УДК 658.012

С. Д. Штовба, д. т. н., проф.; В. В. Мазуренко; Д. А. Савчук

ГЕНЕТИЧНИЙ АЛГОРИТМ ВИБОРУ ПРАВИЛ НЕЧІТКОЇ БАЗИ ЗНАНЬ, ЗБАЛАНСОВАНОЇ ЗА КРИТЕРІЯМИ ТОЧНОСТІ ТА КОМПАКТНОСТІ

Запропоновано генетичний алгоритм пошуку набору правил для формування нечіткої бази знань, збалансованої за критеріями точності та компактності. Відмінністю алгоритму є введення в постановку задачі оптимізації лінійного обмеження, яке задає рівень компенсації точності моделі її компактністю. Це наближує область допустимих розв'язків до парето-фронту.

Ключові слова: *нечітка база знань, точність, компактність, вибір правил, парето-фронт, генетична оптимізація.*

Вступ

Нечіткою базою знань називають сукупність нечітких правил "Якщо – то", які описують взаємозв'язок між входами та виходами деякого об'єкта з використанням лінгвістичних термів [1]. Одним із завдань проектування нечіткої бази знань є вибір правил з деякої, наперед визначеної, множини кандидатів. Правила-кандидати можуть бути сформовані експертом або отримані шляхом оброблення відповідних експериментальних даних.

В ідеальному випадку нечітка база знань має бути і компактною, і адекватною. Досягти цього в реальних задачах неможливо, тому на практиці намагаються обрати базу знань з коректним балансом між цими суперечливими критеріями. Необхідною умовою такого балансу є потрапляння бази знань на парето-фронт у координатах "складність моделі – точність моделі".

Вибір правил нечіткої бази знань можна звести до бінарної задачі про рюкзак. Правилу бази знань відповідає предмет, який може потрапити до рюкзака, точності бази знань – корисність рюкзака, а кількості правил – сумарна об'єм обраних предметів. Відмінність між задачами полягає в різних типах функції корисності, яка є лінійною в задачі про рюкзак та нелінійною в задачі вибору правил бази знань. Аналогічно до класичних постановок задачі про рюкзак [2] сформовано й задачі вибору правил нечіткої бази знань. Основними роботами в цій галузі є статті [3, 4] з формування множини баз знань нечіткого класифікатора, які належать парето-фронту недомінантних альтернатив у координатах "кількість правил – безпомилковість". Для цього застосовують задачі оптимізації з метою: 1) максимізації безпомилковості за обмеженої кількості правил; 2) мінімізації кількості правил за деякого рівня безпомилковості; 3) мінімізації інтегрального критерію якості бази знань у формі лінійної згортки безпомилковості та кількості правил [4] або безпомилковості, кількості правил та сумарної довжини антецедентів правил [5]. Для отримання парето-фронту оптимізацію проводять багаторазово за різних граничних значень в обмеженнях задач 1 і 2 та вагових коефіцієнтів цільової функції в задачі 3. Аналогічні підходи застосовують, обираючи правила нечітких баз знань для об'єктів з неперервним виходом [6].

Задача вибору правил нечіткої бази знань, як і задача про рюкзак, є NP-повною. Відповідно алгоритм точного розв'язання цієї задачі матиме експоненціальну обчислювальну складність, і тому буде прийнятним лише за невеликої кількості правил-кандидатів. На практиці для розв'язання цієї задачі зазвичай застосовують генетичні алгоритми. Кодування варіантів здійснюють за Піттсбурзьким методом [7], представляючи варіант розв'язку хромосомою, кожен ген якої задає належність відповідного правила до бази знань [6].

Граничне обмеження на складність бази знань або на точність нечіткої моделі [3, 4] формує досить велику область допустимих розв'язків, значна частина якої розташована далеко від парето-фронту. Це уповільнює знаходження оптимальних розв'язків, розташованих на парето-фронті. **Метою статті** є скорочення тривалості вибору правил нечіткої бази знань за рахунок розробки нового методу пошуку оптимальних розв'язків в околі парето-фронту. Цей окіл сформуємо лінійним обмеженням, яке описує компенсацію точності моделі її компактністю. Коефіцієнти обмеження оцінимо за крайніми точками парето-фронту, що відповідають майже порожнім та майже заповненим базам знань, а також за його верхньою межею, яку знайдемо жадібним алгоритмом на основі ідей наближеного методу Сахні для задачі про рюкзак [2]. Обчислювальна складність цієї процедури є квадратичною, тому вона суттєво не збільшить тривалість оптимізації. Пошук оптимальних розв'язків здійснимо генетичним алгоритмом.

1. Математичні постановки задач

Вважатимемо відомими:

- вибірку з M пар експериментальних даних про вплив чинників $X = (x_1, x_2, ..., x_n)$ на неперервний вихід y досліджуваної залежності:

$$(X_r, y_r),\ r = \overline{1, M}, \tag{1}$$

де X_r – вхідний вектор у r-ому рядку вибірки; y_r – відповідне вихідне значення;

- множину R з N правил-кандидатів у нечітку базу знань, $N = |R|$.

Позначимо через $y = F(R', X)$ модель на основі нечітких правил $R' \subseteq R$, що пов'язує входи X з виходом y досліджуваної залежності. За критерій точності нечіткої моделі оберемо середню квадратичну помилку на вибірці (1):

$$RMSE(R') = \sqrt{\frac{1}{M} \sum_{r=\overline{1, M}} (y_r - F(R', X_r))^2}. \tag{2}$$

У загальному випадку задача полягає у знаходженні такої множини правил R', що забезпечує:

$$\begin{cases} RMSE(R') \to \min \\ C(R') \to \min \end{cases}, \tag{3}$$

де $C(R')$ – складність нечіткої моделі, яку визначимо кількістю правил $C(R') = |R'|$ або в загальному випадку рівнем наповненості бази знань $C(R') = \frac{|R'|}{N}$.

Багатокритеріальну задачу оптимізації (3) перетворюють у такі скалярні задачі [2, 3]:

$$\begin{cases} RMSE(R') \to \min \\ C(R') \le C^* \end{cases}, \tag{4}$$

$$\begin{cases} C(R') \to \min \\ RMSE(R') \le RMSE^* \end{cases}, \tag{5}$$

де C^* та $RMSE^*$ – максимально допустимі значення складності та помилки.

Постановки (4) та (5) формують велику область допустимих розв'язків, причому значна її частина розташована далеко від парето-фронту (рис. 1а та 1б). Ми пропонуємо обмеження в задачі оптимізації записати так:

$$RMSE(R') \le a \cdot C(R') + b, \tag{6}$$

де $a < 0$ та $b > 0$ – параметри, підбираючи які можна сформувати область допустимих розв'язків в околі парето-фронту (рис. 1в).

Використовуючи обмеження (6), сформулюємо такі задачі вибору правил нечіткої бази знань:

$$\begin{cases} RMSE(R') \to \min \\ RMSE(R') \le a \cdot C(R') + b \end{cases}, \tag{7}$$

$$\begin{cases} C(R') \to \min \\ RMSE(R') \le a \cdot C(R') + b \end{cases}. \tag{8}$$

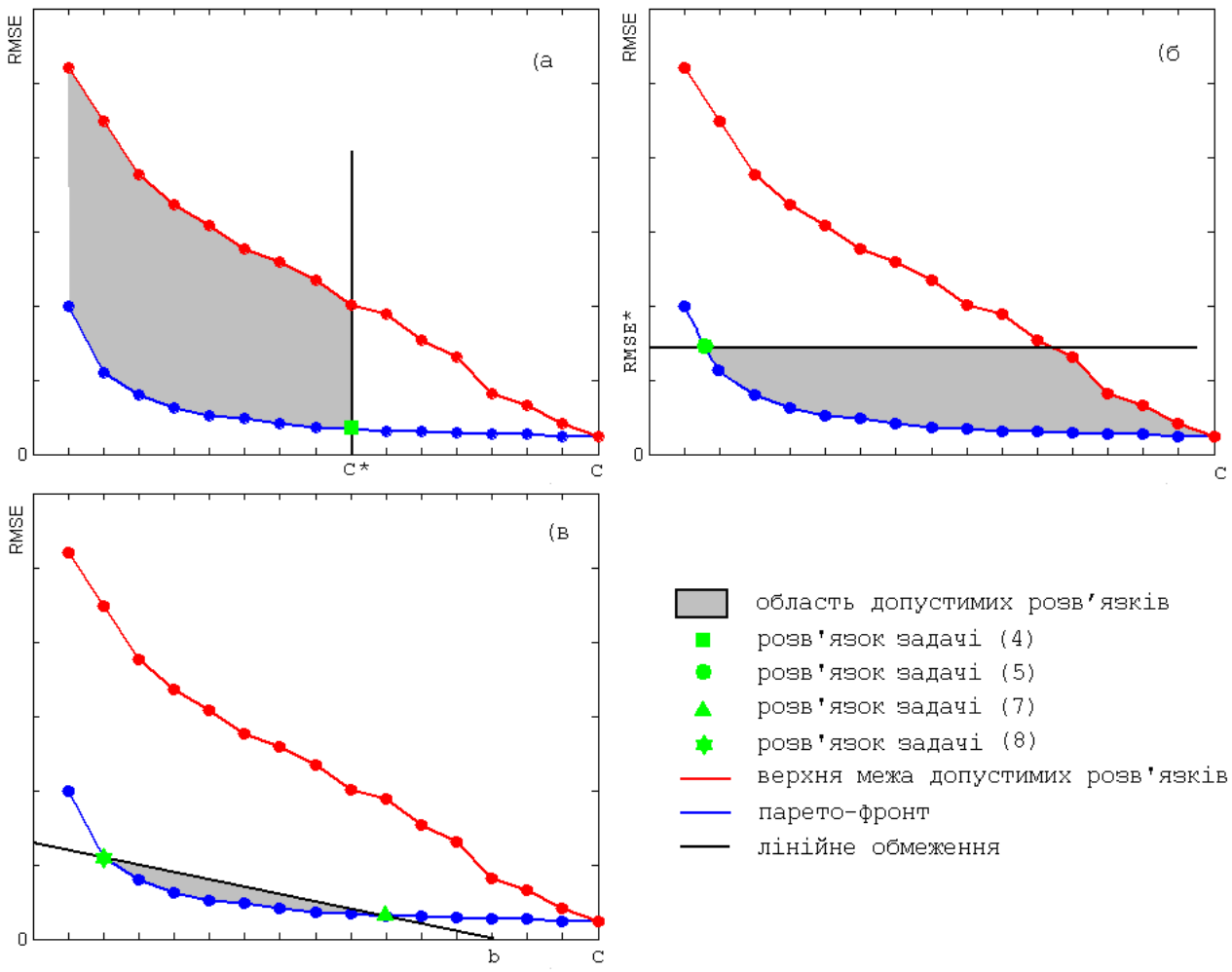

Рис. 1. Область допустимих розв'язків:
а) задача (4); б) задача (5); в) задачі (7) і (8)

2. Оцінювання параметрів лінійного обмеження

Для визначення параметрів a та b обмеження (6) достатньо знати відповідні характеристики двох баз знань, які задовольняють користувача. Позначимо їх $(C_1, RMSE_1)$ та $(C_2, RMSE_2)$. Тоді:

$$\begin{cases} a = \dfrac{RMSE_2 - RMSE_1}{C_2 - C_1} \\ b = RMSE_1 - a \cdot C_1 \end{cases}. \tag{9}$$

Параметр a можна трактувати як коефіцієнт компенсації точності за рахунок компактності. Його можна визначити з відповіді користувача на питання «Наскільки можна зменшити точність моделі за рахунок скорочення числа правил на 1?». Тоді для визначення другого параметра b достатньо знати характеристики однієї прийнятної бази знань.

Параметри a та b можна визначити, провівши лінійне обмеження через будь-які дві крайні точки парето-фронту. Одна з крайніх точок має бути зліва, а інша – справа (рис. 2). Обчислювальна складність повного перебору для визначення 5-ти крайніх точок парето-фронту для баз знань з 1-го, 2-ох, N–2, N–1 та N правил є квадратичною $O(N^2)$, тому такий підхід можна застосовувати і для задач великої розмірності.

Лінійне обмеження можна провести і через дві точки кривої навчання у формі залежності нев'язки від складності нечіткої бази знань. Криву навчання пропонуємо побудувати за результатами виконання жадібного алгоритму вибору правил. Цей алгоритм полягає в додаванні на кожному кроці до бази знань одного правила, яке максимально знижує нев'язку. Отримана крива навчання завжди буде не нижче парето-фронту (див. рис. 2). Початковою базою знань для жадібного алгоритму можна обрати базу знань з парето-фронту, що містить 2 або N–2 правила. Обчислювальна складність жадібного алгоритму є квадратичною.

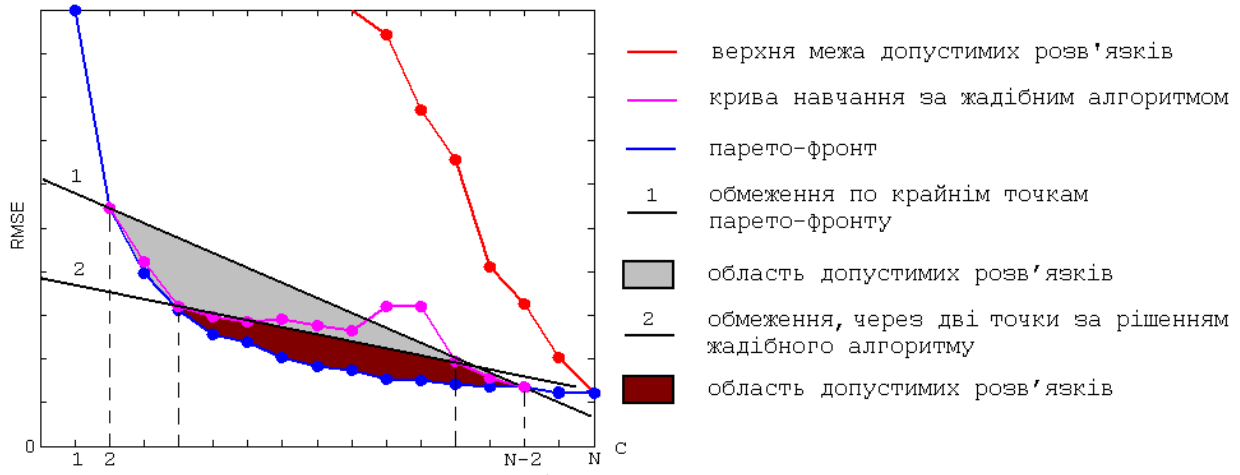

Рис. 2. До розрахунку параметрів лінійного обмеження

3. Генетичний алгоритм розв'язання задачі

Для розв'язання задачі оптимізації (7) та (8) скористаємося генетичним алгоритмом на основі Піттсбурзького підходу. Кожна хромосома популяції задає нечітку базу знань з власним набором правил R'. Кожен із N генів цієї хромосоми може приймати такі значення: 1 (якщо відповідне правило потрапляє до бази знань) та 0 (якщо відповідне правило не потрапляє).

Початкова популяція генерується випадково, але з включенням субоптимальних розв'язків, знайдених за жадібним алгоритмом.

Імовірність відбору хромосоми для схрещення визначають таким чином:

$$p = \frac{n - j}{2n}, \tag{10}$$

де n – розмір популяції; j – ранг хромосоми, який визначають за фітнес-функцією.

Мутації піддається β-частка хромосом, отриманих унаслідок схрещування.

Селекція здійснюється детермінованим вибором n найкращих хромосом.

4. Комп'ютерні експерименти

Експерименти проведемо для синглтонних нечітких баз знань, у яких антецеденти правил задаються нечіткими множинами, а консеквенти – числовими значеннями [1]. Як і в наших попередніх статтях із багатокритеріального формалізованого проектування нечітких баз знань [8 – 11] експерименти здійснимо на трьох еталонних залежностях (рис. 3):

$$\text{зростаючій} - \ y = a\sqrt{b} \quad a \in [2;22],\ b \in [2;14]; \tag{11}$$

$$\text{унімодальній} - \ y = -a^2 - b^2,\ a \in [-7;3],\ b \in [-5;5]; \tag{12}$$

$$\text{багатоекстремальній} - \ y = (1+\sin(a)^2)^b,\ a \in [0;5],\ b \in [0.5;2]. \tag{13}$$

Для кожної залежності (11) – (13) створено повну синглтонну нечітку базу знань на $N = 16$ правил (табл. 1). Фазіфікацію вхідних змінних здійснено гаусовими функціями належності [1] (рис. 4). Консеквенти правил обчислення за функціями (11) – (13) з аргументами, що дорівнюють ядрам нечітких множин з антецедентів правил.

Таблиця 1

Повні набори правил (R) для кожної залежності

№	a	b	y, для залежності (11)	y, для залежності (12)	y, для залежності (13)
1	Very low	Very low	5,04	-71,91	0,95
2	Low	Very low	14,04	-48,94	0,81
3	Medium	Very low	24,84	-45,27	0,94
4	High	Very low	33,84	-62,14	0,79
5	Very low	Low	7,59	-46,08	1,04
6	Low	Low	21,14	-23,11	1,23
7	Medium	Low	37,4	-19,44	1,06
8	High	Low	50,95	-36,3	1,26
9	Very low	Medium	9,82	-31,08	1,17
10	Low	Medium	23,37	-8,1	2,02
11	Medium	Medium	48,42	-4,44	1,25
12	High	Medium	65,96	-21,3	2,17
13	Very low	High	11,36	-31,91	2,29
14	Low	High	31,64	-8,94	3,04
15	Medium	High	55,97	-5,27	1,45
16	High	High	76,25	-22,14	3,40

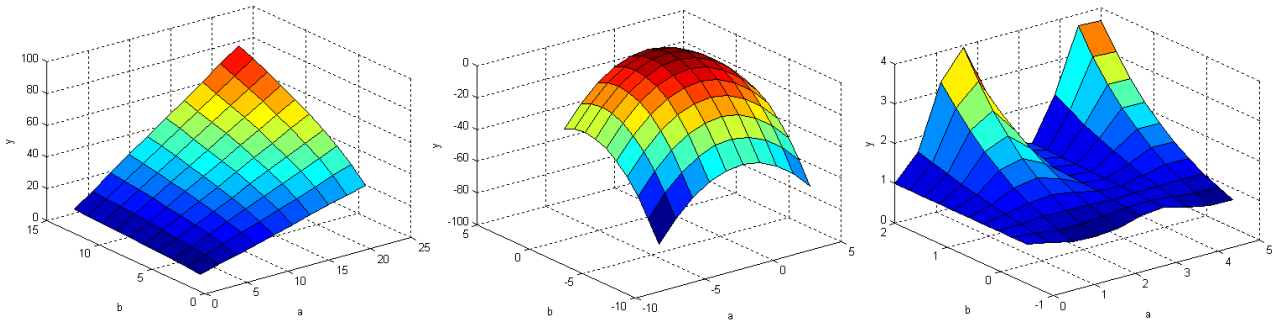

Рис. 3. Поверхні залежностей (11) – (13)

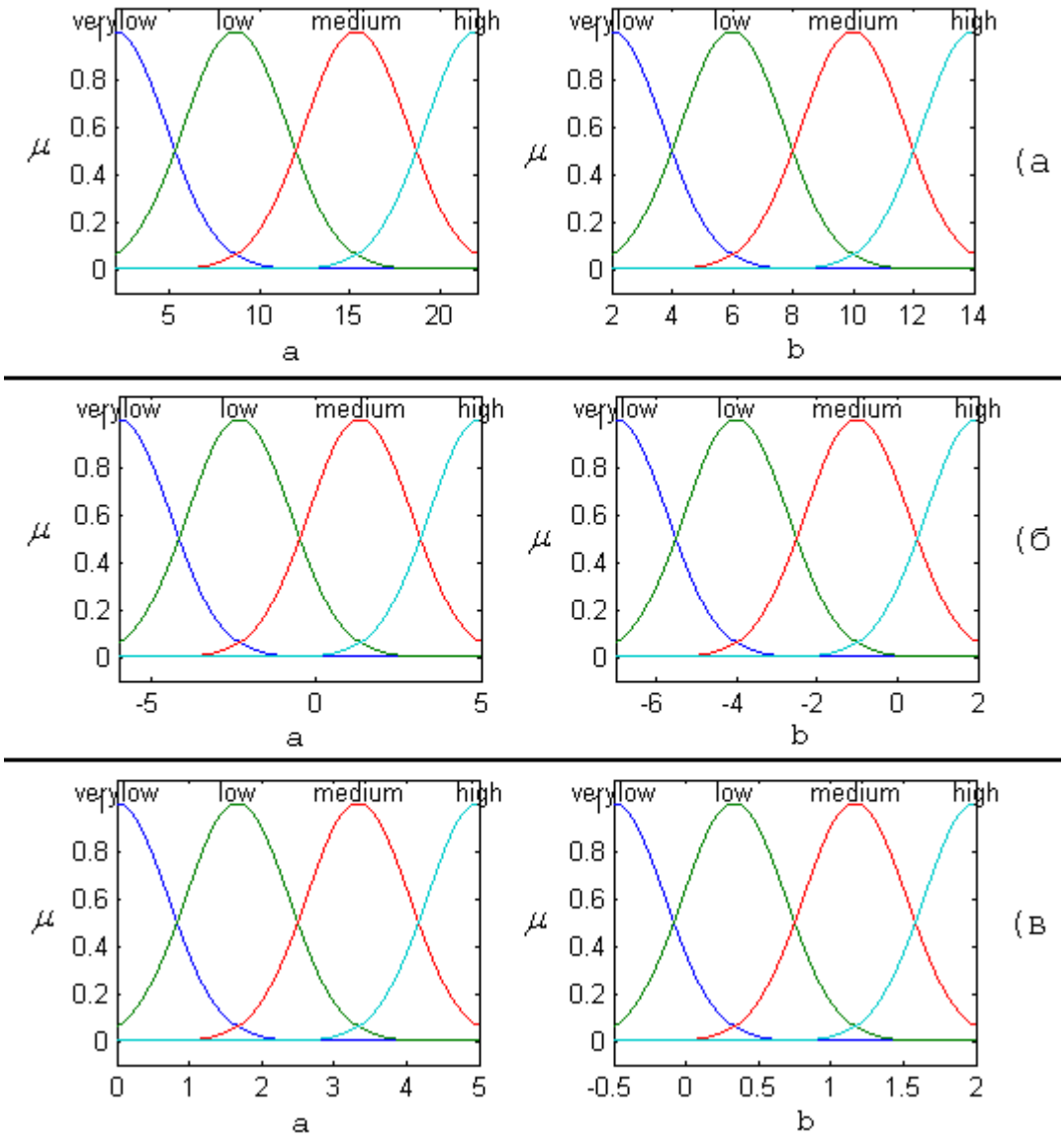

Рис. 4. Функції належності термів вхідних змінних для:
а) залежності (11); б) залежності (12); в) залежності (13)

Параметри лінійного обмеження в (7) – (8) визначимо для кожного експерименту окремо. Для задачі з еталонною залежністю (11) спочатку за жадібним алгоритмом знайдемо найкращі бази знань з різною кількістю правил (рис. 5). Вважаючи отриману криву навчання за орієнтир, установимо бажане значення нев'язки для бази знань із 4-х правил трохи менше, ніж на рис. 5, наприклад, на рівні $RMSE \le 0{,}55$. Другою точкою лінійного обмеження оберемо базу знань із 10 правил з нев'язкою не більшої, ніж для повної бази знань, тобто $RMSE \le 0{,}22$. Підставляючи в (9), отримуємо $a = -0{,}0367$ та $b = 0{,}6968$. Для залежності (12) прийнятними вважатимемо базу знань із 6 правил з нев'язкою $RMSE \le 0{,}75$ та базу знань із 10 правил з нев'язкою $RMSE \le 0{,}58$. Підставляючи ці значення в (9), отримуємо $a = -0{,}0425$ та $b = 1{,}005$. Для залежності (13) прийнятною вважатимемо базу знань із 9-ти правил з нев'язкою $RMSE \le 0{,}0365$ та таким рівнем компенсації нев'язки на одне додаткове правило – $\Delta RMSE \le -0{,}00125$. Звідси $a = -0{,}00125$ та $b = 0{,}0365 + 9 \cdot 0{,}00125 = 0{,}04775$.

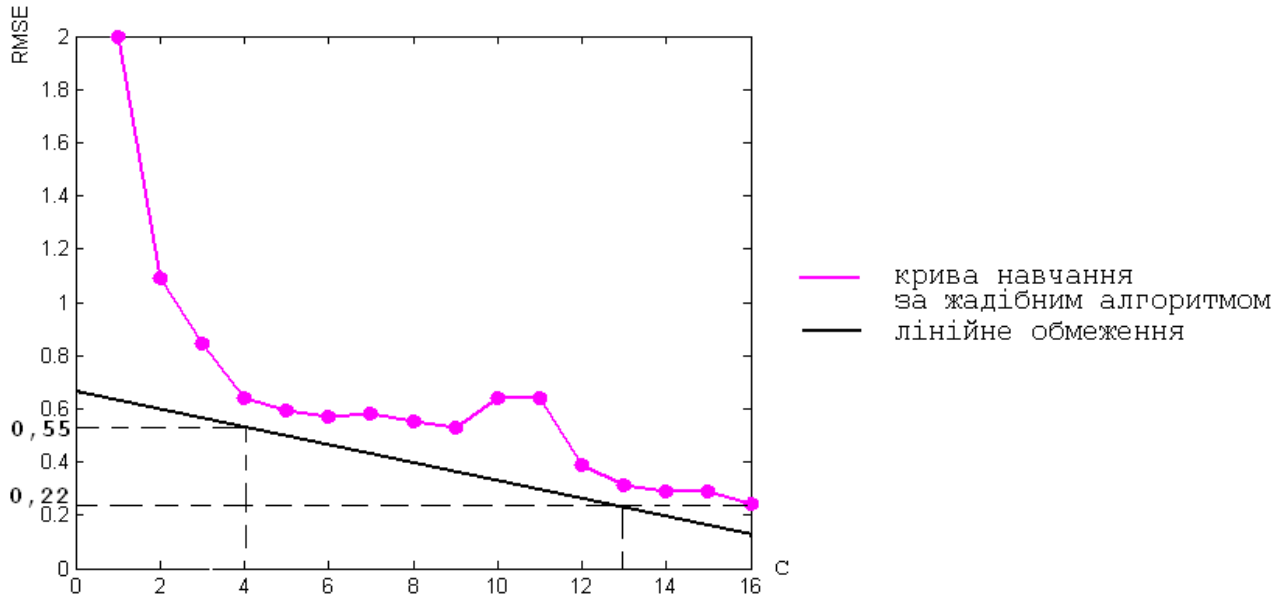

Рис. 5. Оцінка параметрів обмеження для експериментів з еталонною залежністю (11)

Експерименти проведено за таких параметрів генетичного алгоритму: розмір популяції $n = 160$, кількість генів $N = 16$, тиск мутації $\beta = 0{,}3$, кількість епох $k = 10$. Отримані розв'язки задач (7) та (8) зведено в табл. 2. В усіх 6-ти випадках знайдені нечіткі бази знань розташовані на парето-фронті, тобто мають найменшу нев'язку за фіксованого числа правил (рис. 6). Парето-фронт, а також верхню межу області допустимих розв'язків знайдено в наших попередніх роботах [10, 11] за допомогою повного перебору всіх можливих комбінацій правил нечіткої бази знань. Обчислювальна складність повного перебору є експоненційною $O(2^N)$, тому в тих роботах для розв'язання кожної тестової задачі перевірено 65536 варіантів нечіткої бази знань. Запропонований генетичний алгоритм знайшов глобальний розв'язок, перебравши для кожної задачі 1600 варіантів.

Таблиця 2

Результати експериментів

Еталонна залежність	(11)	(11)	(12)	(12)	(13)	(13)
Постановка завдання	(7)	(8)	(7)	(8)	(7)	(8)
Параметри обмеження	a=-0,0367 b=0,6968	a=-0,0367 b=0,6968	a=-0,0425 b=1,005	a=-0,0425 b=1,005	a=-0,00125 b=0,04775	a=-0,00125 b=0,04775
Розв'язок (R')	(1; 2; 5; 6; 7; 9; 10; 12; 16)	(2; 3; 6; 7; 11; 16)	(1; 2; 4; 5; 6; 7; 8; 10; 11; 13; 16)	(1; 3; 6; 11; 12)	(2; 5; 7; 9; 10; 11; 12; 13; 14; 15; 16)	(1; 3; 11; 14; 16)
C(R')	10	5	11	5	11	5
RMSE(R')	0,3098	0,5128	0,5134	0,7235	0,0334	0,0387

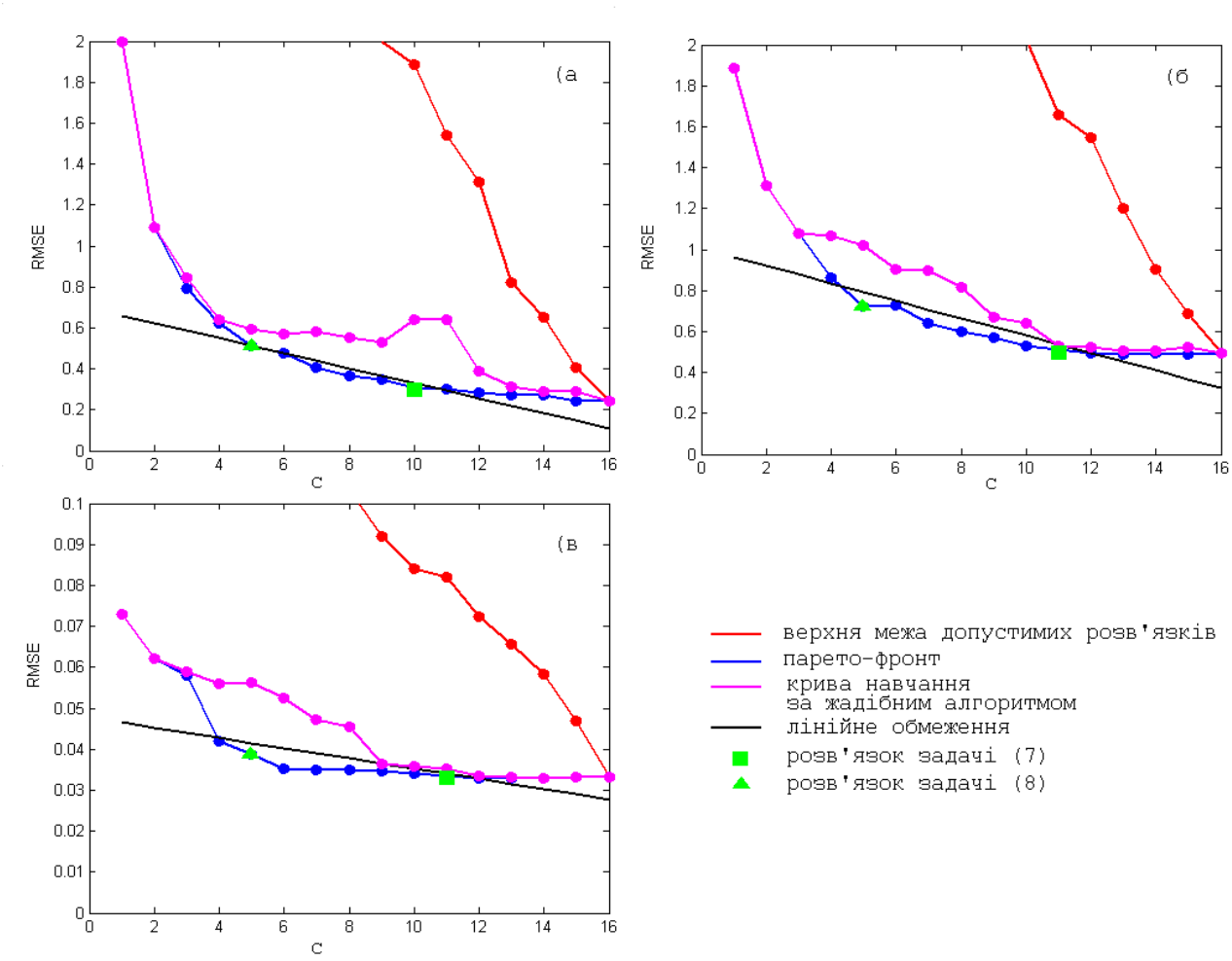

Рис. 6. Криві навчання нечіткої бази знань для:
а) залежності (11); б) залежності (12); в) залежності (13)

Висновки

Розроблено новий метод розв'язання однієї із задач нечіткої ідентифікації, а саме: вибору правил нечіткої бази знань з урахуванням точності та компактності. Новизною методу є використання замість типових граничних обмежень на точність та складність лінійного обмеження, яке задає рівень компенсації між цими суперечливими критеріями. За новим обмеженням вдається суттєво звузити область допустимих розв'язків, стягнувши її до околу парето-фронту. За допомогою комп'ютерів експериментів установлено, що за новою постановкою задачі генетичний алгоритм знаходить глобальний оптимум, згенерувавши в десятки разів менше варіантів нечітких баз знань, ніж за повного перебору варіантів.

СПИСОК ЛІТЕРАТУРИ

1. Штовба С. Д. Проектирование нечетких систем средствами MATLAB / С. Д. Штовба. – М.: Горячая линия. – Телеком, 2007. – 288 с.

2. Martello S. Knapsack problems: algorithms and computer implementations / S. Martello , P. Toth. – New York: John Wiley & Sons, Inc, 1990. – 296 p.

3. Ishibuchi H. Selecting fuzzy if-then rules for classification problems using genetic algorithms / H. Ishibuchi, K. Nozaki, N. Yamamoto, H. Tanaka // IEEE Transactions on Fuzzy Systems. – 1995. – Vol. 3, No. 3. – P. 260 – 270.

4. Ishibuchi H. Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems / H. Ishibuchi, T. Murata, I. B. Turksen // Fuzzy Sets and Systems. – 1997. – Vol. 89, No. 2 – P. 135 – 150.

5. Ishibuchi H. Three-objective genetics-based machine learning for linguistic rule extraction / H. Ishibuchi, T. Nakashima, T. Murata // Inform. Sci. – 2001. – Vol. 136, No. 1. – P. 109 – 133.

6. Cordon O. A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems / O. Cordon // International Journal of Approximate Reasoning. – 2011. – Vol. 52. – P. 894 – 913.

7. Cordon O. Ten years of genetic fuzzy systems: current framework and new trends / O. Cordon, F. Gomideb, F. Herreraa, F. Homannc, L. Magdalenad // Fuzzy Sets and Systems. – 2004. – Vol. 141. – P. 5 – 31.

8. Штовба С. Д. Вплив кількості нечітких правил на точність бази знань Мамдані / С. Д. Штовба , В. В. Мазуренко , О. Д. Панкевич // Вісник Хмельницького національного університету. Технічні науки. – 2011. – № 2. – С. 185 – 188.

9. Штовба С. Д. Дослідження навчання компактних нечітких баз знань типу Мамдані / С. Д. Штовба, В. В. Мазуренко // Штучний інтелект. – 2011. – № 4. – С. 521 – 529.

10. Штовба С. Д. Дослідження навчання компактних нечітких синглтонних баз знань / С. Д. Штовба, В. В. Мазуренко // Вимірювальна та обчислювальна техніка в технологічних процесах. – Хмельницький: ХНУ., 2011 – № 1 – С. 133 – 139.

11. Штовба С. Д. Залежність точності ідентифікації від обсягу нечіткої синглтонної бази знань / С. Д. Штовба , О. Д. Панкевич, В. В. Мазуренко // Інформаційні технології та комп'ютерна інженерія. – 2011. – № 1. – С. 73 – 78.

Штовба Сергій Дмитрович – професор, д. т. н., професор кафедри комп'ютерних систем управління.

Мазуренко Віктор Володимирович – аспірант кафедри комп'ютерних систем управління.

Савчук Дмитро Анатолійович – студент інститут автоматики, електроніки та комп'ютерних систем управління.

Вінницький національний технічний університет.