УДК 004.93

Олійник А. О.

Канд. техн. наук, доцент, Запорізький національний технічний університет, Україна

ВИДОБУВАННЯ ЗНАНЬ НА ОСНОВІ ДЕРЕВ РОЗВ'ЯЗКІВ ТА СТОХАСТИЧНОГО ПОШУКУ

Вирішено завдання розробки математичного забезпечення для автоматизації видобування набору знань у вигляді продукційних правил з навчальних вибірок даних. Об'єктом дослідження є процес побудови моделей неруйнівного контролю якості. Предмет дослідження становлять методи видобування продукційних правил для синтезу моделей контролю якості. Мета роботи: підвищити ефективність процесу виявлення продукційних правил для побудови моделей контролю якості на основі навчальних вибірок. Запропоновано стохастичний метод синтезу дерев розв'язків, який використовує інформацію про інформативність ознак, складність синтезованого дерева, а також точність його розпізнавання, що дозволяє на початковому етапі формувати множину деревовидних структур, що характеризуються простою ієрархією і невисокою помилкою розпізнавання, в процесі пошуку створювати нові множини рішень з урахуванням інформації про значущість ознак та інтерпретабельність створюваних дерев, що, у свою чергу, забезпечує можливість побудови дерев розв'язків з невеликою кількістю елементів (вузлів та зв'язків між ними) і прийнятною точністю розпізнавання, а також видобування на його основі найбільш цінних екземплярів. Розроблено програмне забезпечення, що реалізує запропонований метод. Проведено експерименти з дослідження властивостей запропонованого методу. Результати експериментів дозволяють рекомендувати запропонований метод для використання на практиці.

Ключові слова: вибірка, дерево розв'язків, модель контролю якості, продукційне правило, стохастичний пошук.

НОМЕНКЛАТУРА

$Br(\chi_k)$ – гіллястість дерева розв'язків χ_k;

Cl_c – c-й кластер;

d_{ab} – b-й вузол a-го рівня синтезованого дерева розв'язків;

E – похибка моделі контролю якості;

$G(\chi_k)$ – значення цільової функції k-го розв'язку при стохастичному пошуку;

$Gener(\chi_k)$ – узагальнення дерева розв'язків χ_k;

$GenerM(\chi_k)$ – узагальнення моделі дерева розв'язків χ_k;

$Idist(\chi_k)$ – внутрилистова відстань дерева розв'язків χ_k;

$Int(\chi_k)$ – інтерпретовність дерева розв'язків χ_k;

M – кількість атрибутів;

$N_{branch}(\chi_k)$ – кількість гілок дерева розв'язків χ_k;

N_{er} – кількість неправильно розпізнаних спостережень вибірки S;

$N_{level}(\chi_k)$ – кількість рівнів дерева розв'язків χ_k;

$N_{node}(\chi_k)$ – кількість вузлів дерева розв'язків χ_k;

N_χ – кількість розв'язків на кожній ітерації стохастичного пошуку;

P – набір характеристик (ознак) спостережень;

p_{mab} – ознака-перевірка у вузлі d_{ab} дерева розв'язків;

p_{qm} – значення m-го атрибуту q-го спостереження;

p_{TRab} – граничне значення ознаки-перевірки вузла d_{ab};

Q – кількість спостережень;

$R\sigma_m$ – значення рангу m-ї ознаки p_m;

RB – база правил;

$rule_r$ – r-те правило бази правил;

S – навчальна вибірка;

T – множина значень вихідного параметру;

t_q – значення вихідного параметру q-го спостереження;

V_m – оцінка індивідуальної інформативності m-ї ознаки;

$VE(p_m)$ – ентропія ознаки p_{qm};

σ_{mc} – ширина розкиду значень ознаки p_m в c-му кластері;

χ_k – k-й розв'язок стохастичного пошуку.

ВСТУП

Побудова автоматизованих систем неруйнівного контролю якості пов'язана з необхідністю синтезу моделей прийняття рішень [1]. Як базис для побудови таких моделей ефективно можуть використовуватися нейро-нечіткі мережі [2–4], які є гібридною моделлю обчислювального інтелекту, що характеризується високою інтерпретовністю та сполучає у собі властивості систем, заснованих на знаннях, і однорідних обчислювальних структур.

 DOI 10.15588/1607-3274-2014-2-16

Процес синтезу таких моделей пов'язаний з необхідністю видобування правил на основі заданих вибірок даних. Однак вибірки даних, що описують результати вимірювань характеристик реальних технічних об'єктів і процесів можуть містити дубляж інформації, зокрема, надлишкові для прийняття рішень ознаки й екземпляри [5, 6]. Крім того, можливі ситуації, при яких у вихідній вибірці кількість екземплярів одного класу істотно відрізняється від кількості екземплярів іншого класу (при використанні традиційного навчання екземпляри одного класу можуть пригнічувати екземпляри іншого класу) [5–7]. Отже, застосування відомих методів видобування продукційних правил для синтезу моделей контролю якості на основі нейро-нечітких мереж у деяких випадках є недоцільним.

Тому у цій роботі пропонується на основі заданих вибірок даних синтезувати дерева розв'язків і видобувати на їх основі продукційні правила, що дозволить виділяти найцінніші екземпляри, ранжирувати ознаки за значущістю й, отже, усувати деяку надлишковість інформації, а також скоротити простір пошуку й час синтезу нейро-нечітких моделей контролю якості.

Однак відомі методи синтезу дерев розв'язків [8–11] передбачають використання «жадібного» підходу, що не дозволяє в процесі побудови таких моделей повторно розглядати ознаки p_m, за якими вже було виконано розбиття. Це може привести до низьких узагальнюючих властивостей синтезованої моделі, внаслідок її складності, а, отже, до надлишкового числа правил, витягнутих з неї, що зробить систему правил більше громіздкою та менш інтерпретовною.

Тому в цій роботі для побудови дерев розв'язків пропонується використовувати інтелектуальний стохастичний пошук [11–13], що дозволяє досліджувати різні області пошукового простору й не використовує жадібну стратегію.

Метою роботи є створення методу видобування знань у вигляді продукційних правил на основі дерев розв'язків і стохастичного пошуку.

1 ПОСТАНОВА ЗАДАЧІ

Нехай задана множина спостережень $S =< P, T >$, де P – набір характеристик (ознак) спостережень, T – множина значень вихідного параметру. Набори значень P і T можуть бути подані у вигляді матриці (1) та вектора (2), відповідно:

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} & \cdots & p_{1M} \\ p_{21} & p_{22} & \cdots & p_{2m} & \cdots & p_{2M} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{q1} & p_{q2} & \cdots & p_{qm} & \cdots & p_{qM} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{Q1} & p_{Q2} & \cdots & p_{Qm} & \cdots & p_{QM} \end{pmatrix}, \quad (1)$$

$$T = \begin{pmatrix} t_1 & t_2 & \ldots & t_q & \ldots & t_Q \end{pmatrix}^{\mathrm{tr}}. \quad (2)$$

Тоді задача видобування продукційних правил полягає в пошуку такого набору правил $RB = \{rule_1, rule_2, ..., rule_{NR}\}$, при якому забезпечується прийнятне значення заданого критерію якості G, де кожне r-е правило $rule_r$ являє собою імплікацію антецедента (набору ознак p_m та їх граничних значень p_{TR}) та консеквента (значення t_q вихідного параметру T при виконанні умов, поданих в антецеденті). Як цільовий критерій G при видобуванні продукційних правил можуть бути використані, наприклад:

– похибка розпізнавання (у задачах з дискретним виходом T) [1–3, 14], що обчислюється за формулою: $E = N_{er} / Q$;

– середньоквадратична похибка (у випадку, коли вихідний параметр T може приймати дійсні значення з деякого діапазону $T \in [t_{\min}; t_{\max}]$) [1–3, 14], що розраховується за формулою: $E = \frac{1}{Q}\sum_{q=1}^{Q}\left(t_q - t_{q\,\mathrm{mod}}\right)^2$, де $t_{q\,\mathrm{mod}}$ – значення вихідного параметру q-го спостереження, розраховане за набором правил RB.

2 ОГЛЯД ЛІТЕРАТУРИ

Основні підходи до формування бази правил на основі вибірок даних S для синтезу моделей контролю якості на основі нейро-нечітких систем полягають у наступному [14, 15]:

– копіювання навчальної вибірки в базу знань – для кожного екземпляра навчальної вибірки формується окреме правило. Перевагою даного методу є простота та висока швидкість роботи, недоліком – відсутність узагальнюючих властивостей і громіздкість одержуваної мережі;

– оптимізація кількості продукційних правил – знаходження такого значення кількості продукційних правил NR, при якій значення помилки E є мінімальним, для чого при різних значеннях NR навчають мережу і вимірюють значення помилки, після чого оптимізують функцію $E(NR)$ за параметром S. Недоліком даного методу є дуже високі вимоги до обчислювальних ресурсів, обумовлені необхідністю заново навчати мережу на кожному кроці;

– спільна оптимізація ваг мережі та кількості продукційних правил шляхом вирішення багатоекстремальної оптимізаційної задачі або автоматичне визначення числа кластерів у навчальній вибірці та встановлення центрів функцій приналежності в їхні центри на основі кластер-аналізу;

– скорочення (редукція) правил. При цьому підході виключаються суперечливі правила, які взаємно компенсуються, а також одне з двох співпадаючих правил, як такі, що не несуть нової інформації. При скороченні видаляються ті продукційні правила, вплив яких на точність виявляється мінімальним після оцінки індивідуального

внеску кожного продукційного правила у вихідний сигнал мережі, одержуваної шляхом використання ортогонального методу найменших квадратів. Істотним недоліком методів скорочення є необхідність спочатку працювати зі свідомо надлишковою за розміром базою знань, що обумовлює в ряді випадків повільну роботу методів.

– нарощування (конструювання) правил: формується початкова база продукційних правил (вона може бути і порожньою), що потім послідовно поповнюється нечіткими правилами. Недоліком даного методу є відсутність явного зв'язку між процедурою додавання продукційних правил і точністю апроксимації, що повинна визначатися окремо.

Наявність зазначених недоліків обумовлює необхідність розробки нових методів побудови бази продукційних правил. Тому у цій роботі пропонується видобувати продукційні правила на основі дерев розв'язків [8–11], побудованих на основі заданих вибірок даних S.

Проте відомі методи ідентифікації дерев рішень, зокрема ID3, CART, CHAID, QUEST, C4.5 [8–11], мають певні недоліки, пов'язані з великою обчислювальною складністю, проблемами формування дерева рішень (ріст дерева, відсікання частини дерева) і т. ін. [8–10]. Крім того, такі методи використовують жадібну стратегію пошуку: якщо ознака була обрана один раз, і за нею виконано розбиття на підмножини, то метод не може повернутися назад і вибрати інший атрибут, який привів би до кращого розбиття, внаслідок чого в результаті часто синтезуються дерева розв'язків, що не забезпечують прийнятний рівень апроксимації [9]. Тому в даній роботі запропоновано для синтезу дерев розв'язків використовувати інтелектуальний стохастичних пошук [11–14], що дозволяє досліджувати різні області пошукового простору і не використовує жадібну стратегію.

3 МАТЕРІАЛИ ТА МЕТОДИ

Для синтезу дерев розв'язків пропонується використовується інформацію про інформативність ознак, складність синтезованого дерева, а також точність його розпізнавання. Це дозволить на початковому етапі формувати множину деревоподібних структур, що характеризуються простою ієрархією й невисокою похибкою розпізнавання, у процесі пошуку створювати нові множини розв'язків з урахуванням інформації про значущість ознак й інтерпретовність створюваних дерев, що, у свою чергу, забезпечить можливість побудови дерева розв'язків з невеликою кількістю елементів (вузлів і зв'язків між ними) і прийнятною точністю розпізнавання, а також видобування на його основі найцінніших екземплярів.

На етапі ініціалізації при синтезі дерев розв'язків пропонується генерувати початкову множину розв'язків $R^{(0)} = \left\{\chi_1^{(0)}, \chi_2^{(0)}, ..., \chi_{N\chi}^{(0)}\right\}$, де N_χ – кількість розв'язків у множині $R^{(0)}$.

Кожен k-й розв'язок χ_k являє собою структуру, що відповідає певному дереву розв'язків (рис. 1).

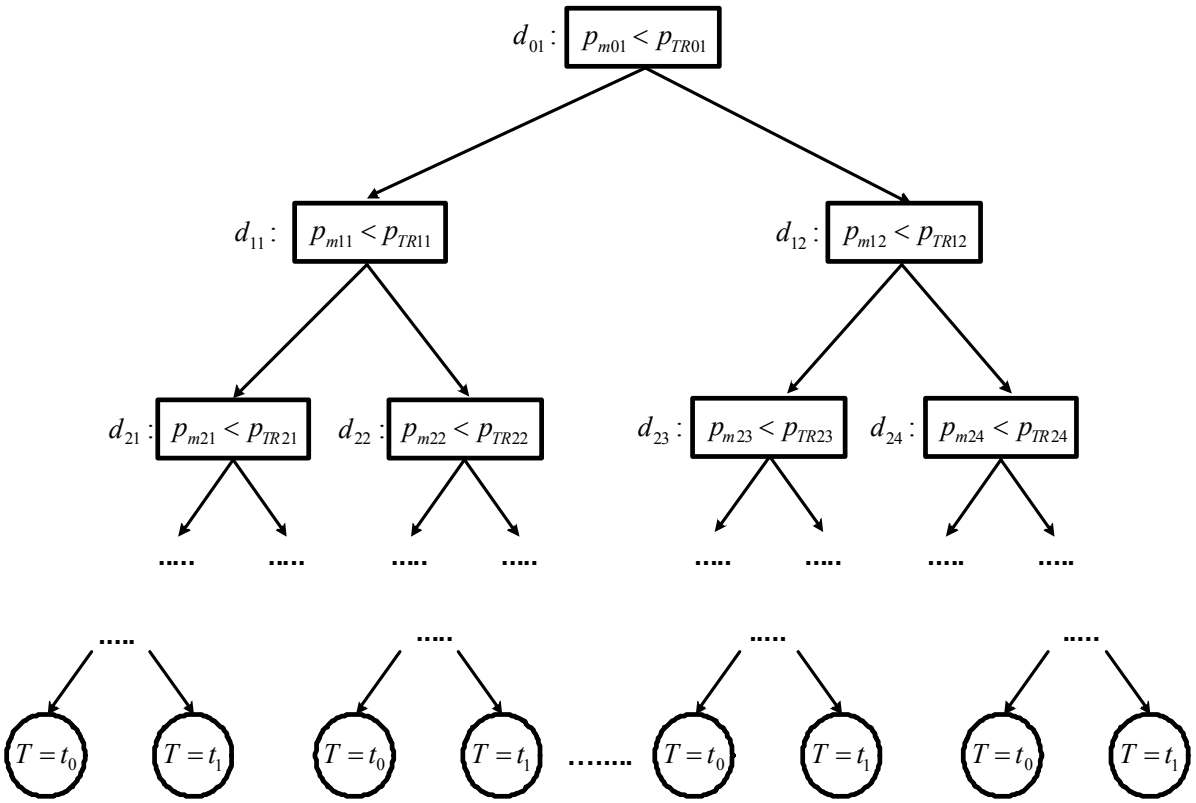

Рисунок 1 – Схематичне подання структури χ_k при синтезі дерев розв'язків

На рис. 1 позначення $T = t_a$ відповідає значенню t_a вихідного параметра T при проході від кореня дерева до одного з його листів (кінцевих вузлів, що містять значення вихідного параметра при виконанні умов, що знаходяться у вищестоящих вузлах-батьках).

Як видно з рис. 1, при синтезі кожного дерева розв'язків χ_k необхідно визначити його структуру, що представляє собою набір взаємозалежних вузлів d_{ab}, які містять інформацію про ознаку p_{mab}, за якою відбувається розбиття, її граничне значення p_{TRab}, а також посилання на лівого $d_{(a+1)(2b-1)}$ й правого $d_{(a+1)(2b)}$ нащадків: $d_{ab} =< p_{mab}, p_{TRab}, d_{(a+1)(2b-1)}, d_{(a+1)(2b)} >$. Отже, для побудови дерева розв'язків χ_k потрібно сформувати множину взаємозалежних вузлів-перевірок d_{ab}, визначивши для кожного з них ознаку-перевірку p_{mab} та її граничне значення p_{TRab}.

При виборі ознаки-перевірки p_{mab} для вузла d_{ab} будемо використовувати апріорну інформацію про значущість ознак V_m. Оскільки дерева розв'язків ефективно застосовуються, як правило, для вирішення задач розпізнавання, які характеризуються кінцевою кількістю класів вихідного параметру T, для оцінювання інформативності V_m ознак доцільно використовувати характеристики, що дозволяють оцінювати значущість ознаки p_m стосовно вихідного параметра T, який приймає дискретні значення t_q з кінцевої множини. Як такі характеристики можуть бути використані наступні критерії [6, 8, 14, 15]:

– ентропія ознаки $VE(p_m)$ – критерій, що відображає ступінь невизначеності стану об'єкта [8, 14], розраховується за формулою:

$$VE(p_m) = -\sum_{n=1}^{N_{\text{int}}(p_{mn})}\left(\rho(p_{mn})\sum_{l=1}^{N_{\text{int}}(T)}\rho(p_{mn},T_l)\log_2\rho(p_{mn},T_l)\right),$$

де $\rho(p_{mn}) = \dfrac{N(p_{mn})}{M}$ – ймовірність того, що значення ознаки p_m екземплярів вибірки S потрапить до n-го інтервалу діапазону її зміни; $N(p_{mn})$ – кількість екземплярів вибірки S, значення m-ї ознаки яких, належать n-му інтервалу діапазону її зміни; $N_{\text{int}}(p_{mn})$ – кількість інтервалів, на які розбивається діапазон значень m-ї ознаки p_m; $N_{\text{int}}(T)$ – кількість інтервалів, на які розбивається діапазон значень вихідного параметра T; $\rho(p_{mn},T_l) = \dfrac{N(p_{mn},T_l)}{N(p_{mn})}$ – умовна ймовірність того, що значення вихідного параметра T потрапить в l-й інтервал T_l за умови, що m-а ознака p_m потрапить в n-й інтервал p_{mn}; $N(p_{mn},T_l)$ – кількість екземплярів вибірки S, значення вихідного параметра T яких належать l-му інтервалу діапазону його зміни T_l за умови, що значення їх m-ї ознаки належить n-му інтервалу p_{mn};

– теоретико-інформаційний критерій $VT(p_m)$ – передбачає використання кількості інформації, що одержує система в процесі розпізнавання об'єктів у результаті використання оцінюваної ознаки:

$$VT(p_m) = \sum_{l=1}^{N_{\text{int}}(T)}\sum_{n=1}^{N_{\text{int}}(p_{mn})}\rho(p_{mn},T_l)\log_2\frac{\rho(p_{mn},T_l)}{\rho(p_{mn})\rho(T_l)},$$

де $\rho(T_l)$ – ймовірність того, що значення вихідного параметра T потрапить в l-й інтервал T_l діапазону зміни його значень.

Однак застосування таких критеріїв передбачає, що відомими є всі значення всіх екземплярів у навчальній вибірці S, що є не завжди можливим при вирішенні практичних задач діагностування й управління якістю продукції (це обумовлено можливими проблемами при вимірюванні параметрів деяких реальних технічних об'єктів або процесів). У випадку, якщо деякі екземпляри навчальної вибірки S містять пропущені значення ознак p_m або вихідного параметра T, пропонується використовувати наступний підхід до оцінювання індивідуальної значущості V_m. Оцінювати інформативність ознак будемо виходячи з їхньої значущості для визначення границь кластерів – груп компактно розташованих екземплярів у просторі ознак. Для цього за допомогою методів кластерного аналізу [7, 14] пропонується виявляти групи екземплярів (кластери) виходячи з їх геометричного розташування.

Будемо вважати, що ознака p_m є тим важливішою для кластера Cl_c ($c = 1,2,...,N_{cl}$), чим меншою є ширина розкиду σ_{mc} її значень у цьому кластері, що розраховується за формулою: $\sigma_{mc} = \sqrt{\sum_{q=1}^{N_{inst}(Cl_c)}\left(p_{mq} - \overline{p_{mc}}\right)^2}$, де $\overline{p_{mc}}$ – середнє значення m-ї ознаки p_m в c-му кластері; $N_{inst}(Cl_c)$ – кількість екземплярів c-го кластеру Cl_c.

Потім найбільш значущій ознаці c-го кластера (ознаці з мінімальним значенням величини σ_{mc}) будемо ставити у відповідність найбільше значення рангу $R\sigma_{mc} = M$, наступній за зростанням величини σ_{mc} ознаці привласнимо значення $R\sigma_{mc} = M - 1$ і т. д. При однакових значеннях величини σ_{mc} ознакам ставляться у відповідність середні значення рангів $R\sigma_{mc}$. Ознакам, що характеризуються низькими значеннями індивідуальної інформативності для екземплярів c-го кластеру ($\sigma_{mc} < \sigma_{mc\min}$), ставиться у відповідність нульове значення величини $R\sigma_{mc}$: $R\sigma_{mc} = 0$.

Значення загального рангу $R\sigma_m$ m-ї ознаки p_m по всіх N_{cl} кластерах визначимо як суму значень рангів $R\sigma_{mc}$ за формулою: $R\sigma_m = \sum_{c=1}^{N_{cl}} R\sigma_{mc}$. Оцінку індивідуальної інформативності V_m m-ї ознаки визначимо за формулою:

$$V_m = \frac{R\sigma_m}{\max\limits_{n=1,2,...,M}\{R\sigma_n\}}.$$

Таким чином, ознака з максимальним значенням рангу $R\sigma_m$ є найбільш інформативною і характеризується одиничним значенням критерію V_m. Використання запропонованого критерію дозволяє ранжирувати ознаки, виходячи з їхньої значущості для опису границь компактного розташування екземплярів, що, у свою чергу, дозволяє оцінювати індивідуальну інформативність ознак у вибірках, у яких деякі екземпляри містять пропущені значення ознак або вихідного параметра.

Після оцінювання інформативності V_m кожної ознаки p_m вибірки S відбувається формування дерев χ_k у вигляді відповідних структур даних. Ознаки p_m з високими оцінками індивідуальної інформативності V_m є більш значущими (істотно впливають на вихідний параметр T), отже, у запропонованому методі такі ознаки будуть мати більшу ймовірність відбору як ознаки-перевірки p_{mab} для відповідного вузла d_{ab}.

Нехай оцінка індивідуальної інформативності m-ї ознаки p_m дорівнює V_m. Тоді нормована оцінка індивідуальної інформативності ознак V_{m_norm} у вибірці S може бути розрахована за формулою: $V_{m\_\text{norm}} = \dfrac{V_m - V_{\min}}{V_{\max} - V_{\min}}$, де $V_{\min}$ та $V_{\max}$ – мінімальне та максимальне значення інформативності V_m ознак p_m у вибірці S, відповідно.

Вибір ознаки p_m для використання у вузлі-перевірці d_{ab} будемо виконувати виходячи зі значення величини V_{m_norm}. Для цього будемо послідовно переглядати ознаки $p_1, p_2, ..., p_M$ й порівнювати значення величини V_{m_norm} з $rand[0;1]$ – випадково згенерованим числом в інтервалі [0; 1]. У випадку, якщо виконується умова $V_{m\_\text{norm}} \geq rand[0;1]$, ознака p_m вважається добре поділяючою екземпляри на класи й включається в поточний вузол d_{ab} дерева розв'язків χ_k як ознака-перевірка.

Далі для ознаки-перевірки p_{mab} у вузлі d_{ab} виконується визначення граничного значення p_{TRab}. Для цього розраховуються значення похибки розпізнавання E_{ab} екземплярів S_{ab}, що попадають у вузол d_{ab}, при різних значеннях $p_{TRab} \in [p_{mab\min}; p_{mab\max}]$, і вибирається таке значення p_{TRab}, при якому значення похибки розпізнавання E_{ab} буде найменшим. У випадку, якщо знайдено значення p_{TRab}, при якому похибка розпізнавання знаходиться в припустимих межах $E_{ab} < E_{TR}$, тоді вважається, що ознака p_{mab} при граничному значенні p_{TRab} дозволяє здійснювати прийнятне розбиття множини екземплярів S_{ab} на класи t_0 й t_1, отже, її нащадками будуть вузли, що представляють собою листи дерева розв'язків χ_k – кінцеві вузли, що містять значення вихідного параметра $T = t_0$ й $T = t_1$. Якщо в результаті розбиття $p_{mab} < p_{TRab}$ в одну з частин дерева ввійдуть екземпляри S_{ab} з єдиним значенням вихідного параметра (наприклад, $T = t_1$), то нащадок, що відповідає даній умові, стає листом, і розбиття триває тільки для другого нащадка вузла d_{ab}.

Формування дерева розв'язків χ_k триває доти, поки не буде досягнуто прийнятної точності розпізнавання ($E < E_{\min}$) або інші критерії, що характеризують складність дерева (досягнення максимально припустимої кількості вузлів, рівнів, гілок та ін.). Аналогічним чином на етапі ініціалізації запропонованого методу формується N_χ дерев розв'язків χ_k.

Потім виконується оцінювання якості синтезованих дерев розв'язків $\chi_k, k = 1,2,...,N_\chi$. Для цього пропонується використовувати цільову функцію $G = G(\chi_k)$, що враховує інтерпретовність дерева $Int(\chi_k)$ і його розпізнаючи властивості (похибку розпізнавання $E(\chi_k)$), і може бути визначена за формулою: $G(\chi_k) = \gamma_1 Int(\chi_k) + \gamma_2 E(\chi_k)$, де γ_1 й γ_2 – коефіцієнти, що дозволяють урахувати важливість критеріїв $Int(\chi_k)$ та $E(\chi_k)$, відповідно.

Для оцінювання інтерпретовності $Int(\chi_k)$ дерев розв'язків пропонується використовувати такі критерії:

– ширина дерева – може бути визначена як кількість його гілок $N_{branch}(\chi_k)$ або кількість вузлів $N_{node}(\chi_k) = N_{branch}(\chi_k) + 1$;

– глибина дерева – визначається як кількість його рівнів $N_{level}(\chi_k)$;

– гіллястість дерева $Br(\chi_k)$ – пропонується обчислювати як відношення кількості вузлів $N_{node}(\chi_k)$ дерева χ_k до максимально можливої кількості вузлів дерева глибини $N_{level}(\chi_k)$: $Br(\chi_k) = \dfrac{N_{node}(\chi_k)}{\max\text{node}(N_{level}(\chi_k))}$, де $\max\text{node}(N_{level}(\chi_k))$ – максимально можлива кількість вузлів дерева χ_k глибини $N_{level}(\chi_k)$ – величина, що може бути визначена за формулою:

$$\max\text{node}(N_{level}(\chi_k)) = \sum_{c=1}^{N_{level}(\chi_k)} 2^{c-1} = 2^{N_{level}(\chi_k)} - 1;$$

– узагальнення рішень $Gener(\chi_k)$ – відношення кількості листів (вузлів-рішень) $N_{leaf}(\chi_k)$ дерева χ_k до кількості екземплярів Q навчальної вибірки S: $Gener(\chi_k) = \dfrac{N_{leaf}(\chi_k)}{Q}$.

– внутрилистова відстань $Idist(\chi_k)$ між екземплярами вибірки, що потрапили в конкретний лист (екземплярами, які відповідають конкретним умовам, поданим у вигляді шляхів від кореня дерева до його листів) – чим менше дана відстань, тим вище компактність рішень у відповідних листах, і, отже, тим кращим є розбиття, що виконується деревом χ_k. Критерій $Idist(\chi_k)$ пропонується обчислювати за формулою:

$Idist(\chi_k)=\frac{1}{N_{leaf}(\chi_k)}\sum_{c=1}^{N_{leaf}(\chi_k)}Idist_c$, де $Idist_c$ – середня відстань між екземплярами вибірки S, що попадають в c-й лист дерева χ_k, – величина, що розраховується за формулою: $Idist_c=\frac{1}{N_{inst}(Leaf_c)}\sum_{q=1}^{N_{inst}(Leaf_c)}\sqrt{\sum_{m=1}^{M}\left(p_{mq}-\overline{p_{mc}}\right)^2}$, де $N_{inst}(Leaf_c)$ – кількість екземплярів вибірки S, що попадають в c-й лист $Leaf_c$ дерева χ_k; $\overline{p_{mc}}$ – середнє значення m-ї ознаки екземплярів, що попадають в c-й лист дерева χ_k. З метою приведення значень показника $Idist(\chi_k)$ до одного інтервалу при аналізі різних вибірок даних, як значення ознак p_{mq} рекомендується використовувати нормовані значення;

– узагальнення моделі дерева $GenerM(\chi_k)$ – відношення кількості всіх настроюваних параметрів $N_{param}(\chi_k)$ моделі на основі дерева розв'язків до розмірності вибірки S: $GenerM(\chi_k)=\frac{N_{param}(\chi_k)}{Q\cdot M}$, де $N_{param}(\chi_k)$ – кількість настроюваних параметрів дерева χ_k – визначається за формулою: $N_{param}(\chi_k)=2N_{node}(\chi_k)$, оскільки кожен d_{ab} вузол характеризується двома параметрами p_{mab} та p_{TRab}.

Як критерій інтерпретовності $Int(\chi_k)$ можна використовувати один із запропонованих вище критеріїв ($N_{branch}(\chi_k)$, $N_{node}(\chi_k)$, $N_{level}(\chi_k)$, $Br(\chi_k)$, $Idist(\chi_k)$) або їх комбінацію.

Після оцінювання якості синтезованих дерев розв'язків χ_k, $k=1,2,...,N_\chi$ виконується перевірка критеріїв зупинення стохастичного пошуку. Як такі критерії можуть бути використані: досягнення прийнятного значення цільової функції $G(\chi_k)$, перевищення максимально припустимої кількості ітерацій N_{It}, відсутність істотних покращень значення цільової функції $G(\chi_k)$ протягом заданої кількості ітерацій.

При невиконанні критеріїв зупинення виконується оператор відбору розв'язків χ_k для формування нової множини $R^{(i)}\to R^{(i+1)}=\left\{\chi_1^{(i+1)},\chi_2^{(i+1)},...,\chi_{N\chi}^{(i+1)}\right\}$. Для цього із множини $R^{(i)}$ відбираються розв'язки $\chi_k^{(i)}$ з метою створення набору розв'язків $RP^{(i)}$, допущених до відтворення, – розв'язків, на основі яких буде згенеровано новий набір $R^{(i+1)}$.

Для відбору розв'язків $\chi_k^{(i)}$ у множину $RP^{(i)}$ кожній структурі $\chi_k^{(i)}$ ставиться у відповідність інтервал $GI(\chi_k)\in[GI_{\min}(\chi_k);GI_{\max}(\chi_k))$. Величини $GI_{\min}(\chi_k)$ та $GI_{\max}(\chi_k)$ обчислюються в такий спосіб: $GI_{\min}(\chi_k)=GI_{\max}(\chi_{k-1})$, $GI_{\max}(\chi_k)=GI_{\min}(\chi_k)+G_O(\chi_k)$, де $GI_{\min}(\chi_k)$ й $GI_{\max}(\chi_k)$ – мінімальне та максимальне значення в інтервалі $GI(\chi_k)$, відповідно; $GI_{\min}(\chi_1)=0$ – мінімальне значення в інтервалі $GI(\chi_1)$ першого розв'язку χ_1 в популяції $R^{(i)}$; $G_O(\chi_k)$ – відносне значення цільової функції $G(\chi_k)$ розв'язку χ_k в множині $R^{(i)}$, визначається за формулою: $G_O(\chi_k)=\frac{GM-G(\chi_k)}{\sum_{K=1}^{N_\chi}(GM-G(\chi_K))}$, де $GM=\max_{k=1,2,...,N_\chi}\{G(\chi_k)\}$ – максимальне значення цільової функції в множині хромосом $R^{(i)}=\left\{\chi_1^{(i)},\chi_2^{(i)},...,\chi_{N\chi}^{(i)}\right\}$.

Як видно з наведених вище формул, $GI_{\min}(\chi_0)=0$ і $GI_{\max}(\chi_{N_\chi})=1$. Отже, кожному розв'язку χ_k ставиться у відповідність деякий інтервал залежно від значення його цільової функції $G(\chi_k)$: чим вище значення $G(\chi_k)$, тим ширше інтервал $[GI_{\min}(\chi_k);GI_{\max}(\chi_k))$. У сукупності інтервали $GI(\chi_k)$, $k=1,2,...,N_\chi$ утворюють інтервал $GI(R^{(i)})\in[0;1)$.

Після цього генерується випадкове число $rnd=rand[0;1)$ з інтервалу $[0;1)$. У множину допущених до відтворення розв'язків $RP^{(i)}$ заноситься розв'язок χ_k, в інтервал $GI(\chi_k)$ якого попадає випадково згенероване число rnd: $rnd\in[GI_{\min}(\chi_k);GI_{\max}(\chi_k))$. Таким чином, чим більше ширина діапазону $GI(\chi_k)$, що визначається значенням цільової функції $G(\chi_k)$, тим вище ймовірність розв'язку χ_k бути відібраним для відтворення. Генерація випадкових чисел rnd і відбір розв'язків χ_k для відтворення триває доти, поки не буде повністю сформовано множину $RP^{(i)}$, $\left|RP^{(i)}\right|=N_\chi$.

Потім на основі розв'язків χ_k із множини $RP^{(i)}$ виконується створення нових розв'язків $\chi_k^{(i+1)}$. Для цього із множини $RP^{(i)}$ вибираються випадковим чином два розв'язки $\chi_{parent1}$ й $\chi_{parent2}$. Після чого в кожному з розв'язків-батьків вибираються вузли $d_{abparent1}$ й $d_{abparent2}$ (при цьому ознака-перевірка $p_{mabparent2}$ вузла $d_{abparent2}$ не повинна бути ідентичною кожній з ознак, що знаходяться у вузлах дерева $\chi_{parent1}$ по напрямку від його кореня до вузла $d_{abparent1}$, включно, аналогічна умова діє й при виборі вузла $d_{abparent1}$), по яких буде відбуватися обмін між відповідними частинами дерев розв'язків

$\chi_{parent1}$ і $\chi_{parent2}$. У результаті такого обміну створюються два нові розв'язки χ_{child1} й χ_{child2}. При цьому перший розв'язок-нащадок χ_{child1} є ідентичним розв'язку $\chi_{parent1}$ за винятком частини (піддерева), вихідним вузлом якої є $d_{abparent1}$. Замість вузла $d_{abparent1}$ в розв'язку χ_{child1} знаходиться вузол $d_{abparent2}$, після якого розташовується відповідна частина дерева $\chi_{parent2}$. Аналогічно формується дерево-нащадок χ_{child2}. Після цього виконується перерахунок граничних значень p_{TRab} ознак-перевірок p_{mab} від вузла $d_{abparent2}$ ($d_{abparent1}$) дерева χ_{child1} (χ_{child2}) до відповідних кінцевих вузлів, що містять значення вихідного параметру. У випадку, якщо в результаті таких перетворень на нижніх рівнях дерев χ_{child1} і χ_{child2} виявляються вузли, перевірки в який пов'язані з ознаками, що вже зустрічаються на більш високих рівнях дерева при проході від кореня до відповідних листів, то такі ознаки-перевірки у вузлах низьких рівнів заміняються на ознаки-перевірки наступного рівня відповідного дерева χ_{child1} або χ_{child2}. Такий підхід дозволяє для однакових батьків створювати множини нащадків. Створення нових розв'язків $\chi_k^{(i+1)}$ за допомогою описаного вище підходу триває доти, поки не буде сформовано $N_{cross} = \beta N_\chi$ розв'язків, де β – коефіцієнт, що визначає значущість формування нової множини розв'язків за допомогою описаної вище процедури схрещування.

Другим способом формування нових розв'язків $\chi_k^{(i+1)}$ є мутація, що припускає виконання деяких змін над структурою χ_k, відібраною із множини $RP^{(i)}$. У розробленому стохастичному методі синтезу дерев розв'язків оператор мутації пропонується виконувати в такий спосіб. Спочатку з множини $RP^{(i)}$ відбирається розв'язок $\chi_{mutated}$, у якому випадковим чином вибирається мутуючий вузол $d_{abmutated}$, потім у цьому вузлі виконується заміна ознаки-перевірки p_{mab} на іншу, що не знаходиться у вузлах дерева $\chi_{mutated}$ по напрямку від його кореня до вузла $d_{abmutated}$, включно. Після визначення нової ознаки-перевірки, виконується перерахунок її граничного значення p_{TRab} й подальше переформатування дерева, починаючи від вузла $d_{abmutated}$ (аналогічно етапу ініціалізації відбувається побудова фрагмента дерева від вузла $d_{abmutated}$ й доти, поки не будуть досягнуті відповідні критерії зупинення процесу формування дерева). Формування нових розв'язків $\chi_k^{(i+1)}$ за допомогою мутації триває, поки не буде сформовано $N_{mutation} = \gamma N_\chi$ розв'язків, де γ – коефіцієнт, що визначає значущість створення нової множини розв'язків за допомогою мутації.

У нову множину розв'язків $R^{(i+1)}$ заносяться N_{cross} й $N_{mutation}$ розв'язків, згенерованих за допомогою схрещування й мутації, а також $N_{elite} = \alpha N_\chi$, елітних розв'язків $\chi_k^{(i)}$ із множини $R^{(i)}$, що характеризуються найкращими значеннями цільової функції $G\left(\chi_k^{(i)}\right)$ в популяції $R^{(i)}$, де α – коефіцієнт, що визначає значущість включення елітних особин у нову множину.

Потім виконується оцінювання значень цільової функції G для розв'язків $\chi_k^{(i+1)}$: $G = G\left(\chi_k^{(i+1)}\right)$, $k = 1,2,...,N_\chi$ і формування нової множини $R^{(i+2)} = \left\{\chi_1^{(i+2)}, \chi_2^{(i+2)}, ..., \chi_{N\chi}^{(i+2)}\right\}$. Даний процес триває до виконання критеріїв зупинення.

Результатом виконання стохастичного методу синтезу дерев розв'язків є дерево χ_{opt} з мінімальним значенням цільової функції $G(\chi_{opt}) = \min\limits_{k=1,2,...,N_\chi} \{G(\chi_k)\}$.

Після цього на основі синтезованого дерева розв'язків χ_{opt} виконується видобування правил RB, що представляють собою найцінніші екземпляри. Для цього обробляється кожний шлях від кореня дерева χ_{opt} до листа, у результаті чого будується відповідне правило, що узагальнює інформацію, подану в деякій множині екземплярів вихідної вибірки S. Використовуючи такий підхід, видобувається N_{RDT} правил, загальна кількість яких дорівнює кількості листів (кінцевих вузлів, що містять значення вихідного параметра) синтезованого дерева розв'язків χ_{opt}.

Таким чином, розроблений стохастичний метод синтезу дерев розв'язків використовує інформацію про інформативність ознак, складність синтезованого дерева, а також точність його розпізнавання, що дозволяє на початковому етапі формувати множину деревоподібних структур, яка характеризуються простою ієрархією й невисокою похибкою розпізнавання, у процесі пошуку створювати нові множини розв'язків з урахуванням інформації про значущість ознак й інтерпретовність генерованих дерев, що, у свою чергу, забезпечує можливість побудови дерева розв'язків із невеликою кількістю елементів (вузлів і зв'язків між ними) та прийнятною точністю розпізнавання, а також видобування на його основі найцінніших екземплярів.

4 ЕКСПЕРИМЕНТИ

Виконаємо експериментальне дослідження розробленого стохастичного методу синтезу дерев розв'язків. Для цього порівняємо його з відомими аналогами – методом C4.5 [8–10], методом CART [8–10], а також еволюційним методом побудови дерев розв'язків, запропонованим в [16].

З метою експериментального порівняння запропонованого й відомих методів на мові C# розроблено програмне забезпечення, що дозволяє на основі заданої вибірки даних $S = < P, T >$ виконувати побудову дерев розв'язків за допомогою різних методів. За допомогою розробленого програмного забезпечення вирішувалася задача прийняття рішень при неруйнівному контролі якості кузовів автотранспортних засобів [17].

При виготовленні автотранспортних засобів важливим етапом є неруйнівний контроль якості кузовів. Виявлення некондиційних виробів (кузовів) на ранніх стадіях виготовлення автомобіля дозволить зменшити витрати на усунення дефектів, і, отже, зменшити собівартість виробництва.

У процесі виробництва кузовів автотранспортних засобів на кожному етапі їхнього виготовлення виміряються деяка група параметрів – контрольних точок, розташованих на кузові та навісних вузлах:

– перша група – контрольні точки на чорному (незабарвленому) кузові;

– друга група – контрольні точки на навісних вузлах (дверях і капоті);

– третя група – зазори й сполучення між навісними вузлами й кузовом.

До аналізованих на першому й другому етапах параметрів (першої й другої групи) відносяться відхилення від номінальних значень контрольних точок. Як правило, більшість таких параметрів знаходиться в областях допуску, і, отже, не впливають на якість кузова транспортного засобу.

Вимірювання параметрів третьої групи пов'язано з необхідністю установки навісних вузлів (дверей і капота) на пофарбований кузов. Однак при установці навісних вузлів на кузов можуть виникнути деякі деформації, обумовлені відхиленнями номінальних розмірів чорного кузова, дверей і капота від еталонних розмірів (при цьому відхилення кожного з вимірюваних параметрів першої й другої групи може знаходиться в межах допуску), а також іншими факторами, що виникають при складанні. Такі деформації приводять до утворення зазорів і сполучень між навісними вузлами й кузовом, є досить частими й помітними для покупців продукції. Усунення таких недоліків пов'язано з необхідністю розбирання кузова й навісних вузлів, а також з повторним фарбуванням і складанням, що ускладнює й здорожує процес виготовлення якісних виробів.

Тому актуальною є задача побудови моделей залежностей показників третьої групи від параметрів першої й другої груп.

Виявлено, що найбільш важливими параметрами перших двох груп є 18 точок, розташованих в області порога кузова й в областях кріплення петель для складання кузова й навісних вузлів [17]. При цьому в шести точках фіксуються відхилення по двох координатах (третя координата є базовою, внаслідок чого відхилення по даній координаті є нульовим), в інших дванадцяти вимірюються всі три координати, отже, вибірка даних містила значення 48 вхідних параметрів. Також виділено 16 істотних параметрів третьої групи (зазори й сполучення між дверима й порогом, капотом і крилом, передніми й задніми дверима й ін.).

Таким чином, необхідно синтезувати 16 моделей залежностей параметрів третьої групи від 48 вхідних ознак (параметрів першої й другої груп). Вихідна вибірка містила інформацію про 172 вироби. Нижче наведено результати побудови однієї з моделей. Для інших параметрів отримано аналогічні результати.

Як критерій оцінювання інтерпретовності $Int(\chi_k)$ дерев розв'язків χ_k при використанні стохастичного й еволюційного методів використовувалося узагальнення розв'язків $Gener(\chi_k) = N_{leaf}(\chi_k)/Q$, оскільки кількість листів N_{leaf} відповідає загальній кількості правил N_{RDT}, що видобуваються з дерева χ_k, і, отже, визначає структуру базу правил, синтезованої на основі дерева розв'язків χ_k. Коефіцієнти, що дозволяють врахувати важливість критеріїв $Int(\chi_k)$ і $E(\chi_k)$, вибиралися рівними 0,5 (кожний критерій $Int(\chi_k)$ і $E(\chi_k)$ в процесі експериментів мав однакову значущість), отже, цільова функція $G(\chi_k) = \gamma_1 Int(\chi_k) + \gamma_2 E(\chi_k)$ при стохастичному і еволюційному пошуку визначалася в такий спосіб: $G(\chi_k) = 0{,}5 Gener(\chi_k) + 0{,}5 E(\chi_k)$.

5 РЕЗУЛЬТАТИ

Результати експериментів по дослідженню різних методів синтезу дерев розв'язків при побудові діагностичної моделі якості кузовів автомобілів наведено в таблиці 1.

6 ОБГОВОРЕННЯ

Як видно з таблиці 1, запропонований стохастичний метод і еволюційний метод синтезу дерев розв'язків дозволяють будувати більш прийнятні діагностичні моделі на основі дерев розв'язків (значення критерію G нижче в порівнянні з моделями, синтезованими на основі методів CART [8–10] і C4.5 [8–10]), оскільки не використовують жадібну стратегію. Розроблений стохастичний метод забезпечив побудову дерева розв'язків, що характеризується незначною кількістю структурних елементів ($N_{level} = 7$, $N_{node} = 73$, $N_{leaf} = 37$), а також високими апроксимаційними (помилка навчальної вибірки склала $E = 0{,}016$) і узагальнюючими (помилка на тестових даних $E_t = 0{,}028$) властивостями, що досягається за рахунок використання в процесі синтезу інформації про інформативність ознак, складність синтезованого дерева та точність його розпізнавання. Скорочення часу пошуку ($t = 7{,}31$) в по-

Таблиця 1 – Результати експериментів

Метод	N_{level}	N_{node}	N_{leaf}	*Gener*	*GenerM*	*Br*	*G*	*Idist*	*t*, мс	*E*	E_t
CART [8–10]	8	95	48	0,279	0,023	0,37	0,166	0,012	15,12	0,052	0,093
C4.5 [8–10]	8	103	52	0,302	0,025	0,4	0,174	0,078	14,82	0,046	0,072
Еволюційний метод [16]	7	83	42	0,244	0,02	0,65	0,131	0,0041	7,42	0,018	0,031
Стохастичний метод	7	73	37	0,215	0,018	0,57	0,116	0,0032	7,31	0,016	0,028

рівнянні з відомими методами забезпечено за рахунок формування на етапі ініціалізації запропонованого методу множини деревоподібних структур, що характеризуються простою ієрархією та невисокою помилкою розпізнавання, а також створення в процесі пошуку нових множин розв'язків із урахуванням інформації про значущість ознак та інтерпретовність генерованих дерев.

Низьке значення критерію узагальнення розв'язків $Gener = 0{,}215$ свідчить про високі узагальнюючі здатності дерева: вибірку з 172 екземплярів перетворено в дерево розв'язків, з якого, у свою чергу виділено $N_{leaf} = 37$ продукційних правил.

Дерева, синтезовані за допомогою еволюційного [16] і запропонованого стохастичного методів є більш гіллястими (значення критеріїв становлять $Br = 0{,}65$ і $Br = 0{,}62$, відповідно) у порівнянні з деревами, побудованими за допомогою методів CART [8–10] і C4.5 [8–10] ($Br = 0{,}37$ і $Br = 0{,}4$), що свідчить про більш компактне розташування вузлів.

Таким чином, результати порівняльного аналізу показали, що запропонований стохастичний метод синтезу дерев розв'язків не уступає по якості побудови деревоподібних моделей прийняття рішень відомим методам, і забезпечує можливість побудови дерев розв'язків з невеликою кількістю структурних елементів і прийнятною точністю розпізнавання.

ВИСНОВКИ

У роботі вирішено актуальну задачу автоматизації видобування знань у вигляді набору продукційних правил з навчальних вибірок даних.

Наукова новизна роботи полягає в тому, що запропоновано стохастичний метод синтезу дерев розв'язків, що використовує інформацію про інформативність ознак, складність синтезованого дерева, а також точність його розпізнавання, що дозволяє на початковому етапі формувати множину деревоподібних структур, яка характеризуються простою ієрархією та невисокою помилкою розпізнавання, у процесі пошуку створювати нові множини розв'язків з урахуванням інформації про значущість ознак і інтерпретовність генерованих дерев, що, у свою чергу, забезпечує можливість побудови дерева розв'язків з невеликою кількістю елементів (вузлів і зв'язків між ними) та прийнятною точністю розпізнавання, а також видобування на його основі найцінніших екземплярів.

Запропоновано систему критеріїв оцінювання моделей на основі дерев розв'язків, що містить у собі критерії оцінювання апроксимаційних властивостей (помилка розпізнавання) і інтерпретовності (ширина, глибина, гіллястість, узагальнення рішень, узагальнення моделі, внутрилистова відстань) синтезованого дерева. Розроблену систему критеріїв можна використовувати для автоматизації аналізу властивостей і порівняння моделей на основі дерев розв'язків при вирішенні задач неруйнівного контролю якості.

Практична цінність отриманих результатів полягає в тому, що: розроблено програмне забезпечення, яке реалізує запропонований метод і дозволяє виконувати побудову моделей контролю якості на основі дерев розв'язків, а також видобувати продукційні правила з вибірок даних; вирішено практичне завдання прийняття рішень при неруйнівному контролі якості кузовів автотранспортних засобів.

Перспективи подальших досліджень полягають у застосуванні запропонованого підходу до видобування знань у вигляді набору продукційних правил з навчальних вибірок даних при синтезі нейро-нечітких моделей для вирішення практичних задач неруйнівного контролю якості.

ПОДЯКИ

Роботу виконано в рамках держбюджетної науково-дослідної теми Запорізького національного технічного університету «Інтелектуальні інформаційні технології автоматизації проектування, моделювання, керування та діагностування виробничих процесів і систем» (номер державної реєстрації 0112U005350) за підтримки міжнародного проекту «Centers of Excellence for young RESearchers» (CERES) програми «Tempus» Європейської Комісії (реєстраційний номер 544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES).

СПИСОК ЛІТЕРАТУРИ

1. Ding S. X. Model-based fault diagnosis techniques: design schemes, algorithms, and tools / S. X. Ding. – Berlin: Springer, 2008. – 473 p.
2. Rutkowski L. Flexible neuro-fuzzy systems : structures, learning and performance evaluation / L. Rutkowski. – Boston : Kluwer, 2004. – 276 p.
3. Нейро-фаззи сети Петри в задачах моделирования сложных систем / [Е. В. Бодянский, Е. И. Кучеренко, А. И. Михалев] – Днепропетровск : Системные технологии. – 2005. – 311 с.
4. Jang J. R. ANFIS: Adaptive-network-based fuzzy inference system / J. R. Jang // IEEE transactions on systems and cybernetics. – 1993. – Vol. 23. – P. 665–685. DOI: 10.1109/21.256541.
5. Mulaik S. A. Foundations of Factor Analysis / S. A. Mulaik. – Boca Raton, Florida: CRC Press. – 2009. – 548 p.
6. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken: John Wiley & Sons, 2008. – 339 p.
7. Abonyi J. Cluster analysis for data mining and system identification / J. Abonyi, B. Feil. – Basel : Birkhäuser, 2007. – 303 p.
8. Rokach L. Data Mining with Decision Trees. Theory and Applications / L. Rokach, O. Maimon. – London : World Scientific Publishing Co, 2008. – 264 p. DOI: 10.1142/9097.
9. Quinlan J. R. Induction of decision trees / J. R. Quinlan // Machine Learning. – 1986. –No. 1. – P. 81–106. DOI: 10.1007/BF00116251.
10. Classification and regression trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. – California : Wadsworth & Brooks, 1984. – 368 p.
11. 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.
12. Yu X. Introduction to Evolutionary Algorithms (Decision Engineering) / X. Yu, M. Gen. – London: Springer, 2010. – 418 p. DOI: 10.1007/978-1-84996-129-5.
13. Gen M. Genetic algorithms and engineering design / M. Gen, R. Cheng. – New Jersey: John Wiley & Sons, 1997. – 352 p. DOI: 10.1002/9780470172254

14 Computational intelligence in fault diagnosis / eds.: V. Palade, C.D. Bocaniala, L. Jain. – London: Springer, 2006. – 362 p. DOI: 10.1007/978-1-84628-631-5.
15. Субботін С. О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень : навч. посібник / С. О. Субботін. – Запоріжжя : ЗНТУ, 2008. – 341 с.
16. Гофман Е. А. Эволюционный метод синтеза деревьев решений / Е. А. Гофман, А. А. Олейник, С. А. Субботин // Штучний інтелект. – 2011. – № 2. – С. 6–14.
17. 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.

Стаття надійшла до редакції 01.09.2014.
Після доробки 26.09.2014.

Олейник А. А.
Канд. техн. наук, доцент, Запорожский национальный технический университет, Украина

ИЗВЛЕЧЕНИЕ ЗНАНИЙ НА ОСНОВЕ ДЕРЕВЬЕВ РЕШЕНИЙ И СТОХАСТИЧЕСКОГО ПОИСКА

Решена задача разработки математического обеспечения для автоматизации извлечения знаний в виде набора продукционных правил из обучающих выборок данных. Объектом исследования являлся процесс построения моделей неразрушающего контроля качества. Предмет исследования составляют методы извлечения продукционных правил для синтеза моделей контроля качества. Цель работы: повысить эффективность процесса извлечения продукционных правил для построения моделей контроля качества по обучающим выборкам. Предложен стохастический метод синтеза деревьев решений, который использует информацию об информативности признаков, сложности синтезируемого дерева, а также точности его распознавания, что позволяет на начальном этапе формировать множество древовидных структур, характеризующихся простой иерархией и невысокой ошибкой распознавания, в процессе поиска создавать новые множества решений с учетом информации о значимости признаков и интерпретабельности генерируемых деревьев, что, в свою очередь, обеспечивает возможность построения дерева решений с небольшим количеством элементов (узлов и связей между ними) и приемлемой точностью распознавания, а также извлечение на его основе наиболее ценных экземпляров. Разработано программное обеспечение, реализующее предложенный метод. Проведены эксперименты по исследованию свойств предложенного метода. Результаты экспериментов позволяют рекомендовать предложенный метод для использования на практике.

Ключевые слова: выборка, дерево решений, модель контроля качества, продукционное правило, стохастический поиск.

Oliinyk A.
Ph.D., Associate Professor, Zaporizhzhya National Technical University, Ukraine

KNOWLEDGE EXTRACTION BASED ON DECISION TREES AND STOCHASTIC SEARCH

The problem of mathematical support development is solved to automate the extraction knowledge as production rules from the training data samples. The object of study is the process of constructing models of non-destructive quality control. The subject of study are methods of production rules extraction for synthesis of quality control models. The purpose of the work is to improve the efficiency of the process of production rules extraction for constructing models of quality control based on training samples. The stochastic method for the decision trees synthesis is proposed, which uses information about the informativeness of features, the complexity of the synthesized tree, as well as the accuracy of its recognition, which allows to form on the initial stage a set of tree structures, characterized by a simple hierarchy and low error recognition, in the process of search to create a new set of solutions with taking into account information about the significance of the features and interpretability of generated trees, which, in turn, provides the possibility of constructing a decision tree with a small number of elements (nodes and branches between them), and an acceptable recognition accuracy and retrieval based on it the most valuable instances. The software implementing proposed method is developed. The experiments to study the properties of the proposed method are conducted. The experimental results allow to recommend the proposed method for use in practice.

Keywords: sample, decision tree, model of quality control, production rule, stochastic search.

REFERENCES

1. Ding S. X. Model-based fault diagnosis techniques: design schemes, algorithms, and tools. Berlin, Springer, 2008, 473 p.
2. Rutkowski L. Flexible neuro-fuzzy systems : structures, learning and performance evaluation. Boston, Kluwer, 2004, 276 p.
3. Bodjanskij E. V., Kucherenko E. I., Mihalev A. I. Nejro-fazzi seti Petri v zadachah modelirovanija slozhnyh system, Dnepropetrovsk, Sistemnye tehnologii, 2005, 311 p.
4. Jang J. R. ANFIS: Adaptive-network-based fuzzy inference system, *IEEE transactions on systems and cybernetics,* 1993, Vol. 23, pp. 665–685. DOI: 10.1109/21.256541
5. Mulaik S. A. Foundations of Factor Analysis. Boca Raton, Florida, CRC Press, 2009, 548 p.
6. Jensen R., Shen Q. Computational intelligence and feature selection: rough and fuzzy approaches. Hoboken, John Wiley & Sons, 2008, 339 p.
7. Abonyi J., Feil B. Cluster analysis for data mining and system identification, Basel, Birkhäuser, 2007, 303 p.
8. Rokach L., Maimon O. Data Mining with Decision Trees. Theory and Applications. London, World Scientific Publishing Co, 2008, 264 p. DOI: 10.1142/9097.
9. Quinlan J. R. Induction of decision trees, *Machine Learning*, 1986, No. 1, pp. 81–106. DOI: 10.1007/BF00116251.
10. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification and regression trees California, Wadsworth & Brooks, 1984, 368 p.
11. Subbotin S. A., Olejnik An. A., Gofman E. A., Zajcev S. A., Olejnik Al. A.; pod red. S. A. Subbotina. Intellektual'nye informacionnye tehnologii proektirovanija avtomatizirovannyh sistem diagnostirovanija i raspoznavanija obrazov : monografija. Har'kov, OOO «Kompanija Smit», 2012, 317 p.
12. Yu X., Gen M. Introduction to Evolutionary Algorithms (Decision Engineering). London, Springer, 2010, 418 p. DOI: 10.1007/978-1-84996-129-5.
13. Gen M., Cheng R. Genetic algorithms and engineering design. New Jersey, John Wiley & Sons, 1997, 352 p. DOI: 10.1002/9780470172254.
14 Palade V., Bocaniala C. D., Jain L., eds. Computational intelligence in fault diagnosis. London, Springer, 2006, 362 p. DOI: 10.1007/978-1-84628-631-5.
15. Subbotin S. O. Podannja j obrobka znan' u sistemah shtuchnogo intelektu ta pidtrimki prijnjattja rishen' : navch. posibnik, Zaporizhzhja, ZNTU, 2008, 341 p.
16. Gofman E. A., Olejnik A. A., Subbotin S. A. Jevoljucionnyj metod sinteza derev'ev reshenij, *Shtuchnij intelekt,* 2011, No. 2, pp. 6–14.
17. Gofman E. A., Olejnik A. A., Subbotin S. A. Ispol'zovanie derev'ev reshenij dlja diagnostirovanija avtotransportnyh sredstv, *Informacionnye upravljajushhie sistemy i komp'juternyj monitoring : II Mezhdunarodnaja nauchno-tehnicheskaja konferencija IUS i KM-2011*, Doneck, 11–13 aprelja 2011 g. : materialy konferencii. Doneck, DonNTU, 2011, Vol. 1, pp. 159–163.