УДК 004.93

Олійник А. О.

Канд. техн. наук, доцент, доцент кафедри програмних засобів, Запорізький національний технічний університет, Запоріжжя, Україна

ВИДОБУВАННЯ ПРОДУКЦІЙНИХ ПРАВИЛ НА ОСНОВІ НЕГАТИВНОГО ВІДБОРУ

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 на основі негативного відбору для синтезу моделей контролю якості. Мета роботи: створення методу синтезу продукційних правил на основі негативного відбору, що полягає в обробці даних навчальної вибірки, яка характеризується істотною відмінністю кількості екземплярів, що відносяться до різних класів. Запропоновано метод синтезу продукційних правил на основі негативного відбору для випадку нерівномірного розподілу екземплярів класів вибірки, який при генерації набору детекторів використовує відому інформацію про екземпляри всіх класів вибірки, враховує інформацію про індивідуальну значущість ознак, як форму детектора використовує гіперкуб максимально можливого об'єму. Розроблений метод дозволяє виключати малозначущі і надлишкові ознаки з вибірки, скоротивши тим самим простір пошуку і час виконання методу, а також формувати набір детекторів з високими апроксимаційними й узагальнюючими здібностями. Запропонований метод за рахунок підвищення узагальнюючих властивостей синтезованих моделей шляхом скорочення числа детекторів і умов антецедентів також підвищує інтерпретабельність моделі, скорочує її розмірність (структурну і параметричну складність), обсяг використовуваної пам'яті і підвищує швидкодію моделі при послідовній реалізації обчислень. Проведено експерименти з дослідження властивостей запропонованого методу. Результати експериментів дозволяють рекомендувати запропонований метод для використання на практиці.

Ключові слова: вибірка, діагностування, модель контролю якості, негативний відбір, продукційне правило.

НОМЕНКЛАТУРА

E – помилка розпізнавання на навчальних даних ($S =< P, T >$);

E_t – помилка розпізнавання на тестових даних;

m – номер ознаки (характеристики) об'єкта;

M – кількість ознак вибірки S;

$N(p_{mn})$ – кількість екземплярів вибірки S, значення m-ї ознаки яких, належать n-му інтервалу діапазону її зміни;

$N_{\text{int}}(p_m)$ – кількість інтервалів, на які розбивається діапазон значень m-ї ознаки p_m;

$N(p_{mn}, t_l)$ – кількість екземплярів вибірки S, значення вихідного параметра T яких дорівнює t_l (належать l-му інтервалу діапазону його зміни t_l) за умови, що значення їх m-ї ознаки належить n-му інтервалу p_{mn};

$N_{\text{int}}(T)$ – кількість можливих значень (інтервалів, на які розбивається діапазон значень) вихідного параметра T;

N_{it} – кількість ітерацій роботи методу;

$N_{t_q = t'_0}$ – кількість екземплярів вибірки $S =< P, T >$, значення вихідного параметра t_q яких дорівнює t'_0;

$N_{t_q = t'_1}$ – кількість екземплярів вибірки $S =< P, T >$, значення вихідного параметра t_q яких дорівнює t'_1;

$N_{t, t_q = t'_1 / t_q = t'_0}$ – кількість екземплярів тестової вибірки, що розпізнані як «свої» ($t_q = t'_1$), але реально відносяться до класу «чужих» ($t_q = t'_0$);

$N_{t, t_q = t'_0}$ – кількість екземплярів тестової вибірки, що відносяться до класу «чужих» ($t_q = t'_0$);

$N_{t, t_q = t'_0 / t_q = t'_1}$ – кількість екземплярів тестової вибірки, що розпізнані як «чужі» ($t_q = t'_0$), але реально відносяться до класу «своїх» ($t_q = t'_1$);

$N_{t, t_q = t'_1}$ – кількість екземплярів тестової вибірки, що відносяться до класу «своїх» ($t_q = t'_1$);

P – набір вхідних характеристик (ознак) об'єктів у вибірці $S =< P, T >$;

p_{qm} – значення m-ї ознаки q-го екземпляра вибірки S;

$p_{m\min}$ – мінімальне значення m-ї ознаки у вибірці;

$p_{m\max}$ – максимальне значення m-ї ознаки у вибірці;

$P_{t, t_q = t'_1 / t_q = t'_0}$ – імовірність помилки віднесення до класу «своїх» ($t_q = t'_1$) за умови, що екземпляр реально відноситься до класу «чужих» ($t_q = t'_0$);

$P_{t, t_q = t'_0 / t_q = t'_1}$ – імовірність помилки віднесення до класу «чужих» ($t_q = t'_0$) за умови, що екземпляр реально відноситься до класу «своїх» ($t_q = t'_1$);

q – номер екземпляра (об'єкта) у вибірці S;

Q – кількість екземплярів вибірки S;

$\rho(p_{mn})$ – імовірність того, що значення ознаки p_m екземплярів вибірки S потрапить у n-й інтервал діапазону її зміни;

$\rho(p_{mn}, t_l)$ – умовна імовірність того, що значення вихідного параметра T буде дорівнює t_l (потрапить у l-й інтервал t_l) за умови, що m-а ознака p_m потрапить у n-й інтервал p_{mn};

$S =< P, T >$ – навчальна вибірка;

t – час роботи методу, мс;

$t_q \in T'$ – значення вихідного параметра q-го екземпляра;

T – множина значень вихідного параметра у вибірці $S =< P, T >$;

T' – множина можливих значень вихідного параметра T.

DOI 10.15588/1607-3274-2016-1-5

ВСТУП

У процесі побудови моделей прийняття рішень для неруйнівного контролю якості, технічного та медичного діагностування, розпізнавання образів [1–4] можуть виникати ситуації, коли велика частина інформації в навчальній вибірці даних відноситься до одного класу (наприклад, переважна більшість виробів відноситься до одного класу придатності) [5, 6].

У таких випадках для формалізації описів досліджуваних об'єктів або процесів доцільно синтезувати моделі на основі штучних імунних систем [7–9], що характеризуються можливостями навчання на основі екземплярів тільки одного класу, а також високим рівнем адаптації. Для вирішення задач, що характеризуються істотною відмінністю кількості екземплярів, що відносяться до різних класів, пропонується використовувати штучні імунні системи, що працюють на основі принципів негативного відбору [10–13], що передбачає побудову набору детекторів (обчислювальних елементів), здатних до розпізнавання невідомих екземплярів [14–16]. Такий підхід дозволяє виявляти аномалії або випадкові зміни в діагностованих об'єктах [7, 10], а також розпізнавати екземпляри чужих класів (класів об'єктів, екземпляри яких не представлені в навчальній вибірці) [8, 12, 15].

Проте відомі методи синтезу штучних імунних систем на основі негативного відбору [8–16] генерують надлишкову кількість детекторів (можливих рішень задачі), висувають значні вимоги до обчислювальних ресурсів ЕОМ, як правило, використовують інформацію тільки про один клас екземплярів («своїх», придатних і т.п.), не враховуючи при цьому дані про екземпляри інших класів. Отже, актуальною є розробка методів синтезу штучних імунних систем на основі негативного відбору, вільних від зазначених недоліків. Крім того, діагностичні моделі на основі штучних імунних систем характеризуються низьким рівнем узагальнення. Не дивлячись на те, що детектори (правила) імунної системи по окремості є легкими в сприйнятті і розумінні людиною, через низький рівень узагальнення, система детекторів має велику розмірність, і, отже, є складною для сприйняття й аналізу людиною, що в цілому призводить до зниження інтерпретабельності діагностичної моделі на основі імунних систем.

Метою роботи є створення методу синтезу продукційних правил на основі негативного відбору, що полягає в обробці даних навчальної вибірки, яка характеризується істотною відмінністю кількості екземплярів, що відносяться до різних класів.

1 ПОСТАНОВА ЗАДАЧІ

Нехай задана навчальна множина $S =< P, T >$. Набір P представляється у вигляді матриці $P = \left(p_{qm}\right)_{QM}$, $m = 1, 2, ..., M$, $q = 1, 2, ..., Q$. Набір значень вихідного параметра представляється у виді вектора $T = \left(t_q\right)_Q$, елементи $t_q \in T'$ якого приймаються значення з множини T'. У задачах неруйнівного контролю якості і розпізнавання образів множина T', як правило, складається з двох елементів $T' = \left\{t'_0, t'_1\right\}$, що визначають клас придатності виробу, наприклад при $t_q = t'_0$ q-й виріб вважається придатним, при $t_q = t'_1$ – некондиційним).

При цьому кількість екземплярів вибірки $S =< P, T >$ одного класу (наприклад, екземплярів класу некондиційних $t_q = t'_1$) істотно відрізняється від кількості екземплярів іншого класу, що визначається нерівністю (1):

$$0 \le N_{t_q = t'_0} << N_{t_q = t'_1}, \qquad (1)$$

де $N_{t_q = t'_0} + N_{t_q = t'_1} = Q$.

Тоді на основі заданої вибірки $S =< P, T >$ необхідно синтезувати набір $RB = \left\{rule_1, rule_2, ..., rule_{NR}\right\}$ продукцій $P_r \to T_r$, що дозволяє забезпечити прийнятний рівень похибки розпізнавання E, розрахованої як відношення кількості N_{er} неправильно розпізнаних за допомогою набору правил RB спостережень вибірки S до загальної кількості екземплярів Q (2):

$$E = \frac{N_{er}}{Q}. \qquad (2)$$

2 ОГЛЯД ЛІТЕРАТУРИ

Методи негативного відбору використовують процеси позитивної та негативної селекції, здійснювані під час дозрівання T-лімфоцитів, використовуються для класифікації та розпізнавання в задачах, де простір станів моделюється на основі наявних знань [10–13, 16]. В основі роботи моделі негативного відбору [10–13, 16] лежить поведінка T-клітин, що забезпечує терпимість імунної системи організму до власних клітин. При цьому T-клітини мають здатність розпізнавати практично будь-які (невідомі їм раніше) антигени. У термінах теорії імунної системи антигеном називають будь-який екземпляр, який може бути розпізнаний системою [7–9].

Основне завдання, яке вирішується за допомогою моделі негативного відбору, полягає у виявленні відмінностей між двома класами об'єктів і в проведенні подальшої двокласової класифікації [11, 16]. З погляду діагностування це завдання можна розглядати як завдання виявлення аномалій або випадкових змін у стані діагностованих об'єктів.

Чисельні детектори використовуються в тих випадках, коли стан системи можна представити у вигляді вектора ознак, значення яких нормалізовані і знаходяться в діапазоні від 0 до 1. Методи негативного відбору [10, 11, 16] засновані на використанні чисельного представлення детекторів, що призводить до більш компактного подання даних та прискорює генерацію детекторів, а також використовують класичні принципи класифікації для визначення належності детектора до придатних або дефектних екземплярів.

Проте використання чисельних детекторів у відомих методах негативного добору приводить до різних проблем, таких як неможливість задати заздалегідь розміри детекторів і їх кількість, неможливість передбачити збіжність методу, в результаті чого можливі ситуації, коли оптимальний набір детекторів так і не буде отриманий.

Потреба усунення недоліків відомих методів обумовлює необхідність розробки нового методу негативного відбору, здатного синтезувати набір детекторів за даними навчальної вибірки, що містить інформацію про екземпляри різних класів.

3 МАТЕРІАЛИ ТА МЕТОДИ

Як відзначено вище, відомі методи негативного відбору [8–16] мають такі недоліки, як генерація надлишкової кількості детекторів, використання інформації про екземпляри тільки одного класу, низька інтерпретабельність синтезованого набору рішень у вигляді детекторів. Крім того, більшість методів, заснованих на принципі негативного відбору, в якості детекторів використовують гіперсфери з фіксованим радіусом, який визначає область простору ознак, що покривається детектором. Вибір величини радіуса гіперсфери-детектора являє собою дуже складну задачу, оскільки при великих значеннях радіуса знижується точність розпізнавання, а при низьких значеннях збільшується кількість генерованих детекторів, що у свою чергу знижує узагальнюючі властивості синтезованої моделі у вигляді набору детекторів штучної імунної мережі.

Наявність зазначених недоліків обумовлює необхідність висування істотних вимог до обчислювальних ресурсів ЕОМ при використанні відомих методів негативного відбору, що, у свою чергу, знижує швидкість пошуку рішення і, у деяких випадках, не дозволяє знайти прийнятне рішення.

Для усунення зазначених недоліків доцільно використовувати метод синтезу продукційних правил на основі негативного відбору для випадку нерівномірного розподілу екземплярів класів вибірки, у якому пропонується:

– при генерації набору детекторів $AB = \{Ab_1, Ab_2, ..., Ab_{N_{Ab}}\}$ використовувати відому інформацію про екземпляри обох класів $T' = \{t'_0, t'_1\}$, що дозволить сформувати набір детекторів з великими апроксимаційними й узагальнюючими властивостями;

– використовувати інформацію про індивідуальну значущість V_m ознак p_m, що дозволить виключити малозначущі та надлишкові ознаки з вибірки $S =< P, T >$;

– як форму детектора використовувати гіперкуб максимально можливого обсягу. На відміну від відомих методів негативного відбору, у яких в якості форми детектора використовується гіперсфера, це дозволить виключити необхідність вирішення ресурсномісткої задачі пошуку оптимальних радіусів гіперсфер детекторів.

У розробленому методі на початковому етапі пропонується оцінювати значущість ознак p_m стосовно вихідного параметра T, що дозволить виявити і виключити з подальшого розгляду малозначущі ознаки, скоротивши тим самим простір пошуку і час виконання методу. Як відзначено вище, у цій роботі розглядається задача, у якій вихідна вибірка $S =< P, T >$ характеризується дискретною кількістю класів $T' = \{t'_0, t'_1\}$. Тому для оцінювання значущості V_m ознак p_m доцільно застосовувати критерії, що дозволяють виконувати оцінку інформативності ознак стосовно дискретного вихідного параметра T [2, 4, 17–22]. Як такий критерій пропонується використовувати ентропію ознаки [4, 17] як характеристику, що відображає ступінь невизначеності стану об'єкта й обчислюється за формулою (3):

$$V_m = -\sum_{n=1}^{N_{\text{int}}(p_m)} \left(\rho(p_{mn}) \sum_{l=1}^{N_{\text{int}}(T)} \rho(p_{mn}, t_l) \log_2 \rho(p_{mn}, t_l) \right), \quad (3)$$

де $\rho(p_{mn}) = \dfrac{N(p_{mn})}{Q}$; $\rho(p_{mn}, t_l) = \dfrac{N(p_{mn}, t_l)}{N(p_{mn})}$.

Ознаки p_m, значення індивідуальної інформативності яких нижче мінімально допустимої ($V_m < V_{\min}$), вважаються малоінформативними і виключаються з вибірки $S =< P, T >$.

Крім того пропонується оцінити взаємозв'язок ознак як інформативність однієї з них стосовно іншої, що дозволить виявити групи взаємозалежних ознак, у кожній з який залишити тільки одну високо інформативну ознаку, а інші ознаки, пов'язані з нею у групі можна виключити з подальшого розгляду, оскільки вони є надлишковими, ускладнюють процес синтезу діагностичних моделей і знижують їх інтерпретабельність. Для оцінювання інформативності ознак між собою V_{md} пропонується також використовувати ентропію ознак, використовуючи формулу (3) і вважаючи при цьому одну з ознак p_d вихідним параметром T (попередньо інтервал значень ознаки, що вважається вихідним параметром p_d, розбивається на $N_{\text{int}}(T)$ дискретних інтервалів).

Після оцінювання інформативності ознак стосовно інших ознак з вибірки S виключаються ті з них, для яких існують аналоги (у випадку, якщо значення взаємної інформативності ознак більше максимально припустимої $V_{md} > V_{\max}$).

Далі виконується побудова множини детекторів – структур, що дозволяють визначити, чи відноситься оцінюваний екземпляр до деякого класу. Важливо відзначити, що при використанні принципів негативного відбору детектори, що будуються на основі класу $T' = t'_1$, дозволяють виявити з невідомих екземплярів такі, які не відносяться до відповідного класу t'_1 [9, 11, 13]. Тому для формування множини детекторів у задачі розпізнавання, в якій вихідний параметр приймає два значення t'_1 (клас «своїх») і t'_0 (клас «чужих»), необхідно з вхідної вибірки $S =< P, T >$ сформувати вибірки S_0 і S_1 з екземплярів, що відносяться до класів t'_1 і t'_0, відповідно: $S_1 =< P, T = t'_1 >$ (вибірка «своїх» екземплярів) і $S_0 =< P, T = t'_0 >$ (вибірка «чужих» екземплярів).

Після цього виконується формування першого кандидата в детектори $Ab_1 =< Ab_{1\min}, Ab_{1\max} > \in AB_1$, де $Ab_{1\min} = \{Ab_{11\min}, Ab_{12\min}, ..., Ab_{1M\min}\}$ і $Ab_{1\max} = \{Ab_{11\max}, Ab_{12\max}, ..., Ab_{1M\max}\}$ – відповідно, набори мінімальних і максимальних значень m-х ознак кандидата в детектори Ab_1 ($Ab_{1m\min} = \min\limits_{q=1,2,...,Q_1} (p_{qm})$,

$Ab_{1m\max} = \max\limits_{q=1,2,...,Q_1}\left(p_{qm}\right)$, $m = 1,2,...,M$, Q_1 – кількість екземплярів у вибірці S_1), представленого у вигляді гіперкуба. Множина AB_1 детекторів Ab_k формується на основі набору «своїх» екземплярів S_1 і дозволяє виявляти серед невідомих екземплярів «чужі», тобто такі, які не відносяться до класу t_1'.

Потім для кожного q-го екземпляра s_q вибірки $S_1 =< P, T = t_1 >$ визначається його відповідність кандидату в детектори Ab_k за формулою (4):

$$eq\left(Ab_k, s_q\right) = \begin{cases} 1, \left(\sum\limits_{m=1}^{M}\left\{1 \mid \left(Ab_{km\min} < p_{qm}\right) \wedge \left(Ab_{km\max} > p_{qm}\right)\right\}\right) = M; \\ 0, \left(\sum\limits_{m=1}^{M}\left\{1 \mid \left(Ab_{km\min} < p_{qm}\right) \wedge \left(Ab_{km\max} > p_{qm}\right)\right\}\right) \neq M, \end{cases} \quad (4)$$

де сума $\sum\limits_{m=1}^{M}\left\{1 \mid \left(Ab_{km\min} < p_{qm}\right) \wedge \left(Ab_{km\max} > p_{qm}\right)\right\}$ визначає кількість відповідностей значень ознак p_{qm} q-го екземпляра кандидату Ab_k. Якщо $\left(\sum\limits_{m=1}^{M}\left\{1 \mid \left(Ab_{km\min} < p_{qm}\right) \wedge \left(Ab_{km\max} > p_{qm}\right)\right\}\right) = M$, то вважається, що екземпляр $s_q =< p_{qm}, t_q >$ відповідає кандидату в детектори Ab_k (знаходиться усередині простору гіперкуба з координатами $Ab_{1\min} = \left\{Ab_{11\min}, Ab_{12\min}, ..., Ab_{1M\min}\right\}$ і $Ab_{1\max} = \left\{Ab_{11\max}, Ab_{12\max}, ..., Ab_{1M\max}\right\}$).

Якщо існує хоча б один екземпляр $s_q =< p_{qm}, t_q = t_1' >\in S_1$, для якого $eq\left(Ab_k, s_q\right) = 1$, то вважається, що кандидат Ab_k активується при співставленні його з екземпляром s_q і, отже, не може бути детектором. Тому при виконанні умови (5)

$$\exists s_q \in S : eq\left(Ab_k, s_q\right) = 1 \quad (5)$$

відбувається етап до навчання кандидата Ab_k. Метою даного етапу є перетворення кандидата в детектори Ab_k таким чином, щоб у вибірці S_1 не існувало екземплярів, при зіставленні з якими відбувалася би активація детектора Ab_k. Для цього вибирається кортеж однієї з ознак $Ab_{km} =< Ab_{km\min}, Ab_{km\max} >$, за якими кандидат у детектори Ab_k збігається з екземпляром s_q. Далі перетворюється одне з граничних значень m-ї ознаки кандидата Ab_k: $Ab_{km\min} = p_{qm} + \eta_n\left(Ab_{km\max} - Ab_{km\min}\right)$, якщо $rnd > 0{,}5$ ($rnd = rand[0;1)$ – випадково згенероване число в діапазоні $[0;1)$), у противному випадку – $Ab_{km\max} = p_{qm} - \eta_n\left(Ab_{km\max} - Ab_{km\min}\right)$, у результаті чого об'єм гіперкуба Ab_k зменшується таким чином, що екземпляр s_q розташовується поза простором, описуваним кандидатом у детектори Ab_k. Коефіцієнт η_n задається користувачем як параметр методу в інтервалі $\eta_n \in (0;1]$. Чим більше значення даного коефіцієнта, тим більше відстань між екземплярами вибірки S_1 і гіперкубом детектора Ab_k.

Після перетворення граничних значень $Ab_{km} =< Ab_{km\min}, Ab_{km\max} >$ однієї з ознак кандидата в детектори Ab_k, він повторно перевіряється з кожним екземпляром вибірки S_1 на виконання умови (5). При виконанні умови (5) аналогічним чином відбувається повторне перетворення граничних значень однієї з ознак кандидата Ab_k. І так доти, поки буде виконуватися умова (5).

Після того, як у множині $S_1 =< P, T = t_1 >$ не залишиться екземплярів s_q, при зіставленні з якими активізується кандидат Ab_k, виконується етап оцінювання пристосованості кандидата Ab_k до узагальнення. При використанні принципів негативного відбору формується множина детекторів $AB_1 = \left\{Ab_1, Ab_2, ..., Ab_{N_{Ab}}\right\}$, що дозволяють з високою точністю визначити належність екземплярів s_q вибірки S до визначеного класу [10–16]. Тому як критерій оцінювання будемо використовувати характеристики (6) і (7), що дозволяють оцінити здатність детектора Ab_k до узагальнення даних:

$$G_1\left(Ab_k\right) = \frac{1}{M}\sum_{m=1}^{M}\frac{Ab_{km\max} - Ab_{km\min}}{p_{m\max} - p_{m\min}}, \quad (6)$$

$$G_2\left(Ab_k\right) = \frac{\prod\limits_{m=1}^{M}\left(Ab_{km\max} - Ab_{km\min}\right)}{\prod\limits_{m=1}^{M}\left(p_{m\max} - p_{m\min}\right)}, \quad (7)$$

де $p_{m\min} = \min\limits_{q=1,2,...,Q_1}\left(p_{qm}\right)$ та $p_{m\max} = \max\limits_{q=1,2,...,Q_1}\left(p_{qm}\right)$.

Критерії (6) і (7) відображають частину детектора, що покривається за допомогою, простору пошуку. Критерій $G_1\left(Ab_k\right)$ показує середню частку простору, що покривається детектором, у кожному з M вимірів простору ознак. Критерій $G_2\left(Ab_k\right)$ відображає об'ємну частину простору, що покривається.

Чим більше значення критеріїв $G_1\left(Ab_k\right)$ і $G_2\left(Ab_k\right)$, тим більш велику частину простору пошуку покриває детектор. Отже, якщо значення критерію оцінювання якості узагальнення $G\left(Ab_k\right)$ вище граничного $G_{\min}$ ($G\left(Ab_k\right) > G_{\min}$), то вважається, що кандидат Ab_k характеризується високими узагальнюючими здібностями, може бути детектором і додається в множину детекторів: $AB_1 = AB_1 \cup \left\{Ab_k\right\}$.

Створення нових кандидатів Ab_k здійснюється доти, поки не будуть досягнуті критерії закінчення пошуку. Як такі критерії можуть бути використані точність розпізнавання $E(S)$, досягнення максимально припустимої

кількості детекторів ($N_{Ab} = |AB| > N_{Ab\max}$), перевищення максимальне припустимого часу пошуку ($t > t_{\max}$).

Отриманий у результаті негативного відбору набір детекторів $AB_1 = \{Ab_1, Ab_2, ..., Ab_{N_{Ab}}\}$ описує область простору пошуку $\overline{S_1}$, комплементарну області простору, у якій розташована множина «своїх» екземплярів S_1. При цьому множина $AB_1 = \{Ab_1, Ab_2, ..., Ab_{N_{Ab}}\}$ характеризується високими апроксимаційними й узагальнюючими здібностями.

Аналогічним чином можна сформувати набір детекторів AB_0 для множини S_0. Проте, у задачах з нерівномірним розподілом екземплярів класів вибірки $S =< P, T >$, коли кількість екземплярів одного класу $Q_1 = N_{t_q = t_1'}$ істотно перевищує кількість екземплярів іншого класу $Q_0 = N_{t_q = t_0'}$ ($Q_0 << Q_1$), можуть виникнути проблеми з генерацією детекторів, що адекватно відображають простір екземплярів S_0 (можуть бути згенеровані детектори Ab_k у виді гіперкубів занадто великого обсягу, що не зможуть узагальнити дані генеральної сукупності). Це обумовлено невеликою (недостатньою) кількістю екземплярів у вибірці S_0 ($Q_0 << Q_1$), а іноді і майже повною їхньою відсутністю.

Тому в цій роботі при генерації детекторів для екземплярів S_0 пропонується використовувати інформацію про розміри детекторів, побудованих на основі вибірки S_1. При цьому детектори будуть відображати інформацію про наявність у гіперкубі екземплярів вибірки S_0 (а не про їх відсутність, як при класичному негативному відборі), і, отже, по суті будуть цілком відповідати детекторам, що побудовані раніше для вибірки S_1 на основі негативного відбору й містять інформацію про області простору пошуку, у яких не розташовуються екземпляри S_1.

Детектори $Ab_k^{(0)}$ вибірки S_0 генеруються таким чином, щоб їх центри відповідали координатам екземплярів $s_k =< p_{km}, t_k = t_0' > \in S_0$ вибірки S_0, а розміри граней їх гіперкубів відповідали аналогічним розмірам детекторів, створених на основі даних вибірки S_1. Отже, координати детектора $Ab_{km}^{(0)} =< Ab_{km\min}^{(0)}, Ab_{km\max}^{(0)} >$ визначаються за формулами (8)–(9):

$$Ab_{km\min}^{(0)} = p_{km} - \frac{1}{2}\Delta Ab_m, \quad (8)$$

$$Ab_{km\max}^{(0)} = p_{km} + \frac{1}{2}\Delta Ab_m, \quad (9)$$

де ΔAb_m – середня довжина граней детекторів $AB_1 = \{Ab_1, Ab_2, ..., Ab_{N_{Ab}}\}$, створених на основі множини S_1. Величину ΔAb_m пропонується обчислювати, виходячи з інформації про детектори $AB_1 = \{Ab_1, Ab_2, ..., Ab_{N_{Ab}}\}$, використовуючи формулу (10):

$$\Delta Ab_m = \frac{1}{N_{Ab}M}\left(\sum_{k=1}^{N_{Ab}}\sum_{m=1}^{M}\left(Ab_{km\max} - Ab_{km\min}\right)\right). \quad (10)$$

Потім виконується зіставлення згенерованих детекторів $Ab_k^{(0)}$ з екземплярами вибірки S_1, використовуючи формулу (4), при виконанні умови (5) відбувається перетворення детекторів $Ab_k^{(0)}$ аналогічно описаному вище етапу донавчання. Після цього обчислюється значення одного з критеріїв $G\left(Ab_k^{(0)}\right)$ оцінювання здатності детектора до узагальнення даних (6)–(7), і, у випадку, якщо його значення вище граничного, детектор $Ab_k^{(0)}$ додається в множину $AB_0 = AB_0 \cup \{Ab_k^{(0)}\}$.

У такий спосіб генерується набір детекторів AB_0, що описує, як і набір AB_1, область простору пошуку $\overline{S_1}$, комплементарну області розташування множини екземплярів S_1. Тому розпізнавальна модель може бути представлена у вигляді множини детекторів $AB = AB_0 \cup AB_1$, що дозволяють розпізнавати належність невідомих екземплярів $s_q' =< p_{qm}', t_q' - ? > \notin S$ до класу «чужих», тобто відносити їх до класу t_0': $t_q' = t_0'$.

З метою підвищення рівня інтерпретабельності отриманої розпізнавальної моделі представленої у вигляді набору $AB = \{Ab_1, Ab_2, ..., Ab_{N_{Ab}}\}$ детекторів, пропонується на основі набору AB сформувати множину продукційних правил $PR_r : P_r \to T_r$, у яких ліва частина P_r імплікації являє собою набір умов вигляду (11)

«Якщо $(p_1 \in [Ab_{k1\min}; Ab_{k1\max}]) \wedge (p_2 \in [Ab_{k2\min};$
$Ab_{k2\max}]) \wedge ... \wedge (p_M \in [Ab_{kM\min}; Ab_{kM\max}])$» (11)

а права частина T_r містить значення вихідного параметра T при виконанні r-го набору умов P_r (11).

При формуванні набору правил PR в антецедент правила P_r будемо включати тільки ті границі ознак, для яких вони не є граничними значеннями, тобто $Ab_{km\min} \neq \min_{q=1,2,...,Q}(p_{qm})$ і $Ab_{km\max} \neq \max_{q=1,2,...,Q}(p_{qm})$. Наприклад, для детектора вигляду $Ab_k = \{< 5,7 >, < 8, p_{2\max} >, < p_{3\min}, p_{3\max} >, < 4,6 >\}$ буде сформоване правило PR_k вигляду: Якщо $(p_1 > 5 \wedge p_1 < 7) \wedge (p_2 > 8) \wedge (p_4 > 4 \wedge p_4 < 6)$, то віднести екземпляр до класу «чужих» ($T \neq t_1$). Як видно, у правило не ввійшли в явному вигляді верхня границя ознаки p_2 і цілком ознака p_3, оскільки відповідні значення детектора знаходяться на мінімальній і максимальній границях ознак і не впливають на якість розпізнавання. Крім того, виключення таких значень із правила PR_k знижує його складність, підвищуючи в такий спосіб інтерпретабельність правила.

Використовуючи такий підхід, на основі кожного *k*-го детектора Ab_k виконується побудова свого правила, формуючи в такий спосіб множину PR, що складається з N_{Ab} продукційних правил $PR_r : P_r \to T_r$.

Таким чином, запропонований метод синтезу продукційних правил на основі негативного відбору для випадку нерівномірного розподілу екземплярів класів вибірки при генерації набору детекторів використовує відому інформацію про екземпляри всіх класів вибірки, враховує інформацію про індивідуальну значущість ознак, як форму детектора використовує гіперкуб максимально можливого обсягу, що дозволяє виключати малозначущі і надлишкові ознаки з вибірки, скоротивши тим самим простір пошуку і час виконання методу, а також формувати набір детекторів з високими апроксимаційними й узагальнюючими здібностями.

Запропонований метод за рахунок підвищення узагальнюючих властивостей синтезованих моделей шляхом скорочення кількості детекторів і умов антецедентів також підвищує інтерпретабельність моделі, скорочує її розмірність (структурну і параметричну складність), обсяг використовуваної пам'яті і підвищує швидкодію моделі при послідовній реалізації обчислень.

4 ЕКСПЕРИМЕНТИ

Для перевірки працездатності запропонованого методу синтезу продукційних правил на основі негативного відбору було розроблено комп'ютерну програму, що реалізує запропонований метод. За допомогою розробленого програмного забезпечення розв'язувалася задача діагностування лопаток газотурбінних авіаційних двигунів [23]. Лопатки характеризувалися значеннями спектрів потужності загасаючих коливань після ударного збудження, які використовувалися як вхідні ознаки. Експертно було визначено класи якості лопаток: придатні (кондиційні) і дефектні (некондиційні). Кожна лопатка була описана 10240 ознаками, що характеризували спектр потужності загасаючих коливань. Для скорочення простору пошуку на основі цих ознак були отримані штучні ознаки-згортки, у результаті чого сформовано набір P, який складається зі штучних 80 ознак.

Отримана вибірка спостережень $S =< P, T >$, вочевидь, не є статистично репрезентативною, оскільки не відображає реального розподілу частот класів (у генеральній сукупності придатних лопаток суттєво більше, ніж дефектних). При цьому дефектні лопатки ($t_q = t_1'$) у наявній вибірці S представляють типові випадки некондиційності, що забезпечує топологічну репрезентативність дефектних лопаток у вибірці. А всі можливі випадки класу придатних ($t_q = t_0'$) неможливо представити у вибірці з практичної точки зору. Тому виникає необхідність на основі наявної вибірки $S =< P, T >$ з нерівномірним розподілом екземплярів по класах побудувати діагностичну модель, що дозволяє виконувати технічне діагностування лопаток авіадвигунів на основі заданого набору вимірювань.

Вибірка $S =< P, T >$ містить 42 екземпляри, що характеризують дефектні лопатки, і 72 екземпляри придатних.

Запропонований метод синтезу продукційних правил на основі негативного відбору порівнювався з існуючими методами негативного відбору, що синтезують набір детекторів тільки на основі «своїх» екземплярів вибірки $S_1 \subseteq S$. Тому за допомогою запропонованого методу задача діагностування лопаток газотурбінних авіаційних двигунів розв'язувалася два рази:

– з використанням підвибірки ($S_1 \subseteq S$), що містить інформацію тільки про некондиційні екземпляри («своїх»);

– з використанням усієї вихідної вибірки $S =< P, T >$.

Як тестова вибірка S_t використовувалася вибірка, що містить інформацію про 273 екземпляри (261 екземпляр придатних виробів, що відносяться до класу $t_q = t_0'$, і 12 екземплярів дефектних виробів, що відносяться до класу $t_q = t_1'$).

Також виконано порівняння запропонованого методу з іншими методами, що дозволяють вирішувати задачі розпізнавання образів. Для цього розв'язувалася описана вище задача діагностування лопаток газотурбінних авіаційних двигунів. Досліджувалися властивості і характеристики таких розпізнавальних моделей:

– модель у вигляді набору продукційних правил, синтезованих за допомогою запропонованого методу на основі негативного відбору з урахуванням інформативності ознак;

– нейромережева модель прямого поширення, що складається з трьох шарів нейронів, побудована на основі методу зворотного поширення помилки. На першому шарі нейромережі знаходилося п'ять нейронів, на другому – три нейрони, на третьому – один нейрон. Нейрони першого і другого шару мали логістичну сигмоїдну функцію активації, третього – граничну функцію активації;

– модель у вигляді набору детекторів, побудована за допомогою методу негативного відбору з маскуванням детекторів [16].

При цьому використовувалася вся навчальна вибірка $S =< P, T >$ обсягом 114 екземплярів (42 екземпляри, що характеризують дефектні лопатки, і 72 екземпляри придатних) при побудові першої і другої моделі, і частина вибірки ($S_1 \subseteq S$) при побудові моделі на основі набору детекторів за допомогою методу негативного відбору з маскуванням, оскільки даний метод припускає роботу з екземплярами тільки одного класу.

5 РЕЗУЛЬТАТИ

Результати експериментів з порівняння запропонованого методу з іншими методами негативного відбору наведено в табл. 1. $P_{t,t_q=t_1'/t_q=t_0'}$ й $P_{t,t_q=t_0'/t_q=t_1'}$ помилки віднесення до класу «своїх» $t_q = t_1'$ («чужих», $t_q = t_0'$) за умови, що екземпляр реально відноситься до класу «чужих» $t_q = t_0'$ («своїх», $t_q = t_1'$) обчислюються за формулами (12) і (13), відповідно:

$$P_{t,t_q=t_1'/t_q=t_0'} = \frac{N_{t,t_q=t_1'/t_q=t_0'}}{N_{t,t_q=t_0'}}, \qquad (12)$$

$$P_{t,t_q=t'_0/t_q=t'_1}=\frac{N_{t,t_q=t'_0/t_q=t'_1}}{N_{t,t_q=t'_1}}. \quad (13)$$

Результати порівняння різних моделей при вирішенні задачі діагностування лопаток газотурбінних авіаційних двигунів [23] наведено в табл. 2, де N_{param} – критерій, що визначає параметричну складність моделі. Критерій N_{param} розраховувався як кількість параметрів моделі: загальна кількість параметрів $Ab_{km\min}$ і $Ab_{km\max}$ в моделях на основі продукцій і на основі набору детекторів [16], загальна кількість настроюваних параметрів (вагових коефіцієнтів) – у нейромережевій моделі.

6 ОБГОВОРЕННЯ

Як видно з табл. 1, прийнятне значення помилки розпізнавання E забезпечив метод з маскуванням детекторів [16] ($E=0{,}018$) і запропонований метод синтезу продукційних правил на основі негативного відбору (МСППНВ). Низькі значення помилки розпізнавання зазначених методів забезпечувалися за рахунок широкого покриття синтезованими детекторами області «своїх» екземплярів вибірки $S_1 \subseteq S$. При цьому запропонований метод МСППНВ, що синтезував набір детекторів на основі екземплярів усіх класів вибірки $S=<P,T>$, забезпечив більш прийнятні результати ($E=0{,}009$) у порівнянні з набором детекторів, синтезованим тільки з використанням екземплярів класу «своїх» $S_1 \subseteq S$ ($E=0{,}026$). Менш прийнятні значення помилки розпізнавання E показали метод з цензуруванням [13] ($E=0{,}070$) і модель V-Detector [14, 15] ($E=0{,}035$), що свідчить про недостатність покриття синтезованими детекторами області «своїх» екземплярів $S_1 \subseteq S$.

За результатами експериментів видно, що при використанні методу з цензуруванням [13] і моделі V-Detector [14, 15] генерується найбільша кількість детекторів N_{Ab} ($N_{Ab}=207$ і $N_{Ab}=41$, відповідно), що негативно впливає на час навчання t і витрати обчислювальних ресурсів комп'ютера. Метод з маскуванням детекторів [16] і запропонований метод синтезу продукційних правил на основі негативного відбору (при використанні вибірки $S_1 \subseteq S$) згенерували істотно меншу кількість детекторів ($N_{Ab}=19$ і $N_{Ab}=20$, відповідно), що свідчить про більш ефективну роботу цих методів. Зокрема, метод МСППНВ використовує апріорну інформацію про значущість ознак на початковому етапі і виключає з подальшого розгляду малозначущі і надлишкові ознаки, що дозволяє скоротити простір пошуку і створювати набір з невеликої кількості детекторів на основі високоінформативних ознак, що характеризується високими апроксимаційними й узагальнюючими здібностями.

Для аналізу узагальнюючих здібностей досліджуваних методів використовувалися критерії E_t, $P_{t,t_q=t'_1/t_q=t'_0}$ і $P_{t,t_q=t'_0/t_q=t'_1}$, що характеризують помилки розпізнавання й імовірності прийняття помилкових рішень на тестових даних. Як видно з табл. 1, помилки розпізнавання на тестових даних E_t у моделей, синтезованих за допомогою запропонованого методу МСППНВ, істотно нижче помилок моделей, побудованих за допомогою відомих методів [13]–[16] ($E_t=0{,}136$, $E_t=0{,}077$, $E_t=0{,}055$ для методів [13], [14, 15] і [16], відповідно). Це пояснюється використанням в якості критеріїв оцінювання кандидатів у детектори характеристик $G(Ab_k)$, що дозволяють оцінювати здатність детекторів до узагальнення даних. Запропонований метод синтезу продукційних правил на основі негативного відбору дозволив досягти рівнів помилки $E_t=0{,}037$ (при використанні частини вибірки $S_1 \subseteq S$) і $E_t=0{,}011$ (при використанні повної вибірки $S=<P,T>$).

Важливо відзначити, що в силу специфіки розв'язуваної задачі діагностування лопаток газотурбінних авіаційних двигунів дуже високу ціну має імовірність по-

Таблиця 1 – Результати експериментів з порівняння запропонованого методу з іншими методами негативного відбору

Метод	N_{Ab}	N_{it}	t, мс	E	$P_{t,t_q=t'_1/t_q=t'_0}$	$P_{t,t_q=t'_0/t_q=t'_1}$	E_t
Метод з цензуруванням [13]	207	50	27,3	0,070	0,126	0,333	0,136
Модель V-Detector [14, 15]	41	50	24,1	0,035	0,069	0,250	0,077
Метод з маскуванням детекторів [16]	19	14	13,2	0,018	0,054	0,083	0,055
Метод синтезу продукційних правил на основі негативного відбору МСППНВ (використовувалася вибірка $S_1 \subseteq S$)	20	12	12,1	0,026	0,038	0	0,037
Метод синтезу продукційних правил на основі негативного відбору МСППНВ (використовувалася вибірка $S=<P,T>$)	31	19	13,7	0,009	0,011	0	0,011

Таблиця 2 – Результати порівняння різних моделей

Модель	N_{param}	E	$P_{t,t_q=t'_1/t_q=t'_0}$	$P_{t,t_q=t'_0/t_q=t'_1}$	E_t
Модель у вигляді набору продукційних правил, синтезованих за допомогою запропонованого методу	652	0,009	0,011	0	0,011
Нейромережева модель прямого поширення	427	0,018	0,065	0,167	0,070
Модель у вигляді набору детекторів, побудована за допомогою методу негативного відбору з маскуванням детекторів [16]	804	0,018	0,054	0,083	0,055

милки $P_{t,t_q=t'_0/t_q=t'_1}$ віднесення до класу «чужих» ($t_q = t'_0$, придатних) за умови, що екземпляр реально відноситься до класу «своїх» ($t_q = t'_1$, дефектних). Це обумовлено тим, що віднесення дефектних лопаток авіадвигунів до класу придатних може коштувати людських життів. Як видно з табл. 1, запропонований метод МСППНВ, на відміну від існуючих, на тестових даних показав нульовий рівень імовірності помилки $P_{t,t_q=t'_0/t_q=t'_1}$, що свідчить про його високу ефективність для розв'язання подібних задач. Нульовий рівень імовірності помилки $P_{t,t_q=t'_0/t_q=t'_1}$ при використанні запропонованого методу пояснюється:

– високим рівнем покриття типових екземплярів класу $t_q = t'_1$ за допомогою згенерованого набору детекторів $AB = \{Ab_1, Ab_2, ..., Ab_{N_{Ab}}\}$, отриманого з використанням апріорної інформації про значущість ознак;

– високими узагальнюючими здібностями синтезованого набору детекторів, що обумовлено застосуванням в якості критерію оцінювання детекторів характеристик (6) і (7), що дозволяють оцінити здатність детекторів Ab_k до узагальнення даних.

Як видно з табл. 2, кількість параметрів N_{param} моделі у вигляді набору продукційних правил, синтезованих за допомогою запропонованого методу ($N_{param} = 652$), менше, ніж в аналогічній моделі, побудованої за допомогою методу негативного відбору з маскуванням детекторів [16] ($N_{param} = 804$). Це обумовлено тим, що при використанні запропонованого методу середній розмір згенерованих детекторів менше, оскільки в процесі негативного відбору використовується апріорна інформація про значущість ознак. Це дозволяє виявляти і виключати з подальшого розгляду малозначущі і надлишкові ознаки, що ускладнюють процес синтезу діагностичних моделей і знижують їх інтерпретабельність. Таким чином, порівняння значень різних критеріїв, представлених у табл. 2, показує, що модель у вигляді набору продукційних правил, синтезованих за допомогою запропонованого методу МСППНВ, є більш простою і зрозумілою в порівнянні з моделлю, створеною за допомогою методу [16]. Апроксимаційні й узагальнюючі здібності моделі, синтезованої на основі методу МСППНВ, також є більш високими, про що свідчать значення критеріїв E, E_t, $P_{t,t_q=t'_1/t_q=t'_0}$ і $P_{t,t_q=t'_0/t_q=t'_1}$.

Порівняння моделі на основі методу МСППНВ і нейромережевої моделі дозволяє зробити висновок про те, що модель, побудована за допомогою запропонованого методу, характеризується більш високими узагальнюючими й апроксимаційними здібностями (критерії E, E_t, $P_{t,t_q=t'_1/t_q=t'_0}$ і $P_{t,t_q=t'_0/t_q=t'_1}$). Однак, кількість параметрів N_{param} моделі на основі методу МСППНВ ($N_{param} = 652$) є дещо більшою, ніж у нейромережевій моделі ($N_{param} = 427$). Це пояснюється тим, що нейромережева модель представляється у вигляді множини нейронів, які пов'язані між собою певним чином і характеризуються ваговими коефіцієнтами як налагоджуваними параметрами. А кожен нейрон, по суті, відповідає деякій функції багатьох аргументів. При цьому така модель є досить складною для сприйняття людиною. Не дивлячись на більше значення критерію N_{param}, модель у виді набору продукційних правил, синтезованих за допомогою запропонованого методу, є більш інтерпретабельною у порівнянні з нейромережевою моделлю, оскільки продукційні правила вигляду «Якщо умова, то дія» є значно більш зрозумілими і зручними для сприйняття людиною, ніж набір коефіцієнтів, що відображаються ступінь зв'язків нейронів у нейромережевій моделі.

Таким чином, результати експериментів показали, що запропонований метод за рахунок використання апріорної інформації і виключення малозначущих і надлишкових ознак з вибірки дозволяє скорочувати простір пошуку і час виконання методу, а також синтезувати розпізнавальні моделі у вигляді набору детекторів з високими апроксимаційними й узагальнюючими здібностями. Крім того за рахунок підвищення узагальнюючих властивостей синтезованих моделей шляхом скорочення кількості детекторів і умов антецедентів підвищує інтерпретабельність моделі, скорочує її розмірність і, отже, обсяг використовуваної пам'яті.

ВИСНОВКИ

У роботі вирішено актуальне завдання автоматизації синтезу продукційних правил на основі негативного відбору для випадку нерівномірного розподілу екземплярів класів вибірки.

Наукова новизна роботи полягає в тому, що запропоновано метод синтезу продукційних правил на основі негативного відбору для випадку нерівномірного розподілу екземплярів класів вибірки, який при генерації набору детекторів використовує відому інформацію про екземпляри всіх класів вибірки, враховує інформацію про індивідуальну значущість ознак, як форму детектора використовує гіперкуб максимально можливого об'єму, що дозволяє виключати малозначущі і надлишкові ознаки з вибірки, скоротивши тим самим простір пошуку і час виконання методу, а також формувати набір детекторів з високими апроксимаційними й узагальнюючими здібностями. Запропонований метод за рахунок підвищення узагальнюючих властивостей синтезованих моделей шляхом скорочення числа детекторів і умов антецедентів також підвищує інтерпретабельність моделі, скорочує її розмірність (структурну і параметричну складність), обсяг використовуваної пам'яті і підвищує швидкодію моделі при послідовній реалізації обчислень.

Практична цінність отриманих результатів полягає в тому, що виконано експериментальне дослідження запропонованого методу і його порівняння з відомими аналогами, а також вирішено практичну задачу діагностування лопаток газотурбінних авіаційних двигунів.

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 при частковій підтримці 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. DOI: 10.1007/978-3-540-76304-8.
2. ASM handbook. – Vol. 17: Nondestructive evaluation and quality control. – Cleveland : ASM International, 1997. – 1607 p.
3. Diagnosis and fault-tolerant control / [M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki]. – Berlin : Springer, 2006. – 672 p. DOI: 10.1007/978-3-662-05344-7.
4. Price C. Computer based diagnostic systems / C. Price. – London : Springer, 1999. – 136 p. DOI: 10.1007/978-1-4471-0535-0.
5. Denton T. Advanced automotive fault diagnosis / T. Denton. – London : Elsevier, 2006. – 271 p. DOI: 10.4324/9780080462585.
6. Ukil A. Intelligent Systems and Signal Processing in Power Engineering / A. Ukil. – Berlin : Springer, 2007. – 372 p. DOI: 10.1007/978-3-540-73170-2.
7. Ishida Y. Immunity-based systems: a design perspective / Y. Ishida. – Berlin : Springer, 2004. – 177 p. DOI: 10.1007/978-3-662-07863-1.
8. Segel L. A. Design principles for immune system and other distributed autonomous systems / L. A. Segel, I. R. Cohen. – New York : Oxford University Press, 2001. – 428 p.
9. Flower D. In silico immunology / D. Flower, J Timmis. – New York : Springer, 2007. – 451 p. DOI: 10.1007/978-0-387-39241-7.
10. A new cluster based real negative selection algorithm / W. Chen, T. Li, J. Qin [et al.] // Information and automation. – 2011. – Vol. 86. – P. 125–131. DOI: 10.1007/978-3-642-19853-3_18.
11. Esponda F. Negative representations of information / F. Esponda, S. Forrest, P. Helman // International Journal of Information Security. – 2009. – Vol. 8, № 5. – P. 331–345. DOI: 10.1007/s10207-009-0078-1.
12. Gonzalez F. The effect of binary matching rules in negative selection / F. Gonzalez, D. Dasgupta, J. Gomez // Genetic and Evolutionary Computation: conference GECCO-2003, Chicago, 9–11 July 2003: proceedings. – Berlin-Heidelberg : Springer-Verlag, 2003. – P. 195–206. DOI: 10.1007/3-540-45105-6_25.
13. Ong A. An adaptive anomaly detection system using data mining and artificial immune system / A. Ong. – London : King's College London, 2007. – 348 p.
14. Gonzalez F. A. Anomaly detection using real-valued negative selection / F. A. Gonzalez, D. Dasgupta // Journal of Genetic Programming and Evolvable Machines. – 2003. – Vol. 4, № 4. – P. 383-403. DOI: 10.1023/a:1026195112518.
15. Chmielewski A. Simple method of increasing the coverage of nonself region for negative selection algorithms / A. Chmielewski, S. T. Wierzchon // Computer Information Systems and Industrial Management Applications: 6th International Conference CISIM'07, Elk, 28–30 June 2007: proceedings. – Los Alamitos : IEEE Computer Society, 2007. – P. 155–160. DOI: 10.1109/cisim.2007.60.
16. Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов : монография / [Субботин С. А., Олейник Ан. А., Гофман Е. А., Зайцев С. А., Олейник Ал. А.; под общ. ред. С. А. Субботина]. – Харьков : Компания СМИТ, 2012. – 318 с.
17. Clarke B. Principles and theory for data mining and machine learning / B. Clarke, E. Fokoue, H. H. Zhang. – New York : Springer, 2009. – 781 p. DOI: 10.1007/978-0-387-98135-2.
18. Bishop C.M. Pattern recognition and machine learning / C. M. Bishop. – New York : Springer, 2006. – 738 p. DOI: 10.1108/03684920710743466.
19. Analysis and design of intelligent systems using soft computing techniques / eds.: P. Melin, O. R. Castillo, E. G. Ramirez, J. Kacprzyk. – Heidelberg : Springer, 2007. – 855 p. DOI: 10.1007/978-3-540-72432-2.
20. Encyclopedia of machine learning / [eds. C. Sammut, G. I. Webb]. – New York : Springer, 2011. – 1031 p. DOI: 10.1007/978-0-387-30164-8.
21. Russel S. Artificial intelligence: a modern approach / S. Russel, P. Norvig. – New Jersey: Prentice Hall, 2009. – 1152 p.
22. Intelligent data analysis: an introduction / [eds. M. Berthold, D. J. Hand]. – New York: Springer Verlag, 2007. – 525 p.
23. Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей : монография / [Дубровин В. И., Субботин С. А., Богуслаев А. В., Яценко В. К.]. – Запорожье: ОАО «Мотор-Сич», 2003. – 279 с.

Стаття надійшла до редакції 09.01.2016.
Після доробки 26.01.2016.

Олейник А. А.

Канд. техн. наук, доцент, доцент кафедры программных средств, Запорожский национальный технический университет, Запорожье, Украина

ИЗВЛЕЧЕНИЕ ПРОДУКЦИОННЫХ ПРАВИЛ НА ОСНОВЕ НЕГАТИВНОГО ОТБОРА

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 на основе отрицательного отбора для синтеза моделей контроля качества. Цель работы: создание метода синтеза продукционных правил на основе множества детекторов, заключающегося в обработке данных обучающей выборки, характеризующейся существенным отличием числа экземпляров, относящихся к разных классам. Предложен метод синтеза продукционных правил на основе отрицательного отбора для случая неравномерного распределения экземпляров классов выборки, который при генерации набора детекторов использует известную информацию об экземплярах всех классов выборки, учитывает информацию об индивидуальной значимости признаков, в качестве формы детектора использует гиперкуб максимально возможного объема. Разработанный метод позволяет исключать малозначимые и избыточные признаки из выборки, сократив тем самым пространство поиска и время выполнения метода, а также формировать набор детекторов с высокими аппроксимационными и обобщающими способностями. Предложенный метод за счет повышения обобщающих свойств синтезируемых моделей путем сокращения числа детекторов и условий антецедентов также повышает интерпретабельность модели, сокращает ее размерность (структурную и параметрическую сложность), объем используемой памяти и повышает быстродействие модели при последовательной реализации вычислений. 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, диагностирование, модель контроля качества, отрицательный отбор, продукционное правило.

Oliinyk A.

PhD., Associate Professor, Associate Professor of Department of Software Tools, Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine

PRODUCTION RULES EXTRACTION BASED ON NEGATIVE SELECTION

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 based on negative selection for synthesis of quality control models. The purpose of the work is to develop a method of production rules synthesis on the basis of a set of detectors is in the handling of data of training sample, characterized by a substantial number of instances of distinction belonging to different classes. A method for the synthesis of production rules on the basis of negative selection in the case of uneven distribution of instances of the sample classes is proposed. The developed method allows to exclude irrelevant and redundant features from the sample, thereby reducing the search space and time of execution of the method, as well as generate a set of detectors with high approximation and generalization capability. The proposed method improves the generalizing properties of synthesized model and its interpretability. 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, diagnostics, model of quality control, negative selection, production rule.

REFERENCES

1. Ding S. X. Model-based fault diagnosis techniques: design schemes, algorithms, and tools. Berlin, Springer, 2008, 473 p. DOI: 10.1007/978-3-540-76304-8.
2. ASM handbook. Vol. 17: Nondestructive evaluation and quality control. Cleveland, ASM International, 1997, 1607 p.
3. Blanke M., Kinnaert M., Lunze J., Staroswiecki M. Diagnosis and fault-tolerant control. Berlin, Springer, 2006, 672 p. DOI: 10.1007/978-3-662-05344-7.
4. Price C. Computer based diagnostic systems. London, Springer, 1999, 136 p. DOI: 10.1007/978-1-4471-0535-0.
5. Denton T. Advanced automotive fault diagnosis. London, Elsevier, 2006, 271 p. DOI: 10.4324/9780080462585.
6. Ukil A. Intelligent Systems and Signal Processing in Power Engineering. Berlin, Springer, 2007, 372 p. DOI: 10.1007/978-3-540-73170-2.
7. Ishida Y. Immunity-based systems: a design perspective. Berlin, Springer, 2004, 177 p. DOI: 10.1007/978-3-662-07863-1.
8. Segel L.A., Cohen I. R. Design principles for immune system and other distributed autonomous systems. New York, Oxford University Press, 2001, 428 p.
9. Flower D., Timmis J. In silico immunology. New York, Springer, 2007, 451 p. DOI: 10.1007/978-0-387-39241-7.
10. Chen W., Li T., Qin J. [et al.] A new cluster based real negative selection algorithm, *Information and automation*, 2011, Vol. 86, pp. 125–131. DOI: 10.1007/978-3-642-19853-3_18.
11. Esponda F., Forrest S., Helman P. Negative representations of information, *International Journal of Information Security*, 2009, Vol. 8, No. 5, pp. 331–345. DOI: 10.1007/s10207-009-0078-1.
12. Gonzalez F., Dasgupta D., Gomez J. The effect of binary matching rules in negative selection, *Genetic and Evolutionary Computation: conference GECCO-2003, Chicago, 9–11 July 2003: proceedings*. Berlin-Heidelberg, Springer-Verlag, 2003, pp. 195–206. DOI: 10.1007/3-540-45105-6_25.
13. Ong A. An adaptive anomaly detection system using data mining and artificial immune system. London, King's College London, 2007, 348 p.
14. Gonzalez F. A., Dasgupta D. Anomaly detection using real-valued negative selection, *Journal of Genetic Programming and Evolvable Machines*, 2003, Vol. 4, No. 4, pp. 383–403. DOI: 10.1023/a:1026195112518.
15. Chmielewski A., Wierzchon S. T. Simple method of increasing the coverage of nonself region for negative selection algorithms, *Computer Information Systems and Industrial Management Applications: 6th International Conference CISIM'07, Elk, 28–30 June 2007: proceedings*. Los Alamitos, IEEE Computer Society, 2007, pp. 155–160. DOI: 10.1109/cisim.2007.60.
16. Subbotin S. A., Olejnik An. A., Gofman E. A., Zajcev S. A., Olejnik Al. A.; pod obshh. red. S. A. Subbotina Intellektual'nye informacionnye tehnologii proektirovanija avtomatizirovannyh sistem diagnostirovanija i raspoznavanija obrazov: monografija. Har'kov, Kompanija SMIT, 2012, 318 p.
17. Clarke B. Fokoue E., Zhang H. H. Principles and theory for data mining and machine learning. New York, Springer, 2009, 781 p. DOI: 10.1007/978-0-387-98135-2.
18. Bishop C. M. Pattern recognition and machine learning. New York, Springer, 2006, 738 p. DOI: 10.1108/03684920710743466.
19. eds.: Melin P., Castillo O. R., Ramirez E. G., Kacprzyk J. Analysis and design of intelligent systems using soft computing techniques. Heidelberg, Springer, 2007, 855 p. DOI: 10.1007/978-3-540-72432-2.
20. eds. Sammut C., Webb G. I. Encyclopedia of machine learning. New York, Springer, 2011, 1031 p. DOI: 10.1007/978-0-387-30164-8.
21. Russel S., Norvig P. Artificial intelligence: a modern approach. New Jersey, Prentice Hall, 2009, 1152 p.
22. eds. Berthold M., Hand D. J. Intelligent data analysis: an introduction. New York, Springer Verlag, 2007, 525 p.
23. Dubrovin V. I., Subbotin S. A., Boguslaev A. V., Jacenko V. K. Intellektual'nye sredstva diagnostiki i prognozirovanija nadezhnosti aviadvigatelej : monografija. Zaporozh'e, OAO «Motor-Sich», 2003, 279 p.