РОЗДІЛ ІІ. ІНФОРМАЦІЙНО-КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

УДК 004.8:004.89:519.7
DOI: 10.25140/2411-5363-2020-3(21)-142-153

Ігор Повхан

ПИТАННЯ СКЛАДНОСТІ ПРОЦЕДУРИ ПОБУДОВИ СХЕМИ АЛГОРИТМІЧНОГО ДЕРЕВА КЛАСИФІКАЦІЇ

Актуальність теми дослідження. *На сучасному етапі розвитку інформаційних систем та технологій, які базуються на математичних моделях теорії штучного інтелекту (методах та схемах алгоритмічних дерев класифікації), виникає принципова проблема вузької спеціалізації наявних підходів та методів у соціально-економічних, екологічних та інших системах первинного аналізу та обробки великих масивів інформації. 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. На сьогодні немає 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. Представлена робота присвячена проблемі моделей логічних та алгоритмічних дерев класифікації (схем ЛДК/АДК), пропонує оцінку складності структур алгоритмічних дерев (моделей дерев класифікації), які складаються з незалежних та автономних алгоритмів класифікації і будуть являти собою певною мірою новий алгоритм розпізнавання (зрозуміло, що синтезований із відомих схем, алгоритмів та методів).*

Постановка проблеми. *Нині актуальні різні підходи до побудови систем розпізнавання у вигляді дерев класифікації (ЛДК/АДК), причому інтерес до методів розпізнавання, які використовують дерева класифікації, викликаний багатьма корисними властивостями, якими вони володіють. З одного боку, складність класу функцій розпізнавання у вигляді моделей дерев класифікації, при визначених умовах, не перевищують складності класу лінійних функцій розпізнавання (простішого з відомих). З іншого – функції розпізнавання у вигляді дерев класифікації дозволяють виділити в процесі класифікації як причинно-наслідкові зв'язки (та однозначно врахувати їх у подальшому), так і фактори випадковості або невизначеності, тобто врахувати одночасно і функціональні, і стохастичні відношення між властивостями та поведінкою всієї системи. При цьому відомо, що процес класифікації нових, таких, що досі не зустрічалися, об'єктів світу багатьох тварин і людей (за винятком об'єктів, інформація про які передається генетичним шляхом (наслідковим), а також в деяких інших випадках), відбувається за так званим логічним деревом рішень (у зв'язку з нейромережевою концепцією). Зрозуміло, що доцільно не розробляти новий алгоритм, а запропонувати деяку концепцію раціонального використання вже накопиченого потенціалу алгоритмів та методів класифікації у вигляді моделей алгоритмічних дерев класифікації (структур АДК). Саме тому ця робота має намір хоча б частково подолати ці обмеження та присвячена оцінці складності процедури побудови моделей алгоритмічних (логічних) дерев класифікації в галузі задач розпізнавання.*

Аналіз останніх досліджень і публікацій. *У дослідженні розглянуті останні наукові публікації у відкритому доступі, які присвячені загальній проблемі підходів, методів, алгоритмів та схем розпізнавання (моделей ЛДК/АДК) дискретних об'єктів (дискретних зображень) у задачах розпізнавання образів (теорії штучного інтелекту).*

Виділення недосліджених частин загальної проблеми. *Можливість простого та економного методу побудови моделі алгоритмічного дерева класифікації (або структур АДК/ЛДК) та оцінка складності такої процедури (моделі структури АДК/ЛДК) на основі початкових масивів дискретної інформації великого об'єму.*

Постановка завдання. *Дослідження актуального питання складності загальної процедури побудови алгоритмічного дерева класифікації (моделі АДК) на основі концепції поетапної селекції наборів незалежних алгоритмів класифікації (можливих їх різнотипних множин та сполучень), яке для заданої початкової навчальної вибірки (масиву дискретної інформації) будує деревоподібну структуру (модель класифікації АДК), з набору алгоритмів оцінених на кожному кроці схеми побудови моделі за даною початковою вибіркою.*

Виклад основного матеріалу. *Пропонується оцінка складності процедури побудови алгоритмічного дерева класифікації для довільного випадку (для умов слабкого та сильного розділення класів навчальної вибірки). Розв'язок цього питання має принциповий характер, щодо питань оцінки структурної складності моделей класифікації (у вигляді деревоподібних конструкцій), структур АДК дискретних об'єктів для широкого класу прикладних задач класифікації та розпізнавання в плані розробки перспективних схем та методів їх фінальної оптимізації (мінімізації) конструкції. Це дослідження має актуальність не лише для конструкцій алгоритмічних дерев класифікації, але й дозволяє розширити саму схему оцінки складності і на загальний випадок структур логічних дерев класифікації.*

Висновки відповідно до статті*. Досліджені питання структурної складності конструкцій ЛДК/АДК, запропонована верхня оцінка складності для процедури побудови алгоритмічного дерева класифікації в умовах слабкого та сильного розділення класів початкової навчальної вибірки.*

Ключові слова*: задачі розпізнавання; алгоритмічне дерево; схема розпізнавання; складність дерева класифікації; дискретна ознака; узагальнена ознака; алгоритм класифікації.*

Рис.: 3. Бібл.: 23.

Актуальність теми дослідження. Сучасні інформаційні системи та технології, які концептуально базуються на моделях розпізнавання (класифікації) образів у вигляді структур ЛДК/АДК (моделей логічних дерев класифікації, алгоритмічних дерев класифікації), широко використовуються в соціально-економічних, екологічних та інших системах аналізу та обробки інформації. Нині фактично немає універсального підходу або концепції до їх вирішення, але запропоновано набір досить загальних теорій та підходів, що дозволяють вирішувати багато типів (класів) прикладних задач. Їх безпосереднє застосування відрізняється досить великою чутливістю до специфіки задачі або предметної галузі застосування [1-7]. Зазначимо, що багато теоретичних результатів отримано для спеціальних випадків та підзадач. Причому вузьким місцем вдалих реальних систем розпізнавання залишається необхідність виконання величезного об'єму обчислень та орієнтація на потужний апаратний інструментарій. Проте велика кількість прикладних задач, у різних галузях природознавства, де вирішуються задачі класифікації з використанням програмних та апаратних систем, визначає інтенсивність та актуальність такого напряму досліджень [8]. Зауважимо, що галузь застосування концепції дерев рішень нині надзвичайно об'ємна, а множина задач та проблем, які вирішуються цим апаратом, може бути зведена до задачі опису структур даних, задачі розпізнавання та класифікації, задачі регресії [9-11]. Саме загальним питанням структурної складності конструкцій АДК у розрізі процедури їх генерації і буде присвячена дана робота.

Постановка проблеми. Нехай на деякій множині G дискретних об'єктів (сигналів) x задане розбиття R на скінчене число k підмножин (класів, образів) $H_i(i = 1, \dots, k)$, $G = \bigcup_{i=1}^{k} H_i$. Відповідні множини $H_1, \dots, H_k$ будемо називати образами, а елементи множини $G -$ зображеннями або представниками образів (класів) $H_1, \dots, H_k$. Об'єкти (зображення) x задаються наборами значень деяких ознак $x_j(j = 1, \dots, n)$. Якщо $x \in H_i$ то будемо рахувати, що даний об'єкт належить образу H_i. Загалом образи $H_1, \dots, H_k$ можуть бути задані імовірнісними розподілами $p(H_1/x), \dots, p(H_k/x)$, де $p(H_i/x)$ – імовірність (або в неперервному випадку щільність імовірності) належності $x(x \in H_i)$ образу H_i. Нехай початковою умовою задачі задана деяка фіксована НВ у вигляді послідовності навчальних пар наступного вигляду:

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

Причому, крім початкової НВ, задана також ТВ (тестова вибірка – 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 розпізнавання (ФР) на цьому наборі. Тоді сукупність значень ознак – це деяке зображення (дискретний об'єкт), а значення функції (ФР) відносить це зображення до відповідного образу [12]. Таким чином, зазвичай стоїть загальна задача побудови моделі АДК з набором деяких параметрів p, структура L якої була би оптимальною $F(L(p, x_i), f_R(x_i)) \to opt$ щодо початкових даних НВ, причому нас у цьому дослідженні буде цікавити складність такої структури на етапі побудови моделі АДК.

Аналіз останніх досліджень і публікацій. Дослідження продовжує цикл робіт, які присвячені проблемі деревоподібних схем розпізнавання (моделей дерев класифікації) дискретних об'єктів [12-14; 17-21]. У них порушуються питання побудови, використання та оптимізації структур дерев класифікації. Так, з [4] відомо, що результуюче правило класифікації (схема), яке побудоване довільним методом або алгоритмом розгалуженого вибору ознак, має деревоподібну логічну структуру. Причому логічне дерево складається з вершин (ознак, атрибутів), які групуються по ярусах і які отримані на певному кроці (етапі) побудови дерева розпізнавання [5]. Важливою задачею, яка виникає з [8], є задача синтезу

дерев розпізнавання, які будуть представлятися фактично деревом (графом) алгоритмів (методи АДК). 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 [12]. Так в роботі [7] піднімаються принципові питання стосовно генерації дерев рішень для випадку малоінформативних ознак, питання оцінки якості побудованих моделей. 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. Така концепція дерев класифікації активно використовується в інтелектуальному аналізі даних, 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 [11].

Так, насьогодні в прикладній площині існує значна кількість алгоритмів, які реалізують концепцію дерев рішень (дерев класифікації), але найбільшого вживання та розповсюдження отримали два їх представника (C4.5/C5.0 та CART). Причому згаданий вище алгоритм логічного дерева C5.0, як критерій відбору вузла (вершини), використовує так званий теоретико-інформаційний критерій. А алгоритм CART базується на розрахунку індексу Gini, який враховує відносні відстані (у межах метрики) між розподілами класів [15].

Оскільки головну ідею методів та алгоритмів розгалуженого вибору ознак (вершин алгоритмів) 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). Саме ці принципові задачі розглядаються в [19], де порушуються питання якісної оцінки окремих дискретних ознак, їх наборів та фіксованих сполучень, що дозволяє запровадити ефективний механізм реалізації розгалуження.

Мета роботи. Метою цієї роботи є верхня числова оцінка складності структури побудованого алгоритмічного дерева класифікації, що забезпечує швидку та якісну схему класифікації дискретних об'єктів у задачах штучного інтелекту.

Виклад основного матеріалу. На цьому етапі дослідження розглянемо принципове питання методів дерев класифікації (моделей класифікації) – питання загальної складності процедури побудови дерева класифікації (методів дерев класифікації). Отже, припустимо, що маємо справу з початковою НВ вигляду (1), яка представлена сукупністю навчальних пар відомої класифікації. Зауважимо, що тут НВ є детермінованою, тобто для неї буде виконуватися така умова:

$$If\ \left(x_k, f_R(x_k)\right)\ and\ \left(x_l, f_R(x_l)\right),\ x_k \neq x_l\ then\ f_R(x_k) \neq f_R(x_l). \qquad (2)$$

Випадок ЛДК. Нехай на кожному кроці в процесі побудови логічного дерева (деякої моделі ЛДК) буде вибиратися лише одна відібрана елементарна ознака з набору фіксованих ознак $(\varphi_1, \varphi_2, \dots, \varphi_n)$. Тоді на $n-$ товому кроці процедури побудови дерева класифікації схема ЛДК буде являти собою деякий предикат p_n (узагальнену ознаку, яка побудована з набору елементарних ознак) [20], який є найбільш ефективною апроксимацією початкової НВ загального вигляду (1) (безперечно, це справедливо і для випадку структури АДК).

Зауважимо, що тут p_n буде являти собою деяку деревоподібну схему (дерево класифікації), яке складається з n вершин, тобто в структуру предикату p_n будуть входити всього n елементарних ознак (атрибутів дискретного об'єкта НВ) із початкового набору.

Зауважимо, що послідовність предикатів $p_1, p_2, \dots, p_j$ (узагальнених ознак) збігається до початкової НВ вигляду (1), якщо починаючи з деякого Q, буде виконуватись умова:

$$f_{Q+m} = f_R(x_i), (i = 1,2, \dots, m), (m \geq 0). \quad (3)$$

Деяка елементарна ознака, яка буде вибиратися (фіксуватися) на $n-$ товому кроці в схемі побудови моделі ЛДК, позначимо через φ_n. Зрозуміло, що ознаці φ_n відповідає деякий фіксований шлях $r_1, r_2, \dots$, який закінчується цим атрибутом (вершиною дерева класифікації – моделі ЛДК). Наприклад, на рис. 1 зображено ЛДК, в якому вершині φ_2 (ознаці) відповідає шлях $\{0\}$, а вершині φ_5 – шлях $\{0,1\}$.

Шлях, який відповідає елементарній ознаці φ_n вказаним чином, позначимо через T_n, а через D_n позначимо множину тих пар $\left(x_i, f_R(x_i)\right)$ початкової НВ загального вигляду (1), для яких об'єкти w_i належать шляху T_n. Наприклад, для структури ЛДК (рис. 1), нехай $\varphi_n = \varphi_4$, тоді шлях T_n буде мати вигляд $\{1,0\}$.

У такому випадку деякий об'єкт w_i належить шляху $\{1,0\}$, якщо виконуються умови $\varphi_1(w_i) = 1$ та $\varphi_3(w_i) = 0$.

Далі будемо вважати, що елементарна ознака φ_n слабо розділяє множину D_n, якщо в D_n існують такі пари $\left(x_i, f_R(x_i)\right)$ та $\left(x_j, f_R(x_j)\right)$, що $\varphi_n(x_i) = 0$ та $\varphi_n(x_j) = 1$ (тобто $\varphi_n(x_i) \neq \varphi_n(x_j)$).

На наступному етапі дослідження введемо поняття кінцевої потужності схеми методу дерева класифікації.

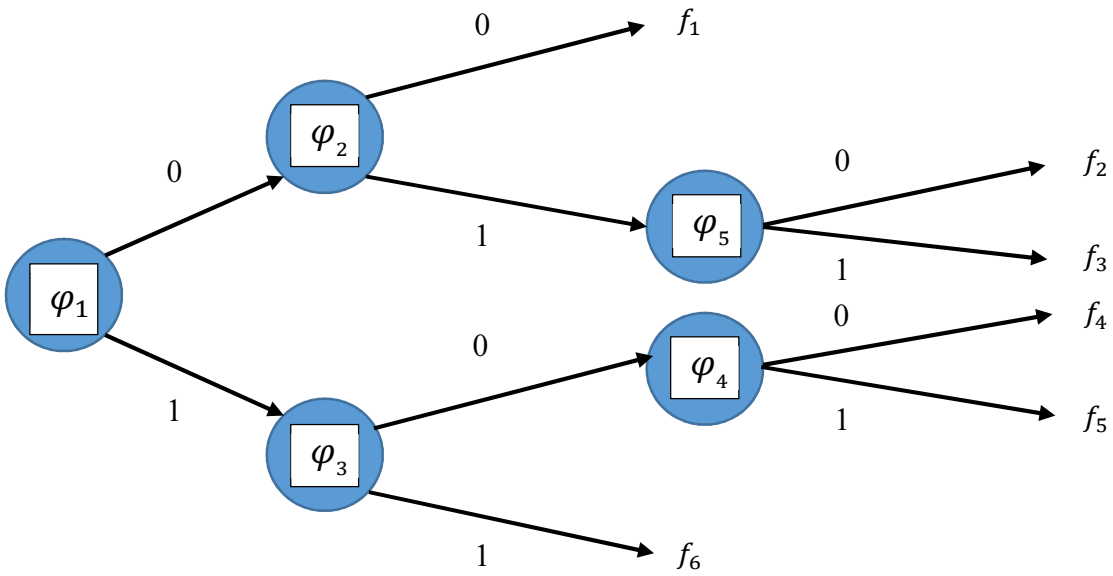

Рис. 1. Приклад структури ЛДК з елементарними ознаками в ролі вершин

Визначення 1. Під кінцевою потужністю схеми методу дерева класифікації (моделей ЛДК/АДК) будемо рахувати кількість усіх кінцевих вершин (визначених листів) цієї схеми. Наприклад, для ЛДК з (рис. 1) кінцева потужність буде дорівнювати 6.

Очевидно, що кінцева потужність схеми методу дерева класифікації також дорівнює кількості всіх кінцевих шляхів в даній схемі. Зрозуміло, що індукцією за n легко довести, що кінцевою потужністю кожної з вищесказаних схем p_n (предикатів), дорівнює $n + 1$. Дійсно, те, що кінцева потужність p_1, до складу якого входить тільки одна ознака або алгоритм (випадків ЛДК/АДК) дорівнює 2, є очевидним.

Нехай кінцева потужність схеми p_n дорівнює $n+1$. Підрахуємо кінцеву потужність p_{n+1}. Зрозуміло, що ця схема будується на основі схеми p_n, коли в деякій кінцевій вершині послідовно додається нова вершина (ознака, алгоритм) з номером $n+1$. Очевидно, що при додаванні цієї ознаки (алгоритму) у схему p_n зникає одна кінцева вершина та додаються дві нові кінцеві вершини. Отже, можна зробити висновок, що кількість усіх кінцевих вершин схеми p_n дорівнює $n+2$.

Припустимо, що на кожному n-вому кроці процедури побудови дерева класифікації (моделі ЛДК) множина D_n слабо розділяється деякою ознакою φ_n. Далі розглянемо схему p_n. У цій схемі маємо відповідно вищесказаному, $n+1$ кінцевих шляхів. Завдяки тому, що D_n на кожному кроці слабо розділяється, кожний такий шлях містить хоча б одну пару початкової НВ загального вигляду (1). Крім того, очевидно, що різні кінцеві шляхи в p_n не мають спільних пар із вибірки (1).

Отже, можна зробити висновок, що схема (предикат) p_n розділяє НВ (на основі базового критерію розгалуження введеного поточним методом дерева класифікації) на $n+1$ непустих частин (підмножин), що не перетинаються. Оскільки в початковій НВ усього знаходиться m навчальних пар, то схема p_{m-1} (або предикат із меншим номером) повністю розділить початкову НВ, тобто p_{m-1} буде повністю розпізнавати вибірку.

Таким чином, якщо на кожному n-вому кроці відібрана елементарна ознака φ_n слабо розділяє множину D_n, то в цьому випадку процес побудови ЛДК збігається відносно початкової НВ та закінчується не більше за $m-1$ кроків, де $m-$ кількість усіх навчальних пар початкової НВ.

Зауважимо, що умова слабого розділення класів є доволі слабою – тому вона забезпечує невисоку збіжність процедури побудови дерева класифікації, отже, важливо розглянути питання збіжності процесу при більш сильній умові. Тому будемо припускати, що маємо справу з випадком, коли НВ містить інформацію про два класи (образи) H_0 та H_1, а сама НВ має детерміновану природу. Нехай n_j- кількість навчальних пар $\left(x_i, f_R(x_i)\right)$ у початковій НВ, які задовольняють співвідношенню $f_R(x_i)=j, (j=0{,}1)$, причому для спрощення та визначеності покладемо, що $n_0 \geq n_1$.

Зафіксувавши $f_R(x) \equiv 0$, буде отримано деяку узагальнену ознаку (схему) f_0, яка апроксимує (повністю або частково) початкову НВ. Очевидно, що в цьому випадку (тобто в ситуації, коли ще не зроблено вибір жодної елементарної ознаки φ_n) узагальнена ознака (схема) f_0 є найкращою апроксимацією початкової НВ. Далі величину n_1 будемо називати безумовною кількістю помилок у початковій НВ.

Нехай на першому кроці побудови дерева класифікації відібрана (довільним шляхом) деяка елементарна ознака φ_1 – причому ця ознака розіб'є початкову вибірку на дві частини (підмножини) H_0 та H_1, де H_j- множина всіх пар $\left(x_i, f_R(x_i)\right)$ початкової НВ, для яких виконується співвідношення $f_1(x_i)=j, (j=0{,}1)$.

Нехай n_m^j- множина всіх пар $\left(x_i, f_R(x_i)\right)$ з вибірки $H_j, (j=0{,}1)$, для яких виконується співвідношення $f_R(x_i)=m, (m=0{,}1)$. Ознаку φ_1 можна рахувати узагальненою ознакою f_1 (схемою), яка побудована на першому кроці процесу побудови ЛДК.

Введемо величину $\rho=\max(n_0^0, n_1^0)+\max(n_0^1, n_1^1)$, яка являє собою кількість правильних відповідей (класифікацій), які реалізуються узагальненою ознакою f_1, а відповідно величина n_0- являє собою кількість правильних відповідей (класифікацій), які реалізуються узагальненою ознакою f_0.

Під кількістю правильних відповідей розуміємо кількість тих навчальних пар $\left(x_i, f_R(x_i)\right)$ у початковій навчальній виборці типу (1), для яких виконується співвідношення рівності $f_R(x_i)=f_1(x_i)$.

Оскільки $n_0^0 + n_0^1 = n_0$ та $n_1^0 + n_1^1 = n_1$, то

$$\rho = \max(n_0^0, n_1^0) + \max(n_0^1, n_1^1) \geq n_0. \quad (4)$$

Таким чином, при виборі ознаки φ_1 кількість правильних відповідей як мінімум не зменшується. Кількість помилок, які дає узагальнений алгоритм f_1, буде дорівнювати:

$$m - \rho = n_1 - (\rho - n_0) \leq n_1. \quad (5)$$

Зауважимо, що (5) випливає з (4). Введемо величину $\lambda_1 = \frac{n_1}{m-\rho}$ та назвемо її якістю елементарної ознаки φ_1 відносно початкової НВ, аналогічно визначається λ_n ознаки φ_n відносно початкової НВ $(n = 1,2,3, \dots)$.

На наступному етапі дослідження зробимо таке припущення – якість λ_n елементарної ознаки φ_n відносно масиву початкової НВ не менше, ніж деяке число y, де $y > 1$.

Проаналізуємо складність процедури побудови дерева класифікації за цієї умови $(y > 1)$, для цього оцінимо кількість кроків, за якими цей процес (процедура) реалізує повне розпізнавання масиву початкової навчальної вибірки.

Розглянемо для визначеності наступну схему побудови дерева класифікації (рис. 2).

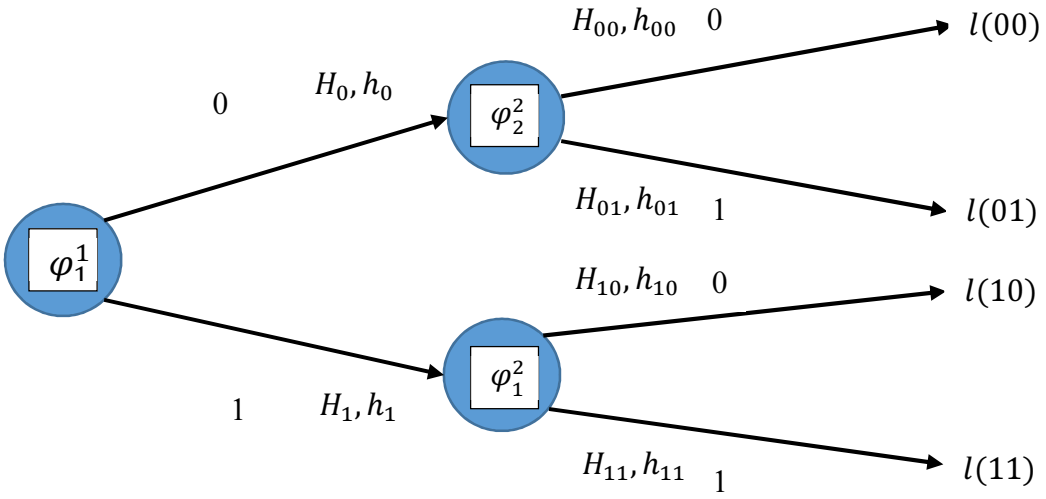

Рис. 2. Схема розбиття на підмножини в структурі дерева класифікації

Нехай n_1 – безумовна кількість помилок початкової НВ. Елементарна ознака φ_1^1 розділяє НВ на дві вибірки H_0 та H_1. Нехай h_0 та h_1, відповідно безумовна кількість помилок у вибірках H_0 та H_1. Ознака φ_1^2 розділить множину H_0 на дві множини H_{00} та H_{01}. Нехай h_{00} та h_{01} – безумовна кількість помилок у вибірках H_{00} та H_{01}. Аналогічно визначимо множини H_{10},H_{11} та кількості h_{10} та h_{11} для елементарної ознаки φ_2^2.

З початкової умови $(y > 1)$ випливає наступна ситуація:

$$\begin{cases} h_0 + h_1 \leq \frac{1}{y} * n_1 \\ h_{00} + h_{01} \leq \frac{1}{y} * h_0. \\ h_{10} + h_{11} \leq \frac{1}{y} * h_1 \end{cases} \quad (6)$$

З (6) отримаємо:

$$h_{00} + h_{01} + h_{10} + h_{11} \leq \frac{1}{y^2} * n_1. \quad (7)$$

Зробимо такі припущення в цьому відношенні: $h_0 \geq 1$, $h_1 \geq 1$, $h_{00} \geq 1$, $h_{01} \geq 1$, $h_{10} \geq 1$ та $h_{11} \geq 1$. Звідси будемо мати таке:

$$2^1 \leq \frac{1}{y} * n_1, 2^2 \leq \frac{1}{y^2} * n_1. \quad (8)$$

Аналогічно для набору ознак $\varphi_1^i, \varphi_2^i, \dots$, які розташовані на i-товому ярусі логічного дерева, будемо мати:

$$2^i \le \frac{1}{y^i} * n_1 \text{ або } (2y)^i \le n_1. \quad (9)$$

Звідси можна зробити висновок, що процес побудови дерева класифікації буде продовжуватися доти, доки в структурі дерева не буде m ярусів (рівнів), де m має такий вигляд:

$$m = R(\frac{\log_2 n_1}{1+\log_2 y}). \quad (10)$$

Під $R(x)$ розуміється заокруглення числа x до найближчого цілого числа, яке перевищує x. Наприклад, $Q(1{,}2) = 2, Q(3{,}7) = 4, Q(4{,}1) = 5$.

Отже, дерево класифікації, яке має m повних ярусів (тобто випадок, коли на $i-$ товому ярусі стоять 2^{i-1} вершин), має $2^{m+1} - 1$ вершин – таким чином розпізнавання початкової НВ за умови $(y > 1)$ за допомогою повного ЛДК відбувається не більше, ніж за $2^{m+1} - 1$ кроків, де m розраховується за допомогою виразу (10).

На наступному етапі дослідження розглянемо питання складності схеми (процедури побудови) дерева класифікації для випадку АДК, ввівши на початку необхідні в подальшому визначення.

Визначення 2. Під потужністю деякої побудованої узагальненої ознаки (УО) або набору УО (для фіксованого кроку схеми АДК) будемо рахувати кількість навчальних пар $\left(x_i, f_R(x_i)\right)$ початкової НВ вигляду (1), які апроксимує (правильно класифікує) дана узагальнена ознака (послідовність узагальнених ознак).

Важливим моментом для схем АДК є те, що при покроковому розбитті НВ на дві вибірки H_0 та H_1 (і так далі) частина вибірки буде повністю покриватися поточним алгоритмом класифікації (узагальненою ознакою або їх набором) – тобто будемо мати справу з випадком сильного розділення класів масиву НВ. Отже, можна зробити припущення, що складність кінцевої схеми АДК (загальна кількість кроків побудови дерева) буде значною мірою залежати від процедури початкової оцінки та відбору набору незалежних алгоритмів класифікації a_i, їх початкових параметрів, параметрів наборів УО f_i, які вони генерують для кожного кроку схеми АДК.

Тоді для схеми АДК важливо розглянути загальну складність процедури побудови дерева класифікації при умові слабкої роздільності класів початкової НВ – при якій генерується не більше однієї УО потужністю в одиницю для кожної вершини дерева та при умові сильної роздільності, коли обмежень на кількість УО та їх потужність не накладається умовами задачі та практичною доцільністю і можливо їх будувати.

Випадок АДК. На першому етапі розглянемо випадок слабкого розділення класів з обмеженнями на набори УО що будуються схемою АДК.

Зазначимо, що процедура побудови алгоритмічного дерева має певні особливості з погляду поетапної апроксимації початкової НВ послідовністю УО. Нехай на кожному кроці побудови деякої моделі АДК, буде вибиратися для роботи один фіксований алгоритм класифікації з набору відібраних алгоритмів $(a_1, a_2, \dots, a_n)$, причому дерево класифікації може бути побудоване одним алгоритмом a_i та послідовністю УО, які він генерує.

Таким чином, після проведення n кроків процедури побудови дерева класифікації структура АДК буде являти собою деяку схему s_n (узагальнену ознаку другого порядку, яка побудована з набору синтезованих алгоритмами класифікації УО), яка є найбільш ефективною апроксимацією початкової НВ загального вигляду (1) набором незалежних алгоритмів класифікації та їх УО. Зрозуміло, що тут також s_n буде представляти деяку деревоподібну схему (структуру ДУО), яка складається з n вершин, тобто в конструкцію

схеми s_n будуть входити всього n алгоритмів класифікації (УО – при умові генерації для кожного кроку процедури побудови дерева не більше однієї узагальненої ознаки мінімальної потужності в одиницю) з початкового набору.

Отже, можна зробити висновок, що послідовність побудованих схем $s_1, s_2, \dots, s_j$ (узагальнених ознак другого порядку) збігається до початкової НВ вигляду (1), не більше ніж за M кроків (де M – загальна потужність початкової НВ), навіть за умов генерації на кожному кроці лише однієї УО, потужність кожної з яких не більше одиниці.

Деякий алгоритм класифікації, який буде вибиратися (фіксуватися) на $n-$ товому кроці в процедурі побудови моделі АДК (для генерації відповідної УО), позначимо через a_n, причому зрозуміло, що цьому алгоритму a_n відповідає деяка схема s_n, яка складається з алгоритмів $a_1, a_2, \dots, a_{n-1}$ та закінчується даним атрибутом (вершиною дерева класифікації – моделі АДК). Наприклад, на рис. 3 зображена деяка модель АДК в якій фіксованій схемі s_2 (вершині дерева класифікації, що будується) відповідає послідовність кроків (схем) $\{s_1\}$, а схемі s_M – послідовний шлях $\{s_1, s_2, \dots, s_{M-1}\}$.

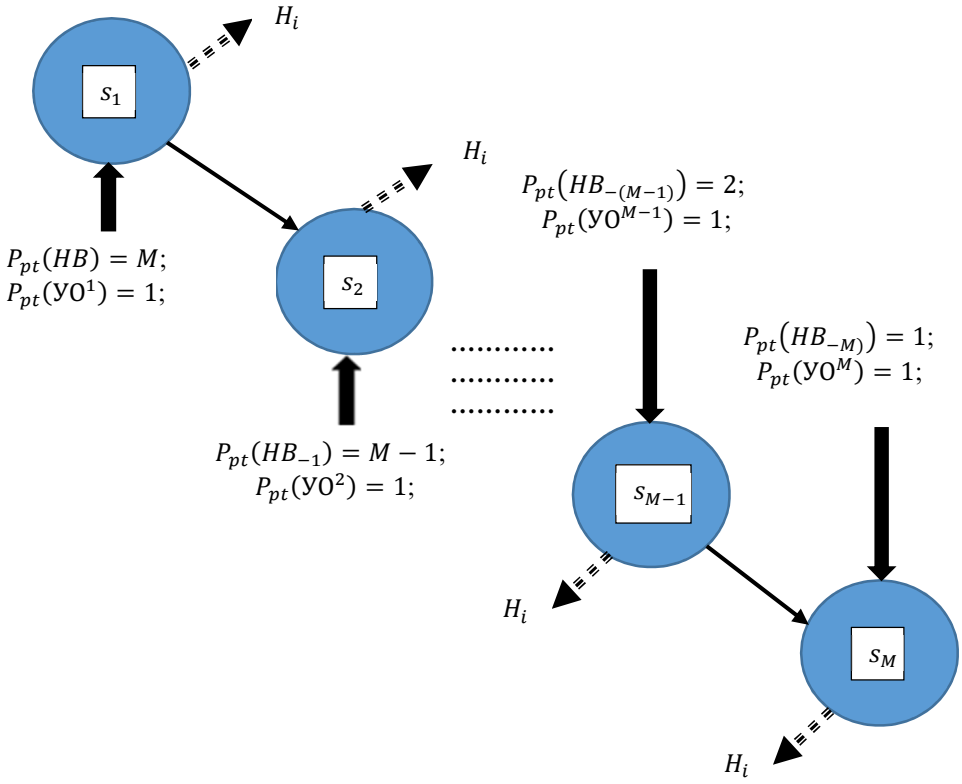

Рис. 3. Приклад структури АДК з УО в ролі вершин

Отже для моделі АДК можна зробити такий висновок: схема s_n (в структурі дерева класифікації) розділяє НВ на n непустих частин (підмножин), що не перетинаються. При цьому, оскільки в початковій НВ усього знаходиться M начальних пар, то схема s_M повністю розділить (апроксимує) початкову НВ (тобто s_M буде повністю розпізнавати вибірку, за умови генерації на кожному кроці по одній УО потужністю один). Таким чином, якщо на кожному $n-$ вому кроці схеми побудови АДК згенерована УО (відібраним алгоритмом класифікації a_n) слабо розділяє множину початкової НВ, то в цьому випадку процес побудови дерева класифікації збігається відносно початкової НВ та закінчується не більше ніж за M кроків, де $M-$ кількість усіх навчальних пар початкової НВ.

На наступному етапі дослідження важливо розглянути випадок сильного розділення класів початкової НВ, коли жодних обмежень на алгоритми a_i щодо генерації УО не накладаються (потужність побудованої УО обмежена лише практичною можливістю самого алгоритму класифікації a_i та структурними параметрами НВ).

Нехай через величину $P(f_j)$ позначимо загальну потужність (апроксимаційну здатність) відповідної УО f_j, $(1 \le j \le s)$, де $s -$ кількість УО в схемі АДК, що будується. Далі на деякому кроці $r, (1 \le r \le M)$ схеми АДК побудована послідовність узагальнених ознак $f_1, \dots, f_r$ з відповідними їм величинами $P(f_i) = z_i$, де $(1 \le z \le M), (1 \le i \le r)$, $M -$ загальна потужність НВ, причому серед них є величини $z^{\max}$ та z^{min} які є для них відповідно максимальними та мінімальними (відносно поточного кроку схеми АДК). Тоді в такому випадку, схема (модель) АДК буде побудована за t кроків, де величина t визначається співвідношенням (11).

$$t \le 2 * \frac{P_{pt}(\text{НВ})}{z^{\max}+z^{\min}} = \frac{2M}{z^{\max}+z^{\min}}. \quad (11)$$

Зауважимо, що у випадку ситуації, коли умовою прикладної задачі на схему АДК, що будується, накладаються обмеження щодо потужності синтезованих УО (не перевищення відповідної величини P) – схема дерева класифікації (модель АДК) буде побудована за t кроків, де величина t визначається співвідношенням (12).

$$t \le \frac{M}{P}. \quad (12)$$

Нагадаємо, що при жорстких обмеженнях схеми АДК на одну генеровану УО (де за умовою $P(f_j) = 1, (1 \le i \le t)$), тобто у випадку слабкого розділення класів поточної задачі схема дерева класифікації (модель АДК) буде побудована за t кроків, де величина $t \le M$.

Висновки відповідно до статті. Отже, зважаючи на все вищезазначене в цьому дослідженні, можна зробити такі висновки:

Для умови слабкого розділення класів у випадку ЛДК, якщо на кожному $n -$ вому кроці відібрана елементарна ознака φ_n слабо розділяє множину (підмножину) об'єктів початкової НВ, то в цьому випадку процес побудови дерева класифікації збігається відносно початкової НВ та закінчується не більше ніж за $m - 1$ кроків, де $m -$ кількість усіх навчальних пар початкової НВ.

Отже, дерево класифікації (структури ЛДК) за умови сильного розділення класів множини об'єктів початкової НВ, яке має m повних ярусів, рівнів (тобто випадок, коли на $i -$ товому ярусі стоять 2^{i-1} вершин), має $2^{m+1} - 1$ вершин. Таким чином розпізнавання масиву початкової НВ при умові $(y > 1)$ за допомогою повного ЛДК відбувається не більш ніж за $2^{m+1} - 1$ кроків, де m розраховується за допомогою виразу $m = R(\frac{\log_2 n_1}{1+\log_2 y})$.

Загальна кількість всіх кінцевих вершин логічної структури (листів дерева розпізнавання) побудованої схеми класифікації буде однозначно визначати кінцеву потужність схеми методу дерева класифікації (моделей ЛДК/АДК).

Потужністю деякої УО (набору побудованих УО), для фіксованого кроку схеми методу АДК – рахується загальна кількість навчальних пар $\left(x_i, f_R(x_i)\right)$ початкової НВ (підмножини початкової НВ) вигляду (1), які апроксимує (правильно класифікує) дана узагальнена ознака (послідовність узагальнених ознак).

У випадку слабкого розділення класів початкової НВ для схеми АДК процес побудови дерева класифікації збігається відносно масиву даних НВ та закінчується не більше ніж за M кроків, де $M -$ кількість всіх навчальних пар початкової НВ.

У випадку сильного розділення класів початкової НВ для схеми АДК, коли потужність побудованої УО (або набору УО) обмежена лише практичною можливістю самого алгоритму класифікації a_i та початковими параметрами НВ – схема (модель) АДК буде побудована за t кроків, де величина t визначається співвідношенням (11).

Список використаних джерел

1. Vtogoff P. E. Incremental Induction of Decision Trees. *Machine Learning.* 2009. № 4. P. 161–186.
2. Whitley D. An overview of evolutionary algorithms: practical issues and common pitfalls. *Information and Software Technology.* 2001. Vol. 43. № 14. P. 817–831.
3. Srikant R., Agrawal R. Mining generalized association rules. *Future Generation Computer Systems.* 1997. Vol. 13, № 2. P. 161–180.
4. Kotsiantis S. B. Supervised Machine Learning: A Review of Classification Techniques. Informatica. 2007. № 31, P. 249–268.
5. 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.
6. Alpaydin E. Introduction to Machine Learning. London: The MIT Press, 2010. 400 p.
7. Суботин С. А. Построение деревьев решений для случая малоинформативных признаков. *Radio Electronics, Computer Science, Control*. 2019. № 1. P. 121–130.
8. 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.
9. Breiman L. L., Friedman J. H., Olshen R. A. Classification and regression trees. Boca Raton : Chapman and Hall/CRC, 1984. 368 p.
10. Dietterich T. G., Kong E. B. Machine learning bias, statistical bias and statistical variance of decision tree algorithms. Corvallis : Oregon State University, 1995. 14 p.
11. Subbotin S.A. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence. *Radio Electronics, Computer Science, Control.* 2014. № 1. Pp. 120–128.
12. Povhan I. General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects. *Збірник наукових праць «Електроніка та інформаційні технології».* 2019. Вип. 11. С. 112–117.
13. 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.
14. 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.
15. Hastie T. Tibshirani R., Friedman J. The Elements of Statistical Learning. Stanford, 2008. 768 p.
16. Mitchell T. Machine learning. New York : McGrawHill, 1997. 432 p.
17. 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.
18. 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.
19. 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.
20. Povhan I. Designing of recognition system of discrete objects. *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)* (Lviv, 2016, Ukraine). Lviv, 2016. Pp. 226–231.
21. 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.
22. Amit Y., Geman D., Wilder K. Joint induction of shape features and tree classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1997. Vol. 19, № 11. P. 1300–1305.
23. Mingers J. An empirical comparison of pruning methods for decision tree induction. *Machine learning.* 1989. Vol. 4. № 2. P. 227–243.

References

1. Vtogoff, P. E. (2009). Incremental Induction of Decision Trees. *Machine Learning,* 4, pp. 161−186.
2. Whitley D. (2001). An overview of evolutionary algorithms: practical issues and common pitfalls. *Information and Software Technology*, *43*(14), pp. 817–831.

3. Srikant, R., Agrawal, R. (1997). Mining generalized association rules. *Future Generation Computer Systems*, *13*(2), pp. 161–180.

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

5. 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.

6. Alpaydin, E. (2010). Introduction to Machine Learning. The MIT Press.

7. Subbotin, S. A. (2019). Construction of decision trees for the case of low-information features [Postroenye derevev reshenyi dlia sluchaia maloynformatyvnыkh pryznakov]. *Radio Electronics, Computer Science, Control*, 1, pp. 121–130.

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

9. Breiman, L. L., Friedman, J. H., Olshen, R. A. (1984). *Classification and regression trees*. Chapman and Hall/CRC.

10. Dietterich, T. G., Kong, E. B. (1995). *Machine learning bias, statistical bias and statistical variance of decision tree algorithms*. Oregon State University.

11. Subbotin, S. A. (2014). Methods and characteristics of locality-preserving transformations in the problems of computational intelligence. *Radio Electronics, Computer Science, Control,* 1, pp. 120–128.

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

13. Vasylenko Yu. A., Vasylenko E. Yu., Povkhan I. F., Vashchuk F. H. (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], pp. 13–15.

14. Vasylenko, Yu. A., Povkhan, I. F., Vashchuk, F. H. (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), pp. 24-28.

15. Hastie T. Tibshirani R., Friedman J. (2008). The Elements of Statistical Learning. Stanford.

16. Mitchell, T. (1997). *Machine learning*. McGrawHill.

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

18. 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], pp. 12–16.

19. 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), pp. 217–222.

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

21. 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]. In *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.

22. Amit, Y., Geman, D., Wilder, K. (1997). Joint induction of shape features and tree classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(11), pp. 1300–1305.

23. Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine learning,* 4(2), pp. 227–243.

UDC 004.8: 004.89: 519.7

Igor Povkhan

QUESTIONS OF COMPLEXITY OF THE PROCEDURE FOR CONSTRUCTING AN ALGORITHMIC CLASSIFICATION TREE SCHEME

Urgency of the research. *At the present stage of development of information systems and technologies that are based on mathematical models of the theory of artificial intelligence (methods and schemes of the classification tree algorithm), there is a fundamental problem of narrow specialization of existing approaches and methods in socio-economic, environmental and other systems of primary analysis and processing of large amounts of information. The problems associated with the theme of pattern recognition are very diverse and arise in the modern world in all areas of Economics and social content of human activity, which leads to the need to build and study mathematical models of the corresponding systems. As of now, there is no universal approach to their solution, several fairly General theories and approaches have been proposed that allow solving many types (classes) of problems, but their applications are quite sensitive to the specifics of the problem itself or the subject area of application. The presented work is devoted to the problems of models of logical and algorithmic classification trees (LCT/ACT schemes), offers an assessment of the complexity of algorithmic structures of trees (classification tree models), which consist of independent and Autonomous classification algorithms and will represent to a certain extent a new recognition algorithm (it is clear that synthesized from known schemes, algorithms and methods).*

Target setting. *Today, various approaches to building recognition systems in the form of classification trees (LCT/ACT) are relevant, and the interest in recognition methods that use classification trees is caused by a number of useful properties that they possess. On the one hand, the complexity of the class of recognition functions in the form of classification tree models, under certain conditions, does not exceed the complexity of the class of linear recognition functions (the simplest known). On the other hand, recognition functions in the form of classification trees allow you to distinguish between causal factors in the classification process – it is known that the process of classification of new objects that have not yet been encountered in the world of many animals and people (with the exception of objects, information about which is transmitted by genetic means (investigative), as well as in some other cases), occurs according to the so-called logical decision tree (in connection with the neural network concept). 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 algorithmic models of classification trees (ACT structures), and that is why this work intends to at least partially overcome these limitations and is devoted to assessing the complexity of the procedure for constructing algorithmic models (logical) classification trees in the field of recognition problems.*

Actual scientific researches and issues analysis. *The study reviewed recent open access publications on the General problem of approaches, methods, algorithms and recognition schemes (LCT/ACT models) for discrete objects (discrete images in image recognition problems.*

The research objective. *The possibility of an efficient and cost-effective scheme for constructing an algorithmic classification tree and evaluating the complexity of such a procedure (ACT structure model) based on the source arrays of large-volume training samples.*

The statement of basic materials. *We propose an estimation of the complexity of the procedure for constructing an algorithmic classification tree for an arbitrary case (for conditions of weak and strong division of classes in the training sample). The solution to this question is of a fundamental nature, in terms of assessing the structural complexity of classification models (in the form of tree structures, ACT structures of discrete objects for a wide class of applied classification and recognition problems in terms of developing promising schemes and methods for their final optimization (minimization) of the design. This research is relevant not only for constructions of algorithmic classification trees, but also allows us to extend the scheme of complexity estimation to the General case of logical classification tree structures.*

Conclusions. *We investigated the structural complexity LCT/ACT, the proposed upper bound for the complexity for building algorithmic classification tree in terms of weak and strong separation of classes of the initial training sample.*

Keywords: *recognition problems, classification trees, algorithmic tree, recognition scheme, discrete object, generalized feature. References: 23.*

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 (89Б Zankovetskoy Str., 88000 Uzhgorod, Ukraine).
E-mail: igor.povkhan@uzhnu.edu.ua
ORCID: http://orcid.org/0000-0002-1681-3466

Повхан І. Питання складності процедури побудови схеми алгоритмічного дерева класифікації. *Технічні науки та технології.* 2020. № 3 (21). С. 142-153.