



І. Ф. Повхан, А. В. Легеза

Ужгородський національний університет, м. Ужгород, Україна

ГРУПОВА СЕЛЕКЦІЯ ЕЛЕМЕНТАРНИХ ОЗНАК У СХЕМАХ ПОБУДОВИ ГІБРИДНИХ СТРУКТУР ДЕРЕВ РІШЕНЬ

Об'єктом дослідження є дерева класифікації. Предметом дослідження є методи, алгоритми та схеми побудови дерев класифікації. Мета роботи – побудова ефективного методу (схеми) синтезу моделей дерев класифікації на основі групової оцінки важливості дискретних ознак у межах розгалуженого вибору атрибутів. Запропоновано метод побудови дерев класифікації, який для заданої навчальної вибірки визначає індивідуальну інформативність (важливість) груп ознак (та їх комбінації) щодо вихідного значення функції класифікації (даних навчальної вибірки). Розроблений метод логічного дерева під час побудови чергового вузла дерева класифікації намагається виділити групу найтісніше взаємопов'язаних дискретних ознак. Це дає змогу знизити загальну структурну складність моделі (кількість рівнів дерева класифікації), прискорити обчислення під час розпізнавання об'єктів на основі моделі, а також підвищити узагальнювальні властивості моделі та її інтерпретабельність. Запропонована схема селекції груп дискретних ознак дає змогу використовувати побудоване дерево рішень для оцінювання інформативності (важливості) ознак. Розроблений метод дерева класифікації реалізований програмно і досліджений під час розв'язання задачі класифікації дискретних об'єктів, представлених набором ознак. Виконані експерименти підтвердили працездатність запропонованого математичного забезпечення і дають підстави рекомендувати його для використання на практиці під час вирішення прикладних завдань класифікації дискретних об'єктів на основі логічних дерев класифікації. Перспективи подальших досліджень можуть полягати у створенні модифікованого методу логічного дерева класифікації за допомогою ефективного перебору та оцінювання наборів елементарних ознак на основі запропонованого методу, оптимізації його програмних реалізацій та експериментальному дослідженні запропонованого методу на ширшому наборі прикладних задач.

Ключові слова: логічне дерево, алгоритмічне дерево, класифікатор, розпізнавання образів, ознака, навчальна вибірка, дерево рішень.

Вступ / Introduction

Комплекс задач класифікації дискретних об'єктів виникає у багатьох прикладних галузях як економічної, так і соціальної діяльності людини. Головною метою створення класичних систем розпізнавання (СР) є необхідність автоматизації групи процесів, які пов'язані зі сприйняттям, пошуком, виділенням та класифікацією образів на основі певних операцій над реальними даними навчальної вибірки (НВ), отриманими тим чи іншим способом [1]. Інформаційні технології, ґрунтовані на математичних моделях розпізнавання образів (а саме моделях дерев рішень різних типів), широко використовують у різноманітних системах збирання та оброблення дискретної інформації, тому що мають можливість усунути недоліки класичних методів нейромережевого розпізнавання та досягти принципово нових результатів, раціонально використовуючи потужності обчислювальних систем [2]. Центральним завданням у теорії розпізнавання образів є побудова на основі певних теоретичних та експериментальних досліджень ефективних обчислювальних

систем для зарахування об'єктів до відповідних класів [3]. Зрозуміло, що для такого класу прикладних задач доволі ефективним інструментом є моделі дерев класифікації, структури логічних дерев класифікації (ЛДК) на основі розгалуженого вибору ознак [14, 21], що являють собою побудовані послідовно моделі, які навчаються на основі масиву даних деякої фіксованої НВ.

Об'єкт дослідження – клас моделей класифікації у вигляді структур ЛДК усіх типів та структур.

Отже, клас задач, які зводяться до теорії розпізнавання образів (РО), надзвичайно різноманітний, і, головне, не існує універсального підходу до їх розв'язання, який забезпечив би порівняно високу швидкість та прийнятну якість (ефективність) під час розв'язання довільних задач. Причиною такого становища є центральна проблема теорії РО – проблема вибору (побудови) ознак (вершин у структурі ЛДК). Універсальне розв'язання цієї проблеми дасть змогу автоматично будувати універсальну СР, ефективну для довільних задач РО [6].

Запропонований у роботі метод логічного дерева класифікації (метод розгалуженого вибору ознак) являє собою перспективну основу для подальших досліджень у галузі математичного конструювання та оптимізації одержаних моделей класифікації у вигляді структур ЛДК із вершинами на основі узагальнених ознак. Саме ця особливість методів розгалуженого вибору ознак і надає їм найважливішу перевагу порівняно з іншими підходами. Гнучкість систем на основі розгалуженого вибору ознак (РВО) полягає у тому, що в ній закладена широка можливість для маніпуляцій із ознаками (елементарними та узагальненими ознаками), які вибирають на кожному кроці генерації результуючого правила класифікації. Здатність до регулювання якості розпізнавання у схемах РВО полягає у тому, що логічне дерево розпізнавання можна будувати як завгодно довго за деякою обмеженою НВ і тим самим добиватися певної міри точності її апроксимації (тобто якості розпізнавання) [7].

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

Мета роботи – створення простого та ефективного методу побудови моделей класифікації на основі дерев рішень, вершинами яких є окремі групи (набори) елементарних (дискретних) ознак.

Для досягнення зазначеної мети визначено такі основні завдання дослідження:

1. Дослідити загальні методи генерації структур логічних дерев класифікації різних типів.
2. Побудувати загальний алгоритм групової селекції наборів (сполучень) елементарних ознак.
3. Розробити загальний метод синтезу логічних дерев класифікації на основі групової оцінки інформативності наборів дискретних ознак.

Аналіз останніх досліджень та публікацій. У [14, 21] розроблено метод синтезу моделей класифікації на основі оцінювання інформативності окремих дискретних ознак (атрибутів). Центральними проблемами такої схеми генерації дерева класифікації є висока структурна складність побудованої моделі та проблема перенавчання на основі даних НВ. У роботах [22] подано вирішення обмежень структур ЛДК за рахунок модульної схеми побудови класифікаторів, яка дає змогу обійти обмеження традиційних методів логічних дерев рішень. Тоді виникає важливе завдання побудови ефективної схеми генерації узагальнених ознак для таких структур – моделей алгоритмічних дерев класифікації (АДК) [24]. Недоліком такого підходу є обмеження щодо обчислювальних ресурсів системи під час побудови моделей АДК та неуніверсальність у прикладному плані.

Принципових обмежень методів дерев рішень, пов'язаних із оцінюванням структурної складності побудованих моделей, та методів їх фінальної обрізки (мінімізації структур ЛДК) стосуються роботи [20, 25]. З робіт [11–14] відомо, що модель ЛДК, побудована довільним методом або алгоритмом РВО, має деревоподібну логічну структуру. Тоді постає центральне

питання теорії дерев рішень – проблема вибору якісного критерію розгалуження. Структура ЛДК складається із вершин, які групують за ярусами і які отримані на певному кроці побудови дерева розпізнавання [15]. Можливим способом модернізації методів дерев рішень (стандартних структур ЛДК) є напрям синтезу дерев розпізнавання, представлених фактично деревом алгоритмів [22, 23].

Важливий випадок, коли дискретні ознаки індивідуально малоінформативні [14]. В такій ситуації відомі методи дерев рішень генеруватимуть дуже глибокі та структурно складні дерева класифікації, які можуть не забезпечити необхідну точність та складність моделі, із низьким рівнем узагальнення даних. Це призведе до великих витрат апаратних ресурсів (пам'яті та процесорного часу системи) [26]. Тобто в багатьох задачах доступні для вимірювання ознаки індивідуально малоінформативні, проте в сукупності характеризуються достатньою груповою інформативністю для побудови деякої узагальненої ознаки (вершини логічного дерева класифікації) [23]. Саме тому і виникає питання побудови структур ЛДК для випадку малоінформативних дискретних ознак. Важливі питання генерації та декомпозиції правил класифікації у структурах дерев рішень розглянуто в роботі [25]. Завдання ефективного оцінювання інформативності ознак [16] під час побудови вершин дерев класифікації залишається центральним. Можливим способом вдосконалення цієї роботи може бути використання сполучень та наборів ознак з метою генерування інформативних вершин структур ЛДК. Цього питання і стосується робота.

Результати дослідження та їх обговорення / Research results and their discussion

Постановка задачі. Нехай $H_1, H_2, \dots, H_k$ – система класів (образів), задана на множині G , що складається з об'єктів x_i , ($i = 1, \dots, m$). Характер розподілу множини G на відповідні класи задають за допомогою навчальної вибірки (НВ) такого вигляду:

$$\left((x_1, f_R(x_1)), (x_2, f_R(x_2)), \dots, (x_m, f_R(x_m)) \right). \quad (1)$$

Зауважимо, що тут $x_h \in G$, $f_R(x) \in \{0, 1, \dots, k-1\}$, ($h = 1, 2, \dots, m$), k – кількість класів НВ, m – загальна кількість об'єктів НВ, а $f_R(x)$ – деяка скінченнозначна функція, що задає розподіл множини G на відповідні образи (базові класи). Співвідношення $f_R(x_h) = l$, ($l = 0, 1, \dots, k-1$) означає, що $x_h \in H_l$. Зауважимо, що кожній НВ вигляду (1) можна поставити у відповідність (за допомогою деякого алгоритму або методу представлення) цілком визначене ЛДК, яке об'єктам x_h , ($h = 1, \dots, m$) НВ (1) у відповіднє значення функції $f_R(x_h)$, що задає розподіл R на множині G . Отже, постає завдання побудови конструкції дерева класифікації (структур логічних / алгоритмічних дерев класифікації), структура якого була би оптимальною $f_R(x_j) \rightarrow opt$ щодо початкових даних НВ [23].

Метод розгалуженого вибору ознак. Загальна схема методу РВО. В основу всіх алгоритмів, побудованих методом розгалуженого вибору дискретних ознак, покладено таку схему: спочатку вибирають деяку елементарну ознаку φ . Необхідно, щоб її важливість щодо вибірки, яка навчає, порівняно з іншими ознаками була максимальною. Під важливістю ознаки розумітимемо таку величину:

$$W(\varphi) = \sum_{i=1}^m (b_i/h) \times \rho_i, \text{ де } \rho_i = \max_{1 \leq m \leq k} \left(\frac{q_i^m}{b_i} \right). \quad (2)$$

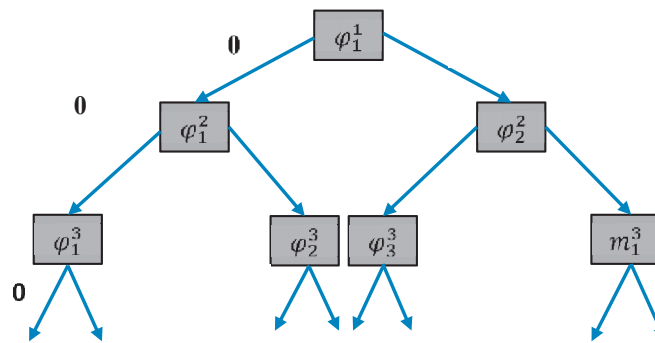
Зрозуміло, що аналогічно можна оцінити важливість інших ознак. Величину (q_i^m/b_i) можна інтерпретувати як імовірність того, що функція $f(x)$ набуде значення O_m , за умови, що значення ознаки φ дорівнює i . Величина ρ_i являє собою максимальну із цих ймовірностей. Можна сказати, що величина ρ_i являє собою інформацію, яку можна отримати про значення функції $f(x)$, знаючи, що на наборі d значення ознаки φ дорівнює i . Величина $W(\varphi)$, яка визначається цією формулою, характеризує ту інформацію, яку можна отримати про функцію $f(x)$, якщо відомо значення ознаки φ на наборі d . Зрозуміло, що ознака, для якої ця інформація найбільша, вважається найважливішою ознакою щодо $f(x)$.

Формулу (2) можна використовувати як критерій розгалуження у методах РВО [21]. Отже, всі кроки схеми розгалуженого вибору ознак можна зобразити за допомогою дерева (див. рисунок).

Зазначимо, що в кожній вершині цього дерева (структури ЛДК) або деяка ознака φ_i^j , або число m_i^j (це число характеризує належність до того чи іншого образу

$H_{m_i^j}$). Вершина із числом m_i^j (значення функції розпізнавання), називається кінцевою вершиною дерева. Від кожної вершини, в якій стоїть ознака φ_i^j , відходять дві стрілки, позначені 0 та 1. Стрілка, позначена 0, відповідає значенню $\varphi_i^j = 0$, а позначена 1 – значенню $\varphi_i^j = 1$ (ознака набуває лише два значення, в іншому випадку кількість стрілок збільшиться). Дерево складається з ярусів. В j -му ярусі містяться ознаки $\varphi_1^j, \varphi_2^j, \dots$. Всі ознаки, вказані в усіх ярусах дерева, починаючи від першого та закінчуючи n -м, являють собою ознаки, отримані після виконання n кроків схеми РВО. Причому ознаки, які містяться у n -му ярусі, являють собою всі ті ознаки, котрі отримані на n -му кроці цього процесу. В цьому прикладі наведено лише три кроки процесу РВО, і $\varphi_1^1, \varphi_2^2, \varphi_3^3, \varphi_1^3, \varphi_2^3, \varphi_3^3$ – всі ознаки, одержані в результаті цих кроків (етапів побудови структури ЛДК). Математичне обґрунтування цієї схеми ЛДК можна знайти у роботі [21].

Схема методу РВО на основі групової селекції елементарних ознак. Аналогічно до наведеної вище схеми РВО можна будувати структуру ЛДК на основі селекції груп елементарних ознак. Для цього введемо оцінку інформативності довільної групи дискретних ознак $\varphi_{i_1}, \varphi_{i_2}, \dots, \varphi_{i_s}$ ($1 \leq i_1, i_2, \dots, i_s \leq h$). Для прикладу зафіксуємо деяку групу елементарних ознак $\varphi_1, \varphi_2, \dots, \varphi_y$ ($2 \leq y \leq h$). Тут так само G – множина елементарних ознак $\varphi_1, \dots, \varphi_h$, на якій задана ФР $f_R(x)$ ($x = \varphi_1, \dots, \varphi_h$). На множину G накладається виконання гіпотези адекватності (G наближається до області визначення $f_R(x)$).



Загальна схема дерева класифікації на основі РВО / General scheme of the classification tree based on the Branched Feature Selection (BFS) method

Зафіксуємо $\Delta = t_1, t_2, \dots, t_y$ як довільний набір значень елементарних ознак. Тоді через U_Δ позначимо кількість всіх наборів $\varphi_1, \dots, \varphi_h$ з G , для яких виконується співвідношення $\varphi_j = t_j$. Тоді через U_Δ^m позначимо кількість наборів $\varphi_1, \dots, \varphi_h$ з G , для яких виконується співвідношення $\varphi_j = t_j$ та $f_R(\varphi_1, \dots, \varphi_h) = O_m$ (де h – потужність множини G), а C – множина всіх наборів ознак $\varphi_1, \varphi_2, \dots, \varphi_y$. Тоді інформативність $W(\varphi_1, \varphi_2, \dots, \varphi_y)$ деякої групи (набору) елементарних ознак визначається за формулою (3).

$$W(\varphi_1, \varphi_2, \dots, \varphi_y) = \sum_{\Delta \in C} (U_\Delta/h) * \theta_\Delta. \quad (3)$$

Зауважимо, що тут $\theta_\Delta = \max_m U_\Delta^m / U_\Delta$. Для обґрунтування (3) можна зафіксувати y ($1 \leq y \leq h$) і серед всіх груп $\varphi_{i_1}, \varphi_{i_2}, \dots, \varphi_{i_y}$ знайти найінформативнішу групу елементарних ознак. Зрозуміло, що за великих y та h обчислення найінформативнішої групи ознак $\varphi_{i_1}, \varphi_{i_2}, \dots, \varphi_{i_y}$ пов'язано з великим обсягом обчислень, а саме переборів, за формулою (4).

$$P_h^y = \frac{h!}{y!(h-y)!} \quad (4)$$

Звернемо увагу, що формула (3) являє собою критерій розгалуження у структурі ЛДК, вершинами якої є деякі узагальнені ознаки (набори або групи дискретних ознак). Ідею узагальненої ознаки можна взяти з роботи [23]. Підкреслимо, що така структура ЛДК з вершинами у вигляді узагальнених ознак великою мірою наближається до структур АДК із робіт [24]. Оцінку інформативності (3) групи (набору) дискретних ознак називатимемо функціоналом важливості групи ознак (ФВГО).

Зауважимо, що (3) забезпечує оцінку інформативності групи дискретних ознак відносно $f_R(\varphi_1, \dots, \varphi_h)$ та має однозначний зв'язок з помилками класифікації об'єктів НВ. Тобто числове значення ФВГО збігається з тією частиною об'єктів НВ, яку можна правильно зарахувати до класів НВ (звичайно, використовуючи тільки відібрану групу елементарних ознак для класифікації дискретних об'єктів). Зауважимо також, що на основі (3) можна виконати ефективну процедуру ранжування підмножин ознак, що дає змогу будувати набори узагальнених ознак (вершин структур ЛДК нового типу).

Важливим моментом ФВГО є те, що довільна підмножина $\varphi_{i_1}, \varphi_{i_2}, \dots, \varphi_{i_s} (1 \leq s \leq h)$ множини дискретних ознак $\varphi_1, \dots, \varphi_h$ утворює деякий тест початкової НВ тоді й тільки тоді, коли:

$$W(\varphi_{i_1}, \varphi_{i_2}, \dots, \varphi_{i_s}) = 1. \quad (5)$$

Тут під тестом розумітимемо деякий набір елементарних ознак $\varphi_{i_1}, \varphi_{i_2}, \dots, \varphi_{i_s}$, який дає змогу однозначно розділити набір класів початкової НВ.

Зазначимо також, що для (3) буде виконуватись таке співвідношення:

$$\frac{1}{k} \leq \text{ФВГО} \leq 1. \quad (6)$$

Зауважимо, що в (6) k – загальна кількість класів початкової НВ.

Схема швидкого обчислення важливості груп дискретних ознак в методі ЛДК. Отже для обчислення ФВГО не потрібно зберігати в пам'яті комп'ютера масив дискретних ознак (запам'ятовувати всі набори множини НВ). Значення $W(\varphi_i)$ можна обчислювати послідовно, подаючи навчальні пари $(x_1, f_R(x_1))$. Тобто можна запропонувати такий алгоритм для обчислення $W(\varphi_i)$.

Нехай на деякому кроці обчислення ФВГО вже подано z пар НВ вигляду $(x_1, f_R(x_1))$ і для цих навчальних пар обчислено значення b_{ij} та q_{ji}^m (тут $j = 0, 1, \dots, k_i - 1$; $i = 1, 2, \dots, h$; $m = 0, 1, \dots, k - 1$).

Величина b_{ij} – кількість усіх наборів з навчальних пар НВ, у яких ознака φ_i набула значення j .

Величина q_{ji}^m – кількість усіх наборів з навчальних пар НВ, у яких ознака φ_i набула значення j , а функція $f_R(x)$ – значення O_m .

Тоді під час отримання наступної $(h + 1)$ пари $(l_1, l_2, \dots, l_n, O_v)$, де $0 \leq l_i \leq k_i - 1$ ($0 \leq v \leq k - 1$), значення z , b_{ij} , q_{ji}^m змінюються так:

1. Значення z , збільшується $(z + 1)$.

2. Значення $b_{ji}^{\cdot} = b_{ij}$, якщо $j \neq l_i$; $b_{ji}^{\cdot} = b_{ij} + 1$, якщо $j = l_i$.

3. Значення $(q_{ji}^m)^{\cdot} = q_{ji}^m$, якщо $j \neq l_i$ або $m \neq v$; $(q_{ji}^m)^{\cdot} = q_{ji}^m + 1$, якщо $j = l_i$ та $m = v$.

Отже, за такої схеми обчислення важливості дискретних ознак у пам'яті комп'ютера ми зберігаємо лише величини v, b_{ij}, q_{ji}^m . Зрозуміло що за таким алгоритмом обчислення цих величин необхідно вести доти, доки значення $W(\varphi_i)$ не почнуть стабілізуватись біля деяких величин l_i . Тоді набір величин l_i можна прийняти за оцінку важливості відповідних елементарних ознак φ_i . Отже, за такою схемою можна обчислювати набори ФВГО. Застосування таких алгоритмів особливо ефективне, коли є деякі обмеження на обсяги оперативної пам'яті.

Звернемо увагу, що функціонал (3) дає змогу порівняно просто (у плані економії обчислювальних ресурсів системи) знаходити набори тестів (нам цікаві насамперед мінімальні тести), а також важливості наборів та груп дискретних ознак в тому випадку, коли початкова НВ складається більш ніж з двох класів. Мінімальні тести принципово важливі у плані побудови оптимальної структури ЛДК (як звичайних, так і модифікованих на основі вершин – узагальнених ознак), які не потребують фінального обрізання, мінімізації структури.

Іноді у прикладних задачах дуже важко (або взагалі неможливо) знайти тест. В такому випадку можливим виходом є пошук такої сукупності дискретних ознак, яка би дала змогу ефективно відмежувати об'єкти одного класу від об'єктів іншого. Таку групу ознак називають квазітестом, а формула (3) дає змогу знаходити квазітести безпосередньо на основі заданої НВ.

Зазначимо, що на основі (3) можна організувати ефективний механізм знаходження наборів елементарних ознак, які забезпечують розпізнавання початкової НВ із заданою наперед точністю. Тобто для структур ЛДК це дає можливість будувати дерева рішень фіксованої складності (структурної складності, структурної глибини) та фіксованої точності розпізнавання. З іншого боку, можна обчислити найінформативнішу групу дискретних ознак (відносно функції, яка задає розподіл початкової НВ) за зафіксованої кількості ознак у групі.

Важливою властивістю функціонала (3) є те, що обчислення важливості дискретних ознак (груп або наборів ознак) та знаходження тестів відбуваються паралельно. Це надає можливість виділяти групу ознак, для якої $W(\varphi_1, \varphi_2, \dots, \varphi_k) < 1$ ($k \geq 2$), як деякий квазітест, що розпізнає об'єкти НВ із деякою точністю W . У різноманітних задачах пошуку тестів підзадачі визначення інформативності елементарних ознак розв'язують послідовно, що збільшує ресурсні витрати та витрати процесорного часу інформаційної системи.

Зазначимо, що за аналогією зі структурами ЛДК функціонал (3) можна використовувати для вибору атрибутів у вершинах граф-схемних структур булевих та багатозначних логічних функцій задачі їх мінімізації.

У задачі оцінки повноти набору елементарних ознак $\varphi_1, \varphi_2, \dots, \varphi_k$ важливе значення має величина $U(\varphi_1, \varphi_2, \dots, \varphi_k) = \min_{\Delta \in C} \theta_{\Delta}$, де сенс θ_{Δ} та Δ той самий, що і в (3). Зауважимо, що ця оцінка жорсткіша, ніж $W(\varphi_1, \varphi_2, \dots, \varphi_k)$, тому її варто застосовувати у задачах із жорсткішими вимогами щодо точності класифікації об'єктів $x_i, (i = 1, \dots, m)$, коли потрібно мінімізувати кількість помилок класифікації (відмов класифікації усіх типів). Аналогічно до функціонала (3) можна ввести функціонал $W'(\varphi_i) = W(\varphi_1, \dots, \varphi_{i-1}, \varphi_{i+1}, \dots, \varphi_k)$. В такому випадку найважливішою дискретною ознакою із обмеження f_R вважається ознака φ , для якої значення $W'(\varphi_i)$ буде мінімальним.

Експериментальна частина. На основі фіксованої НВ, використовуючи розгалужений вибір ознак, можна побудувати структури ЛДК різних типів з набором різних обмежень щодо структурної складності (кількості рівнів структури) та точності фінальної моделі класифікації. Для гібридних структур ЛДК на основі узагальнених ознак як вершин (груп та наборів дискретних ознак) критерієм розгалуження слугує функціонал (3). Зауважимо, що під час побудови таких структур ЛДК витрачається значно більше апаратних ресурсів системи класифікації порівняно з класичними моделями ЛДК [14, 21]. У такому випадку виникає питання оцінювання побудованих моделей класифікації (дерев рішень) з метою вибору найкращої стосовно поточної НВ.

Після процедури фінального обрізання (мінімізації) моделі класифікації розпочинається етап аналізу та порівняння синтезованих деревоподібних моделей розпізнавання (структур ЛДК різних типів та різної структурної складності). На цьому етапі необхідно визначити набори локальних показників (оцінних характеристик), які найінформативніше характеризують базові властивості отриманих моделей класифікації. В абсолютній більшості випадків етап оцінювання побудованих моделей ЛДК реалізується на основі загального інтегрального показника якості – центрального критерію порівняння моделей (структур дерев рішень). Взавши за основу роботи [14], можна виділити такі базові характеристики (показники) синтезованих структур дерев рішень:

V_{TR} – кількість вершин усіх типів у структурі дерева рішень (ЛДК). Логічно припустити, що в такому випадку мінімальне значення цієї величини буде на один більше від потужності початкової НВ.

N_{TR} – кількість елементарних ознак φ_i , на основі яких будують дерево рішень.

C_{TR} – кількість переходів, зв'язків між рівнями в структурі дерева рішень.

O_{TR} – кількість кінцевих вершин (значень функції розпізнавання) у структурі дерева рішень. Також до цього показника в гібридних структурах ЛДК зараховують незакінчені переходи в структурі дерева рішень.

En_{TR} – узагальнена кількість помилок усіх типів моделі класифікації на масиві даних НВ.

Et_{TR} – узагальнена кількість помилок усіх типів моделі класифікації на масиві даних ТВ.

Er_H – кількість помилок класифікації на кожному з класів початкової НВ.

Найважливішими в цьому переліку є показники помилок класифікації всіх типів, які безпосередньо впливають на інтегральну оцінку якості побудованої моделі дерева рішень. Введемо набір показників моделі дерева класифікації відносно його базових характеристик:

$C_{AVG} = (C_{TR}/V_{TR})$ – середня кількість зв'язків (переходів із інших рівнів) на одну вершину в структурі дерева рішень.

$N_{TR}^V = (N_{TR}/n)$ – загальна частка дискретних ознак, використаних у структурі дерева рішень (де n – кількість елементарних ознак φ).

$O_{TR}^V = (O_{TR}/V_{TR})$ – загальна частка результуючих значень функції розпізнавання (листів дерева) в структурі дерева рішень.

$Q_{AVG} = (m/O_{TR})$ – середня кількість наборів початкової навчальної вибірки на результуюче значення функції розпізнавання (листів дерева) в структурі дерева рішень.

На наступному етапі можна ввести загальний показник моделі логічного дерева узагальнення даних початкової навчальної вибірки в такій формі:

$$I_{MAIN} = (m \times n)/(V_{TR} + 2C_{TR}). \quad (7)$$

Найважливішою залишається проблема зменшення структурної складності побудованої моделі дерева класифікації (процедура фінальної обрізки ЛДК). Тобто виділяються показники N_{TR} – кількість ознак у структурі дерева класифікації, V_{TR} – кількість вершин моделі дерева рішень та C_{TR} – загальна кількість переходів у структурі ЛДК, параметри витрат пам'яті μ та процесорного часу τ . Саме мінімізація цих параметрів і дає змогу отримати хороші результати в плані швидкодії класифікації моделі, структурної складності фінальної моделі ЛДК та ресурсних витрат інформаційної системи.

Зважаючи на всі наведені параметри моделі дерева рішень, найбільшу економію ресурсів системи можна отримати у разі збільшенні показників O_{TR} та O_{TR}^V . Такий підхід дасть змогу істотно зменшити час загальної класифікації за цією моделлю логічного дерева (загальний час прийняття рішення за правилом класифікації), заощадити процесорний час.

Зважаючи на характеристики параметра I_{MAIN} (показник узагальнення моделі ЛДК), зрозуміло, що його також необхідно максимізувати, щоб добитися найоптимальнішої (обрізаної) структури дерева рішень. Максимізація параметра I_{MAIN} має забезпечити також максимальне стискання даних початкової навчальної вибірки. Тобто це уможливує подання початкового масиву даних мінімальним за структурною складністю деревом класифікації (структурою ЛДК довільного типу). Зрозуміло, що

показник I_{MAIN} варто розглядати в комплексі з параметром Q_{AVG} (середня кількість наборів НВ на результуючі значення функції розпізнавання), що дає змогу оцінити ступінь узагальнення НВ структурою ЛДК.

На наступному етапі зафіксуємо, що центральним показником якості побудованої моделі дерева класифікації (структури ЛДК) з урахуванням зазначених вище параметрів є інтегральний показник якості моделі в такій формі:

$$Q_{MAIN} = (O_{TR} \times e^{\lambda}) / (N_{TR} \times V_{TR} \times C_{TR}). \quad (8)$$

Зауважимо, що тут:

$$\lambda = - \frac{|En_{TR} \times Et_{TR} - \delta^2|}{M \times M_{TS}}. \quad (9)$$

З наведених формул зрозуміло, що такий інтегральний показник якості моделі дерева рішень (8) має сенс лише у випадку, коли $(En_{TR}/M) \leq \delta$. Якщо $(En_{TR}/M) > \delta$, він дорівнюватиме нулю. Також під M_{TS} розумітимемо загальну потужність тестової вибірки. Отже, можна зробити висновок, що збільшення цього показника характеризує поліпшення якості моделі дерева рішень, і навпаки, зменшення свідчить про погіршення якості класифікації (розпізнавання).

Обговорення результатів дослідження. Перевірку запропонованих методів (схем побудови структур ЛДК

на основі наборів та груп елементарних ознак як вершин) здійснено на базі програмного комплексу “Оріон III” (системи генерації автономних систем розпізнавання та класифікації). Програмний комплекс “Оріон III” розроблено в Ужгородському національному університеті, його бібліотека автономних класифікаторів нараховує 15 алгоритмів розпізнавання. Серед них є як прості геометричні алгоритми розпізнавання, так і запропонована вище алгоритмічна реалізація побудови гібридного ЛДК (на основі узагальнених ознак).

Метод побудови гібридної структури ЛДК перевірено на задачі класифікації геологічних даних – задачі про виділення нафтоносних пластів у геологічних дослідженнях. У задачі класифікації об’єктів використано 13 основних елементарних ознак та 11 додаткових (розмірність ознакового простору 24). Структура НВ характеризується об’єктами двох класів. Загальний об’єм навчальної вибірки – 1395 об’єктів (з них 809 нафтоносних об’єктів), тестової вибірки – 462 об’єкти. Зауважимо, що дані навчальних та тестових вибірок отримано на основі геологічної розвідки на території Закарпатської області в 2001–2021 рр.

Загальні параметричні характеристики задачі класифікації геологічних об’єктів наведено в табл. 1.

Табл. 1. Загальні параметри задачі класифікації / General parameters of the classification task

Набір класів H_i	Розмірність ознакового простору N	Потужність НВ – M	Загальні кількість класів НВ – l	Відношення класів тестової вибірки (H_i/TS)	Відношення класів НВ – (H_i/M)
Нафтоносні пласти (H_1)	(13/11)	1395	2	284/462	809 / 1395
Водоносні пласти (H_2)	(13/11)	1395	2	178/462	586 / 1395

Табл. 2. Порівняльна таблиця побудованих моделей АДК/ЛДК / Comparative table of the constructed ACT/LCT models

Модель класифікації на основі дерева рішень (структура ЛДК/АДК)	Інтегральний показник якості (ЛДК/АДК) Q_{Main}	Показник структурної складності моделі (ЛДК/АДК) S_{Main}	Показник помилок / відмов класифікації моделі (ЛДК/АДК) Er_{All}	Час побудови (моделі) дерева рішень T_{All}
Метод ЛДК на основі однорангового розгалуженого вибору ознак	0,002219	140	11	44 (s.)
Метод ЛДК на основі обмежень ($Z = 5$)	0,003016	99	15	39 (s.)
Algorithmic tree method (type I)	0,005286	51	11	64 (s.)
Algorithmic tree method (type II)	0,003032	63	10	81 (s.)
Метод АДК на основі обмежень ($Z = 4$)	0,002661	57	15	62 (s.)
Метод АДК на основі на основі гіперкуль	0,007219	30	7	51 (s.)
Метод ЛДК на основі селекції наборів (груп) елементарних ознак ($S = 2$)	0,002356	125	9	49 (s.)
Метод ЛДК на основі селекції наборів (груп) елементарних ознак ($S = 3$)	0,002531	112	10	54 (s.)
Метод ЛДК на основі селекції наборів (груп) елементарних ознак ($S = 4$)	0,003014	114	11	58 (s.)

Основні результати наведених вище експериментів (побудованих різнотипних моделей дерев рішень – логічних та алгоритмічних дерев класифікації) наведено відповідно в табл. 2. Зауважимо, що моделі ЛДК, побудовані в межах цієї роботи, забезпечили необхідний рівень точності, заданий умовою задачі (рівень точності класифікації можна ефективно регулювати), швидкодію прийняття рішень та витрати робочої пам'яті інформаційної системи.

Заважаючи на зазначене вище, можна зробити висновки, що запропоновані оцінки якості моделі (інтегральний показник якості структури ЛДК) логічного дерева рішень відображають найважливіші характеристики (параметри) дерева рішень. Отже, запропоновані показники якості моделі ЛДК можна застосовувати як критерій оптимальності в задачах синтезу моделей класифікації на основі логічних дерев рішень (задачах обрізання структур ЛДК, задачах побудови та відбору наборів ЛДК).

Побудову моделей дерев рішень здійснено на апаратній конфігурації Intel i7-13800h. У табл. 2 подана загальна інформація щодо типу моделі дерева класифікації, показника його структурної складності S_{Main} (кількість вершин, атрибутів та структурних переходів), кількості помилок (відмов) класифікації всіх типів Er_{All} , часу генерації класифікаційних моделей, показника загальної інтегральної якості моделі дерева рішень (структури ЛДК/АДК).

Під час експериментів гібридну схему ЛДК на основі наборів (груп) елементарних ознак як вершин (узагальнених ознак) порівнювали з методами ЛДК (схемами розгалуженого вибору ознак як із одинарною, так і з покроковою оцінкою інформативності), а також деревами алгоритмів (структурами АДК двох типів). Вона показала прийнятний результат у межах цієї задачі.

Зауважимо, що в основу методів дерев алгоритмів покладено схему апроксимації набором алгоритмів початкової навчальної вибірки. Для методів ЛДК схема ґрунтується на апроксимації набором елементарних ознак (групою дискретних ознак для гібридних структур ЛДК) початкової НВ.

Як бачимо, порівняно зі структурами ЛДК дерева алгоритмів мають істотні переваги в плані універсальності та порівняно компактну структуру самої моделі. Зрозуміло, що такі переваги моделей класифікації потребують значно більших ресурсних витрат для зберігання узагальнених ознак (наборів класифікаторів) та покрокового оцінювання ефективності алгоритмів класифікації. Однак структури ЛДК усіх типів мають високу швидкодію правил класифікації, зменшені ресурсні витрати інформаційної системи для зберігання структур ЛДК / АДК, точність класифікації, яку можна регулювати.

Наукова новизна отриманих результатів дослідження – вперше розроблено метод побудови гібридних структур ЛДК на основі групової селекції (наборів) елементарних ознак (вершин дерева у вигляді узагальнених ознак).

Практична значущість результатів дослідження – запропонований метод побудови гібридних структур ЛДК (моделей дерев класифікації) дає змогу будувати ефективні класифікатори із заданою наперед точністю. Метод гібридних структур ЛДК був інтегрований в бібліотеку алгоритмів програмної системи “Оріон”, що дає змогу розв'язувати різні прикладні задачі класифікації і демонструє високий ступінь універсальності в різних галузях застосування.

Висновок / Conclusions

Запропонований у роботі метод побудови гібридних структур ЛДК (на основі селекції груп елементарних ознак) та побудовані моделі дерев рішень показали необхідний рівень точності класифікації порівняно з іншими методами дерев класифікації. В основу таких гібридних структур (моделей ЛДК) покладено схему ефективної оцінки інформативності наборів (груп) дискретних ознак. Принциповим недоліком такого підходу залишається необхідність перебору наборів ознак з метою виділення найінформативніших груп (наборів) дискретних ознак. Зрозуміло, що прямий перебір НВ щодо елементарних ознак супроводжується великими ресурсними витратами інформаційної системи і може бути доцільним лише за наявності обмежень на довжину набору елементарних ознак (S).

Серед усіх побудованих моделей дерев рішень потрібно відзначити високу ефективність моделей дерев алгоритмів у плані точності класифікації та компактності структури (S_{Main}). У таких моделях класифікації найефективнішим виявився простий геометричний алгоритм гіперкуль (найтефективніший класифікатор для структур АДК усіх типів). Зауважимо, що побудовані моделі АДК продемонстрували порівняно невелику кількість помилок класифікації всіх типів під час роботи (Er_{All}). Так, дерево алгоритмів першого типу на основі гіперкуль (модель на основі обмежень на вибір класифікаторів), показало хороший результат ($Q_{Main} = 0,007219$). Структура ЛДК на основі розгалуженого вибору ознак (повне дерево класифікації) дала результат ($Q_{Main} = 0,002219$). На відміну від цього, дерево алгоритмів другого типу продемонструвало прийнятну точність та якість ($Q_{Main} = 0,003032$) з одного боку, але, з іншого, зрозумілі великі ресурсні та часові витрати системи ($T_{All} = 81$ (с.)). Перевага структур АДК пояснюється складнішою конструкцією моделі (наборів узагальнених ознак – незалежних класифікаторів), що потребує тривалішого часу генерації (побудови моделі).

Зауважимо, що зазвичай моделі АДК можуть ефективно працювати в задачах структур ЛДК, але інколи економічно доцільне використання саме логічних дерев класифікації (обмеження на складність, обмеження на час генерації моделі, обмеження на обчислювальні ресурси). Помітним недоліком побудованих моделей дерев алгоритмів є доволі високі витрати часу на етапі синтезу моделей дерева класифікації, особливо порів-

няно зі структурами ЛДК. Різниця в часі під час побудови моделей АДК (лише моделі з повною структурою і без обмежень) порівняно з деревами ЛДК була не менше ніж 30%.

Усі побудовані моделі дерев класифікації, а також розроблене програмне забезпечення продемонстрували ефективність у прикладних задачах, що дає змогу рекомендувати розширити сферу застосування гібридних структур ЛДК (дерев рішень на основі групової селекції елементарних ознак).

Майбутнім напрямом досліджень може стати подальший розвиток методів (структур) ЛДК/АДК, урахування впровадження нових типів і схем дерев класифікації. Крім того, оптимізація програмних реалізацій (розпаралелювання алгоритму обчислення інформативності дискретних ознак) запропонованого гібридного методу ЛДК і його практична апробація у різних завданнях точної класифікації та розпізнавання могли б забезпечити цінну інформацію і удосконалення в цій сфері.

References

- Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning*. Springer.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Chapman and Hall/CRC.
- Kintonova, A., Mussaif, M., & Gabdreshov, G. (2023). Improvement of iris recognition technology for biometric identification of a person. *Eastern-European Journal of Enterprise Technologies*, 6(2(120)), 60–69. <https://doi.org/10.15587/1729-4061.2022.269948>
- Bodyanskiy, Y. V., Shafronenko, A. Y., & Pliss, I. P. (2021). Credibilistic fuzzy clustering based on evolutionary method of crazy cats. *System Research and Information Technologies*, 3, 110–119. <https://doi.org/10.20535/SRIT.2308-8893.2021.3.09>
- Miyakawa, M. (1989). Criteria for selecting a variable in the construction of efficient decision trees. *IEEE Transactions on Computers*, 38(1), 130–141.
- 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. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2008)* (pp. 3044–3049). IEEE. <https://doi.org/10.1109/IJCNN.2008.4634228>
- Jaman, S. F., & Anshari, M. (2019). Facebook as marketing tools for organizations: Knowledge management analysis. In S. F. Jaman & M. Anshari (Eds.), *Dynamic perspectives on globalization and sustainable business in Asia* (pp. 92–105). IGI Global. <https://doi.org/10.4018/978-1-5225-7095-0.ch007>
- Strilets, V. E., Shmatkov, S. I., & Ugryumov, M. L. (2020). *Methods of machine learning in the problems of system analysis and decision making*. Karazin Kharkiv National University.
- De Mántaras, R. L. (1991). A distance-based attribute selection measure for decision tree induction. *Machine Learning*, 6(1), 81–92.
- Karimi, K., & Hamilton, H. (2011). Generation and interpretation of temporal decision rules. *International Journal of Computer Information Systems and Industrial Management Applications*, 3, 314–323.
- 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.
- Deng, H., Runger, G., & Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. In *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN 2011)* (Vol. 2, pp. 293–300). Springer-Verlag. https://doi.org/10.1007/978-3-642-21738-8_38
- 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>
- Shyshatskyi, A. (2020). Complex methods of processing different data in intellectual systems for decision support system. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4), 5583–5590. <https://doi.org/10.30534/ijatsec/2020/206942020>
- 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>
- Imamovic, D., Babovic, E., & Bijedic, N. (2020). Prediction of mortality in patients with cardiovascular disease using data mining methods. In *Proceedings of the 19th International Symposium INFOTEH-JAHORINA (INFOTEH 2020)* (pp. 1–4). IEEE. <https://doi.org/10.1109/INFOTEH48170.2020.9066297>
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31, 249–268.
- Zhuravlev, Y. I., & Nikiforov, V. V. (1971). Recognition algorithms based on the calculation of estimates. *Cybernetics*, 3, 1–11.
- Povkhan, I., Mulesa, O., Melnyk, O., Bilak, Y., & Polishchuk, V. (2022). The problem of convergence of classifiers construction procedure in the schemes of logical and algorithmic classification trees. In *Proceedings of the Second International Workshop on Computer Modeling and Intelligent Systems (CMIS-2022)* (Vol. 3137, pp. 1–13). CEUR Workshop Proceedings.
- Povkhan, I. (2020). A constrained method of constructing the logic classification trees on the basis of elementary attribute selection. In *Proceedings of the Second International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020)* (Vol. 2608, pp. 843–857). CEUR Workshop Proceedings.
- Povkhan, I., & Lupei, M. (2020). The algorithmic classification trees. In *Proceedings of the IEEE Third International Conference on Data Stream Mining & Processing (DSMP 2020)* (pp. 37–44). IEEE.
- Povkhan, I., Lupei, M., Kliap, M., & Laver, V. (2020). The issue of efficient generation of generalized features in algorithmic classification tree methods. In *Proceedings of the International Conference on Data Stream Mining and Processing (DSMP 2020)* (pp. 98–113). IEEE.
- Povkhan, I. (2020). Classification models of flood-related events based on algorithmic trees. *Eastern-European Journal of Enterprise Technologies*, 6(4), 58–68. <https://doi.org/10.15587/1729-4061.2020.219525>
- Rabcan, J., Levashenko, V., Zaitseva, E., Kvassay, M., & Subbotin, S. (2019). Application of fuzzy decision tree for signal classification. *IEEE Transactions on Industrial Informatics*, 15(10), 5425–5434. <https://doi.org/10.1109/TII.2019.2904845>
- Bodyanskiy, Y., Vynokurova, O., Setlak, G., & Pliss, I. (2015). Hybrid neuro-neo-fuzzy system and its adaptive learning algorithm. In *Proceedings of the Xth International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)* (pp. 111–114). Lviv.

GROUP SELECTION OF ELEMENTARY TRAITS IN SCHEMES FOR CONSTRUCTING HYBRID DECISION TREE STRUCTURES

The object of research is classification trees. The subject of research is methods, algorithms, and schemes for constructing classification trees. The aim of this work is to build an effective method (scheme) for synthesizing classification tree models based on a group assessment of the importance of discrete features within a branched attribute selection. A method for constructing classification trees is proposed, which for a given training sample determines the individual information content (importance) of groups of features (and their combinations) in relation to the initial value of the classification function (data from the training sample). The developed logical tree method, when constructing the next node of the classification tree, tries to identify a group of the most closely interrelated discrete features. This reduces the overall structural complexity of the model (the number of levels of the classification tree), speeds up calculations when recognizing objects based on the model, and also increases the generalizing properties of the model and its enterprise. The proposed scheme for selecting groups of discrete traits allows using the constructed decision tree to assess the informative value (importance) of traits. The developed classification tree method is implemented programmatically and studied when solving the problem of classifying discrete objects represented by a set of features. The conducted experiments confirmed the operability of the proposed mathematical support and allow us to recommend it for use in practice in solving applied problems of classification of discrete objects based on logical classification trees. Prospects for further research may consist in creating a modified method of the logical classification tree by effectively iterating and evaluating sets of elementary features based on the proposed method, optimizing its software implementations, and experimentally studying the proposed method on a wider set of applied problems.

Keywords: logical tree, algorithmic tree, classifier, pattern recognition, attribute, training sample, decision tree.

Інформація про авторів:

Повхан Ігор Федорович, д-р техн. наук, професор, декан факультету інформаційних технологій.

Email: igor.povkhan@uzhnu.edu.ua; <https://orcid.org/0000-0002-1681-3466>

Легеза Андрій Васильович, викладач кафедри програмного забезпечення систем.

Email: andrii.legeza@uzhnu.edu.ua; <https://orcid.org/0000-0001-8416-8836>

Цитування за ДСТУ: Повхан І. Ф., Легеза А. В. Групова селекція елементарних ознак в схемах побудови гібридних структур дерев рішень. *Український журнал інформаційних технологій*. 2025, т. 7, № 2. С. 09–17.

Citation APA: Povkhan, I. F., & Leheza, A. V. (2025). Group selection of elementary traits in schemes for constructing hybrid decision tree structures. *Ukrainian Journal of Information Technology*, 7(2), 09–17. <https://doi.org/10.23939/ujit2025.02.09>