УДК 001.891:65.011.56
DOI https://doi.org/10.32838/2663-5941/2019.5-1/22

Повхан І.Ф.
ДВНЗ «Ужгородський національний університет»

ОСОБЛИВОСТІ ВИПАДКОВИХ ЛОГІЧНИХ ДЕРЕВ КЛАСИФІКАЦІЇ В ЗАДАЧАХ РОЗПІЗНАВАННЯ ОБРАЗІВ

Робота піднімає важливе питання теорії розпізнавання образів – застосування методів та алгоритмів побудови випадкових логічних дерев класифікації. Розглядаються принципові особливості випадкових дерев класифікації, тобто логічних дерев, у яких вибір (генерація) вершин на довільному етапі побудови дерева відбувається випадково. Алгоритм, запропонований в роботі, дозволяє як будувати випадкові логічні дерева, так і генерувати цілі набори (множини) логічних дерев різної структури (складності), серед яких можна вибирати найбільш оптимальне для даної задачі. Підкреслюється важливість застосування випадкових логічних дерев для розв'язання задач розпізнавання образів, відбору (та можливої перебудови) найбільш ефективного дерева серед множини побудованих випадкових логічних дерев.

У роботі фіксується, що випадкові логічні дерева мають як свої суттєві переваги (програмна простота побудови дерева класифікації, зменшення часу загальної генерації логічного дерева, можливість оцінки та вибору найбільш підходящого дерева класифікації з множини побудованих), так і недоліки (неоптимальність структури, апаратні витрати на генерацію неоптимального дерева класифікації, гарантована складність структури та велика інформаційна місткість, необхідність додаткового етапу оцінки та відбору). Простий, ефективний та економний метод побудови випадкового логічного дерева класифікації навчальної вибірки дозволяє забезпечити необхідну швидкодію, рівень складності схеми розпізнавання, що гарантує проведення простого та повного розпізнавання дискретних об'єктів.

У роботі розглядаються деревоподібні схеми розпізнавання.

Ключові слова: *розпізнавання дискретних об'єктів, логічні дерева класифікації, функція розпізнавання, навчальна вибірка, алгоритм випадкового дерева.*

Постановка проблеми. Задачі, які об'єднуються тематикою розпізнавання образів, дуже різноманітні. Зараз не існує універсального підходу до їх розв'язання, але запропоновано декілька досить загальних теорій, що дозволяють вирішувати багато типів задач, проте їх прикладні застосування відрізняються досить великою чутливістю до специфіки самої задачі. Важливим та перспективним напрямом розв'язання задач розпізнавання є методи та алгоритми, які будуються на основі моделей логічних дерев класифікації (далі – ЛДК) [1; 2; 3]. Багато теоретичних результатів отримано для спеціальних випадків та підзадач, але слід зазначити, що вузьким місцем вдалих реальних систем розпізнавання залишається необхідність виконання величезного об'єму обчислень. Так, станом на сьогодні відомі різні алгоритми побудови логічних дерев класифікації, які зводяться до побудови одного дерева класифікації за даними фіксованої навчальної вибірки (далі – НВ) [4].

Аналіз останніх досліджень і публікацій. Інтерес до методів розпізнавання, які використовують під час побудови ЛДК, викликаний низкою корисних властивостей, якими вони володіють. Так, функції розпізнавання ЛДК дозволяють виділити у процесі класифікації як причинно-наслідкові зв'язки (та врахувати їх), так і фактори випадковості або невизначеності, тобто врахувати водночас і функціональні, і стохастичні відношення між властивостями і поведінкою системи. Було встановлено, що процес класифікації нових, тобто таких, що досі не траплялися, об'єктів світу тварин і людей (за винятком об'єктів, інформація про які передається генетичним шляхом (спадково), а також в деяких інших випадках), відбувається саме за так званим логічним деревом. Відзначимо також основний недолік в питанні побудови ЛДК – відсутність алгоритмів та методів, котрі би дозволили одноманітно описувати різні алгоритми розпізнавання образів у вигляді ЛДК.

Представлення функції розпізнавання (правила класифікації) у вигляді логічного дерева має великі переваги порівняно з іншим представленням схем класифікації [5]. Зауважимо, що запропонований алгоритм генерації випадкових дерев класифікації за даними навчальної вибірки доповнює методологію підходу розгалуженого вибору ознак та дозволяє будувати прості та ефективні

правила класифікації дискретних об'єктів [6]. У роботі зупинимось саме на описі особливостей та алгоритму побудови випадкових ЛДК для масиву даних НВ.

Постановка завдання. Метою статті є вивчення особливостей випадкових логічних дерев класифікації в задачах розпізнавання образів.

Виклад основного матеріалу дослідження. Нехай є задана множина M об'єктів w, а на ній існує розбиття R на кінцеве число підмножин (класів, образів) $|_i$, $(i = 1,\dots m)$, $M = \cup_{i=1}^{m} \Omega_i$. Припустимо, що розбиття M визначене неповністю. Задана тільки деяка інформація I про класи $|_i$. Об'єкти w задаються значеннями ознак x_j, $j = 1,\dots,n$ (цей набір такий самий для всіх об'єктів, тобто існує однакова розмірність об'єктів). Функція $f_R(w)$, яка задає розбиття R, задана на множині об'єктів M та дає на виході номер класу i, яку ми будемо називати функцією розпізнавання (далі – ФР). Зауважимо, що кожний образ (клас) характеризується певною спільністю деяких властивостей його елементів (об'єктів), а елементи з різних образів не мають цієї спільності. Загальна задача розпізнавання полягає в тому, щоб для довільного об'єкта w встановити його належність до певного класу (образу). Множини $|_i$ також називаються компонентами розбиття множини M.

Сукупність значень ознак x_j визначає опис (інформацію) $I(w)$ об'єкта w. Кожна з ознак може приймати значення з різних множин допустимих значень ознак. Задача розпізнавання стандартної інформації полягає в тому, що для фіксованого об'єкта w та набору класів $\Omega_1,\dots,\Omega_m$ за допомогою навчальної інформації $I(\Omega_1,\dots,\Omega_m)$ та опису $I(w)$ необхідно розрахувати значення деяких предикатів $P_i(w),(w \in \Omega_i; i = 1,\dots,m)$.

Нехай є розбиття R та деяка система розпізнавання Q. Система Q може бути представлена людиною або програмно-апаратною системою (системою операцій або логічних елементів). Задача розпізнавання образів буде зводитися до навчання системи Q обчислювати функцію $f_R(x)$, тобто система має реагувати під час подачі на вхід деякого сигналу (об'єкту) x сигналом $f_R(x)$ (фактичним номером класу належності). Основною інформацією під час навчання системи Q є значення функції $f_R(x)$ в деяких точках n-мірного простору (розмірність складає кількість ознак об'єктів множини M). Останнє означає, що при навчанні системи Q їй подаються пари сигналів $(x_i, f_R(x_i))$. На основі даної інформації (апріорної інформації) система Q будує схему обчислення $f_R(x)$.

У роботі ставиться задача дослідження особливостей таких методів розпізнавання, які би давали можливість у процесі навчання побудувати просту деревоподібну схему розпізнавання (схему у вигляді випадкового ЛДК), яка забезпечує необхідну ефективність та складність системи Q.

Питання особливостей випадкових логічних дерев класифікації. У даному дослідженні пропонується підхід, що дозволяє за допомогою єдиної методики одноманітно описувати та автоматично створювати (забезпечувати програмну генерацію) достатньо широкі класи алгоритмів розпізнавання дискретних об'єктів на основі ЛДК. Розглянемо особливості випадкових дерев класифікації, тобто ЛДК, в яких вибір (генерація) вершин на довільному етапі побудови дерева відбувається випадково.

Нехай є задані набори об'єктів НВ, ТВ, якісна оцінка кожної ознаки (ціна, якість), мінімальний час, необхідний для проведення класифікації, максимально допустима складність правил класифікації. У цих умовах потрібно знайти (синтезувати) таке правило класифікації (декілька правил класифікації) у вигляді ЛДК, побудоване за даними НВ, котре дає мінімальну кількість помилок на НВ в умовах дії зовнішнього середовища та здатне швидко адаптуватися до цих умов. Відзначимо, що в даній задачі ніякі обмеження на НВ та ТВ не накладаються, тобто ознаки можуть мати довільну природу (бути різнотипними), об'єм НВ та ТВ також довільний.

Далі залежно від природи (типу) ознак розглядаються два шляхи розв'язку даної задачі:

1) випадок, коли ознаки приймають бінарні значення;

2) випадок, коли ознаки мають різнотипну природу.

Зазначимо, що другий випадок можна звести до першого, застосувавши спеціальний алгоритм кодування. Зауважимо, що при цьому часто виникають досить великі втрати інформації. Це негативно впливає на якість розпізнавання, і взагалі якісне та правильне кодування початкової НВ є нетривіальною задачею, від якої напряму залежать наступні етапи побудови ЛДК.

Розглянемо новий тип алгоритмів кодування початкової інформації $I(l)$, що дозволяють закодувати НВ у такий спосіб, що якщо класи у вихідному описі не перетинаються, то і після кодування вони перетинатися не будуть (умова несуперечливості при кодуванні). Крім того, при цьому відбувається часткове виявлення певних властивостей (логічних закономірностей) об'єктів початкової вибірки.

Запропонований тип алгоритмів кодування базується на такій ідеї: оскільки об'єкти НВ становлять деякі точки n-вимірного простору, то завжди за умови вихідної відмінності між класами ми можемо ввести n-вимірний простір, обмежений можливими значеннями ознак, площинами, у n-вимірні гіперпаралелепіпеди (гіперкуби) так, щоби точки (об'єкти) з різних класів не потрапляли в той самий гіперпаралелепіпед (тобто провести геометричне розділення класів) [7]. На наступному етапі кожному гіперпаралелепіпеду в n-вимірному просторі можна поставити у відповідність певний двозначний код. Ці двозначні коди залежно від того, куди потрапляють об'єкти НВ та ТВ, визначають набір бінарних ознак задачі розпізнавання образів. Замість гіперпаралелепіпедів можна взяти гіпереліпси або гіперсфери, а відповідні ефективні алгоритмічні реалізації можна взяти з роботи [8], де вони пройшли необхідну прикладну апробацію.

Розглянемо загальну методику побудови різних випадкових ЛДК та їхні переваги перед відомими алгоритмами класифікації у вигляді ЛДК.

Під час побудови правил класифікації (схем) у вигляді ЛДК за допомогою відомих алгоритмів, як правило, буде отримано одне фіксоване ЛДК. Вибір ознак в ту або іншу вершину ЛДК відбувається цілеспрямовано, використовуються ті або інші критерії (функціональні оцінки) важливості ознак [9]. Проте найважливіші ознаки, визначені за даними НВ, можуть виявитися не такими важливими в реальному процесі.

Так, у роботі [4] були розглянуті дві алгоритмічні реалізації побудови ЛДК на основі методу розгалуженого вибору ознак. Вибір ознак здійснювався в першому випадку на основі функціональної оцінки якості ознак на кожному кроці, а в другому випаду для економії ресурсів та зменшення часу генерації ЛДК – лише на початку роботи алгоритму з відбором ознак від найбільш до найменш інформативних. Ми пропонуємо алгоритм (схему) побудови випадкового ЛДК, тобто дерева, ознаки в вершинах якого у процесі його побудови відбираються у випадковий спосіб (випадковим програмним генератором PRG). Запропонуємо одну з можливих алгоритмічних реалізацій побудови випадкового ЛДК.

Загальна схема алгоритму побудови випадкового ЛДК. *Крок 1.* Нехай є задана НВ об'ємом m об'єктів розмірності n відомої класифікації. Необхідно за даної початкової інформації та за допомогою деякого випадкового програмного генератора PRG побувати ЛДК, яке дозволяє однозначно класифікувати об'єкти НВ. На першому етапі з використанням PRG вибирається початкова вершина ЛДК та будуються відповідні стрілки (дуги), які виходять з вершини даного дерева. У лічильник об'єктів НВ записується одинця. В ідентифікатор ЛДК, що будується, записується перша згенерована вершина.

Крок 2. За допомогою лічильника НВ аналізується поточний об'єкт, тобто проставляються відповідні значення ознак в гілках дерева за умови, що кількість ярусів не більша від кількості ознак об'єкта, а в іншому разі в кінцевій вершині проставляється значення функції розпізнавання для поточного об'єкта. Перевіряються значення лічильника НВ на m: якщо значення збігаються, то робота алгоритму передається на *Крок 4*.

Крок 3. За допомогою PRG генерується наступна вершина ЛДК за умови, що кількість ярусів не більша від кількості ознак об'єкта, яка також записується в ідентифікатор дерева. Керування передається на *Крок 2*. Якщо кількість ярусів ЛДК дорівнює n, то проводимо інкремент лічильника НВ, потім керування передається на *Крок 2*.

Крок 4. Роботу алгоритму завершено, ЛДК побудовано, а в ідентифікаторі дерева зберігається послідовність вершин, які характеризують дане дерево.

Цей алгоритм досить легко реалізується програмним шляхом. За його допомогою для кожної конкретної задачі розпізнавання ми можемо побудувати деяку множину ЛДК та серед них вибрати саме те дерево (декілька ЛДК), котре задовольнить поставлені вимоги.

Зауважимо, що в пам'яті комп'ютера ми не зберігаємо побудовані випадкові ЛДК, а тільки деякий вектор-характеристику (ідентифікатор) цього ЛДК, наприклад, число, знаючи яке, ми знов можемо побудувати це ЛДК та деякі важливі його характеристики (ефективність на ТВ, складність, максимальний час прийняття рішень у процесі експлуатації тощо). Отже, побудувавши програмно деяку фіксовану множину випадкових ЛДК та сформувавши їх ідентифікатори, ми маємо деяку передісторію. На основі цієї передісторії ми можемо застосовувати ідею цілеспрямованого пошуку для наступної побудови ЛДК та знаходження таких, які задовольняють поставлені вимоги.

Зауважимо, що даний підхід можна застосувати для побудови випадкових k – значних ЛДК. Крім того, можна розглядати випадкові ЛДК, у вершинах яких розміщені не окремі ознаки, а

довільні алгоритми розпізнавання (алгоритмічне дерево класифікації), але у такому разі для успішного програмного моделювання необхідно мати модулі цих алгоритмів [4]. Створюється новий спосіб практичного використання вже відомих алгоритмів (методів) розпізнавання з автоматичним визначенням площин їх компетентності, що є важливою умовою використання правил класифікації [5].

Подібне застосування алгоритмів розпізнавання використовується в програмній системі ОРИОН ІІІ, де алгоритми розпізнавання в вершинах АДК обиралися за деякими критеріями, визначеними на основі початкових даних НВ [10].

Висновки. Випадкові ЛДК мають як суттєві переваги (програмна простота побудови дерева, зменшення часу загальної генерації ЛДК, можливість оцінки та вибору найбільш задовільного ЛДК з множини побудованих), так і недоліки (неоптимальність структури, апаратні витрати на генерацію неоптимального ЛДК, гарантована складність структури та велика інформаційна місткість, наявність додаткового етапу оцінки та відбору).

Алгоритм побудови випадкового ЛДК дозволяє генерувати цілі набори (множини) дерев класифікації різної структури (складності), серед яких можна обирати найбільш оптимальне для даної задачі.

Зауважимо, що важливим питанням в застосуванні випадкових ЛДК є питання вибору (та можливої перебудови) найбільш ефективного дерева серед множини побудованих випадкових ЛДК.

Алгоритм побудови випадкового логічного дерева, який був описаний вище, разом з алгоритмом з покроковою оцінкою важливості дискретних ознак та алгоритмом одноразової оцінки важливості дискретних ознак утворює основну трійку алгоритмів розпізнавання методу розгалуженого вибору ознак під час побудови логічних дерев класифікації [1].

Відзначимо також, що для зберігання фактичної структури побудованого випадкового ЛДК використовується лише програмний ідентифікатор (паспорт) логічного дерева, який містить лише фактичну послідовність вершин (змінних) в даному дереві, що є ресурсно-економним способом представлення таких складних структур даних. Даний алгоритм генерації випадкових ЛДК реалізований у бібліотеці алгоритмів та класифікаторів РО програмного комплексу системи ОРИОН ІІІ.

Список літератури:

1. Повхан І.Ф. Метод розгалуженого вибору ознак в математичному конструюванні багаторівневих систем розпізнавання образів. *Штучний інтелект :* науково-технічний журнал. 2003, № 7. С. 246–249.

2. Quinlan J.R. Induction of Decision Trees. *Machine Learning*. 2008, № 1, P. 1−81. 22.

3. Василенко Ю.А., Повхан І.Ф., Ващук Ф.Г. Проблема оцінки складності логічних дерев розпізнавання та загальний метод їх оптимізації. *European Journal of Enterprise Technologies* : науково-технічний журнал. 2011, 6/4 (54). С. 24–28.

4. Повхан І.Ф., Василенко Ю.А., Василенко Е.Ю. Концептуальна основа систем розпізнавання образів на основі методу розгалуженого вибору ознак. European Journal of Enterprise Technologies : науково-технічний журнал. 2004. № 7 (1). С. 13–15.

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

6. Povhan I. General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects. *Електроніка та інформаційні технології :* збірник наукових праць. Львів. 2019. Випуск 11. С. 112–117.

7. Vasilenko E.Yu., Kuhayivsky A.I., Papp I.O., Vasilenko Yu. Construction and optimization of recongnizing systems. *Інформаційні технології і системи* : науково-технічний журнал. Львів. 1999. № 1 (Т. 1). С. 122–125.

8. Василенко Ю.А., Повхан И.Ф. Апроксимация обучающей выборки гиперпараллелепипедами. *Науковий вісник УжДПЕП.* 1998. № 2. С. 9–17.

9. Повхан І.Ф., Василенко Ю.А. Групова та індивідуальна оцінка важливості бульових аргументів. *Вісник національного технічного університету «ХПІ»*. 2011. № 53. С. 57–64.

10. Повхан І.Ф. Проблема функціональної оцінки навчальної вибірки в задачах розпізнавання дискретних об'єктів. *Вчені записки Таврійського національного університету. Серія «Технічні науки».* 2018. Том 29 (68). № 6. С. 217–222.

Povkhan I.F. FEATURES OF RANDOM LOGICAL CLASSIFICATION TREES IN PATTERN RECOGNITION PROBLEMS

The work raises an important question of pattern recognition theory – the use of methods and algorithms for constructing random logical classification trees. The principal features of random classification trees are considered, i.e. logical trees in which the selection (generation) of vertices at an arbitrary stage of tree construction occurs randomly. The algorithm proposed in this paper allows the construction of a random logical tree, and generate whole sets of logical trees of different structures (complexity), among which you can choose the most optimal for this problem. Emphasizes the importance of the issue in the application of random Boolean trees for the solution of pattern recognition problems – the question of selection (and possible adjustment) is the most efficient tree among the many built of random logical trees.

The work is recorded that random logical tree has its significant advantages (software, the ease of construction of a classification tree, reducing the time of the generation of the logical tree, the ability to evaluate and select the most appropriate classification tree from the set was built) and significant drawbacks (not the optimal structure, the hardware cost of generation is not the optimal classification tree, guaranteed by the complicated structure and large information capacity, the need for additional evaluation phase and selection). A simple, efficient, cost-effective method of constructing a random logical classification tree training sample allows you to provide the necessary speed, the level of complexity of the recognition scheme, which guarantees a simple and complete recognition of discrete objects.

The paper deals with tree-like recognition schemes.

Key words: *recognition of discrete objects, logical classification trees, recognition function, training sample, algorithm is a random tree.*