6Ш

УДК 658.012

С.Д. Штовба
Вінницький національний технічний університет, м. Вінниця, Україна
shtovba@ksu.vstu.vinnica.ua

Навчання нечіткої бази знань за вибіркою нечітких даних

Стаття досліджує навчання нечіткої бази знань за вибіркою з нечіткими значеннями вихідної змінної. Запропоновано 2 способи побудови багатофакторних моделей на нечітких базах знань, на виході яких отримують нечіткі числа. Запропонована постановка задачі навчання таких нечітких баз знань за нечіткою вибіркою даних, вказано методи її розв'язання та наведені відповідні приклади. Проведені обчислювальні експерименти свідчать, що навчання за нечіткими даними покращує точність моделювання як на чіткій, так і на нечіткій тестових вибірках.

Вступ та мета роботи

Навчання нечіткої бази знань являє собою ітераційну процедуру зміни її параметрів для мінімізації відхилення результатів логічного висновку від експериментальних даних. Питання навчання нечітких баз знань різних форматів досліджувалися в багатьох роботах, серед яких виділимо публікації [1-8]. Спільною рисою цих робіт є використання для навчання бази знань вибірки з числових (чітких) даних.

Водночас для багатьох практичних задач доступні для навчання дані містять нечіткі оцінки. В цих випадках настроювання моделі має здійснюватися в умовах невизначеності за нечіткою навчальною вибіркою. У статті [9] запропоновано підхід до навчання нечіткої бази знань типу Мамдані за вибіркою даних, входи якої задано нечіткими числами. Метою цієї статті, яка продовжує роботу [9], є дослідження навчання нечіткої бази знань у випадку навчальної вибірки з нечіткими значеннями вихідної змінної.

Постановка задачі. Розглядається залежність функції належності нечіткого числа $\widetilde{y}$ від факторів $(x_1, x_2, ..., x_n)$. Будемо вважати, що вона описана нечіткою базою знань та репрезентована такою навчальною вибіркою:

$$\left(X_r, \widetilde{y}_r\right),\ r = \overline{1, M}\,, \tag{1}$$

де $X_r = \left(x_{r1}, x_{r2}, ..., x_{rn}\right)$ – вхідний вектор в r-ій стрічці вибірки та $\widetilde{y}_r = \int\limits_{y\in[\underline{y},\, \overline{y}]} \mu_{\widetilde{y}_r}(y) \,/\, y$ – відповідний вихід у вигляді нечіткого числа з функцією належності $\mu_{\widetilde{y}_r}(y)$ на носії $[\underline{y}, \overline{y}]$.

Задачу навчання нечіткої бази знань поставимо як пошук такого вектора P її параметрів, який на нечіткій вибірці (1) забезпечує:

$$\sqrt{\frac{1}{M}\sum_{r=\overline{1, M}} RMSE\left(\widetilde{y}_r, \widetilde{F}\left(P, X_r\right)\right)^2} \to \min\,, \tag{2}$$

де $\widetilde{F}(P, X_r)$ – нечітке число, що отримане для вхідного вектора X_r в результаті логічного виведення за нечіткою базою знань з параметрами P;

RMSE – нев’язка між двома нечіткими числами, які відповідають бажаному та дійсному результату логічного висновку.

Для розрахунку нев’язки між нечіткими числами $\widetilde{A} = \int\limits_{y \in [\underline{y}, \overline{y}]} \mu_{\widetilde{A}}(y) / y$ та $\widetilde{B} = \int\limits_{y \in [\underline{y}, \overline{y}]} \mu_{\widetilde{B}}(y) / y$, що задані на універсальні множині $[\underline{y}, \overline{y}]$, пропонується така формула:

$$RMSE(\widetilde{A},\ \widetilde{B}) = \sqrt{\frac{\int\limits_{\underline{y}}^{\overline{y}} \left(\mu_{\widetilde{A}}(y) - \mu_{\widetilde{B}}(\ y)\right)^2 dy}{\overline{y} - \underline{y}}}\ . \qquad (3)$$

Типові моделі (рис. 1) на основі нечіткої бази знань продукують на виході чітке (числове) значення. Тому перед викладенням підходу до навчання за постановкою (2) спочатку розглянемо, як можна отримати на виході нечіткої моделі нечітке число. Нижче пропонуються два способи синтезу нечітких чисел за нечіткою базою знань.

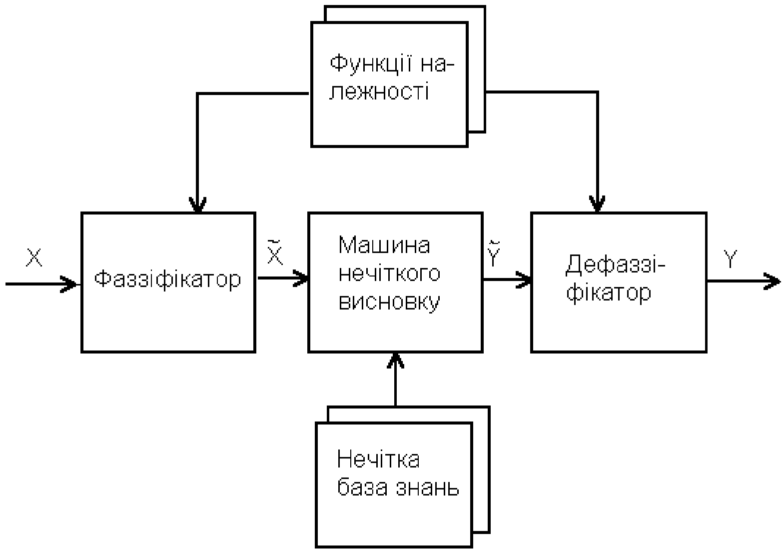

Рисунок 1 – Типова структура нечіткої моделі

Навчання бази знань з нечіткими консеквентами

Як перший спосіб синтезу нечітких чисел пропонується логічне виведення за типовою нечіткою моделлю з рис. 1, але без застосування дефаззіфікації. Для цього способу придатні нечіткі бази знань Мамдані та реляційні нечіткі бази знань Педрича. Крім них можна використовувати і композиційне правило виведення Заде. Ключовою спільною рисою логічного висновку за базами знань з нечіткими консеквентами є продукування на виході нечіткої множини типу:

$$\widetilde{y} = \int\limits_{y \in [\underline{y}, \overline{y}]} \mu_{\widetilde{y}}(y) / y\ . \qquad (4)$$

Звернемо увагу, що результатом нечіткого висновку можуть бути субнормальні чи невипуклі нечіткі множини. Поява субнормальної нечіткої множини не несе проблем, тому що її завжди можна нормалізувати. Задача виправлення невипуклості значно складніша. Розглянемо її детальніше.

Типові приклади невипуклих нечітких множин, які отримані за алгоритмом нечіткого виведення Мамдані, наведені на рис. 2. З нього видно, що в одній нечіткій множині буває декілька зон невипуклості, причому на них може припадати більше половини її носія. Перетворювати невипуклі нечіткі множини $\tilde{y}_{nonc}$ в випуклі $\tilde{y}_{conv}$ пропонується апроксимацією параметричними функціями належності. При апроксимації пропонується підбирати параметри не тільки щоб мінімізувати RMSE між функціями належності, але щоб результати дефаззіфікації нечітких множин збігалися. У нечіткому моделюванні зазвичай використовується дефаззіфікація за центром тяжіння, тому порівнювати нечіткі множини будемо саме за цим методом. Позначимо параметри апроксимувальної функції належності через Z. Тоді задачу перетворення невипуклої нечіткої множини в чітке число поставимо таким чином: *знайти такий вектор Z, щоб*:

$$\left.\begin{array}{l} RMSE(\tilde{y}_{nonc}, \tilde{y}_{conv}) \ \rightarrow \ \min \\ за \ \ умови \ \ defuz(\tilde{y}_{nonc}) = defuz(\tilde{y}_{conv}) \end{array}\right\}, \qquad (5)$$

де defuz – операція дефаззіфікації.

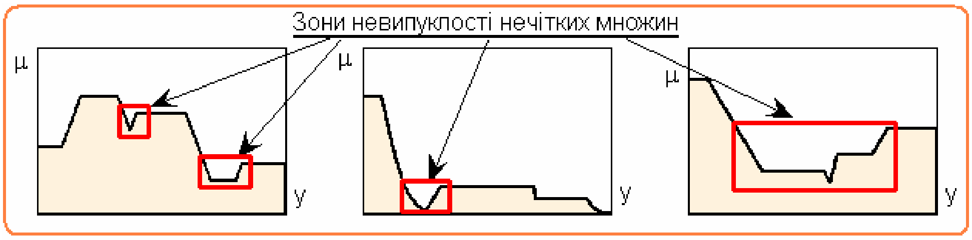

Рисунок 2 – Ефект невипуклості нечітких множин, отриманих за алгоритмом Мамдані

Однією із особливостей функцій належностей з рис. 2 є те, що їх графіки містять декілька протяжних плато. Типовим функціям належності це не властиво (рис. 3а). Тому пропонуються нові моделі функцій належностей, графіки яких можуть містити 2 – 3 плато (рис. 3б). Їх аналітичні вирази зведені в табл. 1.

Таблиця 1 – Аналітичні вирази функцій належності

Назва	Аналітичний вираз
Трикутна (trimf)	$\mu(y) = \begin{cases} 0, & якщо \ \ y < a \\ \dfrac{y-a}{b-a}, & якщо \ \ a \le y < b \\ \dfrac{c-y}{c-b}, & якщо \ \ b \le y \le c \\ 0, & якщо \ \ y > c \end{cases}$
Трикутна з порогами	$\mu(y) = \begin{cases} \min\left(t_l, trimf\,(y,a,b,c)\right), & якщо \ \ y < b \\ \min\left(t_r, trimf\,(y,a,b,c)\right), & якщо \ \ y > b \end{cases}$

Продовження табл. 1

Кускова лінійна	$\mu(y)=\begin{cases} h_1+\dfrac{(h_2-h_1)(y-\underline{y})}{a-\underline{y}}, & \text{якщо } y\le a \\ h_2+\dfrac{(1-h_2)(y-a)}{b-a}, & \text{якщо } y\in(a,\ b) \\ 1, & \text{якщо } y\in[b,\ c] \\ 1+\dfrac{(h_3-1)(y-c)}{d-c}, & \text{якщо } y\in(c,d) \\ h_3+\dfrac{(h_4-h_3)(y-\overline{y})}{\overline{y}-d}, & \text{якщо } y\ge d \end{cases}$
Гаусова (gaussmf)	$\mu(y)=\exp\left(-\dfrac{(y-b)^2}{2\cdot c^2}\right)$
Двостороння гаусова	$\mu(y)=\begin{cases} gaussmf\ (y,b_1,c_1), & \text{якщо } y<b_1 \\ 1, & \text{якщо } y\in[b_1,b_2] \\ gaussmf\ (y,b_2,c_2), & \text{якщо } y>b_2 \end{cases}$
Гаусова з порогами	$\mu(y)=\begin{cases} \min\left(t_l,\ gaussmf\ (y,b_1,c_1)\right), & \text{якщо } y<b_1 \\ 1, & \text{якщо } y\in[b_1,b_2] \\ \min\left(t_r,\ gaussmf\ (y,b_2,c_2)\right), & \text{якщо } y>b_2 \end{cases}$
Дзвіноподібна (bellmf)	$\mu(y)=\dfrac{1}{1+\left\lvert\dfrac{y-b}{c}\right\rvert^{2a}}$
Двостороння дзвіноподібна	$\mu(y)=\begin{cases} bellmf\ (y,a_1,b_1,c_1), & \text{якщо } y<b_1 \\ 1, & \text{якщо } y\in[b_1,b_2] \\ bellmf\ (y,a_2,b_2,c_2), & \text{якщо } y>b_2 \end{cases}$

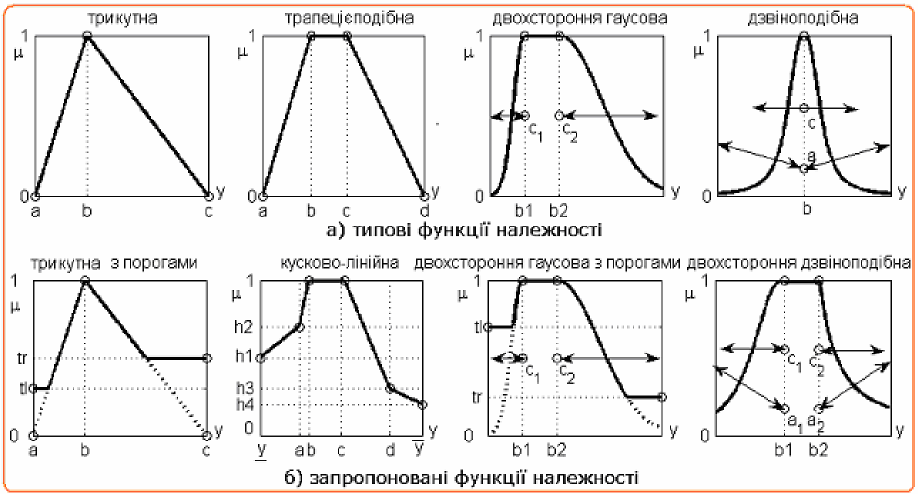

Рисунок 3 – Функції належності для апроксимації невипуклих нечітких множин

Приклад 1. На рис. 4 порівнюються результати апроксимації трьох невипуклих нечітких множин з рис. 2 відомими та запропонованими функціями належності. Апроксимація проведена за постановкою (5). Різниця результатів дефаззіфікації нечітких множин $\widetilde{y}_{nonc}$ та $\widetilde{y}_{conv}$ не перевищує 0.01 %. Рис. 4 свідчить, що точність апроксимації за запропонованими функціями належності значно краща.

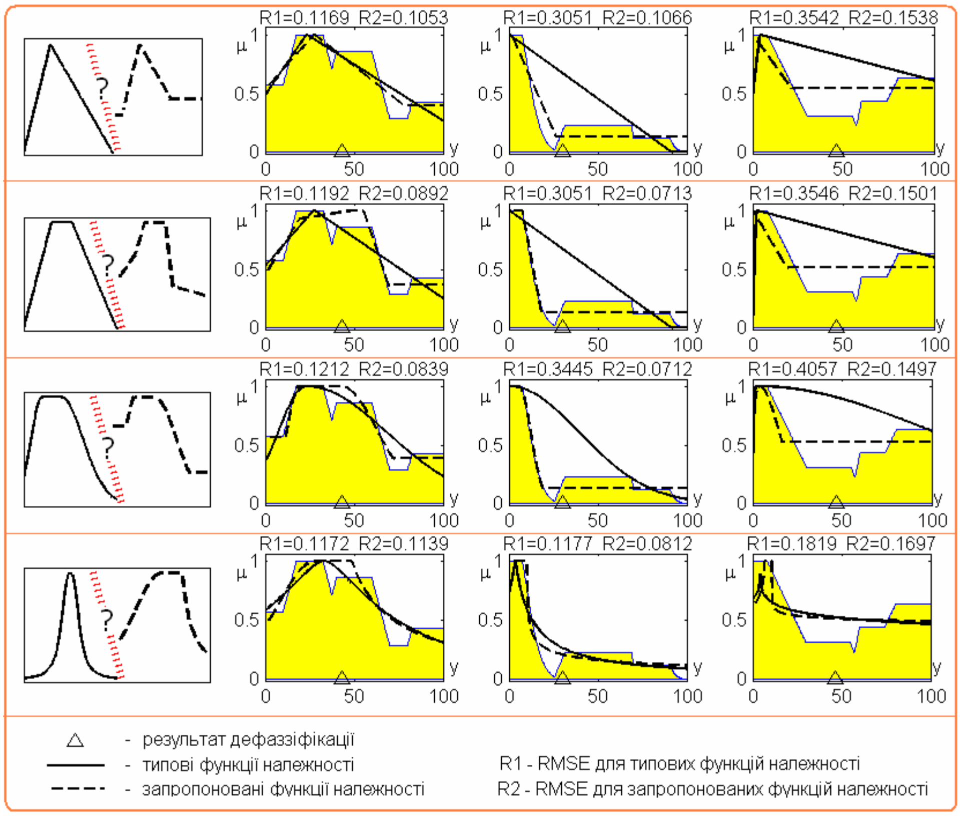

Рисунок 4 – Порівняння апроксимацій невипуклих нечітких множин з прикладу 1

Навчання за постановкою (2) являє собою задачу оптимізації, яку можна вирішити відповідними методами математичного програмування. Як і у випадку чітких навчальних вибірок вектор керованих змінних (P) сформуємо із параметрів функцій належностей нечітких термів.

Приклад 2. В [10] наведені дані 392 експериментів про залежність часу (y) виконання операції «Розігнання автомобіля до швидкості 60 миль за годину» від кількості циліндрів (x_1) та питомої потужності автомобіля (відношення потужності до маси автомобіля, x_2). За цими даними сформуємо для залежності $\tilde{y} = f(x_1, x_2)$ нечіткі навчальну та тестову вибірки та навчимо за ними нечітку базу знань Мамдані.

Навчальну та тестову вибірки сформуємо у такій спосіб. В експериментальних даних фактори впливу приймають значення: $x_1 \in \{3, 4, 5, 6, 8\}$ та $x_2 \in [0.0206, 0.0729]$. Для формування нечітких вибірок округлимо значення фактора x_2 до третього знака після коми. Тоді $x_2 \in \{0.021, 0.022, \ldots, 0.051, 0.054, 0.073\}$. Декартовий добуток $x_1 \times x_2$ складається з 5*33=165 точок. З них для 27 пар (x_1, x_2) в експериментальних даних існує 3 або більше різних значень вихідної змінної y. Для цих 27 пар значень факторів за розподілом вихідної змінної розрахуємо ступені належності, використовуючи ідеї потенціалу точки, на якому базується гірська кластеризація [5]. Потенціал точки – це число, яке показує, наскільки щільно розташовані в її околі експериментальні дані. Чим вище потенціал точки, тим ближче вона до центру кластера. Потенціал точки y_i $(i = \overline{1, v})$ розраховують таким чином [5]:

$$pot_i = \sum_{j=\overline{1, v}} \exp\left(-4\alpha^2 (y_i - y_j)^2\right),$$

де $\alpha > 0$ – коефіцієнт, що визначає розмазаність кластера за гірською кластеризацією.

Перед застосуванням цієї формули дані відмасштабуємо на одиничний відрізок.

Ступені належності нечіткої множини $\tilde{y}$ пропонується розраховувати з потенціалів таким чином:

$$\mu_{\tilde{y}}(y_i) = \frac{pot_i}{\max\limits_{j=\overline{1,v}} (pot_j)}.$$

Ступені належності апроксимуємо двосторонньою гаусовою кривою (табл. 1). У навчальну вибірку оберемо 20 пар даних, а в тестову – 7 (рис. 5).

Експериментальні дані (рис. 6) свідчать, що час виконання операції (y) спадає при збільшенні кількості циліндрів (x_1) та питомої потужності (x_2). Тому цю залежність можна описати такою базою знань:

якщо $x_1 = \tilde{4}$ та x_2 = Слабка, тоді y = Значний,

якщо $x_1 = \tilde{6}$ та x_2 = Середня, тоді y = Середній,

якщо $x_1 = \tilde{8}$ та x_2 = Сильна, тоді y = Малий.

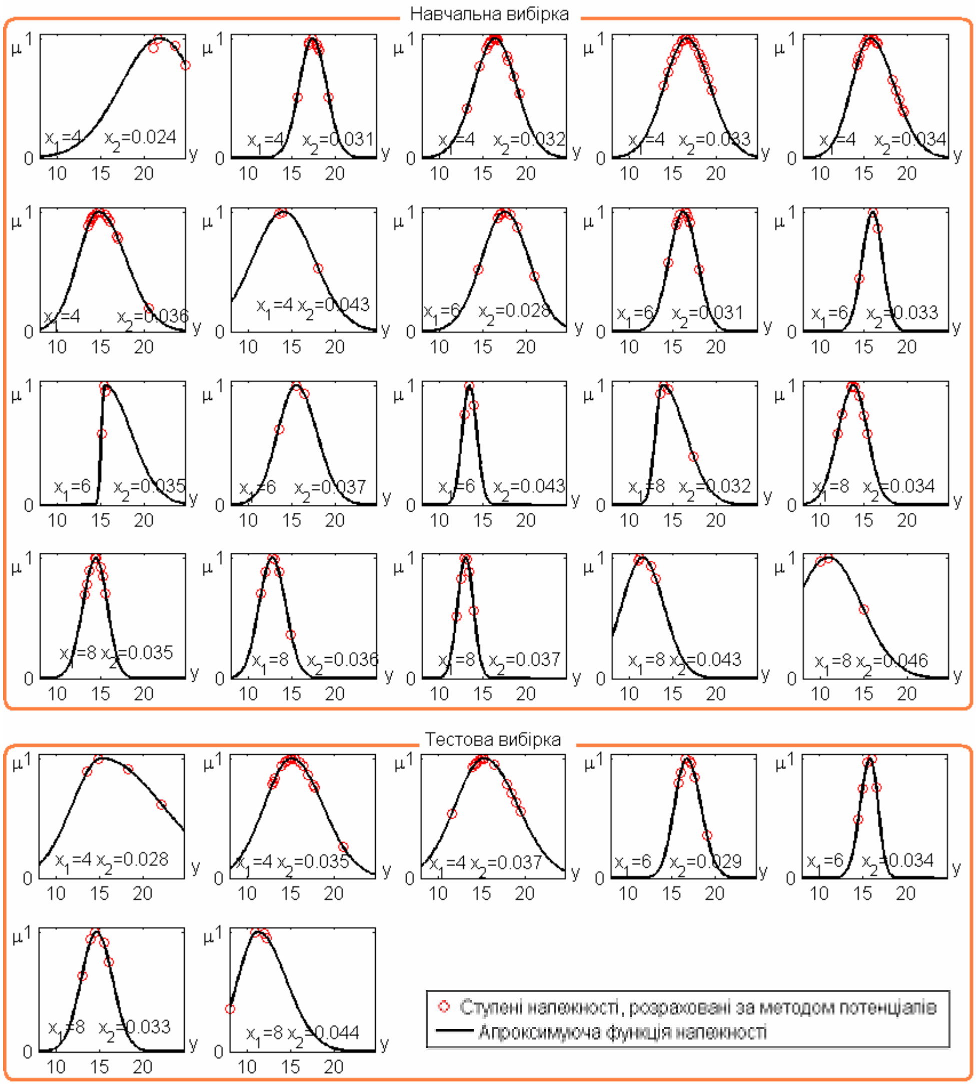

Рисунок 5 – Нечіткі навчальна та тестова вибірки з прикладу 2

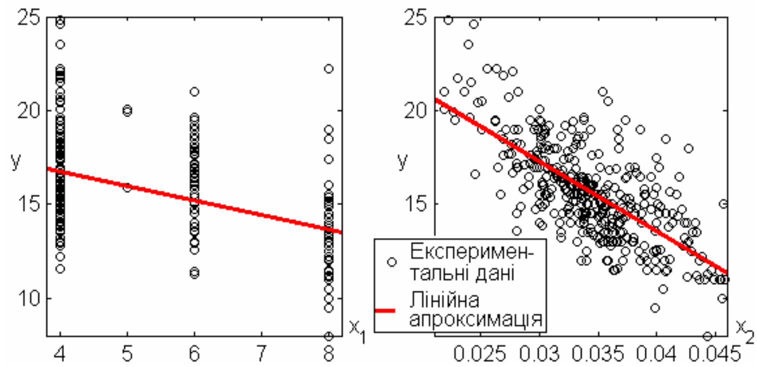

Рисунок 6 – Експериментальні дані з прикладу 2

Як функції належностей нечітких термів цієї бази знань оберемо гаусові криві (рис. 7а). При навчанні настроювалися параметри функцій належності: коефіцієнти концентрації та координати максимумів. Функції належності після навчання зображені на рис. 7б. Тестування (рис. 8) свідчить, що навчання значно покращило точність нечіткої моделі. В результаті навчання нев'язка (2) на нечіткій тестовій вибірці зменшилась з 0.4338 до 0.2194. Зауважимо, що навчання за нечіткою вибіркою зменшило нев'язку і на чіткій тестовій вибірці з RMSE = 1.8883 до RMSE = 1.8041. Для отримання чітких значень на виході нечіткої моделі застосовано дефаззіфікацію за центром тяжіння.

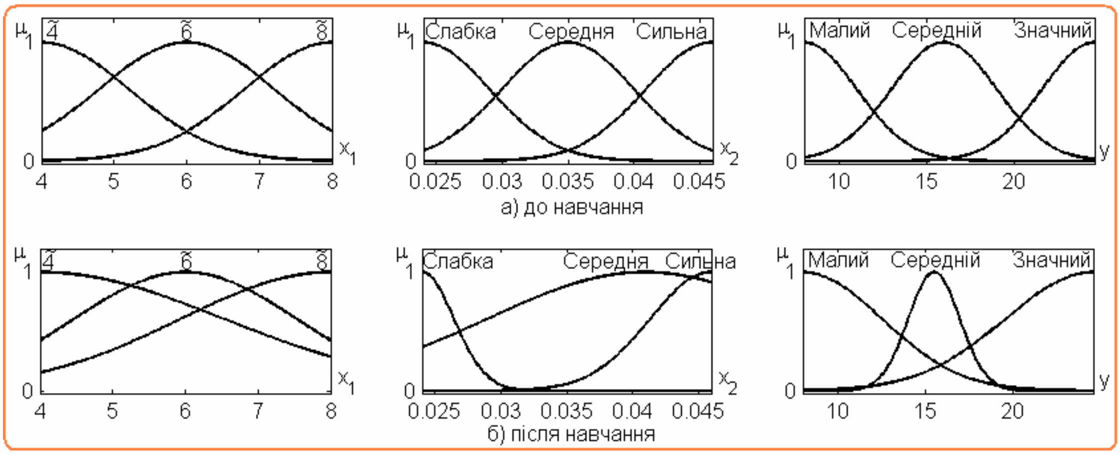

Рисунок 7 – Функції належності нечітких термів з прикладу 2

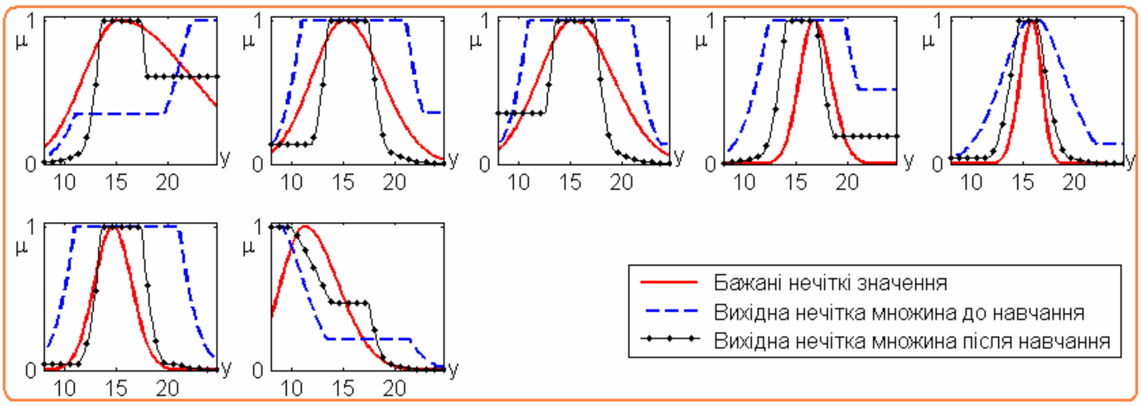

Рисунок 8 – Перевірка нечітких моделей з прикладу 2 на нечіткій тестовій вибірці

Навчання нечіткої бази знань з декількома виходами

Другий спосіб синтезу нечітких чисел розроблено за таких припущень: 1) на усьому факторному просторі шукане нечітке число задається параметричними функціями належності одного типу; 2) залежність параметрів функції належності цього нечіткого числа від факторів впливу описується нечіткою базою знань. Введемо такі позначення:

$\mu(y) = mf(y, Z)$ – модель функції належності з параметрами Z, за якою для елементів у універсальної множини розраховують ступені належності $\mu(y)$;

$Z = f(X, P)$ – нечітка модель з параметрами P, що пов'язує параметри Z функції належності mf з факторами впливу X, тобто $\mu(y) = mf\big(y, f(X, P)\big)$.

Залежність $Z = f(X, P)$ опишемо нечіткою базою знань з декількома вихідними змінними. Кожна вихідна змінна відповідатиме одному з параметрів функції належності шуканого нечіткого числа. Базу знань можна записати в будь-якому з форматів, за якими в результаті логічного виведення отримують чіткі числа. В табл. 2 наведені змістовні інтерпретації вихідних змінних, при яких правила бази знань будуть зрозумілі користувачам. З табл. 2 видно, що для отримання асиметричної функції належності шуканого нечіткого числа база знань повинна мати щонайменше 3 вихідні змінні.

Таблиця 2 – Інтерпретація вихідних змінних нечіткої бази знань, за якою розраховуються параметри функції належності нечіткого числа

Вихідні змінні	Змістовна інтерпретація	Відповідність параметрам функцій належності з табл. 1
y_1, y_2	y_1 – координата максимуму; y_2 – концентрованість.	Для гаусової: $b = y_1$; $c = y_2$. Для трикутної: $b = y_1$; $c = y_1 + y_2$; $a = y_1 - y_2$.
y_1, y_2, y_3	y_1 – координата максимуму; y_2 та y_3 – концентрованість лівої та правої гілок графіка функції належності.	Для трикутної: $a = y_1 - y_2$; $b = y_1$; $c = y_1 + y_3$. Для двохсторонньої гаусової: $b_1 = y_1$; $b_2 = y_1$; $c_1 = y_2$; $c_2 = y_3$.

Приклад 3. За даними прикладу 2 побудуємо нечітку модель з двома виходами, які визначають параметри гаусової функції належності нечіткого часу виконання операції «розігнання автомобіля до швидкості 60 миль за годину».

Гаусова функція належності має 2 параметри. Відповідно, нечітка модель матиме 2 виходи – координату максимуму (b) та коефіцієнт концентрації (c) функції належності нечіткого числа $\tilde{y}$. Початкові функції належності нечітких термів зображені на рис. 7а. База знань типу Сугено цієї моделі сформуємо з 3 правил (табл. 3). Результати її тестування наведені в табл. 3 та на рис. 9.

Таблиця 3 – Нечітка база знань з прикладу 3 до навчання

ЯКЩО		ТОДІ		RMSE на тестовій вибірці
x_1	x_2	b	c	
$\tilde{4}$	Слабка	24.6	1	0.4284 – на нечіткій вибірці 2.8615 – на чіткій вибірці
$\tilde{6}$	Середня	16	1	
$\tilde{8}$	Сильна	8	1	

Функції належності після навчання наведені на рис. 10, а база знань – в табл. 4. Від'ємні коефіцієнти в консеквенті 2-го правила вказують, що при навчанні виявлено тенденцію зменшення часу розгону при зростанні кількості циліндрів та збільшенні питомої потужності автомобіля, що відповідає розподілу експериментальних даних з рис. 6. Результати тестування (табл. 4 та рис. 9) свідчать, що після навчання запропонована модель краще описує залежність «входи – вихід», ніж модель з прикладу 2. Як приклад на рис. 10 зображено залежність нечіткого часу розігнання від питомої потужності чотирициліндрового автомобіля, яка відповідає настроєній нечіткій моделі. Видно, що результати моделювання добре збігаються з експериментальним розподілом.

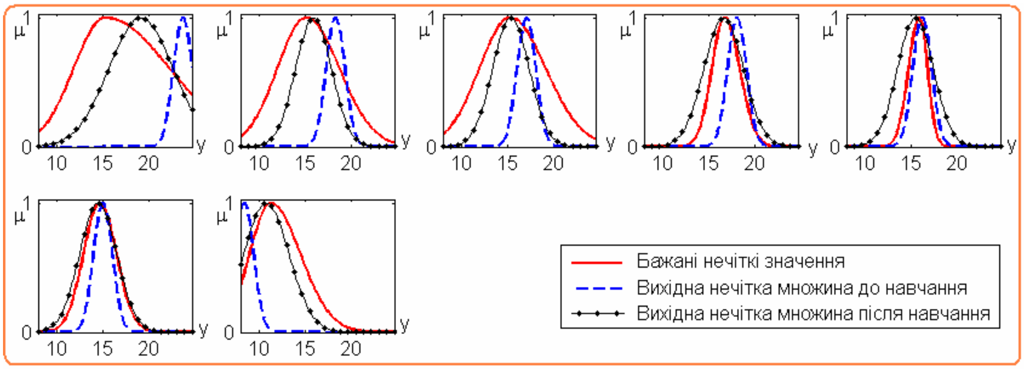

Рисунок 9 – Перевірка нечітких моделей з прикладу 3 на нечіткій тестовій вибірці

Таблиця 4 – Нечітка база знань з прикладу 3 після навчання

ЯКЩО		ТОДІ		RMSE на тестовій вибірці
x_1	x_2	b	c	
$\tilde{4}$	Слабка	24.6	2.81	0.2158 – на нечіткій вибірці 1.5635 – на чіткій вибірці
$\tilde{6}$	Середня	$17.73 - 0.23x_1 - 0.03x_2$	1.84	
$\tilde{8}$	Сильна	8	7.23	

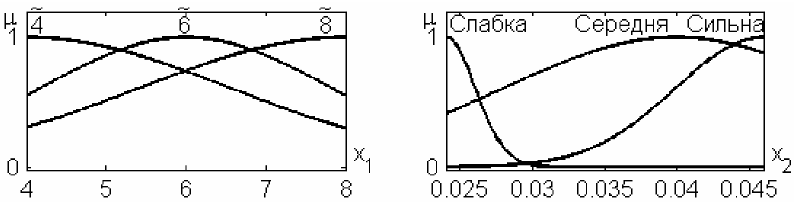

Рисунок 10 – Оптимізовані функції належності нечітких термів з прикладу 3

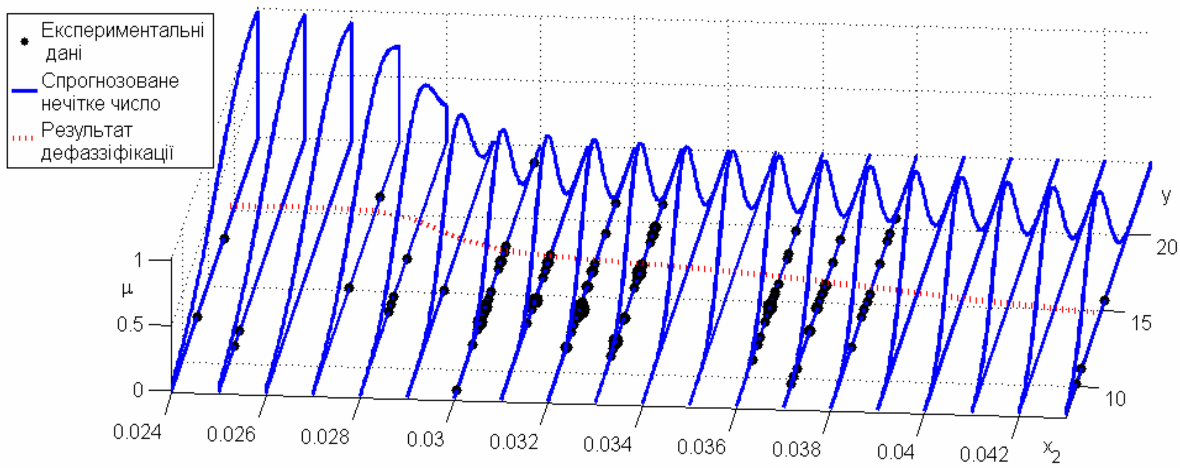

Рисунок 11 – До прикладу 3: ідентифікована залежність нечіткого часу ($\widetilde{y}$) розігнання автомобіля з чотирма циліндрами ($x_1 = 4$) від питомої потужності (x_2)

Висновки

Запропоновано 2 способи побудови багатофакторних моделей на нечітких базах знань, за якими результат логічного висновку являє собою нечітке число. Перший спосіб полягає у вилученні дефаззіфікації з типової нечіткої моделі з подальшою апроксимацією вихідної нечіткої множини параметричною функцією належності. Другий спосіб принципово новий – за ним нечітка оцінка вихідної змінної описується функцією належності, параметри якої залежать від факторів. Вплив факторів на параметри функції належності задається нечіткою базою знань.

Запропонована постановка задачі навчання таких нечітких баз знань за нечіткою вибіркою даних, вказано методи її розв'язання та наведені відповідні приклади. Відмінність запропонованої постановки задачі навчання від інших полягає в тому, що вперше вводиться навчальна вибірка з нечіткими значеннями вихідної змінної. Проведені в роботі обчислювальні експерименти свідчать, що навчання за нечіткими даними покращує точність моделювання як на чіткій, так і на нечіткій тестовій вибірках. Завдяки можливості використання нечіткої навчальної вибірки запропонований підхід може бути корисним для ідентифікації нелінійних залежностей «входи – вихід» в задачах, де експериментальні дані задані нечіткими числами.

Література

1. Ротштейн А.П., Катeльников Д.И. Идентификация нелинейных зависимостей нечеткими базами знаний // Кибернетика и системный анализ. – 1998. – № 5. – С. 53-61.
2. Ротштейн А.П. Интеллектуальные технологии идентификации: нечеткая логика, генетические алгоритмы, нейронные сети. – Винница: УНІВЕРСУМ–Вінниця, 1999. – 320 с.
3. Takagi T., Sugeno M. Fuzzy Identification of Systems and Its Applications to Modeling and Control // IEEE Trans. on Systems, Man, and Cybernetics. – Vol. 15, № 1. – 1985. – P. 116-132.
4. Jang J.–S. R. ANFIS: Adaptive–Network–Based Fuzzy Inference System // IEEE Trans. Systems & Cybernetics. – 1993. – Vol. 23. – P. 665-685.
5. Yager R., Filev D. Essentials of Fuzzy Modeling and Control. USA: John Wiley & Sons, 1994. – 387 p.
6. Ротштейн А.П., Штовба С.Д. Влияние методов дефаззификации на скорость настройки нечеткой модели // Кибернетика и системный анализ. – 2002. – № 5. – С. 169-176.

7. Штовба С.Д. Идентификация нелинейных зависимостей с помощью нечеткого логического вывода в системе MATLAB // Exponenta Pro: Математика в приложениях. – 2003. – № 2. – С. 9-15.
8. Shtovba S., Pankevich O., Dounias G. Tuning the Fuzzy Classification Models with Various Learning Criteria: the Case of Credit Data Classification // Proc. of Inter. Conference on Fuzzy Sets and Soft Computing in Economics and Finance. – Russia, St. Petersburg. – 2004. – Vol. 1. – P. 103-110.
9. Ротштейн А.П., Штовба С.Д. Идентификация нелинейной зависимости нечеткой базой знаний с нечеткой обучающей выборкой // Кибернетика и системный анализ. – 2006. – № 2. – С. 17-24.
10. MPG data base of UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html

С.Д. Штовба

Обучение нечеткой базы знаний по выборке нечетких данных

Статья исследует обучение нечеткой базы знаний по выборке с нечеткими значениями выходной переменной. Предложено 2 способа построения многофакторных моделей на нечетких базах знаний, на выходе которых получают нечеткие числа. Предложена постановка задачи обучения таких нечетких баз знаний по нечеткой выборке данных, указаны методы ее решения и приведены соответствующие примеры. Выполненные вычислительные эксперименты показывают, что обучение по нечетким данным улучшает точность моделирования как на четкой, так и на нечеткой тестовых выборках.

Serhiy Shtovba

Fuzzy Knowledge Base Learning on Fuzzy Training Set

The article devotes to fuzzy knowledge base learning on training set with fuzzy output. Two ways for building of fuzzy rule-based multifactor model with fuzzy output are proposed. The article proposes a problem statement of the fuzzy knowledge base learning on the fuzzy training set. Examples of fuzzy learning are included into the article. Executed computational experiments show that fuzzy training set learning improves the modeling accuracy as on crisp test set as on fuzzy test set.

Стаття надійшла до редакції 19.06.2006.