, що отримана схема буде являти собою певною мірою новий алгоритм розпізнавання (зрозуміло, що синтезований із відомих алгоритмів та методів). Крім того, для роботи даних отриманої моделі АДК немає необхідності зберігати в пам'яті комп'ютера об'єкти вибірки, за якими була сконструйована, тобто великі інформаційні масиви, останнє, у свою чергу, веде до того, що процес побудови системи розпізнавання на основі методів дерева (ЛДК/АДК) значною мірою схожий із процесом стиснення (маються на увазі методи стиснення інформації із втратами) або кодуванням інформації (опис складних структур даних).

На наступному етапі дослідження проведемо оцінювання та порівняння побудованих моделей дерев класифікації (ЛДК та АДК) для представленої вище прикладної задачі класифікації. Так, фрагмент основних результатів, приведених вище експериментів (порівняльних тестів методів побудови моделей ЛДК на масиві даних цієї прикладної задачі), представлений у табл. 2, причому синтезовані моделі різнотипних дерев класифікації забезпечили необхідний рівень точності заданий умовою задачі, швидкодію та витрати робочої пам'яті інформаційної системи, але показували різну структурну складність побудовах дерев класифікації (моделей) та набору узагальнених ознак (у випадку моделі алгоритмічного дерева класифікації).

Таблиця 2

Порівняльна таблиця моделей / методів дерев класифікації

№	**Метод синтезу структури логічного дерева**	**Інтегральний показник якості моделі Q_{Main}**	**Загальна кількість помилок моделі на НВ та ТВ Er_{All}**
1	Метод повного ЛДК на основі селекції елементарних ознак	0,004789	2
2	Модель ЛДК з одноразовою оцінкою важливості ознак	0,002263	3
3	Обмежений метод побудови ЛДК	0,003181	2
4	Метод алгоритмічного дерева (типу I)	0,005234	0
5	Метод алгоритмічного дерева (типу II)	0,002941	0

Зауважимо, що запропонована в дослідженні оцінка якості моделі дерева класифікації (АДК) відображає базові параметри (характеристики) дерев класифікації та може бути застосована як критерій оптимальності в процедурі оцінювання довільної деревоподібної схеми розпізнавання, наприклад у випадку методів побудови та відбору випадкових ЛДК з роботи [16].

Так, запропонований у цьому дослідженні метод алгоритмічного дерева класифікації (методи АДК першого та другого типу) порівнювалися з методом повного ЛДК та обмеженого методу селекції елементарних ознак та показав загалом прийнятний результат.

Висновки відповідно до статті. Практична цінність отриманих результатів полягає в тому, що запропонований метод побудови моделей АДК (першого та другого типу) дає можливість будувати економні та ефективні моделі класифікації заданої точності (цей метод був реалізований у бібліотеці алгоритмів системі «ОРІОН ІІІ» для розв'язку різноманітних прикладних задач класифікації), які характеризуються великим ступенем універсальності відносно широкого кола прикладних задач. Зауважимо також, що практичні застосування підтвердили працездатність побудованих моделей дерев класифікації та розробленого програмного забезпечення. Як перспектива, подальші дослідження можуть бути спрямовані в бік подальшого розвитку методів АДК (введення нових типів дерев класифікації), оптимізації програмних реалізацій запропонованого методу АДК, а також його практичної апробації на множині реальних задач класифікації та розпізнавання.

Отже, зважаючи на все вищезазначене, можна зафіксувати такі пункти.

1) Використання концепції АДК при побудові моделей дерев класифікації дозволяє досягти хороших результатів щодо розширення кола прикладних задач застосування (вимога універсальності), можливості побудови моделей класифікації точність яких можна регулювати в процесі побудови (або будувати системи з наперед заданою точністю), можливості раціонального використання вже накопиченого потенціалу методів та алгоритмів розпізнавання.

2) Концепція АДК дозволяє будувати моделі дерев класифікації різних типів, причому всі вони базуються на простій ідеї апроксимації начальної вибірки набором автономних алгоритмів класифікації та представлення отриманої моделі у вигляді деякої деревоподібної схеми.

3) Важливим моментом є те, що отримана модель АДК з різних алгоритмів та методів класифікації своєю чергою буде являти собою новий алгоритм розпізнавання, тобто концепція АДК – це метод синтезу нових алгоритмів класифікації заданої точності проти НВ на основі набору вже відомих.

Список використаних джерел

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

2. Василенко Ю. А., Василенко Е. Ю., Повхан І. Ф., Ващук Ф. Г. Концептуальна основа систем розпізнавання образів на основі метода розгалуженого вибору ознак. *European Journal of Enterprise Technologies.* 2004. № 7[1]. С. 13–15.

3. Василенко Ю. А., Повхан І. Ф., Ващук Ф. Г. Проблема оцінки складності логічних дерев розпізнавання та загальний метод їх оптимізації. *European Journal of Enterprise Technologies.* 2011. № 6/4(54). С. 24–28.

4. Василенко Ю. А., Повхан І. Ф., Ващук Ф. Г. Загальна оцінка мінімізації деревоподібних логічних структур. *European Journal of Enterprise Technologies.* 2012. № 1/4(55). С. 29-33.

5. Povhan I. General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects. *Електроніка та інформаційні технології.* Львів, 2019. Вип. 11. С. 112–117.

6. Василенко Ю. А., Василенко Е. Ю., Повхан І. Ф., Ковач М. Й., Нікарович О. Д. Мінімізація логічних деревоподібних структур в задачах розпізнавання образів. *European Journal of Enterprise Technologies.* 2004. № 3[9]. С. 12–16.

7. Лавер В. О., Повхан І. Ф. Алгоритми побудови логічних дерев класифікації в задачах розпізнавання образів. *Вчені записки Таврійського національного університету. Серія: технічні науки.* 2019. Т. 30(69), № 4. С.100-106.

8. Vtogoff P. E. Incremental Induction of Decision Trees. *Machine Learning*. 2009. № 4. P. 161−186.

9. Повхан І. Ф. Проблема функціональної оцінки навчальної вибірки в задачах розпізнавання дискретних об‘єктів. *Вчені записки Таврійського національного університету. Серія: технічні науки.* 2018. Т. 29(68). № 6. С. 217-222.

10. Whitley D. An overview of evolutionary algorithms: practical issues and common pitfalls. *Information and Software Technology*. 2001. Vol. 43. №14. P. 817–831.

11. Povhan I. Designing of recognition system of discrete objects. *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)*. Lviv, 2016, P. 226–231.

12. Kotsiantis S. B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica*. 2007. № 31. P. 249–268.

13. Суботин С. А. Построение деревьев решений для случая малоинформативных признаков. *Radio Electronics, Computer Science, Control*. 2019. № 1. P. 121–130.

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

15. Повхан І. Ф. Особливості синтезу узагальнених ознак при побудові систем розпізнавання за методом логічного дерева. *Інформаційні технології та комп'ютерне моделювання ІТКМ-2019 :* матеріали Міжнародної науково-практичної конференції. Івано-Франківськ, 2019. С. 169–174.

16. Повхан І. Ф. Особливості випадкових логічних дерев класифікації в задачах розпізнавання образів. *Вчені записки Таврійського національного університету. Серія : технічні науки.* 2019. Т. 30 (69), № 5. С. 152–161.

References

1. Srikant, R., Agrawal, R. (1997). Mining generalized association rules. *Future Generation Computer Systems*, 13 (2), 61–180.

2. Vasylenko, Yu. A., Vasylenko, E. Yu., Povkhan, I. F., Vashchuk, F. G. (2004). Kontseptualna osnova system rozpiznavannia obraziv na osnovi metoda rozghaluzhenoho vyboru oznak [Conceptual basis of pattern recognition systems based on the method of branched feature selection]. *European Journal of Enterprise Technologies*, 7[1], 13–15.

3. Vasylenko, Yu. A., Vashchuk, F. G, Povkhan, I. F. (2011). Problema otsinky skladnosti lohichnykh derev rozpiznavannia ta zahalnyi metod yikh optymizatsii [The problem of estimating the complexity of the logic trees, recognition, and a general method of optimization]. *European Journal of Enterprise Technologies*, 6/4(54), 24–28.

4. Vasylenko, Yu. A., Povkhan, I. F., Vashchuk, F. G. (2012). Zahalna otsinka minimizatsii derevopodibnykh lohichnykh struktur [General estimation of tree logical structures minimization]. *European Journal of Enterprise Technologies,* 1/4 (55), 29–33.

5. Povkhan, I. (2019). General scheme for constructing the most complex logical tree of classification in pattern recognition of discrete objects. *Electronics and information technology*, 11, 112–117.

6. Vasylenko, Yu. A., Vasylenko, E. Yu., Povkhan, I. F., Kovach, M. Y., Nikarovych, O. D. (2004). Minimizatsiia lohichnykh derevopodibnykh struktur v zadachakh rozpiznavannia obraziv [Minimization of logic tree structures in pattern recognition problems]. *European Journal of Enterprise Technologies,* 3[9], 12–16.

7. Laver, V. O., Povkhan, I. F. (2019). Alhorytmy pobudovy lohichnykh derev klasyfikatsii v zadachakh rozpiznavannia obraziv [Algorithms for constructing logical classification trees in pattern recognition problems]. *Vcheni zapysky Tavriiskoho natsionalnoho universytetu. Seriia: tekhnichni nauky – Scientific notes of Tauride national University. Series: technical Sciences*, 30(69) (4), 100–106.

8. Vtogoff, P. E. (2009). Incremental Induction of Decision Trees. *Machine Learning*, 4, 61−186.

9. Povkhan, I. F. (2018). Problema funktsionalnoi otsinky navchalnoi vybirky v zadachakh rozpiznavannia dyskretnykh obiektiv [The problem of functional evaluation of the training sample in the problems of recognition of discrete objects]. *Vcheni zapysky Tavriiskoho natsionalnoho universytetu. Seriia: tekhnichni nauky – Scientific notes of Taurida national University. Series: technical Sciences*, 29(68) (6), 217–222.

10. Whitley, D. (2001). An overview of evolutionary algorithms: practical issues and common pitfalls. *Information and Software Technology*, 43 (14), 817–831.

11. Povhan, I. (2016). Designing of recognition system of discrete objects, *IEEE First International Conference on Data Stream Mining & Processing (DSMP)* (pp. 226–231). Lviv.

12. Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 31, 249–268.

13. Subbotin, S. A. (2019). Postroenie derevev reshenii dlia sluchaia maloinformativnykh priznakov [Construction of decision trees for the case of low-information features]. *Radio Electronics, Computer Science, Control*, 1, 121–130.

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

15. Povkhan, I. F. (2019). Osoblyvosti syntezu uzahalnenykh oznak pry pobudovi system rozpiznavannia za metodom lohichnoho dereva [Features of synthesis of generalized features in the construction of recognition systems using the logical tree method]. *Informatsiini tekhnolohii ta kompiuterne modeliuvannia ITKM-2019 : materialy Mizhnarodnoi naukovo-praktychnoi konferentsii – Materials of the international scientific and practical conference* "*Information technologies and computer modeling ITKM-2019*" (pp. 169–174). Ivano-Frankivsk.

16. Povkhan, I. F. (2019). Osoblyvosti vypadkovykh lohichnykh derev klasyfikatsii v zadachakh rozpiznavannia obraziv [Features random logic of the classification trees in the pattern recognition problems]. *Vcheni zapysky Tavriiskoho natsionalnoho universytetu. Seriia : tekhnichni nauky – Scientific notes of the Tauride national University. Series: technical Sciences*, 30(69), 5, 152–161.

UDC 004.8:004.89:519.7

Igor Povkhan

A METHOD FOR CONSTRUCTING AN ALGORITHMIC TREE OF THE SECOND TYPE BASED ON THE APPROXIMATION OF THE TRAINING SAMPLE BY A SET OF CLASSIFICATION ALGORITHMS

Urgency of the research. *Modern information technologies based on mathematical models of image recognition in the form of logical classification trees are widely used in socio-economic, environmental and other systems of primary analysis and processing of large amounts of information.it is clear that this is due to the fact that this approach allows you to eliminate a set of existing shortcomings of well-known classical methods and achieve a fundamentally new result. The work is devoted to the topic of classification trees models. It offers an effective method for constructing algorithmic models of classification trees, which consist of independent and autonomous classification algorithms and will represent to a certain extent a new recognition algorithm (it is clear that it is synthesized from known algorithms and methods).*

Target setting. *Currently, there are various approaches and methods for building classification trees models (we know about more than 3600 recognition algorithms based on various concepts that have certain limitations when using them – accuracy, speed, memory, versatility, reliability, etc.), but all of them, as a rule, are reduced to building a single classification tree based on the data of the original training sample. It is clear that it is advisable not to develop a new algorithm, but to offer a concept of rational use of the already accumulated potential of algorithms and classification methods in the form of classification trees models, and that is why this work intends to at least partially overcome these limitations and is devoted to developing a method for constructing algorithmic models of classification trees.*

Actual scientific researches and issues analysis. *The possibility of efficient and economical operation of the proposed method for constructing an algorithmic classification tree based on arrays of large-volume training samples.*

The research objective. *Development of a simple and high-quality method for constructing algorithmic models of classification trees for large arrays of initial samples by synthesizing minimal forms of classification and recognition trees that provide an effective approximation of educational information with a set of Autonomous and independent classification algorithms.*

The statement of basic materials. *Identification of a simple and effective mechanism that could be used to build an algorithmic classification tree (a model of an algorithmic classification tree) based on fixed initial information in the form of an initial training sample. This algorithmic classification tree will accurately recognize the entire training sample for which the classification tree is built. It will have a minimal structure (structural complexity) and consist of Autonomous classification algorithms as construction vertices (tree attributes).*

Conclusions. *The proposed method for constructing algorithmic models of classification trees of the second type allows to work with training samples of a large volume and provides high speed and economy of hardware resources in the process of generating the final classification scheme, to build classification trees with a predetermined accuracy.*

Keywords: *recognition tasks, classification trees, algorithmic tree, recognition scheme, discrete object, generalized model.*

Fig.: 2. Table: 2. References: 16.

Повхан Ігор Федорович – кандидат технічних наук, доцент, доцент кафедри програмного забезпечення систем, ДВНЗ «Ужгородський національний університет» (вул. Заньковецької 89Б, м. Ужгород, 88000, Україна).

Povkhan Igor – PhD in Technical Sciences, Associate Professor, Associate Professor of Department of software, Uzhgorod national University (89B Zankovetskoy Str., 88000 Uzhgorod, Ukraine).

E-mail: igor.povkhan@uzhnu.edu.ua

ORCID: http://orcid.org/0000-0002-1681-3466

Повхан І. Метод побудови алгоритмічного дерева другого типу на основі апроксимації навчальної вибірки набором алгоритмів класифікації. *Технічні науки та технології.* 2020. № 2 (20). С. 126-138.