УДК 004.8:004.89:519.7 DOI:10.31673/2412-4338.2020.021100

Повхан І. Ф. *ДВНЗ "Ужгородський національний університет", Ужгород*

ОЦІНКА ЗАГАЛЬНОЇ СКЛАДНОСТІ ПРОЦЕДУРИ ПОБУДОВИ БІНАРНОГО ЛОГІЧНОГО ДЕРЕВА КЛАСИФІКАЦІЇ ДЛЯ ДОВІЛЬНОГО ВИПАДКУ

Анотація: *Пропонується верхня оцінка складності процедури синтезу бінарного логічного дерева класифікації для довільного випадку (для умов слабкого та сильного розділення класів навчальної вибірки). Розв'язок даного питання має принциповий характер, щодо оцінки структурної складності моделей класифікації (у вигляді деревоподібних конструкцій) дискретних об'єктів для широкого спектру прикладних задач класифікації та розпізнавання в плані розробки перспективних схем та методів їх фінальної оптимізації (мінімізації) структури. Дане дослідження має актуальність не лише для конструкцій логічних дерев класифікації, але дозволяє розширити саму схему оцінки складності і на загальний випадок структур алгоритмічних дерев класифікації (концепції дерев алгоритмів та дерев узагальнених ознак).*

Досліджене актуальне питання складності загальної процедури побудови логічного дерева класифікації на основі концепції поетапної селекції наборів елементарних ознак (можливих їх різнотипних множин та сполучень), яке для заданої початкової навчальної вибірки (масиву дискретної інформації) будує деревоподібну структуру (модель класифікації), з набору елементарних ознак (базових атрибутів) оцінених на кожному кроці схеми побудови моделі за даною вибіркою.

Так сучасні інформаційні технології, засновані на математичних моделях розпізнавання образів (логічних та алгоритмічних дерев класифікації), широко використовуються в соціально-економічних, екологічних та інших системах первинного аналізу та обробки великих масивів інформації. Причому це пояснюється тим фактом, що такий підхід дозволяє усунути набір існуючих недоліків добре відомих класичних методів, схем та досягти принципово новий результат. Робота присвячена проблематиці моделей дерев класифікації (дерев рішень), та пропонує оцінку складності структур логічних дерев (моделей дерев класифікації), які складаються з відібраних та ранжованих наборів елементарних ознак побудованих на основі загальної концепції розгалуженого вибору ознак. Даний метод при формуванні поточної вершини логічного дерева (вузла) забезпечує виділення найбільш інформативних (якісних) елементарних ознак з початкового набору. Такий підхід при побудові результуючого дерева класифікації дозволяє значно скоротити розмір та складність дерева (загальну кількість гілок та ярусів структури) підвищити якість його наступного аналізу (інтерпретабельність).

Ключові слова: *логічне дерево класифікації, розпізнавання образів, класифікація, дискретна ознака.*

Povkhan I. F. *Uzhhorod National University, Uzhhorod*

ESTIMATION OF GENERAL COMPLEXITY OF THE PROCEDURE FOR CONSTRUCTING A BINARY LOGICAL CLASSIFICATION TREE FOR AN ARBITRARY CASE

Abstract: *We propose an upper estimate of the complexity of the binary logical tree synthesis procedure for classifying an arbitrary case (for conditions of weak and strong separation of classes in the training sample). The solution to this question is of a fundamental nature, regarding the assessment of the structural complexity of classification models (in the form of tree structures) of discrete objects for a wide range of applied classification and recognition problems in terms of developing promising schemes and methods for their final optimization (minimization) of the structure. This research is relevant not only for the constructions of logical classification trees, but also allows us to extend the complexity estimation scheme itself to the general case of algorithmic structures of classification trees (concepts of algorithm trees and generalized feature trees).*

The current issue of complexity of the general procedure for constructing a logical classification tree based on the concept of step-by-step selection of sets of elementary features (their possible heterogeneous sets

and combinations), which for a given initial training sample (an array of discrete information) builds a tree structure (classification model), from a set of elementary features (basic attributes) evaluated at each stage of the model construction scheme for this sample.

Thus, modern information technologies based on mathematical models of pattern recognition (logical and algorithmic classification trees) are widely used in socio-economic, environmental and other systems of primary analysis and processing of large amounts of information. This is due to the fact that this approach allows you to eliminate a set of existing disadvantages of well-known classical methods and schemes and achieve a fundamentally new result. The work is devoted to the problems of classification tree models (decision trees), and offers an assessment of the complexity of logical tree structures (classification tree models), which consist of selected and ranked sets of elementary features built on the basis of the General concept of branched feature selection. This method, when forming the current vertex of the logical tree (node), provides the selection of the most informative (qualitative) elementary features from the source set. This approach allows you to 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.

Keywords: *logical classification tree, pattern recognition, classification, discrete attribute.*

Повхан И.Ф. *ГВУЗ "Ужгородский национальный университет", г. Ужгород*

ОЦЕНКА ОБЩЕЙ СЛОЖНОСТИ ПРОЦЕДУРЫ ПОСТРОЕНИЯ БИНАРНОГО ЛОГИЧЕСКОГО ДЕРЕВА КЛАССИФИКАЦИИ ДЛЯ ПРОИЗВОЛЬНОГО СЛУЧАЯ

Аннотация*: Предлагается верхняя оценка сложности процедуры синтеза бинарного логического дерева классификации для произвольного случая (для условий слабого и сильного разделения классов обучающей выборки). Решение данного вопроса имеет принципиальный характер, относительно оценки структурной сложности моделей классификации (в виде древовидных конструкций) дискретных объектов для широкого спектра прикладных задач классификации и распознавания в плане разработки перспективных схем и методов их финальной оптимизации (минимизации) структуры. Данное исследование имеет актуальность не только для конструкций логических деревьев классификации, но позволяет расширить саму схему оценки сложности и на общий случай алгоритмических структур деревьев классификации (концепции деревьев алгоритмов и деревьев обобщенных признаков).*

Исследован актуальный вопрос сложности общей процедуры построения логического дерева классификации на основе концепции поэтапной селекции наборов элементарных признаков (возможных их разнотипных множеств и сочетаний), которое для заданной начальной обучающей выборки (массива дискретной информации) строит древовидную структуру (модель классификации), из набора элементарных признаков (базовых атрибутов) оцененных на каждом этапе схемы построения модели по данной выборке.

Так современные информационные технологии, основанные на математических моделях распознавания образов (логических и алгоритмических деревьях классификации), широко используются в социально-экономических, экологических и других системах первичного анализа и обработки больших массивов информации. Причем это объясняется тем фактом, что такой подход позволяет устранить набор существующих недостатков хорошо известных классических методов, схем и достичь принципиально новый результат. Работа посвящена проблематике моделей деревьев классификации (деревьев решений), и предлагает оценку сложности структур логических деревьев (моделей деревьев классификации), которые состоят из отобранных и ранжированных наборов элементарных признаков построенных на основе общей концепции разветвленного выбора признаков. Данный метод при формировании текущей вершины логического дерева (узла) обеспечивает выделение наиболее информативных (качественных) элементарных признаков из исходного набора. Такой подход при построении результирующего дерева классификации позволяет значительно сократить размер и сложность дерева (общее количество ветвей и ярусов структуры) повысить качество его последующего анализа (интерпретабельность).

Ключевые слова: *логическое дерево классификации, распознавание образов, классификация, дискретный признак.*

Вступ. Інформаційні технології, засновані на математичних моделях розпізнавання образів у вигляді логічних дерев класифікації – ЛДК (деревоподібних моделей), широко використовуються в соціально-економічних, екологічних та інших системах обробки інформації. Це пояснюється тим фактом, що такий підхід дозволяє усунути набір недоліки класичних методів та досягти принципово новий результат, ефективно та раціонально використовуючи потужності обчислювальних систем [1,8,10]. На сьогоднішній день відомо більше ніж 4000 алгоритмів розпізнавання (заснованих на різноманітних підходах та концепціях), які мають певні обмеження при їх використанні (точність, швидкодія, пам'ять, універсальність, надійність, тощо). Крім того, кожний з алгоритмів обмежений певною специфікою задач застосування, а це безумовно є найслабкішим місцем не тільки даних алгоритмів, але й систем розпізнавання, які базуються на відповідних концепціях [2-9]. Відомо, що представлення навчальних вибірок (дискретної інформації) великого об‘єму у вигляді структур логічних дерев має свої суттєві переваги в плані економічного опису даних та ефективних механізмів роботи з ними [5]. Тобто – покриття навчальної вибірки набором елементарних ознак у випадку ЛДК, або покриття навчальної вибірки фіксованим набором автономних алгоритмів розпізнавання та класифікації у випадку АДК (алгоритмічних дерев класифікації), породжує фіксовану деревоподібну структуру даних, яка забезпечує стиск та перетворення початкових даних НВ – а отже дозволяє суттєву оптимізацію та економію апаратних ресурсів інформаційної системи [7].

Постановка завдання. Нехай задана начальна вибірка (НВ) в наступного вигляду:

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

Зауважимо, що тут $x_i \in G$ ($G -$ деяка множина), а функція розпізнавання (ФР) $f_R(x_i) \in \{1,2, \dots, k\}, (i = 1,2, \dots, m)$.

Відповідно $f_R(x_i) = l, (1 \le l \le k)$ означає, що $x_i \in H_l, H_l \subset G$. Тут $f_R -$ деяка скінчено значна функція, яка задає розбиття R множини G, яке складається з підмножин (образів, класів) $H_1, H_2, \dots, H_k$.

Таким чином, НВ – це сукупність (точніше послідовність) деяких наборів, причому кожний набір – це сукупність значень деяких ознак та значень деяких функцій на цьому наборі. Можна ще сказати, що сукупність значень ознак – це деяке зображення, а значення функції відносить це зображення до відповідного образу [2].

Отже, зазвичай стоїть загальна задача побудови моделі логічного дерева класифікації (ЛДК) з набором деяких параметрів p, структура L якої була би оптимальною $F(L(p, x_i), f_R(x_i)) \to opt$ по відношенню до початкових даних НВ, причому нас в межах даного дослідженні буде цікавити складність такої структури на етапі побудови моделі ЛДК.

Аналіз досліджень і публікацій. Аналізуючи проблематику деревоподібних моделей класифікації та розпізнавання можна побачити певний брак поточних досліджень в цьому напрямку, коли головна увага зміщена в бік концепції нейромережевого розпізнавання [14]. В значній мірі це пояснюється особливостями самих моделей ЛДК, складнощами реалізаційних моментів концепції алгоритмічного дерева класифікації (найвищого рівня абстракції концепції ЛДК), набором жорстких правил та обмежень щодо практичної роботи з такими структурами даних [7]. Дане дослідження продовжує цикл робіт, які присвячені проблематиці деревоподібних схем розпізнавання (класифікації) дискретних об‘єктів [2-7,15-18]. В них піднімаються питання побудови, використання, та оптимізації логічних дерев. Так з [2] відомо, що результуюче правило класифікації (схема), яке побудоване довільним методом або алгоритмом розгалуженого вибору ознак, має деревоподібну логічну структуру. Логічне дерево складається з вершин (ознак), які групуються по ярусам і які отримані на певному кроці (етапі) побудови дерева розпізнавання [11]. Важливою задачею, яка виникає з [15] задача синтезу дерев розпізнавання, які будуть представлятися фактично деревом (графом)

алгоритмів. На відміну від існуючих методів, головною особливістю деревоподібних систем розпізнавання є те, що важливість окремих ознак (групи ознак чи алгоритмів) визначається відносно функції, яка задає розбиття об‘єктів на класи [12]. Так в роботі [13] піднімаються принципові питання стосовно генерації дерев рішень для випадку малоінформативних ознак. Здатність ЛДК виконувати одномірне розгалуження для аналізу впливу (важливості, якості) окремих змінних дає можливість працювати зі змінними різних типів у вигляді предикатів (у випадку АДК – відповідними автономними алгоритмами класифікації та розпізнавання). Така концепція логічних дерев активно використовується в інтелектуальному аналізі даних, де кінцева мета полягає в синтезі моделі, яка прогнозує значення цільової змінної на основі набору початкових даних на вході системи [14]. Так, як головну ідею методів та алгоритмів РВО можна визначити як оптимальну апроксимацію деякої початкової НВ 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 [9] де піднімаються питання якісної оцінки окремих дискретних ознак, їх наборів та фіксованих сполучень, що дозволяє запровадити ефективний механізм реалізації розгалуження. Структура ЛДК характеризується компактністю з одного боку та нерівномірністю заповнення (розрядженістю) ярусів з іншого боку в порівнянні з регулярними деревами (алгоритмом з одноразовою оцінкою важливості ознак) [4]. Відмітимо, що важливими питаннями залишаються питання збіжності процесу побудови ЛДК за методами розгалуженого вибору ознак (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) [16]. Зважаючи на цю проблематику і виникає принципове питання деревоподібник моделей класифікації – питання загальної складності структури ЛДК що будується за даними НВ.

Метою дослідження є числова оцінка складності процедури побудови бінарного логічного дерева класифікації (моделі ЛДК) для довільного випадку в умовах слабкого та сильного розділення класів початкової навчальної вибірки.

Результати дослідження. Досліджене актуальне питання загальної складності процедури побудови логічного дерева класифікації на основі концепції поетапної селекції наборів елементарних ознак (можливих їх різнотипних множин та сполучень), яке для заданої початкової навчальної вибірки (масиву дискретної інформації) будує деревоподібну структуру (модель класифікації), з набору елементарних ознак (базових атрибутів) оцінених на кожному кроці схеми побудови моделі за даною вибіркою. Верхня оцінка складності структури ЛДК дозволяє оцінити фінальну складність моделі дерева класифікації на основі початкової НВ та дозволяє ефективно провести процедуру обрізки (оптимізації та мінімізації) побудованої моделі.

Виклад основного матеріалу. На першому етапі даного дослідження припустимо, що на кожному $n-$ вому кроці процедури побудови дерева класифікації (моделі ЛДК) множина D_n (ознакового простору поточної задачі) слабо розділяються деякою ознакою φ_n [2]. Далі розглянемо схему p_n [5]. В цій схемі маємо відповідно $n+1$ кінцевих шляхів. Завдяки тому, що D_n на кожному кроці слабо розділяється [7], кожний такий шлях містить хоча би одну пару початкової НВ загального вигляду (1). Крім того очевидно, що різні кінцеві шляхи в p_n не мають спільних пар з вибірки (1).

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

Таким чином, якщо на кожному $n-$ вому кроці відібрана елементарна ознака φ_n слабо розділяє множину D_n, то в цьому випадку процес побудови ЛДК збігається відносно

початкової НВ та закінчується не більше чим за $m-1$ кроків, де $m-$ кількість всіх навчальних пар початкової НВ [11].

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

Зафіксувавши $f_R(x) \equiv 0$, буде отримано деяку узагальнену ознаку (схему) f_0, яка апроксимує (повністю або частково) початкову НВ [15]. Очевидно, що в даному випадку (тобто в ситуації, коли ще не зроблено вибір жодного елементарної ознаки φ_n) узагальнена ознака (схема) f_0 є найкращою апроксимацією початкової НВ. Далі величину n_1 будемо називати – безумовною кількістю помилок в початковій НВ.

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

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

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

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

Так як $n_0^0+n_0^1=n_0$ та $n_1^0+n_1^1=n_1$, то будемо мати наступне:

$$\rho=\max(n_0^0, n_1^0)+\max(n_0^1, n_1^1) \geq n_0. \quad (2)$$

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

$$m-\rho=n_1-(\rho-n_0) \leq n_1. \quad (3)$$

Зауважимо, що (3) випливає з (2). Введемо величину $\lambda_1=\frac{n_1}{m-\rho}$ та назвемо її якістю елементарної ознаки φ_1 відносно початкової НВ, аналогічно визначається λ_n ознаки φ_n відносно початкової НВ $(n=1,2,3,\dots)$. На наступному етапі дослідження зробимо наступне припущення – якість λ_n елементарної ознаки φ_n відносно масиву початкової НВ не менше чим деяке число y, де $y>1$. Проаналізуємо складність процедури побудови дерева класифікації при даній умові $(y>1)$, для цього оцінимо кількість кроків, за якими даний процес (процедура) реалізує повне розпізнавання масиву початкової навчальної вибірки. Розглянемо для визначеності наступну схему побудови дерева класифікації (рис. 1). Нехай n_1- безумовна кількість помилок початкової НВ. Елементарна ознака φ_1^1 розділяє НВ на дві вибірки H_0 та H_1. Нехай h_0 та h_1, відповідно безумовна кількість помилок в вибірках H_0 та H_1. Ознака φ_1^2 розділить множину H_0 на дві множини H_{00} та H_{01}. Нехай h_{00} та $h_{01}-$ безумовна кількість

помилок в вибірках H_{00} та H_{01}. Аналогічно визначимо множини H_{10},H_{11} та кількості h_{10} та h_{11} для елементарної ознаки φ_2^2.

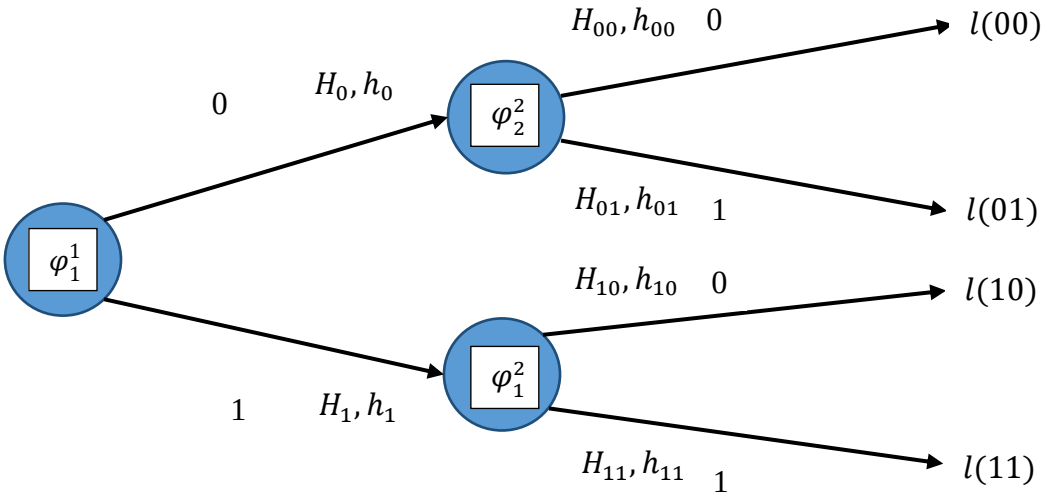

Рис. 1. Схема розбиття на підмножини в структурі дерева класифікації

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

$$\begin{cases} h_0 + h_1 \le \frac{1}{y} * n_1 \\ h_{00} + h_{01} \le \frac{1}{y} * h_0. \\ h_{10} + h_{11} \le \frac{1}{y} * h_1 \end{cases} \quad (4)$$

З (4) отримаємо наступне:

$$h_{00} + h_{01} + h_{10} + h_{11} \le \frac{1}{y^2} * n_1. \quad (5)$$

Зробимо наступні припущення в даному відношенні: $h_0 \ge 1, h_1 \ge 1, \quad h_{00} \ge 1, h_{01} \ge 1$, $h_{10} \ge 1$ та $h_{11} \ge 1$. Звідси будемо мати наступне:

$$2^1 \le \frac{1}{y} * n_1, 2^2 \le \frac{1}{y^2} * n_1. \quad (6)$$

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

$$2^i \le \frac{1}{y^i} * n_1 \text{ або } (2y)^i \le n_1. \quad (7)$$

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

$$m = R(\frac{\log_2 n_1}{1+\log_2 y}). \quad (8)$$

Під $R(x)$ розуміється заокруглення числа x до найблищого цілого числа, яке перевищує x. Наприклад $Q(1.2) = 2, Q(3.7) = 4, Q(4.1) = 5$.

Отже дерево класифікації, яке має m повних ярусів (тобто випадок, коли на $i-$ товому ярусі стоять 2^{i-1} вершин), має $2^{m+1}-1$ вершин – таким чином розпізнавання початкової НВ при умові $(y>1)$ за допомогою повного ЛДК відбувається не більш чим за $2^{m+1}-1$ кроків, де m розраховується за допомогою виразу (8).

Далі повним логічним деревом будемо називати таке ЛДК, яке на кожному $i-$ товому ярусі містить 2^{i-1} вершин. При реалізації процесу побудови дерева класифікації не обов‘язково отримується повне дерево. Покажемо, що якщо виконується умова $y \geq 2$ то довільний процес побудови дерева класифікації містить також не більше $2^{m+1}-1$ кроків, де $m-$ задовольняє попередній умові $(m = R(\frac{\log_2 n_1}{1+\log_2 y}))$. Для цього необхідно показати, що вищевказане повне дерево при умові $(y \geq 2)$ містить найбільшу кількість вершин серед всіх можливих дерев, які реалізують повне розпізнавання початкової НВ. Перед тим, як перейти до доведення цього факту введемо деякого вигляду вирази. Нехай задано фіксована структура (ЛДК) дерево (рис. 2).

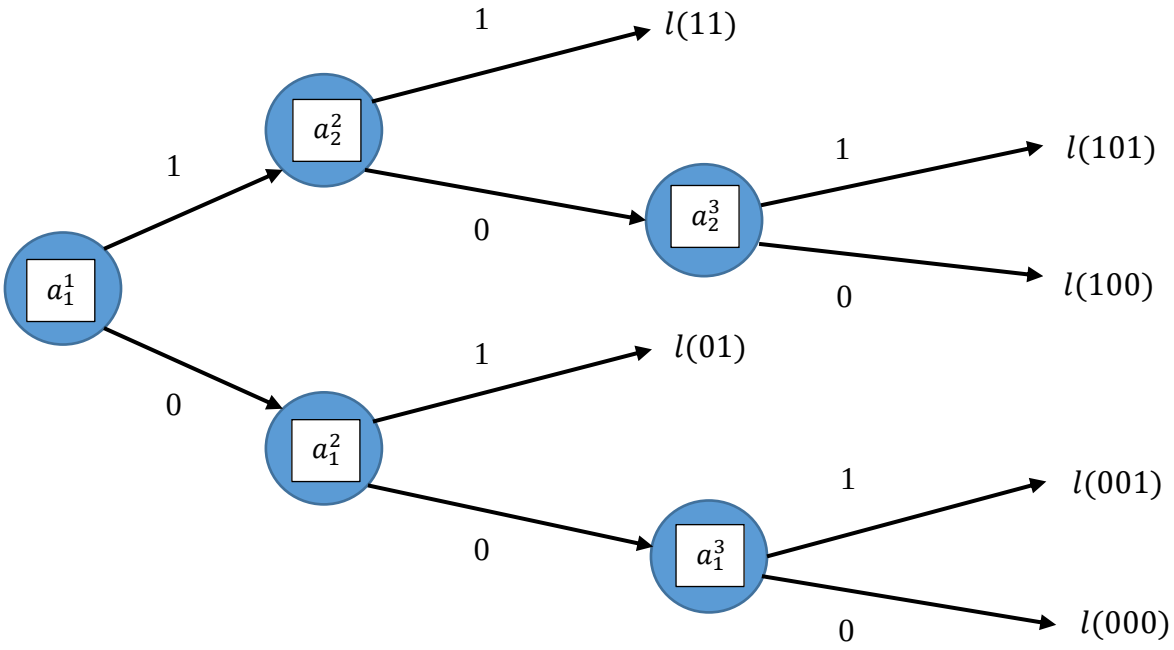

Рис. 2. Початкова структура ЛДК з набором елементарних ознак

Припустимо, що ЛДК (рис. 2) повністю реалізує розпізнавання НВ. Нехай $h_0, h_1, h_{00}, h_{10}-$ безумовна кількість помилок, відповідно в множинах H_0, H_1, H_{00}, H_{10} (множинах розбиття). Підрахуємо кількість всіх помилок, які виправлені ЛДК (рис. 2). Отже, на основі вище сказаного – ця кількість задовольняє наступному співвідношенню:

$$n_1 \geq (th_{00} + th_{10})t \tag{9}$$

Відповідні числа h_{00} та h_{10} задовольняють співвідношенням $h_{00} \geq 1$ та $h_{10} \geq 1$. Дійсно, якщо би наприклад, $h_{10} = 0$, тоді не потрібен в АДК (рис. 2) алгоритм a_2^3. Звідси та з (9) маємо наступне:

$$n_1 \geq (t+t)t \tag{10}$$

Таким чином, для кожного ЛДК, яке реалізує повне розпізнавання НВ, можна підрахувати деяку величину, яка не повинна перевищувати числа n_1. З розглянутого прикладу випливає, підрахунок цієї величини можна проводити наступним чином.

Нехай задано дерево D^0, яке реалізує деякий процес побудови АДК (наприклад логічне дерево, рис. 2). За деревом D^0 будуємо наступний вираз D^*. Спочатку в дереві D^0 прибираємо всі зовнішні стрілки (наприклад для дерева (рис. 2) всі стрілки, які виходять з вершин з ознаками a_1^3, a_2^3, та стрілки в кінці яких стоять l_{01}, l_{11}). Дерево яке отримано наступним чином позначимо через D^1. Якщо в дереві D^1 є не кінцеві вершини, з яких виходить тільки одна стрілка, то від цих вершин проводиться ще одна стрілка, та в її кінці ставимо число 0.

Далі в дереві D^1 прибираємо всі кінцеві вершини, в яких стоять ознаки та на їх місце ставимо 1. Кожній некінцеві вершині дерева D^1 ставимо у відповідність операцію $(x + y)t$, де x та y аргументи цієї операції. Тут слід зауважити, що аргументам x та y відповідають стрілки, які виходять з відповідної вершини. На цьому процес побудови виразу D^* закінчується. Наприклад, вираз D^*, який відповідає АДК (рис. 2) має вигляд (рис. 3).

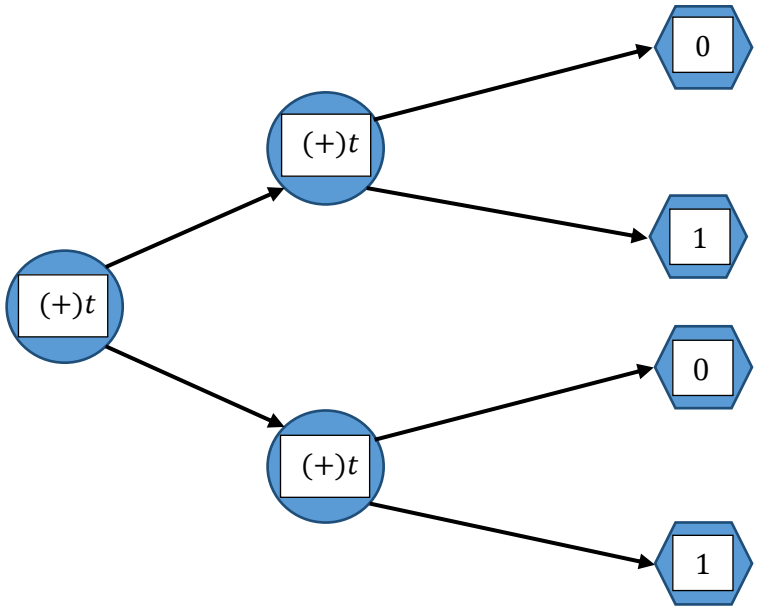

Рис. 3. Структура виразу D^*, який відповідає початковому ЛДК

Результат обчислення виразу D^* на (рис. 3) має вигляд:

$$\big((1 + 0)t + (1 + 0)t\big)t = (t + t)t. \tag{11}$$

Нехай $N -$ деяке ціле додатне число. Зафіксуємо наступне задач: серед всіх дерев, які мають N вершин, знайти таке дерево D^0, для якого відповідний вираз D^* є найменшим. Спочатку розв‘яжемо цю задачу – будемо вважати, що вершина дерева D^* має глибину i, якщо вона стоїть на $i -$ товому ярусі (рахується, що нумерація ярусів іде зверху вниз та самий верхній ярус має нульовий номер). Вираз, який представляє дерево D^*, можна представити у вигляді суми $t^{i_1} + t^{i_2} + \cdots + t^{i_m}$, де $m -$ кількість всіх кінцевих вершин дерева D^*, в яких стоїть одиниця. Причому, якщо вершина в якій стоять одиниця, має глибину $i_s, (1 \le s \le m)$, то їй відповідає в сумі $t^{i_1} + t^{i_2} + \cdots + t^{i_m}$ член t^{i_s}. На наступному етапі ЛДК D^* будемо перетворювати так щоби кількість N вершин в ньому не мінялася, але зменшувався вираз $t^{i_1} + t^{i_2} + \cdots + t^{i_m}$. Нехай $v -$ вершина в дереві D^*, яка має найбільшу глибину та в якій стоїть одиниця, $n -$ глибина вершини v. Нехай в дереві D^* є вершина $v^{'}, (n^{'} < n)$, в якій стоїть нуль. Зауважимо, що при підрахунку вершин в дереві D^*, вершини в яких стоїть нуль не приймаються до уваги. Структура дерева D^* біля вершини v може мати вигляд (рис. 4 (a)) або (рис. 4 (b)).

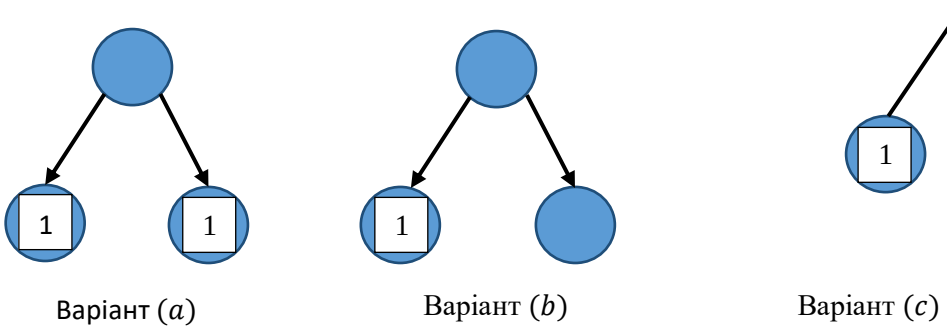

Рис. 4. Початкові варіанти структури дерева D^*

На наступному етапі дослідження проведемо заміну одиниці на нуль в вершині v ЛДК та поставимо одиницю в вершину $v^{'}$. Позначимо ЛДК, яке буде отримано при цьому перетворені через D^{**}. У випадку (рис. 4) при переході від дерева D^* до ЛДК D^{**} в виразі $t^{i_1} + t^{i_2} + \cdots + t^{i_m}$ пропадає член t^n та з'являється замість нього член $t^{n^{'}}$. Так як $n^{'} < n$, то $t^{n^{'}} < t^n$. Таким чином, в цьому випадку вираз $t^{i_1} + t^{i_2} + \cdots + t^{i_m}$ зменшується. У випадку (рис. 4 (b)) при переході від дерева D^* та D^{**} крім переносу одиниці в вершину $v^{'}$ ще структуру (рис. 4 (b)) замінимо на структуру (рис. 4 (c)). Зрозуміло, що при такій заміні ЛДК D^* та D^{**} має однакову кількість вершин. Нагадаємо, що кінцеві вершини, в яких стоїть нуль до уваги не приймаються. Крім того, при вказаній заміні у виразі D^* замість члена t^n з'явиться сума $t^{n-1} + t^{n^{'}}$. З $n^{'} < n$ та $t \geq 2$ безпосередньо випливає наступне:

$$t^{n-1} + t^{n^{'}} \leq t^n. \tag{12}$$

Таким чином у випадку структури (рис. 4 (b)) вираз, який представлений деревом D^{**}, також не збільшується в порівняні з виразом представленим структурою ЛДК D^*.

З приведених вище міркувань можна зробити наступний висновок – серед всіх дерев D^* які мають одне і теж саме число N вершин, мінімальний вираз відповідає тому ЛДК в якому всі яруси крім можливо останнього є повними. Іншими словами, мінімальний в сенсі виразу що представляється ЛДК D^* має наступний вигляд – (рис. 5). Число i визначається наступним відношенням:

$$N = 2^i + 2^{i-1} + \cdots + 1 + \gamma = 2^{i+1} - 1 + \gamma. \tag{13}$$

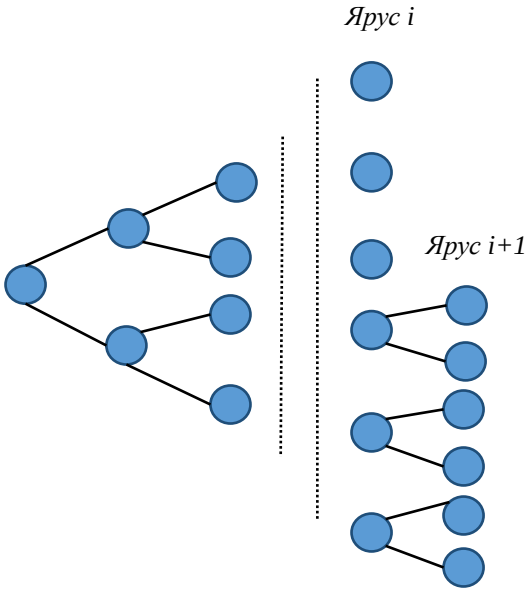

Рис. 5. Мінімальна структура ЛДК D^*

Зауважимо, що тут $\gamma < 2^i$. Число i визначається відповідно (13) позначимо через $r(N)$. Очевидно, що вираз, який відповідає дереву (рис. 5) не менше чим:

$$2^i * t^i. \tag{14}$$

Нехай $n_1 -$ безумовна кількість помилок початкової НВ. Отже можна зробити наступний висновок – що якщо число N задовольняє наступному співвідношенню:

$$(2t)^{r(N)} \geq n_1. \tag{15}$$

Тоді початкова НВ за допомогою процесу побудови ЛДК повністю класифікується не більше ніж за N кроків.

З (15) безпосередньо випливає, що процес побудови ЛДК при тільки що вказаній умові збігається за кількість кроків - $2^{m+1} - 1$, де $m = R(\frac{\log_2 n_1}{1+\log_2 y})$.

Висновки. Отже зважаючи на все вище сказане в даній роботі відносно складності побудови структур ЛДК – можна зафіксувати наступне:

Для умови слабкого розділення класів початкової НВ у випадку ЛДК (дерева класифікації), якщо на кожному $n -$ вому кроці побудови дерева класифікації відібрана елементарна ознака φ_n слабо розділяє множину (підмножину) об‘єктів початкової НВ, то в цьому випадку процес побудови дерева класифікації збігається відносно початкової НВ та закінчується не більше чим за $m - 1$ кроків, де $m -$ кількість всіх навчальних пар початкової НВ. Дерево класифікації (структура ЛДК) при умові сильного розділення класів множини об‘єктів початкової НВ, яке має m повних ярусів, рівнів (тобто випадок, коли на $i -$ товому ярусі стоять 2^{i-1} вершин), має $2^{m+1} - 1$ вершин – таким чином розпізнавання масиву початкової НВ при умові $(y \geq 1)$ за допомогою повного ЛДК відбувається не більш чим за $2^{m+1} - 1$ кроків, де m розраховується за допомогою виразу $m = R(\frac{\log_2 n_1}{1+\log_2 y})$.

Отже в роботі досліджене актуальне питання загальної складності структури логічного дерева класифікації (ЛДК) на основі концепції поетапної селекції наборів елементарних ознак – методів розгалуженого вибору ознак (можливих їх різнотипних множин та сполучень), яке для заданої початкової навчальної вибірки (масиву дискретної інформації) будує деревоподібну структуру (модель класифікації), з набору елементарних ознак (базових атрибутів) оцінених на кожному кроці схеми побудови моделі за даною вибіркою. Числова оцінка складності структури ЛДК дозволяє оцінити фінальну складність моделі дерева класифікації на основі початкової НВ та дозволяє ефективно провести процедуру обрізки (оптимізації та мінімізації) побудованої моделі.

Список використаної літератури

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

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

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

4. Василенко Ю.А Загальна оцінка мінімізації деревоподібних логічних структур / Ю.А. Василенко, Е.Ю. Василенко, І.Ф. Повхан, Ф.Г. Ващук // Науково технічний журнал “European Journal of Enterprise Technologies”. 2012, 1/4(55), – С. 29-33.

5. Povhan I. General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects / I. Povhan // Збірник наукових праць "Електроніка та інформаційні технології", Львів, 2019, Випуск 11, – С. 112-117.

6. Василенко Ю.А. Мінімізація логічних деревоподібних структур в задачах розпізнавання образів / Ю.А. Василенко, Е.Ю. Василенко, І.Ф. Повхан, М.Й. Ковач, О.Д. Нікарович // Науково технічний журнал “European Journal of Enterprise Technologies”. 2004, 3[9], – С. 12-16.

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

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

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

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

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

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

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

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

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

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

17. Povhan I. Generation of elementary signs in the general scheme of the recognition system based on the logical tree / I. Povhan // Збірник наукових праць "Електроніка та інформаційні технології". Lviv, 2019, Vol. 12. – С. 20-29.

18. Povhan I. Question of the optimality criterion of a regular logical tree based on the concept of similarity / I. Povhan // Збірник наукових праць "Електроніка та інформаційні технології". Lviv, 2020, Vol. 13. – С. 19-27.

References

1. Srikant, R., Agrawal, R. (1997) Mining generalized association rules. *Future Generation Computer Systems*, Vol.13, №2, 61–180.

2. Vasilenko, Y.A., Vasilenko, E.Y., Povkhan, I.F, Vashchuk, F.G. (2004) Conceptual basis of pattern recognition systems based on the method of branched feature selection. *Scientific and technical journal “European Journal of Enterprise Technologies”*, №7[1], 13-15.

3. Vasilenko, Y.A., Vashchuk, F.G, Povkhan, I.F. (2011) The problem of estimating the complexity of the logic trees, recognition, and a general method of optimization. *Scientific and*

technical journal "European Journal of Enterprise Technologies", 6/4(54), 24-28.

4. Vasilenko, Y. A., Povkhan, I.F., Vashchuk, F.G. (2012) General estimation of tree logical structures minimization. *Scientific and technical journal "European Journal of Enterprise Technologies",* 1/4 (55), 29-33.

5. Povkhan, I. (2019) General scheme for constructing the most complex logical tree of classification in pattern recognition of discrete objects. *Collection of scientific papers "electronics and information technology"*, Lviv, Issue 11, 112-117.

6. Vasilenko, Y.A., Vasilenko, E.J., Povkhan, I.F., Kovacs, M.J., Nickovic, O.D. (2004) Minimization of logic tree structures in pattern recognition problems. *Scientific and technical journal "European Journal of Enterprise Technologies",* 3[9], 12-16.

7. Laver, V.O., Povkhan, I.F. (2019) Algorithms for constructing logical classification trees in pattern recognition problems. *Scientific notes of Tauride national University. Series: technical Sciences,* Volume 30(69) No. 4 - 2019, 100-106.

8. Vtogoff, P.E. (2009) Incremental Induction of Decision Trees. *Machine Learning,* № 4, 61−186.

9. Povkhan, I.F. (2018) The problem of functional evaluation of the training sample in the problems of recognition of discrete objects. *Scientific notes of Taurida national University. Series: technical Sciences*, Volume 29(68) №.6, 217-222.

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

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

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

13. Subbotin S.A. (2019) Construction of decision trees for the case of low-information features, *Radio Electronics, Computer Science, Control*, No. 1, pp. 121–130.

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

15. Povkhan I.F. (2019) 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, pp. 169–174.

16. 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*, Vol. 30(69), No. 5, pp. 152–161.

17. Povhan I. (2019) Generation of elementary signs in the general scheme of the recognition system based on the logical tree. *Electronics and information technologies*. Lviv, 2019, Vol. 12. P. 20-29.

18. Povhan I. (2020) Question of the optimality criterion of a regular logical tree based on the concept of similarity. *Electronics and information technologies*. Lviv, 2020, Vol. 13. P. 19-27.