УДК 004.8:004.89:519.7 DOI: 10.31673/2412-4338.2020.035566

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

ОЦІНКА СКЛАДНОСТІ ПОБУДОВИ ЛОГІЧНОГО ДЕРЕВА КЛАСИФІКАЦІЇ ДЛЯ ДОВІЛЬНОГО ВИПАДКУ В УМОВАХ СИЛЬНОГО РОЗДІЛЕННЯ КЛАСІВ ПОЧАТКОВОЇ НАВЧАЛЬНОЇ ВИБІРКИ

Анотація: *В роботі пропонується оцінка складності побудованої структури логічного дерева класифікації для довільного випадку в умовах сильного розділення класів початкової навчальної вибірки. Принциповий розв'язок даного питання має визначальний характер, щодо оцінки структурної складності моделей класифікації (у вигляді деревоподібних конструкцій ЛДК/АДК) дискретних об'єктів для широкого спектру прикладних задач класифікації та розпізнавання в плані розробки перспективних схем та методів їх фінальної оптимізації (мінімізації) обрізки (post pruning) структури. Представлене дослідження має актуальність не лише для конструкцій (структур) логічних дерев класифікації, але дозволяє розширити саму схему оцінки складності і на загальний випадок структур алгоритмічних (моделей АДК) дерев класифікації (концепції дерев алгоритмів та дерев узагальнених ознак – ДУО). Досліджене актуальне питання концепції дерев рішень (дерев розпізнавання) – оцінка максимальної складності загальної схеми побудови логічного дерева класифікації на основі процедури 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 для випадку сильного розділення класів.*

Сучасні інформаційні системи та технології, засновані на математичних підходах (моделях) розпізнавання образів (структурах 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. Дослідження 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 (окремих ознак та їх сполучень) 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 інструментального аналізу (фінальної декомпозиції моделі).

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

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

ESTIMATION OF THE COMPLEXITY OF CONSTRUCTING A LOGICAL CLASSIFICATION TREE FOR AN ARBITRARY CASE IN CONDITIONS OF STRONG CLASS SEPARATION OF THE INITIAL TRAINING SAMPLE

The paper offers an estimation of the complexity of the constructed logical tree structure for classifying an arbitrary case in the conditions of a strong class division of the initial training sample. The principal solution to this question is of a defining nature, regarding the assessment of the structural complexity of classification models (in the form of tree-like structures of LCT/ACT) 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 post-pruning structure. The presented research is relevant not only

for constructions (structures) of logical classification trees, but also allows us to extend the scheme of complexity estimation to the General case of algorithmic structures (ACT models) of classification trees (the concept of algorithm trees and trees of generalized features - TGF). Is investigated the actual question of the concept of decision trees (tree recognition) – evaluation of the maximum complexity of the General scheme of constructing a logical tree based classification procedure of stepwise selection of sets of elementary features (they can be diverse sets and combinations) that for given initial training sample (array of discrete information) builds a tree structure (classification model), from a set of elementary features (basic attributes) are estimated at each stage of the scheme of the model in this sample for the case of strong separation of classes.

Modern information systems and technologies based on mathematical approaches (models) of pattern recognition (structures of 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, and this is due to the fact that this approach allows you to eliminate a set of existing disadvantages of well-known classical methods, schemes and achieve a fundamentally new result. The research 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 (individual features and their combinations) 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 instrumental analysis (the final decomposition of the model).

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

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

ОЦЕНКА СЛОЖНОСТИ ПОСТРОЕНИЯ ЛОГИЧЕСКОГО ДЕРЕВА КЛАССИФИКАЦИИ ДЛЯ ПРОИЗВОЛЬНОГО СЛУЧАЯ В УСЛОВИЯХ СИЛЬНОГО РАЗДЕЛЕНИЯ КЛАССОВ НАЧАЛЬНОЙ УЧЕБНОЙ ВЫБОРКИ

Аннотация: *В работе предлагается оценка сложности построенной структуры логического дерева для классификации для произвольного случая в условиях сильного разделения классов начальной обучающей выборки. Принципиальное решение данного вопроса имеет определяющий характер, относительно оценки структурной сложности моделей классификации (в виде древовидных конструкций ЛДК/АДК) дискретных объектов для широкого спектра прикладных задач классификации и распознавания в плане разработки перспективных схем и методов их финальной оптимизации (минимизации) обрезки (post-pruning) структуры. Представленное исследование имеет актуальность не только для конструкций (структур) логических деревьев классификации, но позволяет расширить саму схему оценки сложности и на общий случай алгоритмических структур (моделей АДК) деревьев классификации (концепции деревьев алгоритмов и деревьев обобщенных признаков – ДОП). Исследован актуальный вопрос концепции деревьев решений (деревьев распознавание) – оценка максимальной сложности общей схемы построения логического дерева классификации на основе процедуры поэтапной селекции наборов элементарных признаков (возможных их разнородных 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 для случая сильного разделения классов.*

Современные информационные системы и технологии, основанные на математических подходах (моделях) распознавание образов (структурах логических и алгоритмических деревьев 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. Исследование посвящена проблематике моделей деревьев классификации (деревьев решений), и предлагает оценку

сложности структур логических деревьев (моделей деревьев классификации), которые складываются из отобранных и ранжированных наборов элементарных признаков (отдельных признаков и их сочетаний) 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 инструментального анализа (финальной декомпозиции модели).

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

Вступ. Задачі, які об'єднуються тематикою розпізнавання образів, дуже різноманітні та виникають у сучасному світі в усіх сферах економіки та соціального контенту діяльності людини, що приводить до необхідності побудови та дослідження математичних моделей відповідних систем. Станом на зараз не існує універсального підходу до їх розв'язання, запропоновано декілька досить загальних теорій та підходів, що дозволяють вирішувати багато типів (класів) задач. Причому прикладні застосування відрізняються досить великою чутливістю до специфіки самої задачі або предметної області застосування [1-7]. На сьогоднішній день актуальні різні підходи до побудови систем розпізнавання (СР) у вигляді логічних дерев класифікації (ЛДК), причому інтерес до методів розпізнавання, які використовують ЛДК, викликаний рядом корисних властивостей, якими вони володіють. З одного боку, складність класу функцій розпізнавання (ФР) у вигляді моделей ЛДК, при визначених умовах, не перевищують складності класу лінійних функцій розпізнавання (простішого з відомих). З другого боку, ФР у вигляді дерев класифікації дозволяють виділити в процесі класифікації як причинно - наслідкові зв'язки (та однозначно врахувати їх у подальшому), так і фактори випадковості або невизначеності, тобто врахувати одночасно і функціональні і стохастичні відношення між властивостями та поведінкою всієї системи. Причому відомо, що процес класифікації нових, таких що до сих пір не зустрічалися об'єктів світу багатьох тварин і людей (за виключенням об'єктів, інформація про які передається генетичним шляхом (наслідковим), а також в деяких інших випадках), відбувається за так званим логічним деревом рішень (в зв'язку з нейромережевою концепцією). В даний час існує декілька незалежних загальних підходів (концепцій) для вирішення задачі класифікації в загальній постановці, причому розробку різних концепцій, підходів, методів, моделей, які охоплюють загальну проблематику теорії штучного інтелекту та інформаційних систем [5].

Постановка завдання. Нехай, на початку задачі задана начальна вибірка (НВ) стандартного (недетермінованого) випадку 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 об'єкти $x_i \in G$ ($G -$ деяка множина сигналів x), а функція розпізнавання (ФР) $f_R(x_i) \in \{1,2, \dots, k\}, (i = 1,2, \dots, m)$.

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

Отже можна зробити висновок, що НВ – це сукупність (послідовність) деяких наборів (навчальних наборів), причому кожний набір – це сукупність значень деяких ознак та значень деяких функцій на цьому наборі. Можна зробити висновок, що сукупність значень ознак – це деяке зображення, а значення функції (функції розпізнавання) відносить це зображення до відповідного образу (класу) [3].

Зафіксуємо, що зазвичай стоїть загальна задача побудови моделі логічного дерева класифікації (конструкції 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 в

межах даного дослідженні буде цікавити складність такої структури на етапі побудови моделі ЛДК для умови сильного розділення класів початкової НВ.

Аналіз досліджень і публікацій. Представлене дослідження слідує в циклі робіт, які присвячені загальним питанням деревоподібних схем, конструкцій ЛДК/АДК (моделей) розпізнавання (класифікації) дискретних об‘єктів в теорії штучного інтелекту широкого спектру прикладних задач [1-4,8-12] та являє собою пряме продовження дослідження з роботи [13]. В роботах піднімаються актуальні питання синтезу, використання, аналізу, декомпозиції та оптимізації конструкцій логічних та алгоритмічних дерев класифікації. З робіт [8] відомо, що функція розпізнавання (результуюче правило класифікації), схема яка побудоване довільним методом або алгоритмом розгалуженого вибору ознак, дерев рішень має деревоподібну логічну структуру у вигляді класичного граф - схемного представлення, причому довільна структура ЛДК складається з вершин (ознак або атрибутів), які групуються по ярусам. Відмітимо що всі вершини (розгалуження) отримані на певному кроці (етапі) генерації дерева розпізнавання (моделі ЛДК) [6]. Принципово важливою задачею для концепції алгоритмічних дерев (моделей АДК), яка виникає з роботи [12] є задача побудови дерев розпізнавання, які будуть представлятися фактично деревом (графом) незалежних алгоритмів або деревом узагальнених ознак (ДУО). Відмітимо, що на відміну від інших існуючих методів розпізнавання, головною особливістю деревоподібних схем (граф – схемних класифікаторів), конструкцій ЛДК/АДК є те, що важливість окремих ознак (груп ознак, їх сполучень чи автономних алгоритмів класифікації) визначається відносно функції розпізнавання, яка задає розбиття об‘єктів початкової НВ на класи (образи) [14-17]. В роботі [18] розглядаються принципові питання стосовно побудови структур дерев рішень для випадку малоінформативних ознак та їх сполучень. Відома характеристика структур (моделей дерев класифікації) ЛДК/АДК виконувати одномірне розгалуження структури для аналізу впливу (важливості, якості) окремих змінних дає можливість працювати зі змінними різних типів у вигляді предикатів (наборів предикатів) [19-23]. Відмітимо, що для випадку дерева алгоритмів (структур АДК) – відповідними автономними алгоритмами класифікації та розпізнавання [24]. Домінуючими підходами, є алгоритмічні схеми на основі методів CART (спрямованих для розв‘язку задач класифікації та регресивного аналізу), а також програмні системи (ПС) на основі схеми C4.5 та її сучасних модифікації (для розв‘язку задач розпізнавання та класифікації) та ID3 – причому всі вони характеризуються наступними критеріями розгалуження та зупинки процедури побудови дерева:

1) ID3 схема базується на використанні обмеженого ентропійного критерію – структура ЛДК будується до тих пір, поки для кожної результуючої вершини (листа дерева) не залишаться лише об‘єкти одного фіксованого класу, або доки сама процедура розгалуження в дереві, що будується дає зменшення початкового ентропійного критерію.

2) C4.5/C5.0 схема базується на відомому критерії *Gain-Ratio* (нормативний ентропійний критерій), причому в якості критерії зупинки процедури розгалуження (побудови дерева) використовується обмеження на кількість об‘єктів для результуючої вершини (листа структури ЛДК). Відмітимо, що процедура відсіканні в структурі ЛДК проводиться за схемою Error-Based Pruning, яка базується на загальній оцінці здатності узагальнення для прийняття рішення щодо видалення гілок та вершин конструкції дерева класифікації. В даному підході обробка пропущених атрибутів здійснюється за схемою в якій ігноруються об‘єкти з відсутніми значеннями в процедурі розрахунку критерію розгалуження структури ЛДК, а на наступному етапі відносить такі об‘єкти в обидва піддерева з визначеними атрибутами.

3) CART схема в своїй роботі використовує критерії Джині, причому процедура відсіканні в структурі ЛДК проводиться за схемою *Cost-Complexity Pruning*, а для випадку наявних пропусків атрибутів використовується базова схема сурогатних предикатів. Відмітимо, що алгоритм ID3 є однією з найпростіших схем для отримання дерев рішень з категоріальними класами та атрибутами (на основі ентропійного критерію), причому саме на

основі ID3 і був написаний пізніше алгоритм C4.5. Хоча існує достатньо багато реалізаційних схем різних методів та підходів дерев рішень (дерев класифікації), одною з найбільш вживаних та такою що забезпечує необхідну ефективність є алгоритмічна схема C5.0 (подальший розвиток концепції алгоритму C4.5), причому вона стала по факту галузевим стандартом для побудови моделей дерев класифікації – оскільки підходить для більшості типів задачах безпосередньо прикладного характеру в сегменті аналізу різнотипної інформації. В порівнянні з більш досконалими та складними моделями машинного навчання (наприклад – концепцією нейронних мереж) дерева класифікації в рамках схеми C5.0 зазвичай забезпечують не гіршу ефективність та точність, але в той самий час більш підходять для зовнішнього аналізу та фінальної покрокової інтерпретації (виділенні правил класифікації), тобто є достатньо простими та зрозумілими для експерта та спостерігача [5].

Відмітимо, що методи дерев класифікації допускають принципову можливість в якості ознак, (вершин структури) дерева класифікації використовувати не тільки окремі атрибути (ознаки) об‘єктів їх сполучення та набори, але і окремі незалежні алгоритми розпізнавання (оцінені за даними НВ) – а на виході буде отримане нова структура – АДК [24].

Метою дослідження є оцінка складності процедури (схеми) побудови логічного дерева класифікації (моделі ЛДК) для випадку сильного розділення класів початкової навчальної вибірки методів розгалуженого вибору ознак.

Результати дослідження. Для методів логічних дерев класифікації на основі селекції елементарних ознак (методів розгалуженого вибору ознак) дана оцінка складності процедури синтезу логічного дерева класифікації для умови сильного розділення класів масиву початкової НВ. Верхня оцінка складності структури дерева класифікації дозволяє оцінити загальну складність моделі (конструкції) дерева класифікації на основі початкової НВ та дозволяє ефективно провести процедуру фінальної обрізки (оптимізації та мінімізації) побудованої конструкції.

Виклад основного матеріалу. З першої частини даного дослідження [13] відомо, що 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 $m = R(\frac{\log_2 n_1}{1+\log_2 y})$. Зауважимо, що під величиною y розуміється якість (ефективність) класифікації (розділення множин початкової НВ) фіксованої елементарної ознаки, а під $R(x)$ розуміється заокруглення числа x до найблищого цілого числа, яке перевищує x. Наприклад $Q(1.2) = 2, Q(3.7) = 4, Q(4.1) = 5$.

Дослідимо в даній роботі останній випадок, коли виконується умова $(1 < y < 2)$. Для цього зафіксуємо число N та припустимо, що логічне дерево (структура ЛДК) D^* представляє собою мінімальний вираз серед всіх дерев, які мають N вершин. Покажемо спочатку, що в кожному своєму кінці дерево D^* має вигляд – (рис. 1 (a)).

Припустимо протилежне, а саме, що в деякому кінці дерева D^* має вигляд – (рис. 1 (b)). Тоді від конструкції на (рис. 1 (b)) можна перейти до структури – (рис. 1 (c)). Причому при переході від структури (рис. 1 (b)) до (рис. 1 (c)) загальна кількість вершин не змінюється (вершини, які містять нуль не рахуються). При переході від (рис. 1 (b)) до (рис. 1 (c)) у виразі вигляду $t^i + t^i$ міняється на t^{i+1}, де $i-$ номер останнього ярусу дерева D^*. Так, як за початковою умовою $(y < 2)$, то $t^i + t^i = 2t^i < t^{i+1}$. Таким чином, якщо в дереві D^* була би

конструкція (рис. 1 (b)) тоді вираз, який представляє дане дерево можна зменшити, що протирічить початковому припущенню.

Покажемо тепер, що дерево не містить вершин в яких стоїть одиниця в ярусах з номерами $j \le i-2$ (тут $i-$ номер останнього ярусу дерева D^*). Дійсно, припустимо, що в дереві D^* є вершина v', яка розташована в ярусі $j \le i-2$ і в ній стоїть одиниця. Нехай $v-$ деяка вершина, в якій стоїть одиниця та яка знаходиться на $i-$ товому ярусі. Нехай це буде вершина, яка фігурує на (рис. 1 (a)).

Зробимо наступні перетворення дерева D^*: конструкцію (рис. 1 (a)) замінимо на конструкцію (рис. 2 (a)), а конструкцію (рис. 2 (b)) на конструкцію (рис. 2 (c)).

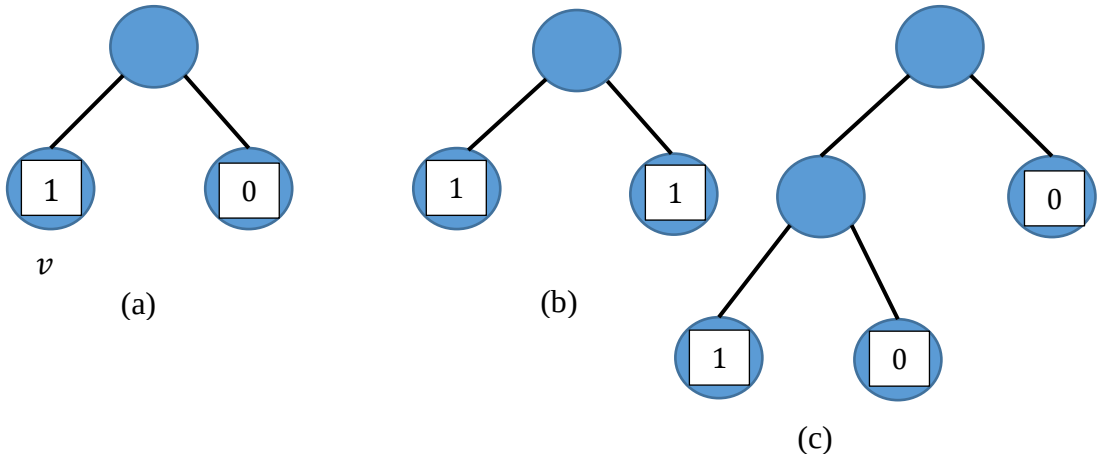

Рис. 1. Кінцева структура дерева D^*

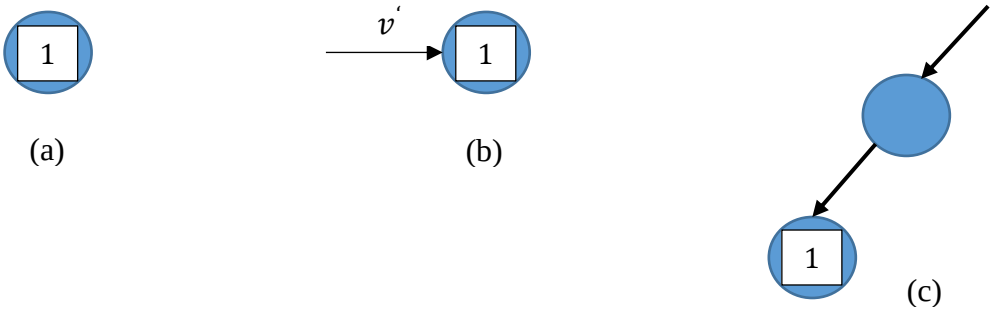

Рис. 2. Перехідна структура дерева D^*

Очевидно, що кількість вершин при такий заміні не міняється. При цій зміні дерева D^*, вираз $t^i + t^j$ замінить ся сумою $t^{i-1} + t^{j+1}$. Але:

$$\left(t^i + t^j\right) - \left(t^{i-1} + t^{j+1}\right) = t^{i-1}(t-1) - t^j(t-1) = (t-1)\left(t^{i-1} - t^j\right) > 0.$$

Останнє випливає з того, що $j \le i-2$. Таким чином, вираз який представляє дерево D^*, можна було би зменшити, що протирічить мінімальності дерева D^*.

З тільки що проведеного випливає, що в кінці дерева D^* має бути наступна структура – (рис. 3 (a)). Причому, кожна кінцева вершина на (рис. 3 (a)) має знаходитись або в останньому, або в передостанньому ярусі D^*. Легко показати , що всі вершини, які стоять на рівні вершини Г (рис. 3 (a)) або вище її, не є нульовими, тобто в цих вершинах нуль не стоїть. Дійсно, нехай вершина Г стоїть в $j-$ товому ярусі, а ліва кінцева з одиницею стоїть в $i-$ товому ярусі (рис. 2 (a)). Припустимо, що в $s-$ товому $(s \le j)$ ярусі стоїть нульова вершина. Цю вершину позначимо через v'. Проведемо процедуру заміни конструкції (рис. 3 (a)) на конструкцію (рис. 3 (b)), а до вершини v' добудуємо таку конструкцію (рис. 3 (c)).

Цю зміну можна інтерпретувати наступним чином – в конструкції (рис. 3 (a)) було прибрано ліву гілку (тобто $(i-j)$ вершини) і цю гілку добудували до вершини v'. Причому,

перша зверху вершина лівої гілки (рис. 3 (а)) суміщається з вершиною v'. При цій зміні виразу t^i, яке стоїть в дереві D^* заміняємо виразом $t^{s+(i-j)-1} = t^{i+(s-j)-1} < t^i$. Таким чином, при існуванні вищевказаної вершини вираз, яке представляється деревом D^*, зменшилося би. Отже, в дереві D^* всі вершини, які стоять в j- товому і вище ярусах, не є нульовими.

Оцінимо тепер різницю $(i-j)$. Припустимо, що в конструкції (рис. 2 (а)) дві гілки мають однакову кількість вершин (яка дорівнює $(i-j)$). Замінимо при цьому припущенні конструкцію (рис. 3 (а)) структурою (рис. 4 (а)).

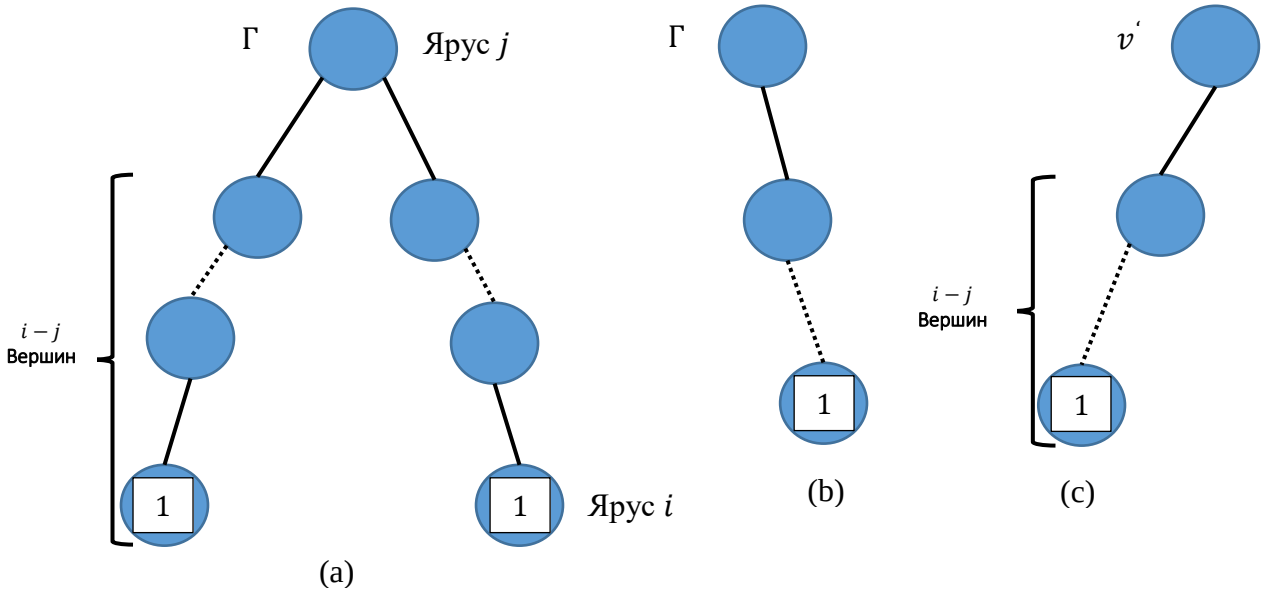

Рис. 3. Другий етап перехідної структури дерева D^*

При переході від конструкції (рис. 2 (а)) до структури (рис. 4 (а)) спочатку видаляємо праву гілку та добудовуємо її до кінцевої вершини лівої гілки. При цьому початок стрілки, яка входить в першу вершину правої гілки структури (рис. 3 (а)) добудовується до кінцевої вершини лівої гілки цієї конструкції. При такому переході вираз $t^i + t^i$ в початковому дереві замінюється виразом $t^{i+(i-1)}$. Зменшення виразу, яке представляє дерево D^*, не відбудеться тоді, коли виконується співвідношення $t^{i+(i-j)} \geq 2t^i$.

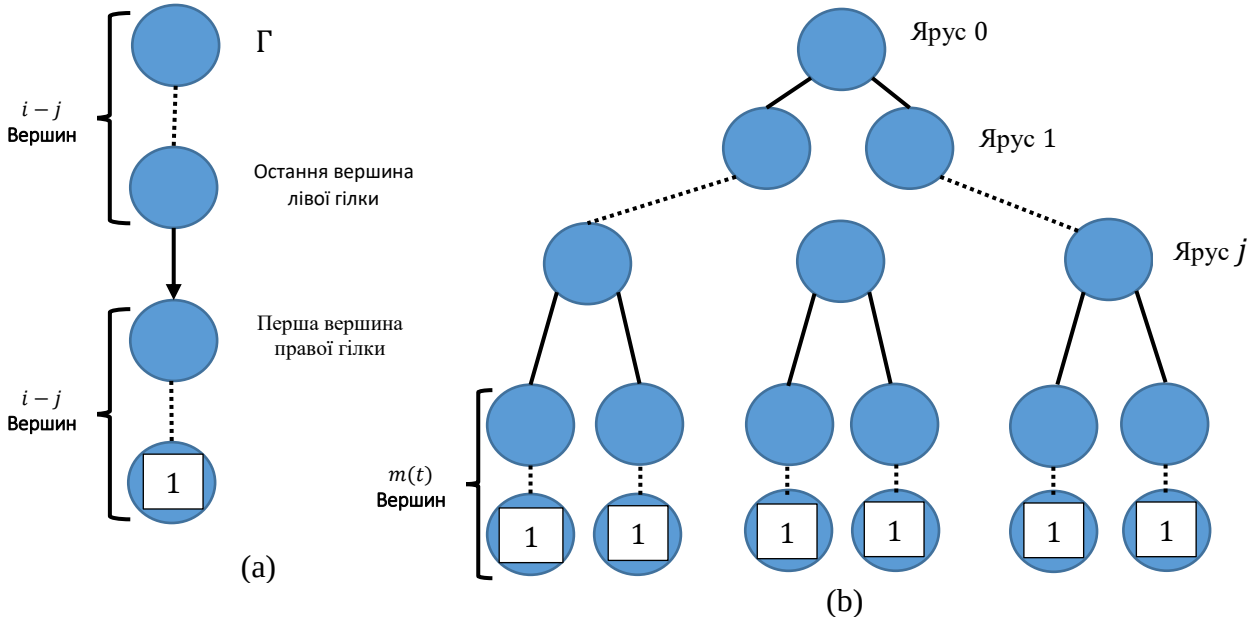

Рис. 4. Третій етап перехідної структури дерева D^*

З останнього можна отримати наступний вираз:

$$t^{i-j} \geq 2. \tag{2}$$

Найменше ціле число q, яке задовольняє співвідношенню $t^q \geq 2$, позначимо через $m(t)$. Число $m(t)$ можна представити в такому вигляді:

$$m(t) = Q\left(\frac{1}{\log_2 t}\right) = \frac{1}{\log_2 t} + \alpha`. \tag{3}$$

Зауважимо, що тут $0 \leq \alpha` < 1$ та $Q(x) -$ заокруглення числа x до найблищого цілого яке не перевищує x.

Таким чином можна показати, що якщо в конструкції (рис. 3 (a)) одна гілка коротша за іншу, то при переході до конструкції (рис. 4 (a)) зменшення виразу, який представляє дерево D^* не відбудеться при умові, що $t^{i-j} \geq 1 + t$. Так як $t > 1$, то з умови $t^{i-j} \geq 1 + t$ випливає умова (2).

Резюмуючи все вище сказане можна зафіксувати основні властивості мінімального логічного дерева:

1) Всі яруси цього логічного дерева починаючи з самого верхнього (тобто нульового) та закінчуючи деяким j- товим ярусом заповнені вершинами. Причому в усіх цих вершинах не стоїть ні нуль ні одиниця.

2) Починаючи від кожної вершини j- тового ярусу дерево D^* має структуру яка зображена на (рис. 3 (a)) або на (рис. 3 (b)) причому довжина кожної гілки структури (рис. 2 (a)), (рис. 3 (b)) не менше чим $m(t)$ та не більше чим $2m(t) + 1$.

3) Кожна вершина в якій знаходиться одиниця стоїть в останньому або передостанньому ярусі дерева D^*.

Розглянемо дерево на (рис. 4 (b)). В ньому всі яруси починаючи з нульового та закінчуючи j- товим заповнені вершинами. Від кожної вершини v деякого j- тового ярусу відходить наступна конструкція – (рис. 5 (a)).

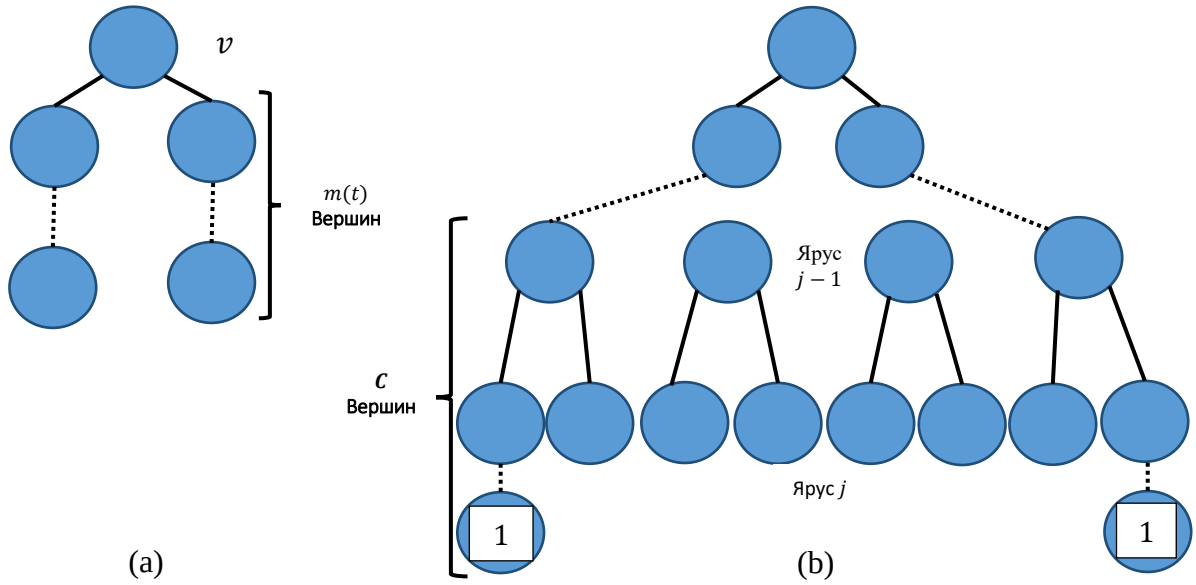

Рис. 5. Четвертий етап перехідної структури дерева D^*

Виходячи з вищеперерахованих властивостей мінімального дерева, можна зробити висновок, що структура мінімального дерева D^*, яке має стільки вершин, скільки їх у дерева на (рис. 4 (b)) має бути наступною – (рис. 5 (b)). Від кожної вершини j- тового ярусу відходить

конструкція вигляду (рис. 2 (a)) або (рис. 2 (b)). Причому на j- тового ярусу знаходиться хоча би одна вершина з якої відходить структура вигляду (рис. 2 (a)). Приймаючи до уваги конструкцію дерева (рис. 5 (a)) та (рис. 5 (b)) робимо наступний висновок – що $c \geq m(t)$. Зауважимо, що тут $c -$ кількість вершин самого короткого шляху, який веде від вершин j-тового ярусу до кінцевих вершин дерева – (рис. 5 (b)).

Очевидно, що дерево (рис. 5 (b)) представляє вираз, який не менше чим вираз $t^{j-1} - t^{j+m(t)-1} = (2t)^{j-1} * t^{m(t)}$. Дерево (рис. 4 (b)) має кількість вершин яке дорівнює:

$$(t^{j+1} - 1) + t^{j+1} * m(t) = 2^{j+1}(m(t) + 1) - 1.$$

З тільки що проведених міркувань можна зробити наступний висновок, що кожне логічне дерево, яке має $t^{j+1}(m(t) + 1) - 1$ вершин обчислює вираз, яке не менше $(2t)^{j-1} * t^{m(t)}$. Нехай $n_1 -$ безумовна кількість помилок початкової навчальної вибірки. Очевидно, що навчальна вибірка буде повністю розпізнана, якщо виконується умова:

$$(2t)^{j-1} * t^{m(t)} \geq n_1. \tag{4}$$

Так як $t^{m(t)} > 2$, то умова (4) виконується при j, яке задовольняє наступній умові:

$$(2t)^{j-1} \geq \frac{n_1}{2}. \tag{5}$$

Найменше ціле число j, яку задовольняє співвідношенню (5), можна представити у вигляді:

$$j = Q\left(\frac{\log_2 \frac{n_1}{2}}{1+\log_2 t}\right) = \frac{\log_2 \frac{n_1}{2}}{1+\log_2 t} + \alpha. \tag{6}$$

Зауважимо, що тут $0 \leq \alpha < 1$.

Приймаючи до уваги (3) та (6), можна записати наступне:

$$\begin{aligned} 2^{j+1}(m(t) + 1) - 1 = 2 * 2^{j}(m(t) + 1) - 1 = \\ = 2 * 2^{\frac{\log_2 \frac{n_1}{2}}{1+\log_2 t}+\alpha} * \left(\left(\frac{1}{\log_2 t} + \alpha^{'}\right) - 1\right) \leq \\ \leq 2 * 2^{\alpha}\left(\frac{n_1}{2}\right)^{\frac{1}{1+\log_2 t}}\left(\frac{1}{\log_2 t} + 1\right) \leq \\ \leq 4 * \left(1 + \frac{1}{\log_2 t}\right)(n_1)^{\frac{1}{1+\log_2 t}}. \end{aligned} \tag{7}$$

Звернемо увагу, що з (7) безпосередньо випливає наступний базовий висновок.

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

Процес побудови структури ЛДК (дерева класифікації), при умові сильного розділення класів множини об‘єктів початкової НВ, яке має m повних ярусів та задовольняє базовій умові $1 < y < 2$, повністю розпізнає початкову навчальну вибірку типу (1) не більше чим за $4 * \left(1 + \frac{1}{\log_2 t}\right)(n_1)^{\frac{1}{1+\log_2 t}}$ кроків, де $n_1 -$ безумовна кількість помилок початкової навчальної вибірки.

Отже в представленій роботі, спираючись на результати з [13] досліджене актуальне питання загальної складності структури логічного дерева класифікації (моделі ЛДК) на основі концепції поетапної селекції наборів елементарних ознак – методів та алгоритмів розгалуженого вибору ознак (можливих їх різнотипних наборів 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 (фінальної оптимізації та мінімізації) побудованої моделі.

Список використаної літератури

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

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

3. 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 2018. С. 217-222.

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

5. Alpaydin E. Introduction to Machine Learning. London: The MIT Prs, 2010. 400p.

6. Painsky A., Rosset S. Cross-validated variable selection in tree-based methods improves predictive performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, Vol. 39, № 11. P. 2142-153.

7. Srikant R. Mining generalized association rules. Future Generation Computer Systems, 1997, Vol. 13, №2. P. 161-180.

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

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

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

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

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

13. Повхан І.Ф. Оцінка загальної складності процедури побудови бінарного логічного дерева класифікації для довільного випадку. Науковий журнал: Телекомунікаційні та інформаційні технології, 2020, №2 (67) 2020. С. 100-111.

14. Vtogoff P.E. Incremental Induction of Decision Trees. Machine Learning, 2009, №4. P. 161-186.

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

16. Povhan I. Designing of recognition system of discrete objects. 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2016, Ukraine. Lviv,

2016. P. 226-231.

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

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

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

20. Subbotin S.A. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence. Radio Electronics, Computer Science, Control, 2014, №1. P. 120-128.

21. Subbotin S.A. Methods of sampling based on exhaustive and evolutionary search. Automatic Control and Computer Sciences, 2013, Vol. 47, №3. P. 113-121.

22. De Mántaras R.L. A distance-based attribute selection measure for decision tree induction. Machine learning, 1991, Vol. 6. №1. P. 81-92.

23. Miyakawa M. Criteria for selecting a variable in the construction of efficient decision trees. IEEE Transactions on Computers, 1989, Vol. 38, №1. P. 130-141.

24. Povkhan I.F. Features of synthesis of generalized features in the construction of recognition systems using the logical tree method. Information technologies and computer modeling ITKM-2019: materials of the international scientific and practical conference, Ivano- Frankivsk, May 20–25, 2019, Ivano-Frankivsk, 2019. P. 169-174.

References

1. 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.

2. 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.

3. 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.

4. 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.

5. Alpaydin E. (2010). Introduction to Machine Learning. London: The MIT Press, 400 p.

6. Painsky A., Rosset S. (2017). Cross-validated variable selection in tree-based methods improves predictive performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Vol. 39, № 11. P. 2142-2153.

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

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. Povkhan I.F. (2020) Question of general complexity of the procedure for constructing a binary logical classification tree for an arbitrary case. *Scientific journal: Telecommunications and information technologies*, No.2 (67), 2020, 100-111.

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

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

16. 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.

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

18. 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.

19. 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.

20. Subbotin S. A. (2014). Methods and characteristics of locality-preserving transformations in the problems of computational intelligence. *Radio Electronics, Computer Science, Control*. № 1. P. 120-128.

21. Subbotin S. A. (2013). Methods of sampling based on exhaustive and evolutionary search. *Automatic Control and Computer Sciences*. Vol. 47, № 3. P. 113-121.

22. De Mántaras R. L. (1991). A distance-based attribute selection measure for decision tree induction. *Machine learning.* Vol. 6, № 1. P. 81-92.

23. Miyakawa M. (1989). Criteria for selecting a variable in the construction of efficient decision trees. *IEEE Transactions on Computers*. Vol. 38, № 1. P. 130-141.

24. Povkhan I.F. (2019). Features of synthesis of generalized features in the construction of recognition systems using the logical tree method. *Information technologies and computer modeling ITKM-2019 : materials of the international scientific and practical conference, Ivano- Frankivsk,* May 20–25, 2019. Ivano-Frankivsk, 2019. P. 169-174.