

**Міністерство освіти і науки, молоді та спорту України
Харківський національний університет радіоелектроніки**

ГОФМАН ЄВГЕНІЙ ОЛЕКСАНДРОВИЧ

УДК 004.89

**МЕТОДИ ПОБУДОВИ ДЕРЕВ РОЗВ'ЯЗКІВ
В ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ ДІАГНОСТУВАННЯ**

05.13.23 – системи та засоби штучного інтелекту

**Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук**

Харків – 2012

Дисертацією є рукопис.

Робота виконана у Запорізькому національному технічному університеті Міністерства освіти і науки, молоді та спорту України.

Науковий керівник – кандидат технічних наук, доцент,
Субботін Сергій Олександрович,
Запорізький національний технічний університет,
доцент кафедри програмних засобів

Офіційні опоненти: доктор технічних наук, професор,
Кучеренко Євген Іванович,
Харківський національний університет радіоелектроніки,
професор кафедри штучного інтелекту;

доктор технічних наук, професор,
Скобцов Юрій Олександрович,
Донецький національний технічний університет,
завідувач кафедри автоматизованих систем управління.

Захист відбудеться «___» _____ 20__ р. о ____ годині на засіданні спеціалізованої вченої ради Д. 64.052.01 у Харківському національному університеті радіоелектроніки за адресою 61166, м. Харків, пр. Леніна, 14.

З дисертацією можна ознайомитись у бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Леніна, 14.

Автореферат розісланий «___» _____ 20__ р.

Учений секретар
спеціалізованої вченої ради

Є. І. Литвинова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Для побудови систем автоматичного розпізнавання образів і діагностування необхідно за набором експериментально отриманих даних синтезувати розпізнавальні моделі, що, як правило, вимагає розв'язання завдань інтелектуального аналізу даних, зокрема відбору інформативних ознак, кластерного аналізу, індукції нечітких правил. Для цього доцільно використовувати дерева розв'язків (ДР), які дозволяють виявляти закономірності в емпіричних даних на основі індуктивного навчання за прецедентами, володіють високими апроксимаційними й узагальнювальними властивостями, а також характеризуються високою інтерпретабельністю, що дозволяє не тільки розраховувати значення вихідного параметра, але й пояснювати, яким чином воно отримано.

Значний внесок у розвиток інтелектуального аналізу даних внесли М. Брамер, Л. Ванг, Л. Заде, Т. Сугено, Є. В. Бодянський, Ю. І. Журавльов, Є. І. Кучеренко, О. Г. Руденко та ін., у створення та дослідження теорії дерев розв'язків – Л. Брейман, В. Й. Донської, Д. Р. Квінлан, Л. Рокач та ін.

Однак існуючі методи побудови ДР, як правило, використовують жадібну стратегію пошуку, застосування якої часто призводить до синтезу моделей, що не забезпечують прийнятний рівень узагальнення. Крім того, при обробці вибірок, що характеризуються великою кількістю ознак, виникають проблеми побудови компактних дерев, що забезпечують прийнятну точність розпізнавання. Тому актуальною є розробка нових і модифікація існуючих методів синтезу ДР, що усувають недоліки існуючих методів та дозволяють підвищити точність, швидкість побудови ДР, а також рівні їх узагальнення й інтерпретабельності.

Зв'язок роботи з науковими програмами, планами, темами. Робота виконана відповідно до координаційних планів Міністерства освіти і науки, молоді та спорту України та Запорізького національного технічного університету (ЗНТУ) відповідно до державної науково-технічної програми “Нові вітчизняні інтелектуальні комп'ютерні засоби”, у рамках науково-дослідних робіт (НДР):

– держбюджетної НДР кафедри програмних засобів ЗНТУ “Інформаційні технології автоматизації розпізнавання образів і прийняття рішень для діагностування в умовах невизначеності на основі гібридних нечіткологічних, нейромережевих і мультиагентних методів обчислювального інтелекту” (№ держ. реєстрації 0109U007673);

– держбюджетної НДР кафедри радіотехніки і телекомунікацій ЗНТУ “Методи, моделі та пристрої прийняття рішень в системах розпізнавання образів” (№ держ. реєстрації 0111U000059);

– госпдоговірної НДР ТОВ “Ітекс” “Методи, моделі та програмні засоби автоматизації побудови систем технічного діагностування” (№ держ. реєстрації 0111U001740);

– госпдоговірної НДР ТОВ “Макфорт” “Інформаційні технології автоматизації підтримки прийняття рішень на основі інтелектуальних обчислень” (№ держ. реєстрації 0111U003965).

В процесі роботи над НДР автор дисертаційної роботи брав участь як виконавець, досліджував і розробив нові методи синтезу дерев розв'язків, а також розв'язав практичне завдання діагностування автотранспортних засобів.

Мета і задачі дослідження. Метою дисертаційної роботи є розробка методів ідентифікації дерев розв'язків шляхом використання стохастичного підходу та додаткової інформації про досліджувані об'єкти, що дозволить підвищити точність, швидкість побудови та роботи, рівні узагальнення й інтерпретабельності синтезованих дерев розв'язків.

Для досягнення поставленої мети в роботі вирішувалися такі завдання:

- аналіз процесу та методів побудови дерев розв'язків в інтелектуальних системах;
- розробка еволюційного методу синтезу дерев розв'язків;
- створення методу побудови дерев розв'язків з використанням теорії функцій довіри для застосування синтезованих моделей в умовах неповноти даних;
- розробка методу індукції нечітких правил на основі дерев розв'язків для побудови інтерпретабельних баз нечітких правил;
- створення методу синтезу нейро-нечітких мереж з використанням дерев розв'язків;
- розробка програмного забезпечення для синтезу моделей на основі дерев розв'язків за допомогою запропонованих методів;
- експериментальне дослідження розроблених методів і програмних засобів на основі розв'язання практичних задач.

Об'єкт дослідження – процес синтезу дерев розв'язків.

Предмет дослідження – методи індуктивного навчання дерев розв'язків в інтелектуальних системах діагностування.

Методи дослідження. Для розв'язання задач, поставлених у роботі, використовувалися методи: еволюційного та мультиагентного пошуку – для формування розв'язків у процесі структурно-параметричного синтезу ДР; теорії дерев розв'язків – як базис для розробки нових методів розв'язання задач інтелектуального аналізу даних; теорії функцій довіри – для подання невизначеності параметрів задач класифікації; математичної статистики – для створення системи критеріїв оцінювання й аналізу ефективності розроблених методів синтезу дерев розв'язків.

Наукова новизна одержаних результатів дисертації полягає у тому, що:

- вперше запропоновано еволюційний метод синтезу дерев розв'язків, який базується на стохастичному підході, не використовує жадібну стратегію пошуку та дозволяє будувати дерева розв'язків з достатніми узагальнювальними й апроксимаційними властивостями за рахунок скорочення вузлів, що підвищує інтерпретабельність синтезованих моделей;
- вперше запропоновано метод синтезу нейро-нечітких мереж на основі дерев розв'язків, який характеризується тим, що для побудованого за навчальною вибіркою ДР автоматично формується нечітке розбиття простору ознак та виділяються правила, на основі яких створюється структура мережі й визначаються значення її параметрів, що дозволяє синтезувати моделі з невеликою кількістю нейроелементів та високим рівнем узагальнення;
- удосконалено метод ідентифікації дерев розв'язків для індукції нечітких правил, який, на відміну від відомих методів, використовує додаткові функції перетворення при розростанні дерева, згладжує дерево для його усікання та використовує критерії сусідства при перетворенні дерев розв'язків, що дозволяє одержувати більш інтерпретабельні бази нечітких правил;

– дістав подальшого розвитку метод синтезу дерев розв'язків ID3 шляхом врахування пігністичних ймовірностей віднесення екземплярів до класів на основі теорії функцій довіри, що дозволяє виконувати класифікацію екземплярів в умовах невизначеності або неповноти даних.

Практичне значення одержаних результатів полягає у тому, що:

– розроблені методи синтезу дерев розв'язків доведено до програмної реалізації, що дозволило створити ефективне програмне забезпечення для побудови простих і зручних для подальшого аналізу моделей у вигляді дерев розв'язків, а також для відбору інформативних ознак, кластерного аналізу, індукції нечітких правил і синтезу нейро-нечітких моделей на основі дерев розв'язків;

– запропоновані методи, які складають математичну основу розробленого програмного забезпечення, істотно (до 30%) підвищують інтерпретабельність синтезованих моделей на основі дерев розв'язків та зменшують помилку розпізнавання (близько 3%), що дозволяє отримувати моделі, зручні для подальшого аналізу та використання на практиці.

Отримані в процесі досліджень наукові висновки та положення є обґрунтованими та достовірними. Обґрунтованість підтверджується результатами експериментальних досліджень, розв'язанням реальних практичних задач на основі запропонованих методів та розробленого програмного забезпечення. Результати експериментів свідчать про високу ефективність запропонованих методів, а також про підвищення точності, швидкості побудови та роботи, рівнів узагальнення й інтерпретабельності синтезованих дерев розв'язків.

Результати дисертації впроваджено в організаціях: 1) ПАТ “Запорізький автомобільний завод” (акт впровадження від 16.02.12); 2) ТОВ “Ітекс” (акт впровадження від 07.02.12); 3) ТОВ “Макфорт” (акт впровадження від 12.03.12); 4) Запорізький національний технічний університет, в навчальному процесі (акт впровадження від 02.02.12).

Особистий внесок здобувача полягає у тому, що наукові положення, висновки й рекомендації, що лежать в основі дисертаційної роботи, були сформульовані, розроблені та досліджені ним самостійно. У публікаціях, написаних у співавторстві, здобувачу належать: [1] – методи побудови дерев розв'язків та їх використання для розв'язання задач розпізнавання образів; [2] – еволюційний метод синтезу дерев розв'язків; [3] – метод ідентифікації дерев розв'язків для індукції нечітких правил; [4] – метод кластерного аналізу з використанням дерев розв'язків; [5] – метод побудови нейро-нечітких мереж на основі дерев розв'язків; [6] – метод побудови дерев розв'язків з використанням мультиагентного підходу; [7] – метод синтезу дерев розв'язків на основі вибірок даних з пропущеними значеннями; [10] – побудова розпізнавальних моделей для діагностування автотранспортних засобів за допомогою дерев розв'язків; [11] – обґрунтування доцільності використання стохастичного підходу до синтезу дерев розв'язків; [12] – розв'язання задачі діагностування автомобільних кузовів; [13] – запропоновано використання дерев розв'язків для вирішення задачі діагностування транспортних засобів; [14] – реалізація мультиагентного методу синтезу дерев розв'язків; [15] – реалізація методу синтезу дерев розв'язків для побудови розпізнавальних моделей на основі вибірок даних з пропущеними значеннями; [16] – метод скорочення баз лінгвістичних правил на основі дерев розв'язків; [17] – обґрунтування доцільності використання дерев розв'язків для кластерного аналізу.

Апробація результатів дисертації. Основні положення та результати дисертаційної роботи доповідалися та обговорювалися на міжнародних конференціях і семінарах: “Математичне і програмне забезпечення інтелектуальних систем” (Дніпропетровськ, 2011), “Радіоелектроніка і молодь в ХХІ сторіччі” (Харків, 2011), “Тиждень науки” (Запоріжжя, 2010, 2011), “Сучасні проблеми прикладної математики та інформатики” (Львів, 2011), “Информационные управляющие системы и компьютерный мониторинг” (Донецк, 2011), “Системний аналіз. Інформатика. Управління” (Запоріжжя, 2011, 2012), “Інформаційні технології: Наука, техніка, технологія, освіта, здоров’я” (Харків, 2011), “Нейроінформатика, её приложения и анализ данных” (Красноярск, 2011).

Публікації. Основні положення та результати дисертації опубліковані в 17 друкованих працях, з них: 1 монографія, 6 статей у виданнях, які входять до переліку наукових фахових видань України (4 статті в журналах, 2 статті в збірниках наукових праць), 10 публікацій у збірниках праць і матеріалів наукових конференцій і семінарів.

Структура дисертації. Дисертаційна робота складається з переліку скорочень, вступу, чотирьох розділів, висновків, додатків та списку використаних джерел. Загальний обсяг дисертації складає 170 сторінок, у тому числі 135 сторінок основного тексту. Робота містить 25 рисунків та 12 таблиць, а також 2 додатки на 15 сторінках та список використаних джерел зі 148 найменувань на 16 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** подано загальну характеристику дисертації: обґрунтовано актуальність теми, визначено зв’язок роботи з науковими програмами, планами, темами, сформульовано мету, задачі та методи дослідження, викладено наукову новизну і практичне значення одержаних результатів дисертаційної роботи, визначено особистий внесок здобувача, наведено інформацію про особистий внесок здобувача, апробації результатів дисертації та публікації, в яких опубліковані основні наукові результати дослідження.

Перший розділ дисертації містить аналіз процесу та методів побудови ДР в інтелектуальних системах. Обґрунтовано використання ДР для розв’язання задач інтелектуального аналізу даних, що виникають при синтезі розпізнавальних моделей. Наведено загальну постановку задачі побудови ДР, описано основні етапи процесу їх синтезу. Виконано порівняльний аналіз методів побудови ДР.

Виявлено основні переваги та недоліки ДР. Зазначено, що ДР є інтелектуальними моделями, у внутрішніх вузлах яких розташовані функції прийняття рішень на основі значень вхідних змінних, а в зовнішніх вузлах (листах) – значення вихідної змінної, що відповідають умовам внутрішніх вузлів. Завдяки деревоподібній структурі такі моделі дозволяють наочно подавати результати обчислень, тому вони добре інтерпретуються людьми-фахівцями в прикладних областях, які не знайомі з моделями штучного інтелекту.

Показано, що більшість методів побудови ДР використовують жадібну стратегію пошуку, застосування якої часто призводить до синтезу моделей, які не забезпечують прийнятний рівень апроксимації. Іншою важливою проблемою при синтезі ДР є проблема обробки даних з пропущеними значеннями, які часто зустрічаються у вибірках, що характеризують реальні об’єкти або процеси. Відзначено, що при обробці великих вибірок, що характеризуються значною

кількістю ознак, виникають проблеми побудови компактних дерев і фрагментації даних. Потреба усунення виявлених недоліків обумовлює необхідність розробки нових та модифікації існуючих методів синтезу ДР.

У **другому розділі** розроблено еволюційний метод синтезу дерев розв'язків та метод синтезу ДР на основі вибірок даних з пропущеними значеннями.

У розробленому *еволюційному методі побудови дерев розв'язків* пропонується інформацію про ДР зберігати у вигляді хромосом, поданих за допомогою деревоподібної структури даних. Для оцінювання хромосом використовується функція, яка відображає інформацію як про якість апроксимації за допомогою ДР, так і про його складність, що дозволяє будувати ДР з достатніми узагальнювальними та апроксимаційними властивостями, які містять при цьому незначну кількість вузлів, що, у свою чергу, підвищує інтерпретабельність синтезованих моделей. Для створення нової множини рішень за допомогою еволюційних операторів схрещування та мутації пропонується використовувати оператори, аналогічні операторам генетичного програмування, оскільки такі оператори ефективно генерують нові рішення, подані у вигляді деревоподібних структур.

Запропонований еволюційний метод синтезу дерев розв'язків містить такі етапи.

Етап 1. Ініціалізація. Створюється початкова множина рішень, поданих у вигляді хромосом $H_j, j = 1, 2, \dots, N$, де N – кількість хромосом. При цьому кожна хромосома H_j відповідає дереву розв'язків T_j .

Для кодування хромосоми (подання дерева розв'язку у вигляді хромосоми) використовується структура даних у вигляді бінарного дерева $T = \{t_k\}$, у вузлах якого $t_k = \langle c_k, l_k, r_k \rangle$ розташовано функції прийняття рішень c_k на основі значень вхідних змінних (у випадку, якщо вузол є внутрішнім) або значення вихідної змінної (для зовнішніх вузлів), а також посилання l_k та r_k на лівого й правого нащадків, відповідно, що представляють собою структури, аналогічні t_k .

Етап 2. Оцінювання поточного набору ДР. У якості цільової функції використовується функція, що враховує: помилку прогнозування або класифікації $E(T_j)$ за допомогою дерева T_j , відповідного до оцінюваної хромосоми H_j ; складність синтезованого дерева $S(T_j)$. Враховуючи обидва критерії, пропонується використовувати цільову функцію (1):

$$f(H_j) = \alpha_1 E(T_j) + \alpha_2 S(T_j) \rightarrow \min, \quad (1)$$

де α_1 та α_2 – коефіцієнти, що враховують значущість кожного з цільових критеріїв ($E(T_j)$ і $S(T_j)$, відповідно), $\alpha_1 + \alpha_2 = 1$.

Для оцінювання складності дерева можна використовувати формулу (2):

$$S(T_j) = N(T_j) + r(T_j) + \eta(T_j), \quad (2)$$

де $N(T_j)$ – кількість вузлів у дереві T_j ; $r(T_j)$ – кількість ребер у дереві T_j ; $\eta(T_j)$ – глибина дерева T_j .

Етап 3. Перевірка критеріїв зупинення пошуку. У якості таких критеріїв можна використовувати наступні: досягнення максимальної кількості ітерацій

пошуку $Iter_{\max}$; досягнення прийняттого значення цільової функції $f^* \leq \varepsilon$, де ε – задане користувачем максимально припустиме значення цільового критерію, f^* – мінімальне значення $f(H_j)$ цільової функції на поточній ітерації. Якщо хоча б один з критеріїв зупинення виконується, тоді пошук завершується. При цьому вважається, що оптимальне ДР подано у вигляді хромосоми H^* з цільовою функцією f^* . У випадку, якщо критерії зупинення не задоволені, виконується перехід до етапу 4.

Етап 4. Відбір. Вибір хромосом для схрещування та мутації пропонується виконувати пропорційно значенню їх цільової функції. Для цього розраховується середнє значення цільової функції в поточній популяції. Після цього для кожної хромосоми H_j обчислюється величина $p_o(H_j)$ як відношення значення цільової функції $f(H_j)$ до середнього значення цільової функції в популяції $f_{\text{ср}}$: $p_o(H_j) = f(H_j) / f_{\text{ср}}$. В залежності від значення $p_o(H_j)$ хромосома H_j копіюється декілька разів у масив особин, допущених до відтворення.

Етап 5. Схрещування. Спочатку обираються випадковим чином на кожному з батьківських дерев одна або декілька точок (вузлів дерева розв'язків). Точки розриву обираються випадковим чином для кожного батька окремо, що призводить до того, що внаслідок вибору різних точок розриву можливі ситуації породження однаковими батьками різних нащадків. Після вибору точок схрещування виконується генерація нащадків. Для цього необхідно обміняти у батьківських дерев фрагменти згідно з обраними точками розриву.

Етап 6. Мутація. Для зміни ДР, обраних для мутації, необхідно випадковим чином обрати мутуючий вузол дерева, після чого виконати мутацію одним з наступних способів: мутація зі збільшенням (в обраному вузлі побудувати нову гілку дерева, збільшивши його в такий спосіб); мутація зі стисненням (вилучити обраний вузол і всі його нащадки, зменшивши тим самим розмір дерева); мутація із заміною (після вибору двох вузлів дерева, виконується обмін відповідних гілок).

Етап 7. Формування нового покоління. Створюється нова множина рішень із хромосом, сформованих у результаті схрещування та мутації, а також з найбільш пристосованих хромосом попереднього покоління (хромосом з найкращими оцінками цільової функції). Після цього виконується перехід до етапу 2.

Запропонований еволюційний метод синтезу дерев розв'язків заснований на стохастичному підході та не використовує жадібну стратегію пошуку, дозволяє будувати інтерпретабельні ДР з достатніми узагальнювальними й апроксимаційними властивостями.

Для забезпечення можливості розпізнавання екземплярів в умовах невизначеності або неповноти даних модифіковано метод синтезу дерев розв'язків ID3 шляхом врахування пігністичених ймовірностей віднесення екземплярів до класів на основі теорії функцій довіри, внаслідок чого запропоновано *метод синтезу дерев розв'язків на основі вибірок даних з пропущеними значеннями*. Розроблений метод ідентифікації дерев розв'язків заснований на застосуванні математичного апарату теорії функцій довіри, характеризується особливостями як на етапі безпосередньої ідентифікації дерева розв'язків, так і на етапі класифікації екземплярів з використанням отриманого дерева розв'язків.

Пропонований метод для ідентифікації ДР використовує теорію функцій довіри та враховує невизначеність деяких параметрів, пов'язаних із завданням класифікації. Таким чином, виділяються деякі відмінності при визначенні гіпотез щодо цих параметрів при їхньому використанні. Навчальна вибірка в рамках цієї невизначеної структури являє собою множину, складену з елементів, поданих парами (ознаки, клас), де для кожного екземпляра, як правило, відомі значення кожної з його ознак і унікальний клас, до якого належить даний екземпляр. Така навчальна вибірка (на відміну від тих, що традиційно використовуються для побудови моделей складних об'єктів і процесів) може містити дані, у яких присутня деяка невизначеність щодо класів. Інакше кажучи, кожний клас навчальної вибірки може бути невизначеним, тоді як значення ознак, що характеризують кожний екземпляр, є однозначно відомі. Невизначеність класу будь-якого екземпляра подається шляхом основного розподілу довіри, що задається експертом та являє собою думку-переконавання цього експерта про фактичне значення класу для кожного екземпляра у навчальній вибірці. Після того, як навчальна вибірка описана, задається другий важливий параметр розроблюваного методу – міра вибору ознаки, яка буде використовуватися для вибору тестової ознаки в кожному вузлі ДР.

Інформаційну значущість ознак пропонується визначати шляхом виконання таких дій:

– розрахувати функцію середньої пігністичної ймовірності $BetP_S$, виходячи з навчальної вибірки S . Потім розрахувати ентропію розподілу класів в S з виразу (3):

$$Info(S) = - \sum_{a=1}^{N_{кл}} BetP_S(C_a) \log_2 BetP_S(C_a), \quad (3)$$

де $N_{кл}$ – кількість класів C_a , що характеризують вихідний параметр Y у вибірці S ;

– ґрунтуючись на отриманих даних, розрахувати приріст інформації, що забезпечується кожною ознакою X_i за формулою (4):

$$Gain(S, X_i) = Info(S) - Info_{X_i}(S), \quad (4)$$

таким чином, для кожного значення x_{ij} ознаки X_i виділяється підмножина S_{ij} , отримана з екземплярів вибірки S , для яких значення відповідної ознаки дорівнює x_{ij} . Потім розраховується середня функція довіри Bel , після чого застосовується пігністичне перетворення для розрахунків пігністичної ймовірності $BetP$. Після отриманих перетворень можна розрахувати $Info(S_{ij})$, при цьому S_{ij} являє собою вибірку, в якій значення ознаки X_i дорівнює x_{ij} ;

– шукане $Info_{X_i}(S)$ буде дорівнювати зваженій сумі $Info(S_{ij})$ щодо розглянутої ознаки;

– після того, як обчислені інформаційні значущості ознак, обирається ознака з найбільшою інформаційною значущістю.

На відміну від традиційного ДР, у якому кожний лист визначає унікальний клас, запропонований метод визначає кожному листу основний розподіл довіри, виражаючи, таким чином, множину переконань про різні класи структури розпізнавання.

Після того, як було виконано ідентифікацію ДР, можна виконувати класифікацію екземплярів, що не відносяться до навчальної вибірки. Класифікація об'єктів з невизначеними параметрами полягає у пошуку листів, які можуть належати до розглянутого екземпляра, відслідковуючи всі можливі шляхи, викликані різними значеннями ознаки. У випадку невідомих значень, беруться до уваги всі розгалуження щодо розглянутої ознаки. Як наслідок, екземпляр може належати декільком листам, кожний з яких характеризується функцією основного розподілу довіри, які поєднуються для одержання переконання щодо можливого класу екземпляра. Отриманий основний розподіл довіри може бути перетворений у функцію ймовірності шляхом застосування пігністичного перетворення. Це дозволяє обчислити ймовірність належності розглянутого екземпляра до окремого класу.

У **третьому розділі** запропоновано метод ідентифікації ДР для індукції нечітких правил, а також метод синтезу нейро-нечітких мереж на основі ДР.

В запропонованому *методі ідентифікації дерев розв'язків для індукції нечітких правил* основними етапами є зростання дерева та його скорочення (згладжування), після чого виконується перетворення ДР у лінгвістичні правила.

На етапі зростання дерева для кожного можливого розбиття (f, v) розраховується оціночна функція, що визначається формулою (5):

$$Q(f, v) = p_{f,v} g(p_{f,v}^1) + (1 - p_{f,v}) g(p_{f,v}^2), \quad (5)$$

де $p_{f,v}^1 = P(y_i = 1 | x_i \in dT_{f,v}^1)$, $p_{f,v}^2 = P(y_i = 1 | x_i \in dT_{f,v}^2)$, $p_{f,v} = P(x_i \in dT_{f,v}^1 | x_i \in dT)$; $g(p)$ – модифікована ентропія для ймовірності віднесення вихідної змінної y до певного класу за умови, що x більше або менше значення v ($p_{f,v}^1$ та $p_{f,v}^2$, відповідно). Величина $g(p)$ розраховується за формулою: $g(p) = -r(p) \ln(r(p)) - (1 - r(p)) \ln(1 - r(p))$, де $r(p)$ перетворює оцінку ймовірності (6):

$$r(p) = \begin{cases} 0,5(1 + \sqrt{2p - 1}), & \text{якщо } p > 0,5; \\ 0,5(1 - \sqrt{1 - 2p}), & \text{якщо } p \leq 0,5. \end{cases} \quad (6)$$

Оціночна функція розраховується для всіх можливих розбиттів, і вибирається розбиття з найменшим значенням оціночної функції. Розбиття починається від кореневого вузла та триває доти, поки не виникне ситуація, коли неможливо виконати нове розбиття.

Після виконання етапу зростання може виникнути ситуація «перенавчання» ДР, що може привести до не зовсім коректної роботи ДР на тестових вибірках.

У зв'язку з цим на другому етапі проводиться скорочення (усікання) великого дерева, щоб дерево менших розмірів давало більш стабільні оцінки ймовірності та було більш інтерпретабельним. Нехай вузли dT_1 та dT_2 є елементами

одного рівня із загальним батьківським вузлом dT . Нехай $p(dT_1)$, $p(dT_2)$ і $p(dT)$ будуть відповідними оцінками ймовірності. Локальна переоцінена ймовірність визначається як $w_{dT}p(dT) + (1 - w_{dT})p(dT_1)$ для dT_1 та $w_{dT}p(dT) + (1 - w_{dT})p(dT_2)$ для dT_2 . Локальна значущість w_{dT} та супутня функція $G(dT)$ розраховуються рекурсивно, виходячи з формул (7) та (8):

$$\frac{w_{dT}}{1 - w_{dT}} = \frac{c \cdot \exp(-|dT| g(p(dT)))}{\exp(-|dT_1| G(dT_1) - |dT_2| G(dT_2))}, \quad (7)$$

$$G(dT) = \begin{cases} g(p(dT)) + \frac{1}{|dT|} \log((1 + \frac{1}{c})w_{dT}), & \text{якщо } w_{dT} > 0,5; \\ \frac{|dT_1|}{|dT|} G(dT_1) + \frac{|dT_2|}{|dT|} G(dT_2) + \frac{1}{|dT|} \log((1 + c)(1 - iw_{dT})), & \text{якщо } w_{dT} \leq 0,5. \end{cases} \quad (8)$$

Етап перетворення ДР у лінгвістичні правила полягає в тому, щоб створити набір правил із правил, кожне з яких відповідає окремому листу дерева шляхом формування логічного об'єднання умов на шляху від кореня дерева до листа. Пропонується підхід, що перетворює ДР у набір логічно еквівалентних правил. Його метою є не одержання доказово мінімального набору правил, а здійснення логічного усікання правил: перевірка умов “>” та “<” у всіх правилах з метою усунення надмірності в описі умов правил та видалення умов, які є логічно надлишковими в контексті всього набору правил, тобто видалення умов, які ідентифікуються виходячи зі структури отриманого ДР. Таке спрощення змінює правило, пов'язане з конкретним листом дерева, при цьому зберігаючи повну адекватність усього набору правил.

Для кожного листа, віднесеного до класу y_k створюється правило про те, що об'єкт відноситься до класу y_k шляхом кон'юнкції умов, які знаходяться на шляху проходження від кореня до y_k , але використовуючи тільки ті умови, які відповідають наступному правилу: для кожного вузла dt на шляху від кореня до листа з міткою y_k , умова, що відповідає батьку dt , є частиною кон'юнкції тільки в тому випадку, якщо спрацює умова сусідства для вузла dt . Умова сусідства для dt вважається успішною, якщо вузол dt не є коренем і сусідній вузол відносно dt не є листом з міткою y_k .

Запропонований метод дозволяє виконувати перетворення та об'єднання правил, що забезпечує можливість одержання більш інтерпретабельних баз правил;

У розробленому методі побудови нейро-нечітких мереж (ННМ) на основі *дерев розв'язків* пропонується використовувати: ДР, синтезоване на основі заданої вибірки даних, – для витягання продукційних правил вигляду «Якщо – то», обчислення параметрів функцій належності, задавання структури і налаштування параметрів нейроелементів ННМ. Синтез ННМ пропонується виконувати у вигляді такої послідовності етапів.

Етап 1. Побудова бази правил типу «Якщо – то». За навчальною вибіркою $< X, Y >$ на основі методів синтезу ДР виконується побудова дерева розв'язків $DT = \{d_k\}$, де $d_k = \langle cd_k, ld_k, rd_k \rangle$ – k -й вузол дерева DT, що є структурою, в якій cd_k – функція прийняття рішень (умова типу $x_i \in A_{ij}$) на основі значень вхідних

змінних (у випадку, якщо вузол є внутрішнім) або значення вихідної змінної (для зовнішніх вузлів), ld_k та rd_k – посилання на лівого та правого нащадків k -го вузла, відповідно, що є структурами, аналогічними d_k ; $A_{ij} = [l_{ij}; r_{ij}]$ – інтервал значень ознаки x_i в умові cd_k , що обмежується значеннями l_{ij} та r_{ij} .

Для побудови набору правил R шляхом вилучення правил з ДР виконується прямий обхід дерева: від кореня дерева до кожного вузла. При цьому для кореня дерева як першого вузла-батька викликається рекурсивно процедура обходу лівого та правого піддерев. В процесі обходу дерева DT при кожному відвідуванні нового вузла d_k відбувається модифікація поточного антецедента $Antecedent_j$ правила R_j шляхом додавання умови cd_k , що знаходиться у вузлі d_k : $Antecedent_j = Antecedent_j \cap cd_k$.

У випадку, коли в умові cd_k значення деякої ознаки X_i виходить за межі діапазону її можливих значень $[x_{\min i}; x_{\max i}]$, виконується коригування лівої l_{ij} та правої r_{ij} меж значень ознаки X_i в умові cd_k (9):

$$\begin{cases} l_{ij} = x_{\min i}, & \text{якщо } l_{ij} < x_{\min i}; \\ r_{ij} = x_{\max i}, & \text{якщо } r_{ij} > x_{\max i}. \end{cases} \quad (9)$$

Коли вузол d_k виявляється листом, виконується поповнення бази правил R новим правилом R_j . При цьому як антецедент приймається $Antecedent_j$, а в якості консеквента – значення cd_k листа дерева: $Antecedent_{j+1} = Antecedent_j / cd_k$.

Обхід дерева виконується до тих пір, поки не будуть пройдені всі його листя (кожен вузол дерева відвідується до тих пір, поки не відвідані всі його нащадки).

Таким чином, в результаті виконання першого етапу пропонованого методу виконується формування бази правил R типу «Якщо – то» за формулою (10), де кожне j -те правило формується на основі умов (перевірок) типу $x_i \in A_{ij}$, що відносяться до гілок ДР, розташованих на шляху від кореня до j -го листа:

$$\left\{ \begin{array}{l} \text{Правило 1: Якщо } \bigcap_{i=1}^L x_i \in A_{i1}, \text{ то } y = y_1; \\ \text{Правило 2: Якщо } \bigcap_{i=1}^L x_i \in A_{i2}, \text{ то } y = y_2; \\ \dots\dots\dots \\ \text{Правило } m: \text{Якщо } \bigcap_{i=1}^L x_i \in A_{im}, \text{ то } y = y_m, \end{array} \right. \quad (10)$$

де A_{ij} – інтервал (множина) значень ознаки X_i , при яких виконується j -те правило.

Етап 2. Синтез блоків першого шару ННМ, що визначають належність розпізнаваного екземпляра до термів ознак. Для цього обчислюються ліва l_{ij} і права

r_{ij} межі діапазонів значень кожної ознаки X_i в правилі R_j . Далі для кожної i -ї ознаки X_i визначається кількість інтервалів розбиття діапазону її значень N_i по всій множині правил R .

Потім задаються функції належності розпізнаваного екземпляра до кожного з термів. При цьому використовуються межі чітких інтервалів $\Delta_{ij} = [l_{ij}, r_{ij})$, що були знайдені раніше.

В якості функцій належності доцільно використовувати такі функції, які дозволяють обмежувати інтервал значень ознак:

– трапецієподібну функцію (11):

$$\mu_{ik} = \begin{cases} 0, & \text{якщо } x_i < a_{ik}; \\ (x_i - a_{ik}) / (l_{ik} - a_{ik}), & \text{якщо } a_{ik} \leq x_i < l_{ik}; \\ 1, & \text{якщо } l_{ik} \leq x_i \leq r_{ik}; \\ (b_{ik} - x_i) / (b_{ik} - r_{ik}), & \text{якщо } r_{ik} < x_i \leq b_{ik}; \\ 0, & \text{якщо } x_i > b_{ik}, \end{cases} \quad (11)$$

де $a_{ik} = 0,25(3r_{ik-1} + l_{ik})$, $b_{ik} = 0,25(r_{ik} + 3l_{ik+1})$, $r_{i0} = l_{i1} - 2$, $l_{iN_i+1} = l_{iN_i} + 2$;

– П-подібну функцію (12):

$$\mu_{ik} = \mu S_{ik} \cdot \mu Z_{ik}, \quad (12)$$

$$\text{де } \mu S_{ik} = \begin{cases} 0, & \text{якщо } x_i < l_{ik}; \\ (x_i - l_{ik}) / (r_{ik} - l_{ik}), & \text{якщо } l_{ik} \leq x_i \leq r_{ik}; \\ 1, & \text{якщо } x_i > r_{ik}, \end{cases} \quad \mu Z_{ik} = \begin{cases} 1, & \text{якщо } x_i < l_{ik}; \\ (r_{ik} - x_i) / (r_{ik} - l_{ik}), & \text{якщо } l_{ik} \leq x_i \leq r_{ik}; \\ 0, & \text{якщо } x_i > r_{ik}. \end{cases}$$

Етап 3. Побудова другого шару ННМ, що складається з нейронів, які реалізують рядки-кон'юнкції антецедентів правил нечіткої бази знань. Виходи кожного нейрона цього шару визначають ступені виконання умов правил для розпізнаваного екземпляра.

При обчисленні ступенів виконання антецедентів правил доцільно враховувати достовірність (вагу, надійність, впевненість) кожного правила. Тому при розрахунку виходів нейронів другого шару пропонується використовувати також коефіцієнт упевненості кожного правила, що дозволить посилити внесок більш надійних правил у розв'язок. Для обчислення коефіцієнта упевненості j -го правила γ_j пропонується по всій вибірці $\langle X, Y \rangle$ знайти кількість екземплярів n_j , для яких спрацювало j -те правило, а також визначити $n_{\text{пом } j}$ – кількість помилкових рішень при використанні даного правила. Коефіцієнт упевненості j -го правила γ_j будемо розраховувати за формулою: $\gamma_j = \alpha_j P_{\text{прав } j}$, де $\alpha_j = n_j / m$ – ймовірність спрацювання j -го правила; $P_{\text{прав } j} = 1 - P_{\text{пом } j} = 1 - n_{\text{пом } j} / n_j$ – ймовірність правильних рішень при використанні j -го правила.

Тоді ступінь відповідності виконання антецедента j -го правила для розпізнаваного екземпляру може бути розрахована за формулою (13):

$$\mu_{Aj} = \alpha_j \min \left\{ \max_{i,k} \{w_{ik}^{(2j)}; \mu_{ik}\} \right\}, \quad (13)$$

де μ_{Aj} – ступінь виконання антецедента j -го правила для розпізнаваного екземпляру; $w_{ik}^{(2j)}$ – ваги нейронів другого шару, що визначають наявність зв'язку від нейронів першого шару до другого і відображають наявність k -го терму i -ї ознаки в антецеденті j -го правила. Ваги $w_{ik}^{(2j)}$ визначаються за синтезованим ДР наступним чином: $w_{ik}^{(2j)} = 0$, якщо на шляху від кореня ДР до j -го листа є умова, за якої i -та ознака потрапляє в k -й інтервал діапазону її розбиття (k -й терм i -ї ознаки присутній в описі умов j -го правила), в іншому випадку: $w_{ik}^{(2j)} = 1$.

Етап 4. Визначення параметрів нейронів третього шару ННМ, які обчислюють ступені належності вхідного вектора до відповідних термів вихідної змінної, використовуючи формулу (14):

$$\mu_{y^{(t)}} = \min \left\{ \max_j \{w_j^{(3t)}; \mu_{Aj}\} \right\}, \quad (14)$$

де $\mu_{y^{(t)}}$ – ступінь належності вхідного вектора до t -го терму вихідної змінної y ; $w_j^{(3t)}$ – ваги нейронів третього шару, що визначають наявність зв'язку від нейронів другого шару до третього та відображають відповідність t -го терму вхідного параметру до j -го правила нечіткої бази знань. Ваги $w_j^{(3t)}$ визначаються за синтезованим ДР: $w_j^{(3t)} = 0$, якщо значення j -го листа ДР дорівнює t -му значенню вхідного параметра $y^{(t)}$ з множини його значень D_y , в іншому випадку: $w_j^{(3t)} = 1$.

Етап 5. Дефазифікація – здійснюється за допомогою нейрона четвертого шару ННМ, вихід якого може бути обчислений за допомогою центроїдного методу.

Запропонований метод синтезу ННМ на основі дерев розв'язків дозволяє будувати моделі з відносно невеликою кількістю нейроелементів, які володіють високим рівнем узагальнення. На відміну від градієнтних методів настроювання параметрів мережі, запропонований метод не відноситься до методів локальної оптимізації та не має проблеми невизначеності вибору початкової точки пошуку.

У четвертому розділі виконується експериментальне дослідження розроблених методів та порівняння їх з відомими аналогами. Для дослідження різних методів синтезу ДР при розв'язанні практичних задач розроблено методику проведення експерименту, пропонується система критеріїв порівняння ДР, створюється програмне забезпечення, яке за даними навчальної вибірки на основі розроблених методів дозволяє будувати ієрархічні моделі у вигляді ДР.

Для оцінювання якості синтезованих ДР і методів їх побудови пропонується використовувати три групи критеріїв: критерії оцінювання апроксимаційних та узагальнювальних властивостей, часові критерії (час синтезу ДР та час розпізнавання за допомогою синтезованого ДР), критерії інтерпретабельності

побудованого ДР (загальна кількість вузлів, загальна кількість ребер, кількість листів, глибина дерева, кількість використовуваних у ДР ознак).

З метою дослідження запропонованих методів та порівняння їх з відомими аналогами розроблено методику проведення експерименту та комп'ютерну програму “Автоматизована система синтезу дерев рішень на основі інтелектуальних обчислень для неруйнівного діагностування та класифікації” (свідоцтво про реєстрацію авторського права № 41237), призначену для розв’язання задач інтелектуального аналізу даних на основі використання апарата теорії ДР. Розроблене програмне забезпечення реалізує відомі та запропоновані в роботі методи побудови ДР, а також методи, що дозволяють виконувати відбір інформативних ознак, кластерний аналіз, індукцію нечітких правил і синтез нейро-нечітких моделей на основі ДР.

За допомогою запропонованих методів і програмних засобів розв’язано задачу діагностування кузовів автотранспортних засобів. Виявлено набір інформативних ознак, виділено кластери та побудовано моделі досліджуваних об’єктів як у вигляді ДР, так і у вигляді нейро-нечітких мереж. Вибірка даних формувалася за результатами вимірів контрольних точок на автомобілях. З 1400 вимірюваних точок експертами (фахівцями в автомобілебудуванні) виділено істотні (рис. 1), які найбільшим чином впливають на якість кузова й у яких фіксувалися відхилення від номінального значення по базових координатах. Експериментально отримані характеристики методів побудови ДР наведено у табл. 1, де E – помилка розпізнавання за синтезованою моделлю; t – час побудови ДР; N , Ls , η – загальна кількість вузлів, листів, використовуваних ознак в побудованому ДР, відповідно.

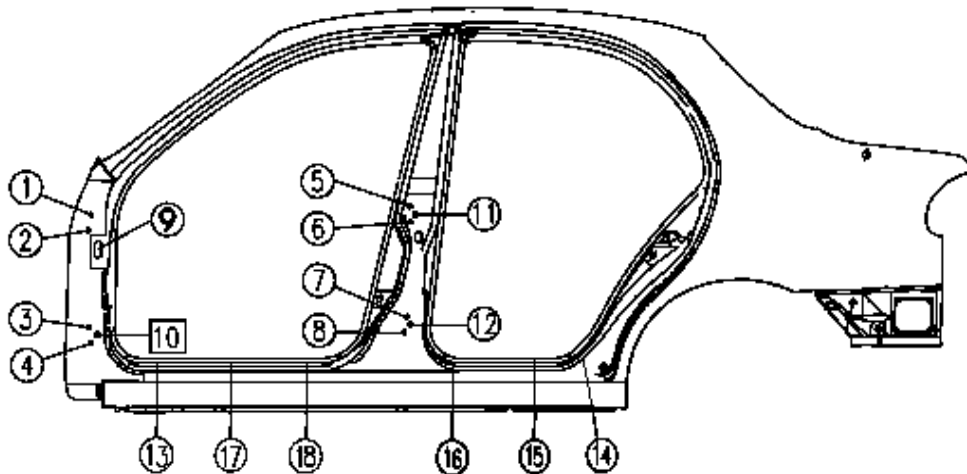


Рисунок 1 – Контрольні точки кузова автомобіля

Таблиця 1 – Характеристики методів побудови ДР

Метод	E	t	N	Ls	η
CART	0,052	15,12	120	24	17
C4.5	0,046	14,82	98	22	17
Мультиагентний метод	0,021	9,82	86	18	14
Еволюційний метод	0,018	7,42	88	19	13

Як видно з табл. 1, запропонований еволюційний метод синтезу ДР дозволяє синтезувати ДР, що забезпечують більш високу точність розпізнавання. Це пояснюється його стохастичною природою, що дозволяє виключити можливість зациклення в локальних оптимумах. Крім того, ДР, синтезовані за допомогою запропонованого методу характеризуються меншою кількістю вузлів, ребер, листів.

Результати експериментів показали, що запропоновані методи дозволяють швидше синтезувати ДР (в середньому на 20%), які володіють при цьому більш високими рівнями узагальнення й інтерпретабельності (в середньому на 30%), а також точністю розпізнавання (близько 3%).

Висновки містять основні наукові та практичні результати дисертаційної роботи.

Додатки містять вхідні дані для експериментальних досліджень, а також копії документів про впровадження та належність результатів роботи.

ВИСНОВКИ

У дисертаційній роботі вирішено актуальну наукову задачу побудови інтелектуальних моделей багатовимірних об'єктів, що полягає в розробці методів індуктивного навчання дерев розв'язків для підвищення рівнів інтерпретабельності й узагальнення, а також швидкості роботи та синтезу розпізнавальних моделей. Отримано такі результати.

1. Новий еволюційний метод синтезу дерев розв'язків, що заснований на стохастичному підході та не використовує жадібну стратегію пошуку. В запропонованому методі інформація про синтезовану модель на основі дерев розв'язків зберігається у вигляді хромосом, представлених за допомогою деревоподібної структури даних, для оцінювання хромосом використовується функція, яка відображає інформацію як про якість апроксимації за допомогою ДР, так і про його складність, що дозволяє будувати дерева розв'язків з достатніми узагальнювальними та апроксимаційними властивостями, які містять при цьому незначну кількість вузлів, що, у свою чергу, підвищує інтерпретабельність синтезованих моделей.

2. Новий метод синтезу нейро-нечітких мереж на основі ДР, який не вимагає розв'язання завдань оптимізації для настроювання значень параметрів моделі, для побудованого за навчальною вибіркою ДР автоматично створює нечітке розбиття простору ознак й виділяє правила, на основі яких формує структуру мережі та визначає значення її параметрів. Це дозволяє синтезувати моделі з невеликою кількістю нейроелементів і високим рівнем узагальнення.

3. Удосконалений метод ідентифікації дерев розв'язків для індукції нечітких правил, що використовує додаткові функції перетворення при зростанні ДР, згладжує дерево для його усікання та використовує критерії сусідства при перетворенні ДР. Запропонований метод дозволяє виконувати перетворення й об'єднання правил, що забезпечує можливість розробки експертних систем на основі більш інтерпретабельних баз лінгвістичних правил.

4. Розвинений метод синтезу дерев розв'язків ID3, у якому розраховуються пігністичні ймовірності віднесення екземплярів до класів на основі теорії функцій довіри, що дозволяє виконувати класифікацію екземплярів в умовах невизначеності

або неповноти даних. Розвинений метод відрізняється від існуючих методів фазою побудови ДР, оскільки враховується невизначеність, яка характеризує класи навчальних екземплярів за рахунок застосування функцій довіри, і використовує особливу процедуру класифікації нових екземплярів, значення ознак яких можуть бути не визначені, що дозволяє виконувати класифікацію неоднозначно заданих об'єктів.

5. Розроблені методи побудови дерев розв'язків доведено до програмної реалізації, що дозволило створити ефективне програмне забезпечення для синтезу простих і зручних для подальшого аналізу моделей у вигляді дерев розв'язків, а також для відбору інформативних ознак, кластерного аналізу, індукції нечітких правил і побудови нейро-нечітких моделей на основі дерев розв'язків.

6. Запропоновані методи складають математичну основу розробленого програмного забезпечення, істотно (до 30%) підвищують інтерпретабельність синтезованих моделей на основі дерев розв'язків та зменшують помилку розпізнавання (близько 3%), що дозволяє отримувати моделі, зручні для подальшого аналізу та використання на практиці.

За допомогою запропонованих методів і програмних засобів вирішено завдання технічного діагностування кузовів автотransпортних засобів. Проведено експерименти по порівнянню запропонованих методів з відомими аналогами. Результати експериментів показали, що запропоновані методи дозволяють швидше синтезувати дерева розв'язків, які володіють при цьому більш високими рівнями узагальнення й інтерпретабельності, а також точністю розпізнавання.

Отримані в роботі теоретичні та практичні результати впроваджено у виробництво на ПАТ “Запорізький автомобільний завод” (акт впровадження від 16.02.12), ТОВ “Ітекс” (акт впровадження від 07.02.12) та ТОВ “Макфорт” (акт впровадження від 12.03.12), а також у навчальний процес Запорізького національного технічного університету (акт впровадження від 02.02.12).

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні наукові результати дисертації

1. Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов : монография / С. А. Субботин, Ан. А. Олейник, Е. А. Гофман, С. А. Зайцев, Ал. А. Олейник ; под ред. С. А. Субботина. – Харьков : ООО “Компания Смит”, 2012. – 317 с.

2. Гофман Е. А. Эволюционный метод синтеза деревьев решений / Е.А. Гофман, А. А. Олейник, С. А. Субботин // Искусственный интеллект. – 2011. – № 2. – С. 6–14.

3. Гофман Е. А. Индукция лингвистических правил с использованием деревьев решений / Е. А. Гофман, А. А. Олейник, С. А. Субботин // Бионика интеллекта. – 2011. – № 3. – С. 126–130.

4. Гофман Е. А. Кластер-анализ с использованием деревьев решений / Е. А. Гофман, А. А. Олейник, С. А. Субботин // Радиоэлектроника и информатика. – 2011. – № 2. – С. 21–25.

5. Гофман Є. О. Метод синтезу нейро-фаззі мереж з використанням дерев розв'язків / Є. О. Гофман, С. О. Субботін, А. О. Олійник // Реєстрація, зберігання і обробка даних. – 2012. – Т. 14, № 1. – С. 57–65.

6. Гофман Є. О. Агентний метод синтезу дерев рішень / Є. О. Гофман, А. О. Олійник, С. О. Субботін // Вестник НТУ “ХПИ”. – Харьков : НТУ “ХПИ”. – 2011. – № 17. – С. 16–25.

7. Гофман Є. О. Синтез дерев рішень на основі вибірок даних з пропущеними значеннями / Є. О. Гофман, А. О. Олійник, С. О. Субботін // Вестник НТУ “ХПИ”. – Харьков : НТУ “ХПИ”. – 2011. – № 36. – С. 35–45.

Наукові праці апробаційного характеру

8. Гофман Е. А. Семантический анализ естественно-языковых текстовых конструкций / Е. А. Гофман // Тиждень науки : науково-практична конференція, Запоріжжя, 12–16 квітня 2010 р. : збірник тез доповідей. – Запоріжжя: ЗНТУ, 2010. – Т. 2. – С. 139–141.

9. Гофман Е. А. Интеллектуальный анализ данных на основе деревьев решений / Е. А. Гофман // Радіоелектроніка і молодь в ХХІ сторіччі : 15-й міжнародний молодіжний форум, Харків, 18–20 квітня 2011 р. : матеріали форуму. – Харків : ХНУРЕ, 2011. – Т. 9. – С. 78–79.

10. Гофман Е. А. Диагностирование автотранспортных средств на основе деревьев решений, обучаемых с помощью эволюционного поиска / Е. А. Гофман, С. А. Субботин // Нейроинформатика, её приложения и анализ данных: XIX Всероссийский семинар, Красноярск, 1–3 октября 2011 г. : материалы / под ред. А. Н. Горбаня, Е. М. Миркеса. – Красноярск: Сиб. федер. ун-т, 2011. – С. 39–44.

11. Гофман Є. О. Стохастичний підхід до синтезу дерев рішень / Є. О. Гофман, С. О. Субботін // Тиждень науки : науково-практична конференція, Запоріжжя, 11–15 квітня 2011 р. : збірник тез доповідей. – Запоріжжя: ЗНТУ, 2011. – Т. 2. – С. 53–55.

12. Субботин С. А. Применение аппарата деревьев решений в задаче диагностирования автомобильных кузовов / С. А. Субботин, Е. А. Гофман // Системний аналіз. Інформатика. Управління : II Всеукраїнська наукова-практична конференція САІУ-2011, Запоріжжя, 10–11 березня 2011 р. : матеріали конференції. – Запоріжжя : КПУ, 2011. – С. 211–212.

13. Гофман Е. А. Использование деревьев решений для диагностирования автотранспортных средств / Е. А. Гофман, А. А. Олейник, С. А. Субботин // Информационные управляющие системы и компьютерный мониторинг : II Международная научно-техническая конференция ИУС и КМ-2011, Донецк, 11–13 апреля 2011 г. : материалы конференции. – Донецк : ДонНТУ, 2011. – Т. 1. – С. 159–163.

14. Гофман Є. О. Синтез дерев рішень на основі агентного підходу / Є. О. Гофман, А. О. Олійник, С. О. Субботін // Інформаційні технології: Наука, техніка, технологія, освіта, здоров'я : XIX Міжнародна наукова-практична конференція MicroCAD-2011, Харків, 01–03 червня 2011 р. : матеріали конференції. – Харків : НТУ «ХПІ», 2011. – Ч. IV. – С. 17.

15. Гофман. Є. О. Побудова дерев рішень на основі вибірок даних з пропущеними значеннями / Є. О. Гофман, А. О. Олійник, С. О. Субботін // Сучасні проблеми прикладної математики та інформатики : XVII Всеукраїнська наукова конференція, Львів, 6–7 жовтня 2011 р. : тези доповідей. – Львів : ЛНУ ім. І. Франка, 2011. – С. 39.

16. Гофман Є. О. Аналіз та перетворення лінгвістичних правил на основі дерев рішень / Є. О. Гофман, О. О. Олійник, С. О. Субботін // Математичне і програмне забезпечення інтелектуальних систем : матеріали IX міжнародної науково-технічної конференції, Дніпропетровськ, 23–25 листопада 2011 р. : тези доповідей. – Дніпропетровськ : ДНУ, 2011. – С. 68–69.

17. Гофман Е. А. Кластерный анализ в информационных технологиях поддержки принятия решений / Е. А. Гофман, А. А. Олейник, С. А. Субботин // Системный анализ. Информатика. Управление : 3-я Міжнародна науково-практична конференція, Запоріжжя, 14–16 березня 2012 р. : матеріали конференції. – Запоріжжя : КПУ, 2012. – С. 74–76.

АНОТАЦІЯ

Гофман Є. О. Методи побудови дерев розв'язків в інтелектуальних системах діагностування. – На правах рукопису.

Дисертація на здобуття наукового ступня кандидата технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту. – Харківський національний університет радіоелектроніки Міністерства освіти і науки, молоді та спорту України, Харків, 2012.

Дисертацію присвячено розробці методів індуктивного навчання дерев розв'язків для підвищення рівнів інтерпретабельності й узагальнення, а також швидкості роботи та синтезу розпізнавальних моделей в інтелектуальних системах.

У роботі виконано аналіз процесу побудови та методів синтезу дерев розв'язків в інтелектуальних системах. Розроблено еволюційний метод синтезу дерев розв'язків, який заснований на стохастичному підході та не використовує жадібну стратегію пошуку, що дозволяє будувати дерева розв'язків з достатніми узагальнювальними й апроксимаційними властивостями з незначною кількістю вузлів. Запропоновано модифікацію методу синтезу дерев розв'язків ID3, у якій розраховуються пігністичні ймовірності віднесення екземплярів до класів на основі теорії функцій довіри, що дозволяє виконувати класифікацію екземплярів в умовах невизначеності або неповноти даних. Розроблено метод побудови дерев розв'язків, що дозволяє виконувати індукцію лінгвістичних правил і забезпечує можливість розробки експертних систем на основі більш інтерпретабельних баз лінгвістичних правил. Запропоновано метод синтезу нейро-нечітких мереж на основі дерев розв'язків, який не вимагає вирішення задач оптимізації для настроювання значень параметрів моделі. Створено автоматизовану систему синтезу дерев розв'язків, що дозволяє виконувати побудову простих і зручних для подальшого аналізу моделей у вигляді дерев розв'язків.

За допомогою запропонованих методів і програмних засобів вирішено завдання технічного діагностування кузовів автотранспортних засобів. Проведено експерименти по порівнянню запропонованих методів з відомими аналогами.

Ключові слова: дерево розв'язків, нейро-нечітка модель, база правил, інтерпретабельність, навчальна вибірка, розпізнавання, еволюційний метод.

АННОТАЦИЯ

Гофман Е. А. Методы построения деревьев решений в интеллектуальных системах диагностирования. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.23 – системы и средства искусственного интеллекта. – Харьковский национальный университет радиоэлектроники Министерства образования и науки, молодежи и спорта Украины, Харьков, 2012.

В диссертационной работе решена актуальная научная задача построения интеллектуальных моделей многомерных объектов, заключающаяся в разработке методов индуктивного обучения деревьев решения для повышения уровней интерпретабельности и обобщения, а также скорости работы и синтеза распознающих моделей. Были получены следующие результаты.

Выполнен анализ процесса построения деревьев решений в интеллектуальных системах. Проанализированы основные методы синтеза деревьев решений. Выявлены основные преимущества и недостатки деревьев решений и методов их синтеза. Показано, что большинство методов построения деревьев решения использует жадную стратегию поиска, применение которой часто приводит к синтезу моделей, не обеспечивающих приемлемый уровень аппроксимации. Другой важной проблемой при синтезе деревьев решений является проблема обработки данных с пропущенными значениями, которые часто встречаются в выборках, характеризующих реальные объекты или процессы. Отмечено, что при обработке больших выборок, характеризующихся значительным количеством признаков, возникают проблемы построения компактных деревьев и фрагментации данных. Результаты анализа показали необходимость модификации существующих и разработки новых методов синтеза ДР.

Разработан эволюционный метод синтеза деревьев решений, который основан на стохастическом подходе и не использует жадную стратегию поиска. Предложенный метод позволяет строить деревья решений с достаточными обобщающими и аппроксимационными свойствами, содержащие при этом незначительное количество узлов, что повышает интерпретабельность синтезированных моделей.

Предложена модификация метода синтеза деревьев решений ID3, в которой рассчитываются пигнистические вероятности отнесения экземпляров к классам на основании теории функций доверия, что позволяет выполнять классификацию экземпляров в условиях неопределенности или неполноты данных. Модифицированный метод отличается от существующих методов идентификации деревьев решений фазой построения дерева решений, поскольку учитывается неопределенность, характеризующая классы обучающих экземпляров за счёт использования функций доверия, и характеризуется особой процедурой классификации новых экземпляров, значения признаков которых могут быть не определены, что позволяет выполнять классификацию неоднозначно заданных объектов.

Разработан метод построения деревьев решений, позволяющий выполнять индукцию лингвистических правил, что достигается за счёт введения дополнительных функций преобразования при росте ДР, путём сглаживания дерева

для его усечения и за счёт введения критерия соседства при преобразовании ДР. Предложенный метод позволяет выполнять преобразование и объединение правил, что обеспечивает возможность разработки экспертных систем на основании более интерпретабельных баз лингвистических правил.

Предложен метод синтеза нейро-нечетких сетей на основе деревьев решений, который не требует решения задач оптимизации для настройки значений параметров модели, для построенного по обучающей выборке ДР автоматически формирует нечеткое разбиение признакового пространства и выделяет правила, на основе которых формирует структуру сети и определяет значения её параметров. Это позволяет синтезировать модели с небольшим числом нейроэлементов, обладающие высоким уровнем обобщения.

Создана автоматизированная система синтеза деревьев решений, позволяющая выполнять построение простых и удобных для дальнейшего анализа моделей в виде ДР, а также по данным обучающей выборки выделять набор информативных признаков, выполнять кластерный анализ, индукцию нечетких правил и синтез нейро-нечетких моделей на основе ДР.

С помощью предложенных методов и программных средств решена задача технического диагностирования кузовов автотранспортных средств. Проведены эксперименты по сравнению предложенных методов с известными аналогами. Результаты экспериментов показали, что предложенные методы позволяют быстрее синтезировать деревья решений, обладающие при этом более высокими уровнями обобщения и интерпретабельности, а также точностью распознавания.

Полученные в работе теоретические и практические результаты внедрены в производство на ПАО “Запорожский автомобильный завод” (акт внедрения от 16.02.12), ООО “Итэкс” (акт внедрения от 07.02.12) и ООО “Макфорт” (акт внедрения от 12.03.12), а также в учебный процесс Запорожского национального технического университета (акт внедрения от 02.02.12).

Ключевые слова: дерево решений, нейро-нечеткая модель, база правил, интерпретабельность, обучающая выборка, распознавание, эволюционный метод.

ABSTRACT

Gofman Ye. A. Methods of decision trees construction in intelligent systems of diagnosis. – Manuscript.

The thesis for the candidate degree in technical sciences on the specialty 05.13.23 – Artificial Intelligence Systems and Means. – Kharkiv National University of Radio Electronics Ministry of Education and Science, Youth and Sport of Ukraine, Kharkiv, 2012.

The thesis is devoted to development of methods for inductive learning of decision trees to improve the levels of interpretability and generalization, as well as the speed of synthesis of recognizing models in intelligent systems.

In the thesis the process of construction and methods for the synthesis of decision trees in intelligent systems are analyzed. The evolutionary method for the synthesis of decision tree is developed. It is based on a stochastic approach and does not use a greedy search strategy that allows to construct decision trees with sufficient generalization and

approximation properties with the small amount of nodes. The modification of the decision tree synthesis method ID3 is proposed. It is calculated the pignistic probabilities of instances referring to classes based of the theory of belief functions that allows to classify the instances under the uncertainty or incompleteness of the data. The method of decision trees constructing is created. It allows to induct the linguistic rules and to develop the expert systems based on a more interpret linguistic rules databases. A method for the synthesis of neuro-fuzzy networks based on decision trees is proposed. It does not require solving optimization to adjust the values of model parameters. The automated system for the synthesis of decision trees is created. It allows to construct simple and convenient model for further analysis in the form of decision trees.

The problem of technical diagnostics of vehicle bodies is solved using the proposed methods and software. The experiments of comparison of the proposed methods with known analogues are carried out.

Keywords: decision tree, neuro-fuzzy model, rule base, interpretability, training sample, recognition, evolutionary method.

Відповідальний випусковий Бодянський Є. В.