ISSN 2707-1898 (print)
Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2022.02.025
Article received 15.10.2022 p.
Article accepted 17.10.2022 p.
UDK 004.[8+89]:519.7

Correspondence author
I. F. Povkhan
igor.povkhan@uzhnu.edu.ua

І. Ф. Повхан
Ужгородський національний університет, м. Ужгород, Україна

МЕТОД СИНТЕЗУ ЛОГІЧНИХ ДЕРЕВ КЛАСИФІКАЦІЇ НА ПІДСТАВІ СЕЛЕКЦІЇ ЕЛЕМЕНТАРНИХ ОЗНАК

Розглянута загальна задача побудови логічних дерев класифікації та розпізнавання дискретних об'єктів. Об'єктом даного дослідження є логічні дерева класифікації. Предметом дослідження є актуальні методи та алгоритми побудови логічних дерев класифікації. Метою роботи є створення простого та ефективного методу побудови моделей розпізнавання на підставі дерев класифікації для навчальних вибірок дискретної інформації, який характеризується елементарними ознаками в структурі синтезованих логічних дерев класифікації. Запропоновано загальний метод побудови логічних дерев класифікації, який для заданої початкової навчальної вибірки будує деревоподібну структуру, яка складається з набору елементарних ознак, оцінених на кожному кроці побудови моделі за даною вибіркою. Розроблено метод побудови логічного дерева, основна ідея якого полягає в апроксимації начальної вибірки довільного об'єму набором елементарних ознак. Під час формування поточної вершини логічного дерева, його вузол забезпечує виділення найбільш інформативних, якісних елементарних ознак з початкового набору. Такий підхід при побудові остаточного дерева класифікації дає змогу значно скоротити розмір та складність дерева, загальну кількість гілок та ярусів структури, підвищити якість його подальшого аналізу. Запропонований метод побудови логічного дерева класифікації дає змогу будувати деревоподібні моделі розпізнавання для широкого класу задач теорії штучного інтелекту. Розроблений та наведений в роботі метод отримав програмну реалізацію та був досліджений під час розв'язання задачі класифікації даних геологічного типу. Проведені в роботі експерименти підтвердили працездатність запропонованого математичного забезпечення та показують можливість його використання для розв'язання широкого спектру практичних задач розпізнавання та класифікації. Перспективи подальших досліджень полягають в створенні обмеженого методу логічного дерева класифікації, який полягає у введенні критерію зупинки процедури його побудови за глибиною структури, оптимізації його програмних реалізацій, а також проведення експериментального дослідження цього методу на більш широке коло практичних задач.

Ключові слова: логічне дерево; селекція ознак; критерій розгалуження; дискретний об'єкт.

Вступ / Introduction

Інформаційні технології, засновані на математичних моделях розпізнавання образів у вигляді логічних дерев класифікації (ЛДК), широко використовують в соціально-економічних, екологічних і інших системах оброблення інформації. Це пояснюється тим фактом, що такий підхід дає змогу усунути набір недоліків класичних методів і досягти принципово новий результат, ефективно та раціонально використовуючи потужності обчислювальних систем [19].

На сьогодні відомо понад 3500 алгоритмів розпізнавання (заснованих на різноманітних підходах та концепціях), які мають певні обмеження при їх використанні (точність, швидкодія, пам'ять, універсальність, надійність, тощо). Окрім цього, кожний з алгоритмів обмежений певною специфікою задач застосування, а це безумовно є найслабкішим місцем не тільки даних алгоритмів, але й систем розпізнавання, робота яких побудована на відповідних концепціях [31].

Відомо, що подання навчальних вибірок (дискретної інформації) великого об'єму у вигляді структур логічних дерев має свої істотні переваги щодо економічного опису даних і ефективних механізмів роботи з ними [14]. Тобто, покриття навчальної вибірки набором елементарних ознак у випадку ЛДК, або покриття навчальної вибірки фіксованим набором автономних алгоритмів розпізнавання та класифікації у випадку алгоритмічних дерев класифікації (АДК), породжує фіксовану деревоподібну структуру даних, яка якоюсь мірою забезпечує навіть стиск та перетворення початкових даних навчальної вибірки (НВ), а також дає істотну оптимізацію та економію апаратних ресурсів інформаційної системи [20].

Відзначимо, що галузь застосування концепції ЛДК в даний час надзвичайно обширна, а множина задач та проблем, які розв'язуються цим апаратом, можна звести до наступних трьох базових сегментів – задачі опису структур даних, задачі розпізнавання та класифікації, задачі регресії [5].

Так, здатність ЛДК виконувати одновимірне 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) [13]. В даному випадку структура логічного дерева подана у вигляді гілок та вузлів, причому на гілках дерева розташовують 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. Відзначимо, що при побудові ЛДК центральними питаннями залишаються питання вибору критерію атрибута (вершини ЛДК), за якою відбудеться поділ початкової навчальної вибірки, критерію зупинки навчання (побудови структури ЛДК) та критерію відкидання гілок логічного дерева (піддерев ЛДК). Саме на цьому етапі виникає принципове пи-

тання теорії ЛДК – питання можливої побудови всіх варіантів логічних дерев, які відповідають початковій навчальній вибірці, та відбору мінімального за глибиною (кількістю ярусів) логічного дерева [15]. Дана задача є *NP* – повною (це було зафіксовано ще Л. Хайфілем та Р. Рівестом), тому немає простих і ефективних методів її розв'язання.

Основні наявні методи оброблення навчальних вибірок при побудові функції розпізнавання не дають змогу досягнути потрібного рівня точності системи розпізнавання та регулювати їх складність у процесі конструювання даних систем [2]. Цей недолік відсутній у методах побудови систем розпізнавання, робота яких побудована на методах логічних дерев класифікації (дерев рішень). При цьому особливістю методу логічного дерева (методу алгоритмічного дерева класифікації) – є можливість комплексного використання для розв'язання кожної конкретної задачі побудови схеми розпізнавання багатьох відомих алгоритмів (методів) розпізнавання. В основі знаходиться єдина методологія – оптимальна апроксимація навчальної вибірки набором узагальнених ознак (автономних алгоритмів), які входять в деяку схему (оператор), побудовану в процесі навчання [30].

Об'єкт дослідження – процеси синтезу логічних дерев класифікації (структур ЛДК).

Предмет дослідження – методи, алгоритми та схеми побудови логічних дерев класифікації різних типів.

Мета роботи – створення простого та ефективного методу побудови моделей розпізнавання на підставі дерев класифікації для навчальних вибірок дискретної інформації, який характеризується структурою отриманих логічних дерев класифікації з елементарних ознак, оцінених на підставі функціоналу розрахунку їх інформативності.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*: аналіз базової схеми селекції елементарних ознак; побудова методу синтезу структури ЛДК на підставі селекції елементарних ознак.

Аналіз останніх досліджень та публікацій. Аналізуючи проблематику деревоподібних моделей класифікації та розпізнавання можна побачити певну відсутність поточних досліджень в цьому напрямі, коли головна увага зміщена в бік концепції нейромережевого розпізнавання [1]. Значною мірою це пояснюється особливостями самих моделей ЛДК, труднощами реалізаційних моментів концепції алгоритмічного дерева класифікації (найвищого рівня абстракції концепції ЛДК), набором жорстких правил та обмежень щодо практичної роботи з такими структурами даних [10].

Дане дослідження продовжує цикл робіт [26], [27], [28], які присвячені проблематиці побудови деревоподібних схем розпізнавання (класифікації) дискретних об'єктів. В них піднято питання побудови, використання та оптимізації логічних дерев. Так, з роботи [26] відомо, що остаточне правило класифікації (схема), яке побудоване довільним методом або алгоритмом розгалуженого вибору ознак, має деревоподібну логічну структуру. Логічне дерево складається з вершин (ознак), які групуються ярусами і отримані на певному кроці (етапі) побудови дерева розпізнавання [31]. Важливою задачею, яка виникає з роботи [13], задача синтезу дерев розпізнавання, які будуть представлені фактично деревом (графом) алгоритмів. На відміну від наявних методів, головною особливістю деревоподібних систем розпізнавання є те, що важливість окремих ознак (групи ознак чи алгоритмів) визначають відносно функції, яка задає поділ об'єктів на класи [32].

Так, в роботі [24] піднято принципові питання стосовно генерування дерев рішень для випадку малоінформативних ознак. Здатність ЛДК виконувати одновимірне 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). Таку концепцію логічних дерев активно використовують в інтелектуальному аналізі даних, де остаточна мета полягає в синтезі моделі, яка прогнозує значення цільової змінної на підставі набору початкових даних на вході системи [4], [6], [7].

На сьогодні існує значна кількість алгоритмів, які реалізують концепцію дерев рішень – CART, C4.5/C5.0, Sparc, NewID, ITrule, CHAID, CN2, Oris та інші. Однак, найбільшого вживання та розповсюдження отримали два їхні перші представника. Згаданий вище алгоритм логічного дерева C4.5/C5.0 як критерій відбору вузла (вершини) використовує так званий теоретико-інформаційний критерій, а робота алгоритму CART базується на розрахунку індексу Gini, який враховує відносні відстані між розподілами класів [9].

Важливими елементами методів розгалуженого вибору ознак є робота [31], де запропоновано та проаналізовано схему побудови ЛДК на підставі алгоритму логічного дерева з покроковою оцінкою важливості дискретних ознак за даними навчальної вибірки. В роботі [30] запропоновано модифікований алгоритм розгалуженого вибору ознак (РВО) з одноразовою оцінкою набору ознак, а в роботі [18] – алгоритм генерування набору (множини) випадкових логічних дерев класифікації з фінальним етапом відбору оптимального.

Так, як головну ідею методів і алгоритмів РВО можна визначити як оптимальну апроксимацію деякої початкової навчальної вибірки набором елементарних ознак (атрибутів об'єкту), то на перший план виходить їх центральна проблема – питання вибору ефективного критерію розгалуження (відбору вершин, атрибутів, ознак дискретних об'єктів). Саме ці принципові задачам розглядаються в роботі [1], де піднято питання якісного оцінювання окремих дискретних ознак, їх наборів і фіксованих сполучень, що дає змогу запровадити ефективний механізм реалізації розгалуження.

Структура ЛДК характеризується компактністю з одного боку та нерівномірністю заповнення (розрідженістю) ярусів з іншого боку порівняно з регулярними деревами (алгоритмом з одноразовою оцінкою важливості ознак) [17]. Важливими питаннями залишаються питання збіжності процесу побудови ЛДК за методами РВО та питання вибору критерію зупинки процесу синтезу логічного дерева (наприклад, обмеження за глибиною або складністю дерева, обмеження за точністю або кількістю помилок структури, що будується) [31].

Концепції логічних дерев не суперечать можливості, як ознаки (вершин) ЛДК, використовувати не тільки окремі атрибути (ознаки) об'єктів їх сполучення (ідея узагальненої ознаки, розглядалась в роботі [29]) та набори. Однак, якщо не розглядати як розгалуження атрибути

об'єктів (ознаки), а відбирати окремі незалежні алгоритми розпізнавання (оцінені за даними навчальної вибірки), то на виході буде отримана нова структура – АДК.

Результати дослідження та їх обговорення / Research results and their discussion

Нехай задана навчальна вибірка в такому вигляді:

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

де: $x_i \in G$, G – деяка множина; ФР $f_R(x_i) \in \{0,1,2,...,k-1\}$, $i = \overline{1,M}$. Відповідно $f_R(x_i) = l, (0 \le l \le k-1)$ означає, що $x_i \in H_l, H_l \subset G$, де f_R – деяка скінчено значна функція, яка задає поділ R множини G, що складається з підмножин (образів, класів) $H_0, H_1, H_2, ..., H_{k-1}$.

З довідкової літератури відомо [16], що навчальна вибірка – сукупність (точніше послідовність) деяких наборів, причому кожний набір – сукупність значень деяких ознак та значень деяких функцій на цьому наборі. Це сукупність значень ознак – деяке зображення, а значення функції відносить це зображення до відповідного образу. Отже, буде наше завдання – побудувати ЛДК, оптимальна структура якого $f_R(x_j) \to opt$ за відношенням до початкових даних навчальної вибірки.

Основна схема алгоритмів методу РВО на підставі концепції ЛДК полягає в тому, щоби максимізувати величину $W_M(f)$ [31]. Це означає, що в алгоритмах логічного дерева має бути знайдена для навчальної вибірки (1) така узагальнена ознака f, для якої величина $W_M(f)$ є за можливості найбільшою. Під важливістю ознаки будемо розуміти таку величину

$$W(\phi_1^1) = \sum_{i=1}^{m} \frac{b_i}{M} \rho_i, \text{ де } \rho_i = \max_{1 \le m \le k} \frac{q_i^m}{b_i}. \quad (2)$$

Зрозуміло, що аналогічно можна оцінити важливість інших ознак. Величину q_i^m / b_i можна інтерпретувати як імовірність того, що функція $f_R(x)$ набуде значення $O_m, (1 \le m \le k)$, за умови, що значення ознаки ϕ_1^1 становить i $(1 \le i \le k)$. Величина ρ_i представляє максимальну з цих ймовірностей. Можна стверджувати, що величина ρ_i представляє ту інформацію, яку можна отримати про значення функції $f_R(x)$, знаючи, що на наборі z значення ознаки ϕ_1^1 становить i. Величина $W(\phi_1^1)$, яку визначають цією формулою, характеризує ту інформацію, яку можна отримати про функцію $f_R(x)$, якщо відомо значення ознаки ϕ_1^1 на наборі z. Зрозуміло, що ознака, для якої ця інформація є найбільшою, вважається найбільш важливою ознакою за відношенням до $f_R(x)$.

Вибірка (1) може мати імовірнісний характер, тобто пари $(x_i, f_R(x_i))$, $i = \overline{1,M}$ можуть в ній з'являтися згідно з деяким імовірнісним розподілам $p(x/H_0),...,p(x/H_{k-1})$, але узагальнена ознака є детермінованою. Отже, потрібно здійснити оптимальну апроксимацію ймовірнісної вибірки (1) за допомогою деякої детермінованої функції, яка загалом подана узагальненою ознакою f. Очевидно, що така задача має сенс тоді, коли характер образів (класів) $H_0, H_1, ..., H_{k-1}$ достатньо близький до детермінованого. Це означає, що основну частку займають ті точки (об'єкти) x, для яких величина $\max(p(x/H_0),...,p(x/H_{k-1}))$ близька до одиниці. Ця величина може істотно мінятись тільки в точках (об'єктах), які знаходяться на межі декількох класів $H_0, H_1, ..., H_{k-1}$.

На практиці алгоритми РВО, в основному, працюють з задачами, де образи (класи) $H_0, H_1, ..., H_{k-1}$ мають характер, близький до детермінованого випадку. Такі алгоритми мають наступну особливість – кожний алгоритм відображає процес, який містить певні кроки $d_0, d_1, ..., d_i$. Кожний крок d_j, водночас, містить два етапи (режими) – навчання та перевірочного тесту.

В режимі навчання на кроці d_i формується деяка узагальнена ознака f_i. В режимі тесту для цієї узагальненої ознаки розраховують ефективність $W_M(f_i)$ відносно навчальної вибірки (1). Якщо $W_M(f_i) \ge \delta$, то на цьому процес навчання завершується, якщо $W_M(f_i) < \delta$, тоді здійснюють перехід до кроку d_{i+1}.

Відзначимо особливості подачі вибірки (1) на етапі навчання. На практиці можливі два випадки:

a) вибірка (1) фіксована, тобто вся вона подається на кожному кроці навчання d_i;

b) вибірка (1) залежить від кроку d_i, тобто на кожному кроці навчання d_i подається своя вибірка.

Випадок (*a*) має місце тоді, коли вибірка (1) представляє данні деякого експерименту (наприклад комп'ютерних замірів), які записані в постійну пам'ять. Алгоритм навчання в цьому випадку виконує багатократну обробку вибірки (1), яка може мати дуже великий об'єм. Тому алгоритми оброблення вибірки мають бути такими, що би при їх роботі вибірка (1) не заносилася в оперативну пам'ять.

Якщо відсутній випадок (*a*) та не потрібно зберігати дані в постійній пам'яті, то маємо випадок (*b*). В цьому випадку всі пари, які обробляють на кроці d_i, не запам'ятовуються, і тому на кроці d_{i+1} подається вже деяка інша серія навчальних пар виду (1). Для визначеності далі будемо вважати, що має місце випадок (*a*), тобто на кожному кроці d_i подається одна і та сама навчальна вибірка (1).

В даній схемі побудови ЛДК спочатку вибираємо деяку елементарну ознаку ϕ_1^1. Від цієї ознаки вимагається, щоби величина $W_M(\phi_1^1)$ відносно (1) була за можливості найбільшою. Величину $W_M(\phi_1^1)$ розраховують відповідно до методики [16], [23], [25], [31]. Наступні кроки методу логічного дерева зручно інтерпретувати за допомогою дерева – (рис. 1).

В кожній вершині ЛДК – (рис. 1) знаходиться або деяка ознака ϕ_i^j, або число m_i^j, яке належить множині $\{0,1,...,k-1\}$. Вершину, в якій знаходиться m_i^j, називають остаточною вершиною дерева. Від кожної вершини, в якій знаходиться ознака ϕ_i^j, відходять дві направляючі (стрілки) які позначені 0 та 1. Направляючій, яка позначена 0, відповідає значення $\phi_i^j = 0$, а позначеній 1 – значення $\phi_i^j = 1$. Дерево розбите умовно за ярусами. В j-му ярусі знаходяться ознаки $\phi_1^j, \phi_2^j, ...,$.

Всі ознаки, які знаходяться в усіх ярусах, починаючи з першого та завершуючи n-им, представляють ті ознаки, які отримані після проведення n кроків (етапів) процесу побудови дерева класифікації (ЛДК). Причому ознаки, які знаходяться на n-му ярусі, представляють ті

ознаки, які отримані на n кроці (етапі) процесу побудови логічного дерева.

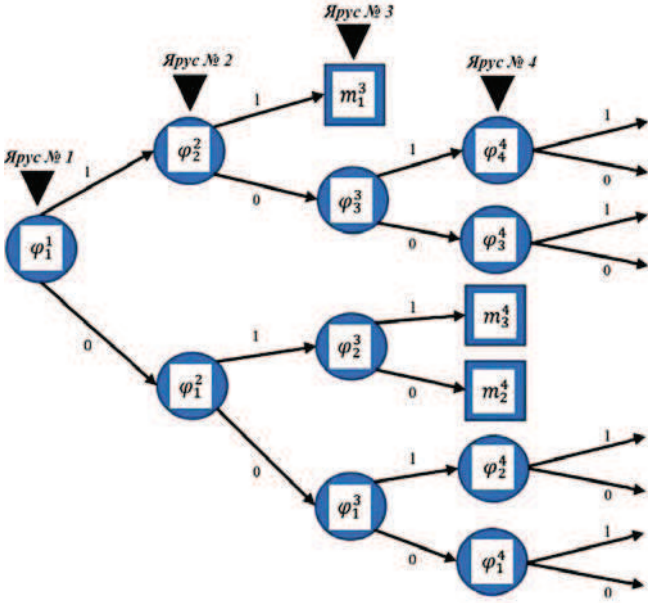

Рис. 1. ЛДК на підставі методу розгалуженого вибору ознак / LCT based on the branched feature selection method

Припустимо, що проведено тільки три кроки побудови ЛДК та $\phi_1^1,\phi_1^2,\phi_2^2,\phi_1^3,\phi_2^3,\phi_3^3$ – всі ознаки, які отримані внаслідок цих крокі. Логічне дерево, яке отримаємо за ці три кроки, буде мати такий вигляд (рис. 2). Кожній зв'язаній парі $(x_i, f_R(x_i))$, $(1 \le i \le M)$ навчальної вибірки (1) буде відповідна певний шлях ЛДК. Цей шлях реалізують в такий спосіб. Спочатку розраховують $\phi_1^1(x_i) = r_i$. Далі від вершини ϕ_1^1 спускаємося вниз за стрілкою, яка позначена через r_i. Нехай, наприклад, $\phi_1^1(x_i) = r_1 = 0$. Тоді спускаємось у вершину, в якій знаходиться ознака ϕ_1^2. Далі розрахуємо $\phi_1^2(x_i) = r$ та спускаємося за стрілкою, яка виходить з вершини ϕ_1^2 та позначеної значенням r_2 і т.д.

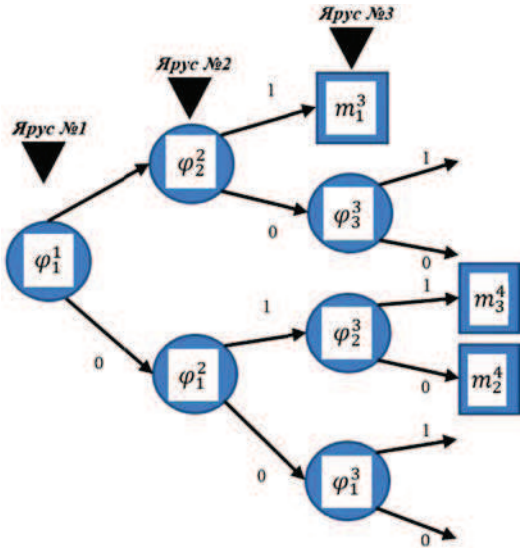

Рис. 2. Логічне дерево після трьох кроків процедури генерування ЛДК / Logical tree after three steps of the LCT generation procedure

Шлях, якій відповідає парі $(x_i, f_R(x_i))$ (він повністю визначають значенням x_i), позначимо через T_i. В цій ситуації можливі два випадки:

a) Нехай шлях T_i завершується деякою напрямною стрілкою. Наприклад, якщо маємо ситуацію, коли $\phi_1^1(x_i) = 0$, $\phi_1^2(x_i) = 0, \phi_1^3(x_i) = 0$, то шлях T_i завершується стрілкою, яка виходить з вершини ϕ_1^3 та позначається символом 0 (рис. 2).

b) Нехай шлях T_i завершується деякою вершиною, в якій знаходиться значення – m_i^j. Наприклад, коли $\phi_1^1(x_i) = 1, \phi_2^2(x_i) = 1$, то шлях T_i завершується вершиною, в якій знаходиться значення m_1^3 (рис. 2).

Шляхи у випадку (*a*) будемо називати незавершеними, а шляхи (*b*) – завершеними. Якщо шлях T_i, який відповідає парі $(x_i, f_R(x_i))$, є завершеним і в кінці знаходиться значення $m_i^j, (m_i^j \in \{0,1,...,k-1\})$ то це означає, що $f_R(x_i) = m_i^j$. Наприклад на (рис. 2) для всіх пар $(x_i, f_R(x_i))$, які задовольнять умову $\phi_1^1(x_i) = \phi_2^2(x_i) = 1$, виконується ще така умова: $f_R(x_i) = m_1^3$. Це означає, що для значення x_i, якому відповідає завершений шлях T_i, реалізують повне розпізнавання за деревом (рис. 2). Тобто, пара $(x_i, f_R(x_i))$ належить відповідному шляху T_i.

При проведенні наступних етапів побудови дерева класифікації розглядаються тільки незавершені шляхи.

Далі кожний шлях на дереві, що будується, будемо позначати двійковим набором $r_1, r_2, r_3 ....(r_i \in \{0,1\})$. Наприклад, двійковий набір 010 на дереві (рис. 2) позначає шлях, який завершується кінцевою стрілкою, з вершини ϕ_2^3 та позначеною символом 0. Очевидно, що сукупність 000, 001, 010, 011, 100, 101 представляє множину всіх незавершених шляхів на ЛДК (рис. 2).

Нехай M_{r_1,r_2,r_3} – кількість всіх пар $(x_i, f_R(x_i))$ вибірки (1), які належать незавершеному шляху r_1, r_2, r_3 логічного дерева (рис. 2), та $M^j_{r_1,r_2,r_3}$, $(0 \le j \le k-1)$ – кількість всіх пар, які належать шляху r_1, r_2, r_3 та, окрім цього, для них виконується таке співвідношення: $f_R(x_i) = j$. Для кожного незавершеного шляху r_1, r_2, r_3 дерева (рис. 2) розрахуємо величини:

$$t^j_{r_1,r_2,r_3} = \frac{M^j_{r_1,r_2,r_3}}{M_{r_1,r_2,r_3}}, \ (j = 0,1,...,k-1). \qquad (3)$$

Далі знайдемо таку величину $l(r_1, r_2, r_3)$, що:

$$l(r_1,r_2,r_3) \in \{0,1,...,k-1\}, \ t^{l(r_1,r_2,r_3)}_{r_1,r_2,r_3} = \max_j t^j_{r_1,r_2,r_3}.$$

Підставивши в кінці кожного шляху r_1, r_2, r_3 на ЛДК (рис. 2) величину $l(r_1, r_2, r_3)$, отримаємо наступне ЛДК (рис. 3). Це дерево реалізує деяку узагальнену ознаку $f_3(x)$, яка визначена на множині G та приймає значення з множини $\{0,1,...,k-1\}$. Ознаку $f_3(x)$ розраховують в такий спосіб. Спочатку знаходимо для x весь шлях T_x, якій відповідає даному елементу (об'єкту навчальної вибірки). Наприклад, якщо $\phi_1^1(x) = 0, \ \phi_1^2(x) = 1, \ \phi_2^3(x) = 1$, то $T_x = 011$. Як значення $f_3(x)$ беремо той елемент з $\{0,1,...,k-1\}$, яким завершує шлях T_x. Наприклад, якщо $T_x = 011$, то $f_3(x) = l(011)$. Якщо об'єкту x відповідає завершений шлях T_x, в кінці якого знаходиться число m_x, $(0 \le m_x \le l)$, тоді вважаємо, що $f_3(x) = m_x$.

Після побудови ознаки $f_3(x)$ починають етап перевірочного тесту. В режимі тесту підраховують кількість S

всіх тих пар $(x_i, f_R(x_i))$ з вибірки (1), для яких виконується співвідношення $f_R(x_i) = f_3(x)$.

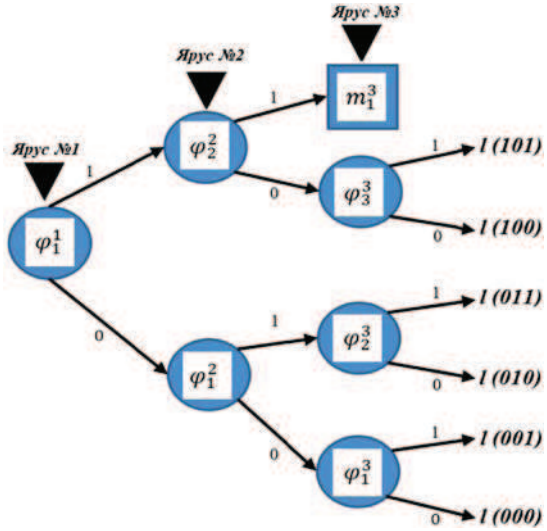

Рис. 3. Логічне дерево після трьох кроків процедури генерування з фіксованими шляхами / Logical tree after three steps of the generation procedure with fixed paths

Далі перевіряємо умову $S / M \geq \delta$. Якщо ця умова виконана, то на цьому процес побудови дерева класифікації завершується, а узагальнена ознака $f_3(x)$, яка видається логічним деревом (рис. 3), є такою, що забезпечує апроксимацію вибірки вигляду (1). Якщо $S / M < \delta$, то процес побудови дерева продовжується. При побудові дерева розпізнавання спочатку виділяємо на ЛДК (рис. 3) всі ті значення $l(r_1, r_2, r_3)$, для яких виконується співвідношення $t_{r_1,r_2,r_3}^{l(r_1,r_2,r_3)} = 1$. Шляхи r_1, r_2, r_3, для яких виконується тільки що вказане співвідношення, можна вважати завершеними.

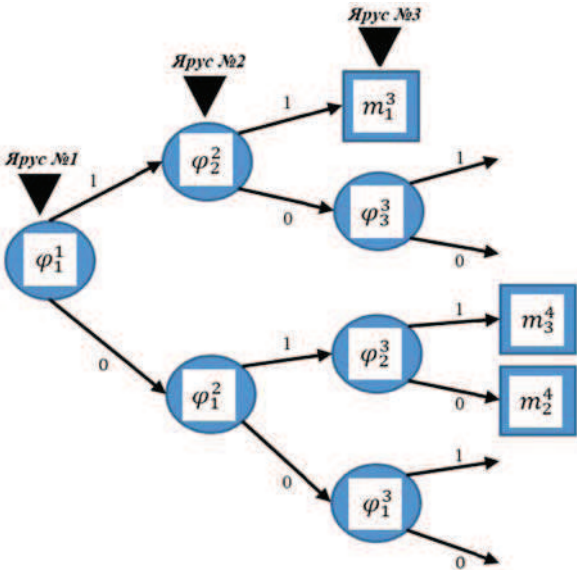

Рис. 4. Фінальне ЛДК / Final LCT structure

Наприклад нехай:

$$t_{010}^{l(010)} = t_{011}^{l(011)} = 1;\ \ t_{000}^{l(000)} < 1;\ \ t_{001}^{l(001)} < 1;\ \ t_{100}^{l(100)} < 1;\ \ t_{101}^{l(101)} < 1.$$

В цьому випадку отримаємо дерево вигляду (рис. 4), де $m_2^4 = l(010),\ m_3^4 = l(011)$. Всі шляхи 000, 100 та 101 на цьому дереві є незавершеними. Для кожного з цих шляхів r_1, r_2, r_3 розглядаємо множини H_{r_1,r_2,r_3}, де H_{r_1,r_2,r_3} – множина всіх тих пар $(x_i, f_R(x_i))$ навчальної вибірки, які належать шляху r_1, r_2, r_3. Множини H_{r_1,r_2,r_3} можна вважати деякими вибірками. У випадку логічного дерева – (рис. 4) будемо мати такі вибірки $H_{000}, H_{001}, H_{100}, H_{101}$.

Для кожної вибірки H_{r_1,r_2,r_3} вибираємо таку елементарну ознаку ϕ_{r_1,r_2,r_3}, для якої величина $W_{H_{r_1,r_2,r_3}}(\phi_{r_1,r_2,r_3})$ є за можливості найбільшою. Величина $W_{H_{r_1,r_2,r_3}}(\phi_{r_1,r_2,r_3})$ – ефективність розпізнавання вибірки H_{r_1,r_2,r_3} за допомогою ознаки ϕ_{r_1,r_2,r_3}. Після вибору ознак ϕ_{r_1,r_2,r_3} отримаємо нове ЛДК, яке зображено на (рис. 1). Тут будемо мати наступне:

$$\phi_1^4 = \phi_{000}; \phi_2^4 = \phi_{001}; \phi_3^4 = \phi_{100}; \phi_4^4 = \phi_{101}.$$

Далі до дерева (рис. 1) застосовуємо той самий процес, що і до дерева (рис. 2). Для реалізації кожної вибірки H_{r_1,r_2,r_3} не потрібно формувати окрему множину навчальних пар. Всі ці вибірки можна реалізувати в такий спосіб: послідовно подають пари вибірки (1) та до уваги приймають тільки ті навчальні пари, які належать шляху r_1, r_2, r_3. Внаслідок реалізації цього процесу і буде реалізовуватись вибірка H_{r_1,r_2,r_3}.

Зрозуміло, що на підставі однієї початкової навчальної вибірки можна побудувати набір ЛДК, використовуючи відповідні методи та алгоритми. Так, використовуючи як критерій розгалуження формулу (2) можна побудувати мінімум два ЛДК залежно від того, чи проводити оцінку якості елементарних ознак на кожному кроці генерування логічного дерева, чи зробити це один раз на початку побудови і, цим самим, заощадити апаратні ресурси системи. Тому виникає актуальна задача порівняння отриманих моделей для вибору найкращої відносно поточної навчальної вибірки.

Важливим етапом порівняння побудованих деревоподібних моделей розпізнавання є етап визначення показників, які характеризують базові властивості отриманих моделей. Причому порівняння моделей проводять на підставі інтегрального показника якості – інтрального критерію порівняння моделей ЛДК.

За аналогією з роботами [8], [21], [23], [24] зафіксуємо наступні базові характеристики синтезованих ЛДК в такому переліку:

- V_{tr} – загальна кількість вершин побудованого ЛДК. Мінімальна кількість вершин ЛДК становить $(M+1)$;
- N_{tr} – загальна кількість ознак, які використовують в структурі побудованої моделі ЛДК;
- C_{tr} – загальна кількість переходів (зв'язків) в структурі моделі ЛДК;
- O_{tr} – загальна кількість остаточних значень ФР (листків дерева) в структурі моделі ЛДК;
- En_{tr} – помилка моделі ЛДК на масиві даних навчальної вибірки;
- Et_{tr} – помилка моделі ЛДК на масиві даних тестової вибірки (ТВ);
- Er_H – полика на кожному з класів дискретних об'єктів, причому $(H = 1,...,k)$.

На наступному етапі зафіксуємо основні параметри моделі ЛДК відносно його характеристик:

- $C_{avg} = (C_{tr} / V_{tr})$ – середня кількість переходів на одну вершину в ЛДК;
- $N_{tr}^V = (N_{tr} / n)$ – частка елементарних ознак, які використовують в структурі ЛДК;

- $O_{tr}^{V} = (O_{tr} / V_{tr})$ – частка остаточних значень ФР в загальній структурі ЛДК;
- $Q_{avg} = (M / O_{tr})$ – середня кількість наборів навчальної вибірки на остаточне значень ФР (листків дерева) в структурі ЛДК.

Загальний показник моделі ЛДК узагальнення даних початкової навчальної вибірки розраховують:

$$I_{Main} = \frac{M \cdot n}{V_{tr} + 2C_{tr}}.$$

Зрозуміло, що критично важливими параметрами моделі ЛДК, які необхідно мінімізувати, є помилки моделі En_{tr}, Et_{tr}, Er_H (відповідно на даних навчальної вибірки, ТВ та для кожного з класів початкового поділ множини G поточної задачі).

Принциповим моментом залишається питання зменшення складності структури ЛДК (тут мається на увазі параметри N_{tr} – кількість ознак в структурі ЛДК, V_{tr} – кількість вершин моделі ЛДК та C_{tr} – загальна кількість переходів в структурі ЛДК), параметри витрат пам'яті λ та процесорного часу τ.

Доречно в структурі ЛДК збільшити параметри O_{tr} та O_{tr}^{V}, позаяк це дасть змогу зменшити тривалість прийняття рішень за даною моделлю логічного дерева та заощаджує процесорний час. Також необхідно максимізувати параметр I_{Main} (показник узагальнення моделі ЛДК), що дає можливість домогтися найбільш оптимальної структури ЛДК та забезпечити фактично максимальний стиск даних початкової навчальної вибірки (подати масив початкових даних мінімальним за структурною складністю логічним деревом) та параметр Q_{avg} (середня кількість наборів навчальної вибірки на остаточні значення ФР – лист ЛДК) [3], [4], [12], [22].

Важливим показником якості побудованої моделі у вигляді ЛДК з врахуванням зазначених вище параметрів є інтегральний показник якості моделі, а саме:

$$Q_{Main} = \frac{O_{tr}}{N_{tr} \cdot V_{tr} \cdot C_{tr}} \cdot Exp\left[-\frac{\left| En_{tr} \cdot Et_{tr} - \delta^2 \right|}{M \cdot M_{ts}} \right].$$

Цей інтегральний показник якості моделі ЛДК має сенс тільки у випадку, коли буде виконана умова, що $En_{tr} / M \le \delta$, інакше він буде дорівнювати нулю. Збільшення цього показника характеризує зростання якості моделі ЛДК і навпаки – зменшення свідчить про погіршення якості класифікації.

Обговорення результатів дослідження. Отже, поклавши в основу моделей розпізнавання метод дерева класифікації та принцип модульності, в Ужгородському національному університеті було розроблено програмний комплекс "Оріон ІІІ" для генерування автономних систем розпізнавання. Алгоритмічна бібліотека системи нараховує 11 алгоритмів розпізнавання, серед яких запропонована вище алгоритмічна реалізації побудови ЛДК.

Базовою задачею, для якої було проведене конструювання автономної системи, є розпізнавання на підставі геологічних даних (задача про розділення нафтоносних пластів). Для розпізнавання об'єктів використовувалися 12 основних елементарних ознак та 10 додаткових.

В навчальній вибірці наведена інформація про об'єкти двох класів. На етапі екзамену побудована система класифікації має забезпечити ефективне розпізнавання об'єктів невідомої класифікації відносно цих двох класів. Перед початком роботи навчальна вибірка була автоматично перевірена на коректність (пошук та видалення однакових об'єктів різної належності – помилки першого роду), хоча в системі й реалізована схема донавчання та виправлення помилок у дереві класифікації (алгоритм ДВП) – оскільки генерування проходило в автоматичному режимі, то даний алгоритм не використовувався.

Навчальна вибірка містила 1250 об'єктів (з них нафтоносні 756 об'єктів), причому ефективність сконструйованої системи розпізнавання оцінювалася на тестовій виборці обсягом 240 об'єктів. Дані навчальних і тестових вибірок отримані на підставі геологічної розвідки на території Закарпатської області в період з 2001 року по 2011 рік.

Фрагмент основних результатів, наведених вище експериментів, подано в (табл. 1). Причому побудовані моделі ЛДК забезпечили необхідний рівень точності, заданий умовою задачі, швидкодію та витрати робочої пам'яті системи. Запропоновані оцінки якості моделі ЛДК фіксують найважливіші характеристики логічних дерев, які можна застосувати як критерій оптимальності в процедурі побудови ЛДК та фінального відбору з множини моделей ЛДК.

Табл. 1. Порівняння методів синтезу дерев класифікації / Comparison of classification tree synthesis methods

Метод синтезу структури логічного дерева	Початкові дані				Інтегральний показник якості моделі Q_{Main}	Загальна кількість помилок моделі на ТВ, Et_{tr}
	Кількість:		Загальна потужність:			
	класів в НВ	ознак об'єктів НВ	НВ	ТВ		
Запропонований метод селекції елементарних ознак (метод ЛДК).	2	22	1250	240	0,00231412	0
Метод дерева класифікації на підставі автономних алгоритмі класифікації та розпізнавання (метод АДК).	2	22	1250	240	0,00183608	2

Запропонована схема побудови ЛДК порівнювалась з методом алгоритмічного дерева класифікації (АДК) та показала прийнятний результат. Головна ідея АДК полягає в апроксимації початкової навчальної вибірки набором алгоритмів. Отримана структура АДК характеризується високою універсальністю та відносно компактною структурою самої моделі, однак вимагає великих апаратних витрат для зберігання узагальнених ознак та початкової оцінки якості алгоритмів класифікації за навчальною вибіркою. Порівняно з нею, ЛДК має високу швидкодію правил класифікації, незначні апаратні витрати для зберігання та роботи самої структури дерева та високу якість класифікації.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – вперше розроблено метод побудови ЛДК на підставі селекції елементарних ознак з постійною оцінкою їх важливості на кожному кроці генерування дерева класифікації. Причому на кожному кроці розгалуження враховують вплив того чи іншого значення ознаки на остаточне значення ФР в структурі дерева.

Практична значущість результатів дослідження – запропонований метод побудови ЛДК був реалізований в бібліотеці алгоритмів універсальної програмної системі "ОРІОН ІІІ" для розв'язування різноманітних практичних задач класифікації (розпізнавання) масивів дискретних об'єктів.

Висновки / Conclusions

В роботі вирішена задача автоматизації побудови ЛДК на підставі апроксимації навчальної вибірки набором елементарних ознак. За результатами виконаної роботи можна зробити такі основні висновки.

1. Запропонований в роботі функціонал можна використовувати не тільки для оцінювання інформативності окремих елементарних ознак, але й для розрахунку важливості наборів ознак і їх сполучень, що в перспективі дає змогу досягти більш оптимальної структури синтезованого ЛДК за початковими даними навчальної вибірки.

2. Запропонований в роботі набір загальних показників, який дає змогу ефективно представити загальні характеристики моделі ЛДК, можна використовувати для відбору найбільш оптимального ЛДК з множини побудованих (наприклад у випадку алгоритмів побудови випадкових ЛДК).

3. Проведені практичні випробовування підтвердили працездатність запропонованих моделей ЛДК та розробленого програмного забезпечення, що дає змогу рекомендувати їх для використання під час розв'язання широкого спектру прикладних задач класифікації та розпізнавання дискретних об'єктів.

4. Перспективи подальших досліджень можна спрямувати в бік розвитку методів побудови алгоритмічних дерев класифікації, оптимізації програмних реалізацій запропонованого методу побудови ЛДК, а також його практичної апробації на множині реальних задач класифікації та розпізнавання.

References

[1] Bodyanskiy, Y., Vynokurova, O., Setlak, G. & Pliss, I. (2015). Hybrid neuro-neo-fuzzy system and its adaptive learning algorithm, *Xth Scien. and Tech. Conf. "Computer Sciences and Information Technologies"* (*CSIT*), Lviv, 111–114. https://doi.org/10.1109/STC-CSIT.2015.7325445

[2] Breiman, L. L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Boca Raton, Chapman and Hall/CRC, 368 p.

[3] De Mántaras, R. L. (1991). A distance-based attribute selection measure for decision tree induction. *Machine learning*, *6*(1), 81–92. https://doi.org/10.1023/A:1022694001379

[4] 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*), 293–300. https://doi.org/10.1007/978-3-642-21738-8_38

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2008). The Elements of Statistical Learning. Berlin, Springer, 768 p. https://doi.org/10.1007/978-0-387-84858-7

[6] Kamiński, B., Jakubczyk, M., & Szufel, P. (2017). A framework for sensitivity analysis of decision trees. *Central European Journal of Operations Research*, *26* (1), 135–159. https://doi.org/10.1007/s10100-017-0479-6

[7] Karimi, K., & Hamilton, H. J. (2011). Generation and Interpretation of Temporal Decision Rules. *International Journal of Computer Information Systems and Industrial Management Applications*, *3*, 314–323.

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

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

[10] Laver, V. O., & Povkhan, I. F. (2019). The algorithms for constructing a logical tree of classification in pattern recognition problems. *Scientific notes of the Tauride national University. Series: technical Sciences*, *30*(*69*)(4), 100–106. https://doi.org/10.32838/2663-5941/2019.4-1/18

[11] Miyakawa, M. (1989). Criteria for selecting a variable in the construction of efficient decision trees. *IEEE Transactions on Computers, 38*(1), 130–141. https://doi.org/10.1109/12.8736

[12] Painsky, A., & Rosset, S. (2017). Cross-validated variable selection in tree-based methods improves predictive performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2142–2153. https://doi.org/10.1109/TPAMI.2016.2636831

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

[14] Povhan, I. (2019). General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects . *Electronics and Information Technologies, 11*, 112–117. https://doi.org/10.30970/eli.11.7

[15] Povhan, I. F. (2019). The problem of general estimation of the complexity of the maximum constructed logical classification tree. *Bulletin of the national technical University Kharkiv Polytechnic Institute*, 13, 104−117. https://doi.org/10.20998/2411-0558.2019.13.10

[16] Povkhan, I. F. (2018). The problem of functional evaluation of a training sample in discrete object recognition problems. *Scientific notes of the Tauride national University. Series: technical Sciences*, *29*(*68*)(6), 217–222.

[17] Povkhan, I. F. (2019). Features of synthesis of generalized features in the construction of recognition systems using the logical tree method, Materials of the international scientific and practical conference "*Information technologies and computer modeling ITKM-2019*". Ivano-Frankivsk, 169–174.

[18] Povkhan, I. F. (2019). Features random logic of the classification trees in the pattern recognition problems. *Scientific notes of the Tauride national University. Series: technical Sciences*, *30*(*69*)(5), 152–161. https://doi.org/10.32838/2663-5941/2019.5-1/22

[19] Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, 1*, 81-106. https://doi.org/10.1007/BF00116251

[20] Srikant, R., & Agrawal, R. (1997). Mining generalized association rules. *Future Generation Computer Systems, 13*(2), 161-180. https://doi.org/10.1016/S0167-739X(97)00019-8

[21] Subbotin, S. (2013). The neuro-fuzzy network synthesis and simplification on precedents in problems of diagnosis and pattern recognition. *Optical Memory and Neural Networks* (*Information Optics*), *22*(2), 97–103. https://doi.org/10.3103/S1060992X13020082

[22] Subbotin, S. A. (2013). Methods of sampling based on exhaustive and evolutionary search. *Automatic Control and Computer Sciences, 47*(30), 113–121. https://doi.org/10.3103/S0146411613030073

[23] Subbotin, S. A. (2014). Methods and characteristics of localitypreserving transformations in the problems of computational intelligence. *Radio Electronics, Computer Science, Control*, (1), 120–128. https://doi.org/10.15588/1607-3274-2014-1-17
[24] Subbotin, S. A. (2019). Construction of decision trees for the case of low-information features. *Radio Electronics, Computer Science, Control*, (1), 121–130. https://doi.org/10.15588/1607-3274-2019-1-12
[25] Subbotin, S., & Oliinyk, A. (2017). The dimensionality reduction methods based on computational intelligence in problems of object classification and diagnosis, Recent Advances in Systems, Control and Information Technology, [eds.: R. Szewczyk, M. Kaliczyńska]. Cham, Springer, 11–19. (Advances in Intelligent Systems and Computing, *543*. https://doi.org/10.1007/978-3-319-48923-0_2
[26] Vasilenko, Y. A., Vashuk, F. G., & Povkhan, I. F. (2011). The problem of estimating the complexity of logical trees recognition and a general method for optimizing them. *Scientific and technical journal "European Journal of Enterprise Technologies"*, *6/4*(54), 24–28.
[27] Vasilenko, Y. A., Vashuk, F. G., & Povkhan, I. F. (2012). General estimation of minimization of tree logical structures. *European Journal of Enterprise Technologies*, *1/4*(55), 29–33.
[28] Vasilenko, Y. A., Vashuk, F. G., Povkhan, I. F., Kovach, M. Y., & Nikarovich, O. D. (2004). Minimizing logical tree structures in image recognition tasks. *European Journal of Enterprise Technologies, 3*(9), 12–16.
[29] Vasilenko, Y. A., Vasilenko, E. Y., & Povkhan, I. F. (2002). Defining the concept of a feature in pattern recognition theory. *Artificial Intelligence*, 4, 512–517.
[30] Vasilenko, Y. A., Vasilenko, E. Y., & Povkhan, I. F. (2003). Branched feature selection method in mathematical modeling of multi-level image recognition systems. *Artificial Intelligence*, 7, 246−249.
[31] Vasilenko, Y. A., Vasilenko, E. Y., & Povkhan, I. F. (2004). Conceptual basis of image recognition systems based on the branched feature selection method. *European Journal of Enterprise Technologies,* 7(1), 13–15.
[32] Vtogoff, P. E. (1989). Incremental Induction of Decision Trees. *Machine Learning*, (4), 161−186.

I. F. Povkhan

Uzhhorod National University, Uzhhorod, Ukraine

METHOD FOR SYNTHESIZING LOGICAL CLASSIFICATION TREES BASED ON THE SELECTION OF ELEMENTARY FEATURES

The general problem of constructing logical recognition and classification trees is considered. The object of this study is logical classification trees. The subject of the research is current methods and algorithms for constructing logical classification trees. The aim of the work is to create a simple and effective method for constructing recognition models based on classification trees for training samples of discrete information, which is characterized by elementary features in the structure of synthesized logical classification trees. A general method for constructing logical classification trees is proposed, which builds a tree structure for a given initial training sample, which consists of a set of elementary features evaluated at each step of building a model for this sample. A method for constructing a logical tree is proposed, the main idea of which is to approximate the initial sample of an arbitrary volume with a set of elementary features. When forming the current vertex of the logical tree, the node provides selection of the most informative, qualitative elementary features from the original set. This approach, when constructing the resulting classification tree, can significantly reduce the size and complexity of the tree, the total number of branches and tiers of the structure, and improve the quality of its subsequent analysis. The proposed method for constructing a logical classification tree makes it possible to build tree-like recognition models for a wide class of problems in the theory of artificial intelligence. The method developed and presented in this paper received a software implementation and was investigated when solving the problem of classifying geological data. The experiments carried out in this paper confirmed the operability of the proposed mathematical support and show the possibility of using it to solve a wide range of practical recognition and classification problems. Prospects for further research may consist in creating a limited method of the logical classification tree, which consists in maintaining a criterion for stopping the procedure for constructing a logical tree according to the depth of the structure, optimizing its software implementations, as well as experimental studies of this method for a wider range of practical tasks.

***Keywords**:* logical tree; feature selection; branching criterion; discrete object.

Інформація про автора:

Повхан Ігор Федорович, д-р техн. наук, доцент, кафедра програмного забезпечення. **Email:** povkhan.igor@uzhnu.edu.ua; https://orcid.org/0000-0002-1681-3466

Цитування за ДСТУ: Повхан І. Ф. Метод синтезу логічних дерев класифікації на підставі селекції елементарних ознак. *Український журнал інформаційних технологій.* 2022, т. 4, № 2. С. 25–32.

Citation APA: Povkhan, I. F. (2022). Method for synthesizing logical classification trees based on the selection of elementary features. *Ukrainian Journal of Information Technology, 4*(2), 25–32. https://doi.org/10.23939/ujit2022.02.025