МЕТОДИЧНІ ВКАЗІВКИ

до самостійної роботи з дисципліни

"Дискретна математика"

для студентів спеціальності 125 «Кібербезпека»
спеціалізації «Безпека інформаційних і комунікаційних систем»

Міністерство освіти і науки України
Вінницький національний технічний університет
Факультет інформаційних технологій та комп'ютерної інженерії
Кафедра захисту інформації

МЕТОДИЧНІ ВКАЗІВКИ

до самостійної роботи з дисципліни

"ДИСКРЕТНА МАТЕМАТИКА"

для студентів спеціальності 125 «Кібербезпека»
спеціалізації «Безпека інформаційних і комунікаційних систем»

Вінниця
ВНТУ

2018

Рекомендовано до друку М[illegible]о радою Вінницького національного технічного університету Міністерства освіти і науки України (Протокол № від 2018 р.)

Рецензенти:

О. М. Ткаченко, кандидат технічних наук, доцент
О. П. Войтович, кандидат технічних наук, доцент

Методичні вказівки до самостійної роботи з дисципліни «Дискретна математика» / Уклад. Н. Р. Кондратенко, О. О. Снігур. – Вінниця: ВНТУ, 2018. – 60 с.

Методичні вказівки призначені для надання допомоги при виконанні самостійної роботи з дисципліни "Дискретна математика", дано рекомендації щодо коректного розв’язання задач з самостійної роботи.

ЗМІСТ

Вступ

Основи теорії нечітких множин і нечіткої логіки були закладені наприкінці 1960-х років у працях американського математика Лотфі Заде. Його праця “Fuzzy Sets”, опублікована у 1965 р. в журналі “Information and Control”, стала поштовхом до розвитку нової математичної теорії. Він дав назву і новій галузі науки – “fuzzy sets” (fuzzy – нечіткий, розмитий). Основною причиною появи нової теорії стали нечіткі і наближені міркування, що використовувались для опису людиною процесів, систем, об'єктів. Математична теорія нечітких множин (fuzzy sets) і нечітка логіка (fuzzy logic) є узагальненнями класичної теорії множин і класичної формальної логіки.

Основною характеристикою нечіткої множини є її функція належності, яка ставить у відповідність кожному елементу універсальної множини число з інтервалу [0, 1], що означає ступінь належності. Поняття функції належності є узагальненням поняття характеристичної функції чіткої множини, яка оперує значеннями {0, 1}. Тому основні властивості та операції над нечіткими множинами являють собою узагальнення відповідних властивостей та операцій класичної теорії множин.

Подальше узагальнення поняття функції належності привело до появи нечітких множин типу 2 та множин вищих порядків. Узагальнена нечітка множина визначається функціями належності, в ролі значень яких також виступають нечіткі множини. Проте побудова моделей на основі узагальнених нечітких множин пов’язана зі значною обчислювальною складністю, тому на практиці використовується їх інтервальне подання. Апарат інтервальних нечітких множин оперує лише крайніми точками

інтервалу зміни значення функції належності, і не враховує особливостей розподілу, що виникає в межах цього інтервалу. Таке спрощення значно знижує кількість обчислювальних ресурсів, необхідних для побудови нечіткого логічного висновку; при цьому на якості функціонування системи це майже не відбивається.

У США розвиток нечіткої логіки йде шляхом створення систем, що потрібні великому бізнесу і військовим. Нечітка логіка застосовується при аналізі нових ринків, біржовій грі, оцінці політичних рейтингів, виборі оптимальної цінової стратегії, оцінці рівня зрілості процесів захисту інформації і т. ін. З'явилися і комерційні системи масового застосування.

Що стосується вітчизняного ринку комерційних систем на основі нечіткої логіки, то його формування почалося в середині 1995 року. Найбільш популярні в замовників такі пакети:

- CubiCalc 2.0 RTC - одна з найбільш могутніх комерційних експертних систем на основі нечіткої логіки, що дозволяє створювати власні прикладні експертні системи;

- CubiQuick - дешева «університетська» версія пакету CubiCalc;

- RuleMaker - програма автоматичного витягу нечітких правил із вхідних даних;

- FuziCalc - електронна таблиця з нечіткими полями, що дозволяє робити швидкі оцінки при неточно відомих даних без нагромадження похибки;

- OWL - пакет, що містить вихідні тексти усіх відомих видів нейронних мереж, нечіткої асоціативної пам'яті і т.д.

Основними споживачами нечіткої логіки на ринку є банкіри та фінансисти, а також фахівці в області політичного й економічного аналізу. Вони використовують CubiCalc для створення моделей різних економічних, політичних, біржових ситуацій. Що ж стосується легкого в освоєнні пакета FuziCalc, то він зайняв своє місце на комп'ютерах великих банкірів і фахівців з надзвичайних ситуацій - тобто тих, для кого найбільше важлива швидкість проведення розрахунків в умовах неповноти і неточності вхідної інформації. Однак можна з упевненістю сказати, що епоха розквіту прикладного використання нечіткої логіки на вітчизняному ринку ще попереду.

Сьогодні елементи нечіткої логіки можна знайти в десятках промислових виробів - від систем керування електропоїздами і бойовими вертольотами до пилососів і пральних машин. Рекламні кампанії багатьох фірм (переважно японських) підносять успіхи у використанні нечіткої логіки як особливу конкурентну перевагу. Без застосування нечіткої логіки немислимі сучасні ситуаційні центри керівників західних країн, у яких приймаються ключові політичні рішення і моделюються всілякі кризові ситуації. Одним із вражаючих прикладів масштабного застосування нечіткої логіки стало комплексне моделювання системи охорони здоров'я і соціального забезпечення Великої Британії (National Health Service - NHS), що вперше дозволило точно оцінити й оптимізувати витрати на соціальні програми.

Серед лідерів нового ринку виділяється американська компанія Hyper Logic, заснована в 1987 році Фредом Уоткінсом (Fred Watkins). Спочатку компанія спеціалізувалася на нейронних мережах, однак незабаром цілком сконцентрувалася на нечіткій логіці. Недавно вийшла на ринок друга версія

пакета CubiCalc фірми HyperLogic, яка є однієї з найбільш могутніх експертних систем на основі нечіткої логіки. Пакет містить інтерактивну оболонку для розробки нечітких експертних систем і систем керування, а також run-time модуль, що дозволяє оформляти створені користувачем системи у виді окремих програм.Крім Hyper Logic серед "патріархів" нечіткої логіки можна також назвати такі фірми як IntelligenceWare, InfraLogic, Aptronix. Усього ж на світовому ринку представлено більш 100 пакетів, які тим чи іншим видом використовують нечітку логіку. У трьох десятках СУБД реалізована функція нечіткого пошуку. Власні програми на основі нечіткої логіки анонсували такі гіганти як IBM, Oracle та інші.

1 Нечіткі множини типу 1

1.1 Основні відомості. Операції над нечіткими множинами типу 1

Завданням нечітких множин є визначення належності деякого об'єкта чи елемента до заданої множини. Нехай E – деяка множина, а A – підмножина E, тобто $A \subset E$. Той факт, що елемент x множини E належить і множині A в теорії множин позначають так: $x \in A$. Щоб виразити цю належність, можна скористатися й іншим поняттям – характеристичною функцією $\mu_A(x)$, значення якої вказують, чи є (так або ні) x елементом A:

$$\mu_A(x) = \begin{cases} 1, \text{ якщо } x \in A, \\ 0, \text{ якщо } x \notin A. \end{cases}$$

Згідно з теорією нечітких множин, характеристична функція належності може набувати будь-якого значення в інтервалі *[0, 1]*, а не тільки два – 0 або 1. Відповідно до цього, елемент x_i множини E може не належати A ($\mu_A(x) = 0$), бути елементом A невеликою мірою (значення $\mu_A(x)$ близьке до нуля), бути елементом A значною мірою ($\mu_A(x)$ близьке до *1*) або бути елементом A ($\mu_A(x)=1$). Отже, поняття належності узагальнюється. Нечітку підмножину A універсальної множини E позначають $A_н$ і визначають упорядкованими парами:

$$A_н = \{(x, \mu_A(x)) \mid x \in E\}.$$

Характеристична функція належності (або просто функція належності) $\mu_A(x)$ набуває значень у деякій упорядкованій множині M (наприклад, $M = [0, 1]$). Ця функція належності вказує ступінь (або рівень) належності елемента x до підмножини A. Множину M називають множиною належностей. Якщо $M = \{0, 1\}$, то нечітку підмножину A можна розглядати як звичайну або чітку множину, функція належності якої набуває лише бінарних значень.

Тому основні операції над нечіткими множинами також являють собою узагальнення відповідних властивостей та операцій класичної теорії множин.

Рівність. Дві нечіткі множини $\tilde{A}$ і $\tilde{B}$, що задані на U, є рівними, якщо вони складаються з одних і тих же елементів для всіх $x \in U$ та $\mu_A(x) = \mu_B(x)$. Позначаються як $\tilde{A} = \tilde{B}$.

Доповнення. Дві нечіткі множини $\tilde{A}$ і $\tilde{B}$, доповнюють одна одну, якщо $\mu_A(x) = 1 - \mu_B(x)$ для всіх $x \in U$.

Перетин. Перетином двох нечітких множин $\tilde{A}$ і $\tilde{B}$ є така множина $\tilde{D} = \tilde{A} \cap \tilde{B}$, що складається з елементів для всіх $x \in U$, що одночасно належать $\tilde{A}$ і $\tilde{B}$ з функцією належності $\mu_D(x) = min(\mu_A(x), \mu_B(x))$.

Об'єднання. Об'єднанням двох нечітких множин $\tilde{A}$ і $\tilde{B}$ є така множина $\tilde{D} = \tilde{A} \cup \tilde{B}$, що складається з елементів для всіх $x \in U$, що належать $\tilde{A}$ або $\tilde{B}$ з функцією належності $\mu_D(x) = max(\mu_A(x), \mu_B(x))$.

Різниця. Різницею двох нечітких множин $\tilde{A}$ і $\tilde{B}$ є така множина $\tilde{D} = \tilde{A} - \tilde{B}$, що складається з елементів для всіх $x \in U$, що мають функцію належності $\mu_D(x) = \mu_A(x) - \mu_B(x)$, якщо $\mu_A(x) > \mu_B(x)$, інакше $\mu_D(x) = 0$.

На рис. 1.1 графічно показано основні операції над нечіткими множинами: а) $\tilde{A} \subset \tilde{B}$, б) $\tilde{A} = 1 - \tilde{B}$, в) $\tilde{A} \cap \tilde{B}$, г) $\tilde{A} \cup \tilde{B}$.

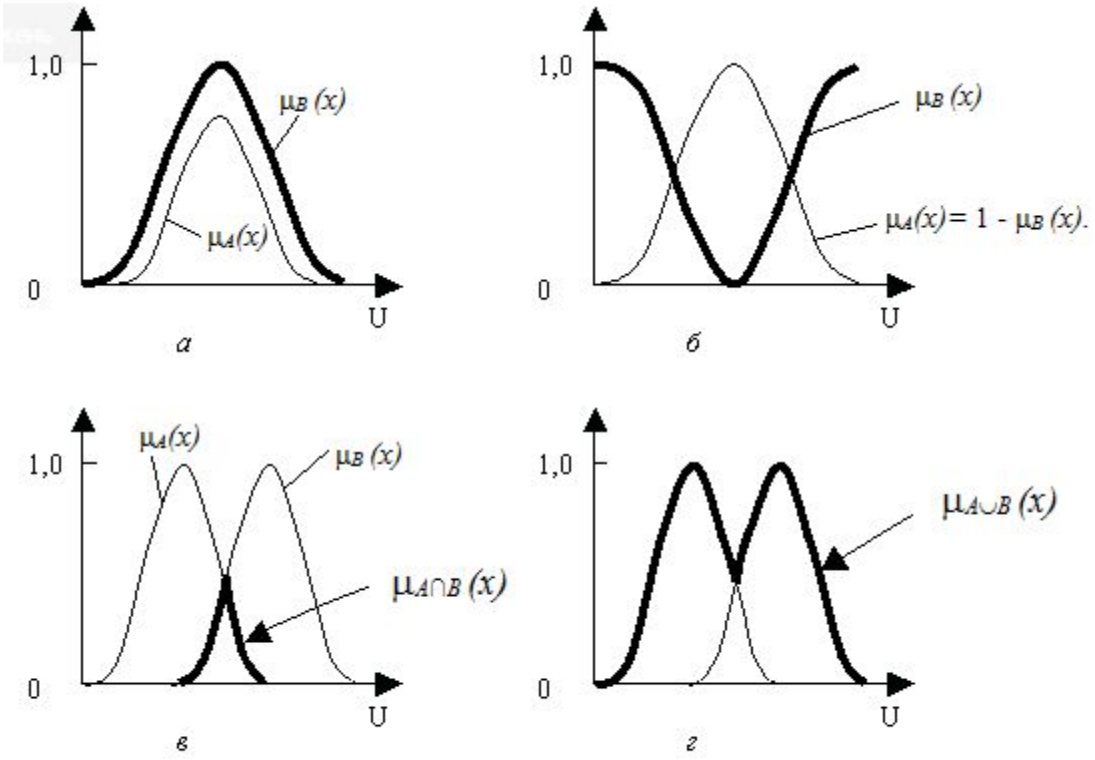


Рисунок 1.1 – Операції над нечіткими множинами

Декартовим добутком $\tilde{A} \times \tilde{B}$ двох нечітких множин $\tilde{A}$ і $\tilde{B}$, визначених відповідно на універсумах U_1 та U_2, є нечітка множина $\tilde{D}$ пар (кортежів), визначених на U_1 та U_2 з функцією належності $\mu_D(x) = min(\mu_A(x),\ \mu_B(x))$. Тобто $\tilde{D} = \tilde{A} \times \tilde{B} = \{(min\,(\mu_A(x),\ \mu_B(x)),\ (a,b)) :\ a \in U_1,\ b \in U_2\}$.

Види функцій належності. Існує багато кривих для визначення функцій належності. Найбільш розповсюдженими функціями належності є трикутна, трапецієвидна, функція Гаусса та *Z*- (або *S*-) подібні.

Трикутна функція належності визначається трійкою чисел *(a, b, c)*, а її значення в довільній точці *x* обчислюється за формулою

$$\mu(x) = \begin{cases} \dfrac{x-a}{b-a}, & a \le x \le b \\ \dfrac{c-x}{c-b}, & b \le x \le c \\ 0, & \textit{в інших випадках} \end{cases}$$

При *(b-a)=(c-b)* маємо симетричну трикутну функцію належності, яка однозначно задається двома параметрами з трійки *(a, b, c)*.

Для визначення *трапецієвидної функції належності* потрібні чотири числа *(a, b, c, d)*, а її значення в заданій точці *x* обчислюється за формулою

$$\mu(x) = \begin{cases} \dfrac{x-a}{b-a}, & a \le x \le b \\ 1, & b \le x \le c \\ \dfrac{d-x}{d-c}, & c \le x \le d \\ 0, & \textit{в інших випадках} \end{cases}$$

При *(b-a)=(d-c)* ця функція належності приймає симетричний вигляд.

Функція належності Гауса (рис. 1.2), зазвичай описується формулою $\mu(x) = e^{-\frac{(x-c)^2}{2\sigma^2}}$ та визначається параметрами *(c,* σ*)*.

Z- та *S*-подібні функції належності одержали свою назву в зв'язку з виглядом кривих, які зображують їхні графіки (рис. 1.3).

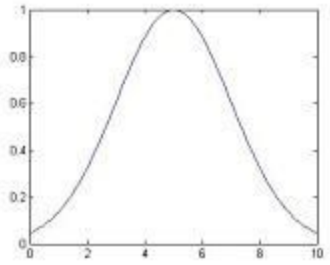
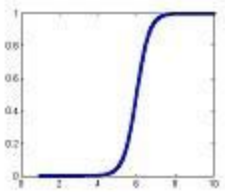
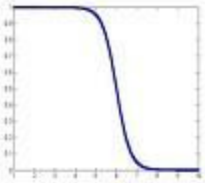


Рисунок 1.2 – Графік функції належності Гауса при $c=5,\ \sigma=2$



а)



б)

Рисунок 1.3 - Графіки сигмоїдальної функції належності при (а) $a=3,\ b=6$; (б) $a=-3,\ b=6$

До типу *Z*- і *S*-подібних функцій належить сигмоїдальна функція належності, яка в загальному випадку описується формулою $\mu(x)=\frac{1}{1+e^{-a(x-b)}}$, де $a,\ b$ – деякі числові параметри, такі, що $a < b$. При цьому, якщо $a > 0$, маємо *S*-подібну функцію належності, якщо $a < 0$ – *Z*-подібну.

1.2 Нечіткі логічні системи типу 1

Нечіткі логічні системи часто виступають як основна складова систем підтримки прийняття рішень. Задачу прийняття рішень в загальному випадку можна розглядати як задачу ідентифікації, що володіє такими властивостями [16]:

- для прийняття рішення необхідно встановити залежність між вхідними та вихідною змінною;
- вихідна змінна асоціюється з об'єктом ідентифікації, тобто з видом рішення, що приймається;
- вхідні змінні асоціюються з параметрами стану об'єкта ідентифікації;
- вихідна і вхідні змінні можуть мати кількісні та якісні оцінки;
- структура взаємозв'язку між вихідною та вхідними змінними описується правилами ЯКЩО <входи>, ТОДІ <вихід>. Правила використовують якісні (лінгвістичні) оцінки змінних та в своїй сукупності складають нечітку базу знань.

У розробці моделей та методів ідентифікації багатовимірних залежностей на основі нечітких баз знань використовується ряд принципів лінгвістичного моделювання.

Принцип лінгвістичності вхідних та вихідних змінних передбачає, що входи об'єкта та його вихід розглядаються як лінгвістичні змінні, які оцінюються якісними термами. Лінгвістичною називається така змінна, значеннями якої є слова або висловлювання природної мови, тобто якісні терми [8]. Використовуючи поняття функції належності, кожен із термів, що оцінюють лінгвістичну змінну, можна формалізувати у вигляді нечіткої множини, заданої на відповідній універсальній множині.

Прицип формування структури залежності «вхід-вихід» у вигляді нечіткої бази знань. Нечітка база знань являє собою сукупність правил ЯКЩО <входи>, ТОДІ <вихід>, які відображають досвід експерта та його розуміння причинно-наслідкових зв’язків у розглядуваній задачі прийняття рішень. Особливість подібних висловлювань полягає в тому, що їхня адекватність не змінюється при незначних коливаннях умов експерименту. Тому формування нечіткої бази знань можна трактувати як аналог етапу структурної ідентифікації, на якому будується наближена модель об'єкта з параметрами, що вимагають подальшого налаштування. У випадку лінгвістичних систем налаштуванню підлягають форми функцій належності нечітких термів, за допомогою яких оцінюються входи та виходи об'єкта.

Крім того, сукупність правил ЯКЩО-ТО можна розглядати як набір експертних точок в просторі входів та виходів. Застосування апарату нечіткого логічного висновку дозволяє за цими точками відновити багатовимірну поверхню, що дозволяє отримувати значення виходу за різних комбінацій значень вхідних змінних.

Принцип ієрархічності баз знань полягає в рівневому поданні експертних знань. В умовах роботи з багатовимірними даними побудова системи висловлювань, що описують залежність між входами та виходом вистеми, пов’язана з труднощами. З огляду на особливості людської пам’яті не рекомендується будувати логічні висловлювання, що складаються більш як з 9 ознак. Натомість пропонується виконати класифікацію вхідних ознак та на її основі побудувати дерево, що визначає систему вкладених висловлювань меншої розмірності. Можливі також інші шляхи подолання «прокляття розмірності», такі як оптимізація простору вхідних ознак. Крім

того, в деяких застосуваннях кваліфікація експерта дозволяє оперувати більшою кількістю вхідних ознак одночасно, але ця кількість також обмежена.

Принцип двохетапного налаштування нечітких баз знань передбачає побудову моделі об’єкта в два етапи, які відповідають етапам структурної та параметричної ідентифікації (рис. 1.4). Налаштуванню підлягають ваги правил та параметри функцій належності. На першому етапі здійснюється формування та початкове налаштування моделі об’єкта шляхом побудови бази знань за доступною експерту інформацією. Для початкового налаштування ваг правил та форм функцій належності застосовується метод парних порівнянь Сааті. Чим вищий професійний рівень експерта, тим вища адекватність нечіткої моделі, побудованої на етапі початкового налаштування.

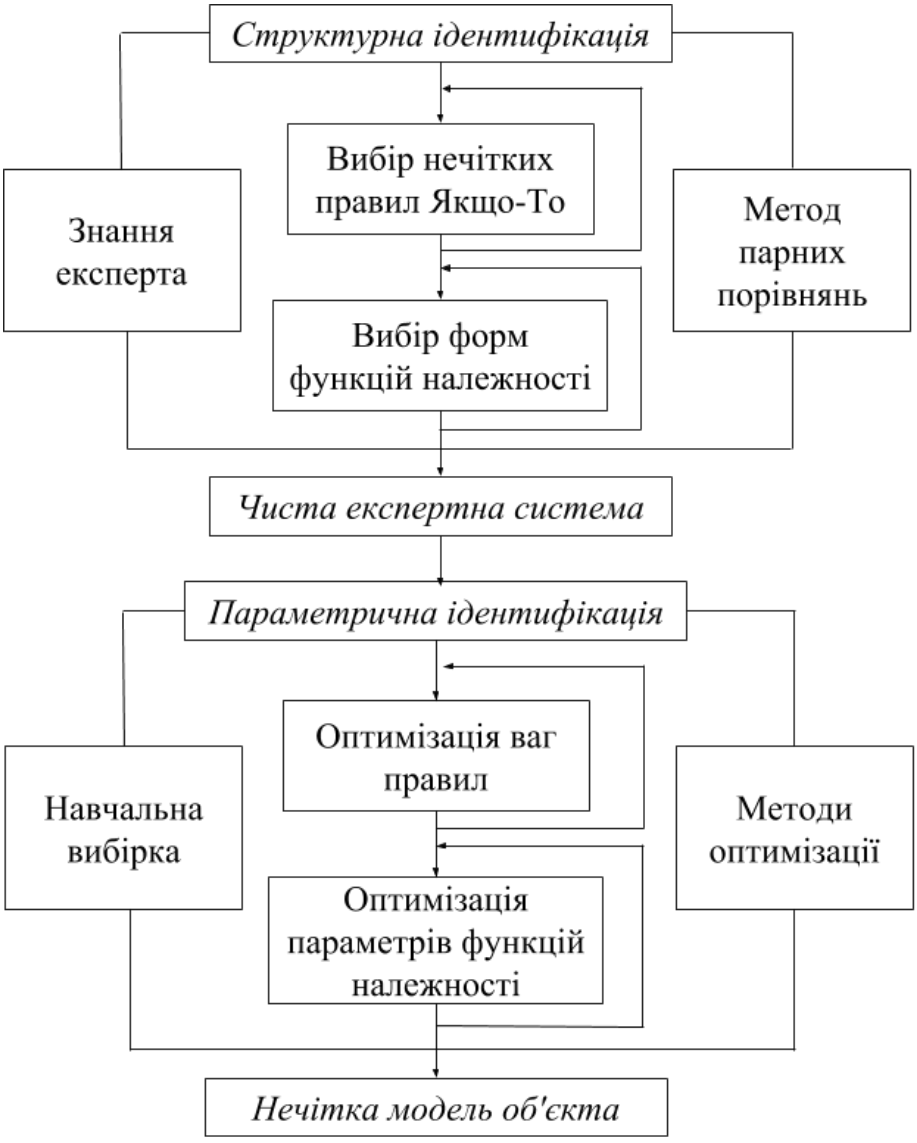


Рисунок 1.4 – Етапи побудови нечіткої логічної системи

Проте результати побудови нечіткого логічного висновку такою системою, побудованою суто на теоретичних знаннях, можуть не узгоджуватись із експериментальними даними. Тому необхідний другий етап, на якому відбувається тонке налаштування нечіткої моделі шляхом її навчання на експериментальних даних. Суть етапу тонкого налаштування полягає в підборі таких ваг нечітких правил ЯКЩО-ТО і таких параметрів

функцій належності, які мінімізують розходження між бажаним (експериментальним) і модельним (теоретичним) виходом об'єкта. Етап тонкого налаштування формулюється як задача нелінійної оптимізації, яка може вирішуватися різними методами, серед яких метод найшвидшого спуску та генетичні алгоритми оптимізації.

Основні компоненти нечіткої логічної системи зображено на рис. 1.5.

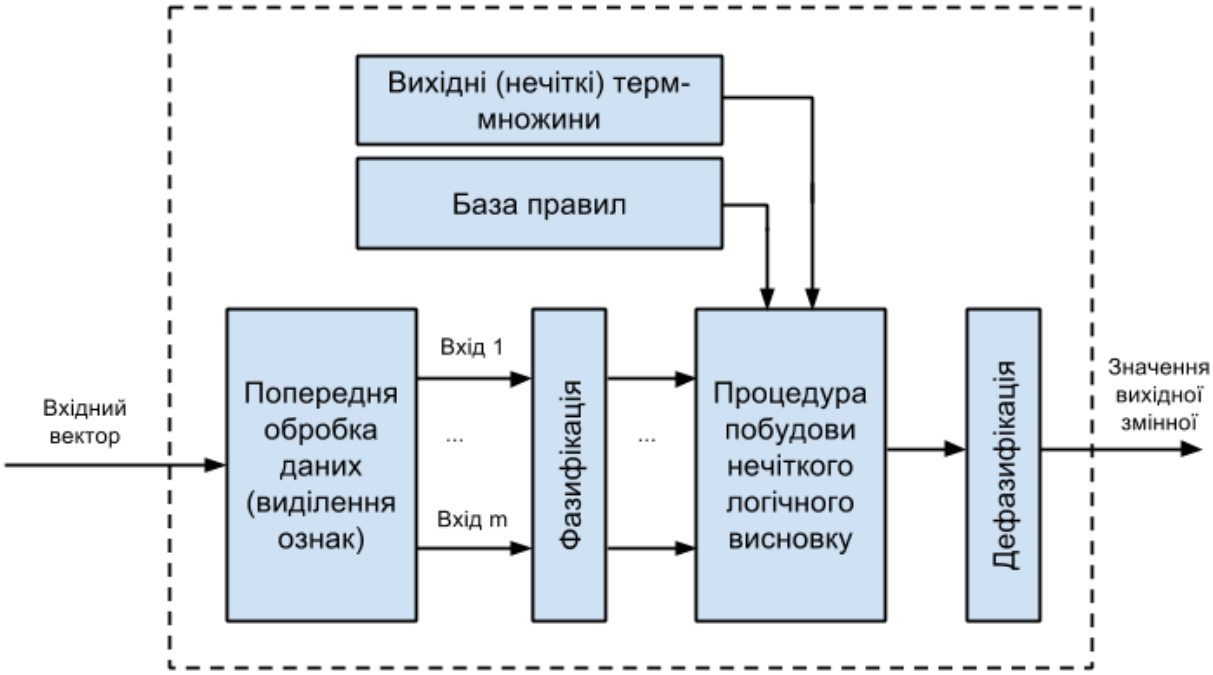


Рисунок 1.5 – Нечітка логічна система класифікації даних

Перед надходженням на вхід нечіткої логічної системи дані можуть піддаватися попередній обробці, наприклад можливе виключення частини ознак із розгляду експертом із предметної галузі.

Практично будь-яка нечітка логічна система має блоки фазифікації, побудови нечіткого логічного висновку, дефазифікації, а також базу знань. Вхідний вектор являє собою набір значень вхідних лінгвістичних змінних

$X^i = \{x_1^i, ..., x_m^i\}$. *Лінгвістична змінна* – це множина нечітких змінних, вона використовується для того, щоб дати словесний опис нечіткому числу, отриманому в результаті деяких операцій.

Лінгвістична змінна визначається як *<x, L, U, G, M>*, де *x* – найменування змінної, *L* – множина її значень (базова терм-множина), що складається з найменувань нечітких змінних, областю визначення кожної з яких є множина *U*; *G* – синтаксична процедура (граматика), що дозволяє оперувати елементами терм-множини *L*, зокрема генерувати нові осмислені терми; *L′=L ∪ G(L)* задає розширену терм-множину (∪ – знак об'єднання); *M* – семантична процедура, що дозволяє приписати кожному новому значенню лінгвістичної змінної нечітку семантику, шляхом формування нової нечіткої множини.

Терм-множина – це множина всіх можливих значень лінгвістичної змінної.

Терм – будь-який елемент терм-множини. В теорії нечітких множин терм формалізується нечіткою множиною за допомогою функції належності.

Наприклад, змінна «швидкість автомобіля» може набувати значень «низька», «середня», «висока» і «дуже висока». В цьому випадку лінгвістичною змінною є «швидкість автомобіля», термами – лінгвістичні оцінки «низька», «середня», «висока» і «дуже висока», які і складають терм-множину.

Нечіткий терм – це нечітка множина, яка має властивість, якій відповідає певне поняття.

Фазифікація – це процедура перетворення чітких даних у нечітку множину. Процедура фазифікації полягає в формалізації даних як

лінгвістичних змінних, введенні всіх можливих термів, що їх характеризують, та побудові функцій належності для кожного терма.

Дефазифікацією (в нечіткій логіці) називається процедура перетворення нечіткої множини в чітке число. В теорії нечітких множин процедура дефазифікації аналогічна знаходженню математичного сподівання, моди або медіани випадкових величин в теорії ймовірностей. Наприклад, таким числом може стати максимум функції належності, центр мас функції належності, або щось інше.

Нечіткою базою знань називається сукупність нечітких правил вигляду "ЯКЩО – ТО", що визначають взаємозв'язок між входами і виходами досліджуваного об'єкта. Узагальнений формат нечітких правил такий: – "ЯКЩО <антецедент правила>, ТО <консеквент правила>".

Антецедент правила являє собою твердження вигляду "*x* є *низький*", де "*низький*" – це терм (лінгвістичне значення), заданий нечіткою множиною на універсальній множині лінгвістичної змінної *x*. Квантифікатори "дуже", "більш-менш", "не", "майже" і т.п. можуть використовуватися для модифікації термів антецедента.

У випадку, коли база знань будується на основі експериментальних даних, кожна пара (X_i, Y_i), де вхідному вектору поставлено у відповідність лінгвістичну оцінку значення вихідної змінної *Y*, дану експертом, генерує одне правило. Антецеденти правил утворюються заміною значення x_j^i відповідним йому нечітким термом $A_{x_j}^i$, консеквентами є терм лінгвістичної змінної *y*, визначений експертом для вектора X^i:

$$R^i : IF\ x_1 \in A^i_{x_1} \wedge x_2 \in A^i_{x_2} \wedge ... \wedge x_m \in A^i_{x_m}\ THEN\ y \in L^k_y \in \left\{ L_1, ..., L_p \right\},$$

де x_i – вхідні змінні, y – вихідна змінна, $L_y \in \left\{ L_1, ..., L_p \right\}$ – терм-множини вихідної змінної. Терми вхідних та вихідної змінної описуються гаусовими функціями належності вигляду

$$\mu(x_i^j) = e^{-\left(\frac{x - b_i^j}{c_i^j}\right)^2}$$

Значення математичного сподівання b_i^j при цьому приймаються рівними експериментальним значенням, середньоквадратичне відхилення c_i^j вибирається довільним чином.

У разі виявлення невідповідності результатів роботи системи, побудованої на емпіричних значеннях параметрів функцій належності, очікуваним значенням можна додатково провести оптимізацію параметрів. Вважається, що результат відповідає очікуваному, якщо терм-множина на виході відповідає висновку експерта.

1.3 Методи побудови нечітких логічних систем типу 1

Мендель та Ванг [18] виділяють 3 основні підходи до синтезу нечітких логічних систем та побудови нечітних баз знань.

1. *Архітектура нечіткої моделі встановлюється дослідником, експериментальні дані використовуються лише для оптимізації параметрів*. Такі властивості моделі як форма функцій належності, метод фазифікації та дефазифікації, t-норма, метод побудови нечіткого логічного висновку, кількість правил, кількість антецедентів (і в загальному випадку консеквентів) правила задає розробник моделі. При цьому невідомі величини, тобто параметри моделі, оптимізуються шляхом навчання на експериментальних даних. Емпіричний підхід до задання форми функцій належності та оптимального числа правил не завжди коректний, звідси висока ймовірність отримати логічно прозору модель, яка тим не менш недостатньо адекватно відображає досліджуваний процес.

2. *Центри нечітких множин термів лінгвістичних змінних визначаються експериментальними даними*. Решту параметрів нечітких множин задає розробник моделі. Правила будуються виключно на основі експериментальних даних, при цьому число нечітких множин термів кожної вхідної та вихідної змінної строго дорівнює числу правил та числу навчальних векторів даних. Отримана таким чином база правил надлишкова навіть для моделей невеликої розмірності; що стосується розв’язання задач вищої розмірності, то обсяг бази правил, побудованої за таким підходом, суттєво ускладнює обчислення. Цей підхід вимагає застосування додаткових методів зменшення надлишковості бази правил.

3. *Нечіткі множини термів лінгвістичних змінних початково визначаються розробником*; в подальшому з цими нечіткими множинами пов’язуються експериментальні дані (власне метод Менделя-Ванга, якому присвячено роботу). Цей підхід дозволяє зберегти розмірність бази знань в

розумних межах за рахунок безпосередньої участі розробника в процесі формування нечітких множин кожного антецедента та консеквента. База правил в цьому підході містить число правил, менше або рівне числу навчальних векторів даних. При цьому модель від самого початку адекватна предметній галузі, оскільки побудована на експериментальни даних, отриманих в процесі функціонування реальної системи.

В залежності від вибраного підходу до синтезу нечіткої моделі дослідники схильні обирати ті чи інші методи побудови функцій належності нечітких змінних. Для першого підходу найбільш очевидним є *прямий метод побудови функцій належності*, який полягає в тому, що експерт безпосередньо задає правила визначення значень функції належності, що характеризує дане поняття. Ці значення узгоджуються з його уявленнями на множины об'єктів *U* таким чином:

1. Для будь-яких $u_1, u_2 \in U$, $\mu_A(u_1) < \mu_A(u_2)$ тоді і тільки тоді, якщо u_2 має перевагу над u_1, тобто більшою мірою характеризується поняттям *A*.

2. Для будь-яких $u_1, u_2 \in U$, $\mu_A(u_1) = \mu_A(u_2)$ тоді і тільки тоді, якщо u_1 та u_2 однаковою мірою характеризуються поняттям *A*.

Експерт може безпосередньо задавати значення функції належності таблицею, формулою, перерахуванням.

Очевидно, що такий спосіб задання функцій належності вимагає від експерта не лише знання предметної галузі, а й фундаментальних математичних знань. Для спрощення роботи експерта існує ряд непрямих методів, у яких значення функції належності вибираються таким чином, щоб

задовільняти заздалегідь сформульованим умовам. Експертна інформація є лише вихідними даними для подальшої обробки.

Так, окрему групу складають методи на основі парних порівнянь, які полягають в обробці матриці оцінок, що відображає думку експерта про відносну належність елементів множині або ступеня вираженості в них властивості, що формалізується множиною [3]. Ступінь належності елементів множині визначається за допомогою оцінок, наведених в таблиці 1.1.

Таблиця 1.1

Шкала для визначення матриці суджень

Оцінка важливості	Якісна оцінка	Пояснення
1	Однакова значимість	За даним критерієм альтернативи мають однаковий ранг
3	Незначна перевага	Твердження про перевагу однієї аль-тернативи над іншою малопереконливі
5	Суттєва перевага	Наявні надійні докази суттєвої переваги однієї альтернативи
7	Очевидна перевага	Існують переконливі свідчення на користь однієї альтернативи
9	Абсолютна перевага	Свідчення на користь надання переваги однієї альтернативи над іншою надзвичайно переконливі
2, 4, 6, 8	Проміжні значення	Необхідний компроміс

Кожен із експертів заповнює матрицю парних порівнянь розміром NxN, де N – кількість альтернатив; для кожної пари альтернатив експерт вказує, яка з альтернатив має більшу перевагу (є кращою, важливішою тощо.) Існує ряд алгоритмів, що реалізують метод парних порівнять. Вони розрізняються за кількістю експертних оцінок, що використовуються (індивідуальті та колективні оцінки), за шкалами порівняння альтернатив та деякими іншими особливостями.

Найбільш широке розповсюдження з методів цієї групи на сьогоднішній день має метод аналізу ієрархій (Analytic Hierarchy Process, AHP). Його особливість у відсутності обмеження на суму вагових коефіцієнтів, а отже і необхідності тримати в полі зору всі фактори (ознаки), присутні в системі, при роботі з кожним окремим фактором. Ідея методу полягає в структуризації задачі прийняття рішення шляхом побудови багаторівневої ієрархії, що об'єднує всі альтернативи, що існують в даній задачі. Альтернативи далі порівнюються між собою за допомогою того чи іншого різновиду процедури парних порівнянь. В результаті стає можливим отримання кількісних оцінок інтенсивності взаємовпливу елементів ієрархії, на основі яких оцінюється міра переваги альтернатив відносно цільового критерію задачі.

В результаті вдається отримати детальне уявлення про те, як саме взаємодіють фактори, що впливають на пріоритети альтернативних розв'язків, а також власне розв'язки. На противагу цьому, формування структури моделі прийняття рішення в методі AHD - досить трудомісткий процес. Шкали суджень, що використовуються для визначення ваг

альтернатив, в більшості своїй не дозволяють адекватно відобразити всі можливі відношення між альтернативами, та фактично працюють коректно лише на якісних показниках, без уточнення кількісних мір переваги однієї альтернативи над іншою. Крім того, метод AHP не враховує невизначеностей, пов'язаних із поданням судження у вигляді числа; з іншого боку, суб'єктивний характер суджень експертів має суттєвий вплив на отримані результати. Що стосується його потенційних застосувань для моделювання природних процесів, часто неможливо зрозуміти, як саме взаємодіють фактори, та вибудувати семантично адекватну ієрархію ознак, адже далеко не всі реальні природні об'єкти і явища характеризуються набором ознак, який можна розбити на змістовні групи приблизно однакової потужності. З огляду на це, на сьогоднішній день метод AHP використовується переважно в задачі чіткого прийняття рішень за припущення повної відсутності невизначеностей в досліджуваній системі.

У методах побудови функцій належності лінгвістичних термів з використанням статистичних даних за ступінь належності елемента множині приймається оцінка частоти використання поняття, що задається нечіткою множиною, для характеристики елемента. Припустимо, що спостерігаючи за об'єктом протягом деякого часу, людина n разів фіксує свою увагу на тому, має місце факт A чи ні. Подія, що полягає в n перевірках наявності факту A, називаються оціночною. Нехай у k перевірках мав місце факт A. Тоді оператор реєструє частоту $p=k/n$ появи факту A та оцінює її за допомогою лінгвістичних висловлювань на зразок «часто», «рідко» тощо.

Для обробки таких статистичних даних можна послуговуватись так званою матрицею підказок. Це та застосування методів апроксимації

дозволяють отримати гладкі функції належності. Слід відзначити, що для коректної роботи методів цього класу вибірка спостережень відповідної події повинна мати достатню статистичну значущість.

При побудові функцій належності на основі експертних оцінок використовуються результати експертного опитування для синтезу наближених точкових (наприклад, X ПРИБЛИЗНО ДОРІВНЮЄ Y) або інтервальних (X ЗНАХОДИТЬСЯ ПРИБЛИЗНО В ІНТЕРВАЛІ ВІД Y ДО Z).

Для побудови нечіткого числа, приблизно рівного деякому числу *K*, можна використовувати функцію

$$\mu_{K(u)} = e^{-\alpha(K-u)^2}, \qquad (1.1)$$

де α залежить від потрібного ступеня нечіткості $\mu_{K(u)}$ та визначається з виразу

$$\alpha = \frac{4\ln 0.5}{\beta^2},$$

β - відстань між точками переходу для $\mu_{K(u)}$, тобто точками *a* і *b*, в яких функція вигляду (1.1) приймає значення *0.5*.

Для визначення множини вигляду ЧИСЛО, ПРИБЛИЗНО РІВНЕ *K*, слід виявити, як експерти уявляють собі границі класів таких чисел. Для цього іноді проводяться статистичні дослідження. Опитуваним пропонують назвати такі *a(K)* та *b(K)*, які на їхню думку відділяють числа, приблизно

рівні *K*, від чисел, що такими не є. Таким чином, задача побудови $\mu_{K(u)}$ для деякого числа зводиться до відшукання параметрів *a* і *b*, щоб потім можна було визначити $\beta(x)$, за допомогою $\beta(x)$ - α, та використовуючи α, побудувати $\mu_{K(u)}$.

Параметричний підхід до побудови функцій належності полягає в побудові модифікованих нечітких термів на основі існуючих. При цьому визначаються параметри дробово-лінійного перетворення, що відповідає нечіткому модифікатору, та з його допомогою перетворюється вихідний терм. Метод базується на припущенні, що експерт, характеризуючи лінгвістичне значення деякої ознаки, з мінімальним зусиллям може вказати три точки універсальної шкали: *A, B, C*, з яких *B* та *C* – точки, які, на його думку, ще (або вже) не належать описуваному лінгвістичному значенню, *A* – точка, що беззаперечно йому належить.

Побудова функцій належності на основі інтервальних оцінок спирається на припущення, що експерт може вказати інтервал $[h^*, h^{\square}]$ значень критерію *h*, що відповідає побажанню вибрати, скажімо, «хороший» об’єкт. При цьому граничні значення інтервалу мають таку інтерпретацію. Нехай h^a - результат вимірювання значення характеристики *h* для об’єкта *a*. Тоді h^* є границею «ідеальної» області, тобто якщо $h^a \geq h^*$, об’єкт слід вважати таким, що ідеально відповідає поняттю «хороший». Можливість (з точки зору теорії можливостей) такого твердження $\pi(Q) = 1.0$ (Q – суб’єктивна подія, яка полягає в тому, що об’єкт, з точки зору експерта, знаходиться в стані «хороший».) Якщо $h^a \leq h^{\square}$, ситуація інтерпретується так:

можливість того, що об'єкт *a* – «хороший», $\pi(Q)=0$. Очевидно, що при $h^{□} < h^{a} < h^{*}$ відповідні можливості мають значення $0<\pi(Q)<1.0$.

Відчуття експерта про характер зміни ступеня відповідності об'єкта *a* поняттю «хороший» з наближенням значення h^{a} до границі h^{*}, та апроксимуючи дані оцінок експертів для $h^{□}$ та h^{*} ще в кілької точках *z*, отримують аналітичні вирази двох функцій $h^{*}=f^{*}(z)$ та $h^{□}=f^{□}(z)$, які називаються рівневими обмеженнями. Ці функції шляхом експертного опитування можна побудувати таким чином, щоб охопити весь діапазон реальної зміни параметра *Z*.

Для отримання повного уявлення про альтернативу *a* необхідно провести ряд експериментів із визначення оцінки h^{a} за різних значень *z*. За допомогою апроксимації отримують функцію $h^{a}=f^{a}(z)$; для ряду значень *z* розраховуються значення $\pi_{z}(Q)$, апроксимуючи які отримують ступінь відповідності альтернативи поняттю експерта «хороша альтернатива» на множині значень параметра *Z*. Отримана функція називається розподілом можливостей і являє собою нечітке обмеження на значеннях параметра *Z*.

Всі ці методи тим чи іншим чином використовуються в рамках першого з трьох підходів, визначених на початку цього огляду. Проте робота експерта з підготовки до прийняття рішень часто вимагає завеликих затрат часу для однієї людини та вимагає спеціальної математичної підготовки, що суттєво звужує коло потенційних фахівців, яких можна залучити до розробки моделі. Крім того, в межах цього підходу всі компоненти системи цілковито

залежать від суджень експерта, які незалежно від його кваліфікації можуть розходитись із фактичними спостереженнями об'єктів реального світу.

Тому зважаючи на це та на очевидні недоліки другого підходу (надлишковість формату правил та бази знань), за наявності достатніх експериментальних даних надають перевагу третьому підходу до побудови нечітких моделей, який наряду з урахуванням експертних знань дозволяє побудову правил на основі об'єктивних даних функціонування реальної системи.

1.4 Нечіткий логічний висновок в системах на основі нечітких множин типу 1

Нечітким логічним висновком прийнято називати процес прийняття рішення про значення вихідної лінгвістичної змінної на основі значень *n* вхідних змінних. При цьому зв'язки, що існують між вхідними та вихідною змінними, описуються природною мовою у формі матриці (бази) знань, що задається експертом:

$y = f_y(x_1, x_2, \ldots, x_n)$,

де – вихідна змінна,

$x_1, x_2, \ldots, x_n$ – вхідні змінні.

f_y – деяка апріорно невідома функція, що описує залежність значення вихідної змінної від вхідних.

Всі вхідні та вихідна змінні характеризуються наборами якісних термів: $A_i = \{a_i^1, a_i^2, \ldots, a_i^{l_i}\}$ для вхідних змінних x_i, $i=1\ldots n$, $Y = \{y_1, y_2, \ldots, y_m\}$ для вихідної змінної y, m – кількість можливих класів вихідної змінної. Функції

належності нечітких множин-термів можуть задаватися експертом (прямий метод), визначатися за одним із методів парних порівнянь, за допомогою параметричного підходу (побудова модифікованих нечітких термів на основі існуючих шляхом визначення параметрів дробово-лінійного перетворення відповідно до нечіткого модифікатора [17]) або на основі інтервальних оцінок. Також для визначення центрів нечітких терм-множин можуть використовуватись експериментальні дані [1].

Найчастіше для опису змінних у нечітких логічних системах використовують трикутні, трапецієподібні, дзвоноподібні та гаусові функції. Дзвоноподібні та гаусові функції гладкі, мають ненулеве значення в будь-якій точці області визначення, та записуються одним математичним співвідношенням. Кусково-лінійні функції належності всіх цих переваг не мають. У практичних застосуваннях найчастіше використовуються гаусові функції належності через поширеність функції Гауса, або нормального розподілу, в природі.

Правила нечіткої бази знань подаються у вигляді логічних висловлювань вигляду «ЯКЩО – ТО», які встановлюють відповідність між набором значень вхідних змінних $x_1, x_2, \ldots, x_n$ та висновком y_j:

$$ЯКЩО\ x_1 \in a_i^1 \wedge x_2 \in a_i^2 \wedge \ldots \wedge x_l \in a_i^l, ТО\ y \in y_j\ ,$$

Механізм побудови нечіткого логічного висновку в системах на основі нечітких множин типу 1 можна отримати, записавши набір таких правил,

пов’язаних між собою логічним оператором об’єднання, у формі функцій належності:

$$\mu^{d_j}(x_1, x_2, ..., x_n) = \bigvee_{p=1}^{k_j} \left[\bigwedge_{i=1}^{n} \mu^{jp}(x_i) \right], \; j = 1, m$$

Таким чином, нечіткий логічний висновок являє собою апрокимацію залежності між входами і виходами системи за допомогою нечіткої бази знань та операцій над нечіткими множинами.

1.5 Нечіткі множини в задачах інтелектуального аналізу даних

Data Mining. Методи кластеризації

Задачі прийняття рішень в умовах невизначеності вимагають роботи з великими масивами багатовимірних даних, а тому тісно пов’язані з колом задач Data Mining, інтелектуального аналізу даних. Data Mining – це процес аналізу даних із різних точок зору та подання їх у вигляді корисної інформації, тобто такої, яку можна безпосередньо використовувати для розв’язання конкретних, зокрема оптимізаційних, задач. Технічно Data Mining – це процес виявлення кореляцій, зв’язків чи закономірностей в масштабах десятків полів у великих реляційних базах даних.

Математичний інструментарій методів Data Mining складають методи класифікації, моделювання, прогнозування, що ґрунтуються на використанні штучних нейронних мереж, генетичних алгоритмів, еволюційного програмування, асоціативної пам’яті, нечіткої логіки. Одним із призначень Data Mining також є наочне подання (візуалізація) результатів обчислень, що

дозволяє використовувати технології Data Mining людьми, що не мають спеціальної математичної підготовки.

Задача Data Mining ставиться таким чином. Нехай є деяка досить велика база даних, та висувається припущення, що в базі даних знаходяться деякі «приховані знання». Необхідно розробити методи виявлення знань, прихованих у великому обсязі «сирих» даних. Знайдені закономірності в контексті загальної орієнтації сучасного програмного забезпечення на використання концепції Big Data можуть стати джерелом додаткових прибутків чи конкурентних переваг для компаній. «Приховані знання» повинні володіти рядом ознак. Це повинні бути знання

- раніше невідомі, нові;
- нетривіальні (такі, що їх не можна отримати будь-яким іншим простішим шляхом, наприклад, при безпосередньому візуальному аналізі даних або обчисленні простих статистичних характеристик);
- практично корисні, тобто такі, що мають цінність для дослідника чи споживача;
- доступні для інтерпретації, тобто знання, які легко можна подати в наочній для користувача формі та легко пояснити в термінах предметної галузі.

Знання, що добуваються за допомогою методів Data Mining, прийнято подавати у вигляді закономірностей, в ролі яких можуть виступати асоціативні правила, дерева розв’язків, кластери, математичні функції тощо.

Одним із найпотужніших інструментів Data Mining є кластерний аналіз. Він широко використовується для виділення прихованих закономірностей та внутрішніх взаємозв’язків у великих масивах багатовимірних даних.

За допомогою кластеризації успішно розв’язуються задачі обробки та сегментації зображень, зокрема медичних, розпізнавання образів, розпізнавання та впорядкування мультимедійного трафіку, обробки та категоризації документів, сегментації множини абонентів провайдера телекомунікаційних послуг, будуються соціальні дослідження, економічні прогнози тощо. Кластерний аналіз також може використовуватись для виділення інформативних ознак при роботі з надлишковими даними.

Застосування методів кластерного аналізу дозволяє розв’язувати задачу класифікації об’єктів у випадку відсутності будь-якої інформації про кількісний або якісний склад кластерів.

В загальному вигляді алгоритм кластеризації – це функція a: $X \to Y$, яка будь-якому об'єкту $x \in X$ ставить у відповідність мітку кластера $y \in Y$. Множина міток Y у деяких випадках відома заздалегідь, однак частіше ставиться завдання визначити оптимальне число кластерів з погляду того або іншого критерію якості кластеризації [5]. Для розв’язання цієї задачі універсальних методів, що дозволяють швидко знайти абсолютно точні розв’язки, не існує. Розв’язання задачі кластеризації принципово неоднозначне з декількох причин [7]. По-перше, не існує однозначно найкращого критерію якості кластеризації. Відомий ряд критеріїв та алгоритмів, що не мають чітко вираженого критерію, але які здійснюють досить змістовну кластеризацію "за побудовою". Всі вони можуть давати різні результати. По-друге, число кластерів, як правило, невідоме заздалегідь і встановлюється відповідно до деякого суб'єктивного критерію. По-третє, результат кластеризації істотно залежить від метрики, вибір якої, як правило, також суб'єктивний і визначається експертом [10].

Точні методи кластерного аналізу. Найбільш прямий спосіб розв'язання задачі кластеризації полягає в повному переборі всіх можливих розбиттів на кластери та знаходженні такого розбиття, яке веде до оптимального (мінімального) значення цільової функції. Методи повного перебору – єдина група методів, які дають змогу завжди відшукати оптимальний розв'язок. Проте така процедура практично нездійсненна за винятком тих випадків, коли *n* (число об'єктів) та *m* (число кластерів) невелике. Очевидно, що основним недоліком алгоритмів повного перебору є настільки великий обсяг обчислень, що, не зважаючи на високу швидкодію сучасних обчислювальних машин, в задачах із великою кількістю даних або суттєвою розмірністю вони непридатні. Тому використовують алгоритми граничного перебору, в яких кількість варіантів розбиття, що перебираються, скорочується. Найефективнішим із них є метод на основі динамічного програмування, який завжди забезпечує точний результат при значному зменшенні обчислювальної складності. Але навіть за такого підходу обсяг необхідних обчислень все одно залишається надзвичайно великим. За таких умов єдиним прийнятним методом на сьогодні є застосування наближених алгоритмів кластерного аналізу.

Методи ієрархічної кластеризації. За способом розбиття на кластери наближені алгоритми бувають двох типів: ієрархічні та неієрархічні.

Класичні ієрархічні алгоритми працюють тільки з категорійними атрибутами, коли будується повне дерево вкладених кластерів. Тут поширені агломеративні методи побудови ієрархій кластерів - в них проводиться послідовне об'єднання вихідних об'єктів і відповідне зменшення числа кластерів. Ієрархічні алгоритми мають ряд суттєвих обмежень: вони

працюють лише з числовими даними, чутливі до викидів та вимагають задання порогових значень. Якщо не використовуються спеціально розроблені алгоритми скорочення розмірності, то застосування ієрархічних методів може вимагати обчислення та зберігання матриці подібності великої розмірності. Іншим недоліком є те, що об’єкти розподіляються по кластерах лише за один прохід, а погане початкове розбиття множини даних уже не може бути змінене на наступних кроках кластеризації. Третій недолік всієї родини ієрархічних алгоритмів полягає в тому, що вони можуть породжувати різні розв’язки в результаті простого перемішування об’єктів у матриці подібності. Результати також змінюються, якщо деякі об'єкти виключаються з розгляду. Необхідно відмітити, що стійкість є важливою властивістю кластеризації, якій ієрархічні алгоритми не задовольняють.

Хоча й існують досить ефективні ієрархічні алгоритми (*AGNES*, *CURE*, *DIANA*, *BIRCH* [14]), наведені недоліки роблять практично неможливим подальше підвищення їхньої швидкодії та точності отримуваних результатів. Із цих причин в даній роботі ієрархічні алгоритми кластерного аналізу не розглядаються.

Методи неієрархічної (оптимізаційної) кластеризації. Неієрархічні алгоритми базуються на оптимізації деякої цільової функції, що визначає оптимальне в певному сенсі розбиття множини об'єктів на кластери. У цій групі популярні алгоритми родини *k*-середніх *(k-means)*, які в ролі цільової функції використовують суму квадратів зважених відхилень координат об'єктів від центрів шуканих кластерів. Кластери шукаються сферичної або еліпсоїдної форми. У канонічній реалізації мінімізація функції здійснюється на основі методу множників Лагранжа і дозволяє знайти тільки найближчий

локальний мінімум. Серед неієрархічних алгоритмів, не заснованих на відстані, слід виділити *EM*-алгоритм (*ExpectationMaximization*). У ньому замість центрів кластерів передбачається наявність функції щільності ймовірності для кожного кластера з відповідним значенням математичного сподівання і дисперсією.

Всі описані вище методи розбивають вихідну множину об'єктів X на декілька підмножин, що не перетинаються. При цьому будь-який об'єкт із X належить лише одному кластеру. Таке формулювання задачі кластеризації не позбавлене обмежень, серед яких обмеження на геометрію кластерів та повна відсутність врахування невизначеностей та викидів, присутніх у вхідних даних.

Методи нечіткої кластеризації

Відмінністю методів нечіткої кластеризації від чіткої є область визначення ступенів належності. Якщо за чіткої кластеризації ступені належності можуть приймати лише значення 0 або 1, таким чином відносячи об’єкт до одного з класів, то у випадку нечіткої кластеризації ступені належності можуть набувати будь-яких значень із інтервалу *[0, 1]*. Таким чином, нечітка кластеризація дозволяє віднести об’єкт до кількох, а в загальному випадку до всіх кластерів, але з різними ступенями належності. Зокрема ця особливість дозволяє більш достовірно описувати об’єкти, що знаходяться на межі кластерів, або об’єкти, рівновіддалені від центрів двох або більше кластерів. Крім того, вона дає змогу врахувати невизначеності, що існують у даних, що також є суттєвою перевагою даного класу методів.

Інтерпретація результатів нечіткої кластеризації може бути дуже спірною, але в той же час створює найповніше уявлення про зв'язки між об'єктами. До того ж, в деяких випадках не можна виділити підмножини, що не перетинаються.

В цілому, основна перевага нечітких моделей в порівнянні з традиційними математичними моделями пов'язана з можливістю використання для їх розробки значно менших обсягів інформації про систему, причому вона може носити наближений характер.

Найпотужнішим засобом для розв’язання задачі в такій постановці є алгоритм *FCM* (*Fuzzy c-means)*, що являє собою модифікацію *EM*-алгоритму.

В його основі лежить процедура нечіткої самоорганізації, а критерієм оптимальності є середньозважене відхилення точок-об’єктів від центрів кластерів:

$$E = \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij}^{m} \left\| x_j - c_i \right\|^2,$$

де μ_{ij} - ступінь належності об’єкта j до кластера i,

x_j – вектор ознак об’єкта j,

c_i – центр кластера i,

$m \geq 1$ - експоненційна вага, що визначає нечіткість, розсіяність кластерів.

Початкові значення центрів кластерів c_j вибираються випадковим чином із областей допустимих значень відповідних компонент векторів x_j .

Основна ідея полягає в тому, що на кожній ітерації переобчислюється центр мас для кожного кластера, отриманого на попередньому кроці, після чого точки розбиваються на кластери знову відповідно до того, який із нових центрів виявився ближче за обраною метрикою. Алгоритм завершується, коли на деякій ітерації не відбувається зміни кластерів.

Перерахування координат центрів мас кластерів та ступенів належності відбувається за формулами:

$$c_i = \frac{\sum_{j=1}^{p} \mu_{ij}^m x_j}{\sum_{j=1}^{p} \mu_{ij}^m}, \qquad \mu_{ij} = \frac{1}{\sum_{k=1}^{K} \left(\frac{d_{ij}^2}{d_{kj}^2} \right)^{\frac{1}{m-1}}},$$

де d_{ij} - Евклідова відстань між центром c_i та вектором x_j, $d_{ij} = \|x_j - c_i\|$.

Багаторазове повторення ітераційної процедури веде до досягнення мінімуму функції *E*, який необов'язково буде глобальним мінімумом. Проте якість знайдених центрів суттєво залежить від попереднього вибору як значень μ_{ij}, так і центрів c_j. Крім того, *FCM* використовує обмеження, подібне до того, що накладає на шуканий розв'язок теорія ймовірностей: сума ступенів належності *i*-тої точки до всіх кластерів $j = \overline{1, N}$ становить *1*: $\sum_{i=1}^{c} \mu_{ij}$ для всіх *j*. Таке обмеження має на меті уникнути тривіального розв'язку, коли всі ступені належності виявляються рівними *0*, і дає змістовні

результати в тих прикладних застосуваннях, де припущення про «ймовірнісну» природу ступенів належності має практичний сенс.

Але, оскільки ступені належності, отримані при такому обмеженні, відносні, вони непридатні в тих задачах, у яких ступінь належності точки до кластера повинен відображати її типовість, характерність саме для цього кластеру. Це повністю узгоджується з теорією нечітких множин Заде, адже ступінь належності точки до класичної нечіткої множини є абсолютною величиною, незалежною від ступенів належності цієї ж точки до інших нечітких множин, визначених на тій самій універсальній множині. Таке формулювання є більш придатним для більшості задач кластеризації, оскільки ступінь належності точки до кластера є мірою того, наскільки ця точка є носієм спільних характеристик кластеру, її типовості, і не повинен залежати від того, як вона розташована відносно інших кластерів. Можливісний підхід до визначення природи ступенів належності *PCM* оперує ступенями належності, що володіють цією властивістю. Його авторами було переглянуто цільову функцію методу *FCM* таким чином, щоб при досягненні її мінімуму ступені належності для репрезентативних точок кластерів були високими, а для нерепрезентативних – низькими, незалежно від взаємного положення точок та кластерів. Результуючий функціонал має вигляд

$$E = \sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}^{m} d_{ij}^{2} + \sum_{i=1}^{c}\eta_i \sum_{j=1}^{N}(1-\mu_{ij})^{m}, \qquad (1.2)$$

де η_i – додатне число.

Значення η_i визначає відстань, на якій значення ступеня належності точки до кластера стає рівним *0,5*.

За такої цільової функції відповідним чином змінюються також і формули для перерахунку змінних величин методу:

$$\mu_{ij} = \frac{1}{1+\left(\frac{d_{ij}^2}{\eta_j}\right)^{\frac{1}{m-1}}} \quad ; \qquad \eta_{ij} = \frac{\sum_{j=1}^{N} \mu_{ij}^m d_{ij}^2}{\sum_{j=1}^{N} \mu_{ij}^m}$$

Співвідношення, що використовується для перерахунку координат центрів кластерів, порівняно з *FCM* залишається без змін:

$$c_i = \frac{\sum_{j=1}^{p} \mu_{ij}^m x_j}{\sum_{j=1}^{p} \mu_{ij}^m},$$

Розв’язки, отримані при такому підході, більше відповідають дійсності та інтуїтивному уявленню про природу кластерів.

2 Нечіткі множини типу 2

2.1 Основні відомості

Подальше узагальнення поняття функції належності привело до появи нечітких множин типу 2 та множин вищих порядків. Узагальнена нечітка множина визначається функціями належності, в ролі значень яких також виступають нечіткі множини. Моделі на основі узагальнених нечітких множин типу 2, тобто множин, значення функції належності яких являють собою повноцінну нечітку множину, були запропоновані ще в 70-х роках XX століття, але не набули значного поширення через надмірну обчислювальну складність та неможливість строго визначити закон, за яким змінюються вторинні функції належності. Натомість широко використовуються інтервальні нечіткі множини типу 2, які являють собою спрощене подання узагальнених нечітких множин. Апарат інтервальних нечітких множин оперує лише крайніми точками інтервалу зміни значення функції належності, і не враховує особливостей розподілу, що виникає в межах цього інтервалу. Таке спрощення значно знижує кількість обчислювальних ресурсів, необхідних для побудови нечіткого логічного висновку; при цьому на якості функціонування системи це майже не відбивається [15].

Для задання інтервальних нечітких множин типу 2 використовуються функції належності вигляду

$$\mu_{\tilde{A}}(x,u) = \int_{x\in X} \mu_{\tilde{A}}(x)/x = \int_{x\in X}\left[\int_{u\in J_x} f_x(u)/u\right]\Big/x = \int_{x\in X}\left[\int_{u\in\left[\underline{\mu}_{\tilde{A}}(x),\overline{\mu}_{\tilde{A}}(x)\right]} 1/u\right]\Big/x,$$

де J_x – первинний ступінь належності x інтервальній нечіткій множині типу 2 $\tilde{A}$;

$\tilde{A}$ – область визначення вторинної функції належності;

$\overline{\mu}_A(x)$, $\underline{\mu}_A(x)$ – верхня і нижня функції належності, нечіткі множини типу 1, які задають границі зони невизначеності інтервальної нечіткої множини 2-го типу $\tilde{A}$; $\overline{\mu}_A(x) = \overline{J}_x$, $\underline{\mu}_A(x) = \underline{J}_x$, $\forall x \in X$.

$f_x(u)$ – вторинний ступінь належності, амплітуда вторинної функції належності. У випадку інтервальної нечіткої множини $f_x(u) = 1, \forall u \in J_x \subseteq [0,1]$

2.2 Нечіткі логічні системи типу 2

Одним із найбільш розповсюджених застосувань математичного апарату інтервальних нечітких множин при побудові нечітких логічних систем є розширення можливостей *методів класу АНР*. Більшість досліджень у цьому напрямку зосереджується на пошуку нових та вдосоналенні існуючих шкал порівняння альтернатив та побудові більш збалансованих шкал. При цьому залишаються в силі такі недоліки методу АНР як неврахування невизначеностей, пов'язаних із поданням суб'єктивного судження у кількісному вигляді та неможливість побудувати семантично адекватну ієрархію ознак в деяких задачах. У ряді робіт фігурують дев'ятиточкові шкали, що являють собою варіації оригінальної шкали Сааті, наведеної в таблиці 1.1. Всі вони мають трапецієвидні верхні та нижні

функції належності; їхню висоту, як правило, приймають на рівні 1 та 0.8 відповідно, як наприклад для шкали в табл. 1.2.

Таблиця 1.2

Інтервальна шкала суджень

Якісна оцінка	Оцінка важливості, інтервальна нечітка множина
Однакова значимість	((0; 0.1; 0.1; 0.1; 1;1), (0; 0.1; 0.1; 0.05; 0.9; 0.9))
Проміжне значення	((0.1; 0.2; 0.2; 0.3; 1;1), (0.05; 0.2; 0.2;0.25; 0.9; 0.9))
Незначна перевага	((0.2; 0.3; 0.3; 0.4; 1;1), (0.25; 0.3; 0.3;0.35; 0.9; 0.9))
Проміжне значення	((0.3; 0.4; 0.4; 0.5; 1;1), (0.35; 0.4; 0.4;0.45; 0.9; 0.9))
Суттєва перевага	((0.4; 0.5; 0.5; 0.6; 1;1), (0.45; 0.5; 0.5;0.55; 0.9; 0.9))
Проміжне значення	((0.5; 0.6; 0.6; 0.7; 1;1), (0.55; 0.6; 0.6;0.65; 0.9; 0.9))
Очевидна перевага	((0.6; 0.7; 0.7; 0.8; 1;1), (0.65; 0.7; 0.7;0.75; 0.9; 0.9))
Проміжне значення	((0.7; 0.8; 0.8; 0.9; 1;1), (0.75; 0.8; 0.8;0.85; 0.9; 0.9))
Абсолютна перевага	((0.8; 0.9; 0.9;1; 1; 1), (0.85; 0.9; 0.9;0.95; 0.9; 0.9))

Подібні лінгвістичні змінні також використовуються для визначення оцінок ваг альтернатив; з них найбільшою популярністю користується лінгвістична змінна, наведена в таблиці 1.3. Її так чи інакше використовує переважна більшість розробників методів на основі AHP.

Таблиця 1.3

Інтервальна лінгвістична змінна для оцінювання ваг альтернатив

Якісна оцінка	Інтервальна нечітка множина
Дуже низький	((0; 0;0; 0.1;1; 1), (0; 0;0; 0.05;0.9; 0.9))

Низький	((0; 0.1; 0.1; 0.3;1; 1), (0.05; 0.1; 0.1; 0.2; 0.9; 0.9))
Нижче середнього	((0.1; 0.3; 0.3; 0.5;1; 1), (0.2; 0.3; 0.3; 0.4; 0.9; 0.9))
Середній	((0.3; 0.5; 0.5; 0.7;1; 1), (0.4; 0.5; 0.5; 0.6; 0.9; 0.9))
Вище середнього	((0.5; 0.7; 0.7; 0.9;1; 1), (0.6; 0.7; 0.7; 0.8; 0.9; 0.9))
Високий	((0.7; 0.9; 0.9; 1; 1;1), (0.8; 0.9; 0.9; 0.95;0.9; 0.9))
Дуже високий	((0.9;1; 1;1; 1; 1), (0.95;1; 1;1; 0.9; 0.9))

Ряд сучасних досліджень присвячено адаптації *методів класу TOPSIS* – Technique for Order Preference by Similarity to Ideal Solution – групи методів багатокритеріальної оптимізації, які також ґрунтуються на порівнянні альтернатив – для інтервальних нечітких розрахунків. За допомогою цього підходу розв'язують задачі вибору оптимальної структури веб-додатків, оцінювання ризиків забруднення довкілля та вибору кращої серед альтернатив утилізації небезпечних відходів, вибору постачальника, вибору оптимального розташування виробничих потужностей підприємства . Авторами цих підходів також відзначається зручність використання інтервальних нечітких множин для подання значень оцінок та ваг критеріїв в умовах невизначеності, проте вони зазначають, що для побудови лінгвістичних змінних як правило залучаються експерти, які мають досвід у предметній галузі та володіють знаннями, необхідними для кваліфікованого оцінювання критеріїв та альтернатив. Такого роду обмеження є спільними для всіх моделей, побудованих виключно на основі експертних знань. Наряду з вичерпними знаннями з предметної галузі експерт повинен володіти базовими концепціями математичного моделювання, що значно звужує коло доступних експертів. Не зважаючи на цей недолік, автори досліджень

стверджують, що саме застосування інтервальних нечітких множин дозволяє отримувати результати, що більш адекватно відображають реальність.

Інтервальні форми *методу PROMETHEE* (Preference Ranking Organization METHod for Enrichment Evaluations) дозволяють оцінити та вибрати альтернативу з деякого набору, вважаючи за основу критерії, що відображають переваги та недоліки альтернатив, та виконати ранжування альтернатив. *Метод QUALIFLEX* (QUALItative FLEXible) вважається перспективним для використання в контексті теорії інтервальних нечітких множин завдяки його можливостям обробки кількісної та якісної інформації одночасно. В основі методу лежить процедура парних порівнянь альтернатив відповідно до кожного критерію для всіх можливих перестановок альтернатив. Після цього обчислюються індекси відповідності та невідповідності для кожної пари альтернатив перестановок і визначається оптимальна перестановка альтернатив, яка максимізує значення індексу відповідності/невідповідності, та найкраща альтернатива.

Суть *методу DEMATEL* (Decision-Making Trial and Evaluation Laboratory) полягає в розрахунку непрямих відношень між змінними на основі оцінки прямих зв'язків для структуризації складних проблем. Виходячи з оцінки природи відношень між змінними, метод DEMATEL дозволяє поділити змінні на причини та наслідки. Змінна «причина» має значно більший вплив на інші змінні в порівнянні з впливом на неї саму.

Метод COPRAS (Complex Proportional Assessment) використовується для багатовимірного аналізу даних на основі значень максимізуючого та мінімізуючого критеріїв. При цьому обидва критерії розглядаються незалежно один від одного. Основна ідея методу – оцінка *j*-тої альтернативи

Z^j прямо пропорційна впливу, що його чинить максимізуючий критерій S^{+j}, та оберенено пропорційна сумі зважених нормалізованих значень мінімізуючого критерію S^{-j} . Серед його переваг відзначається обчислювальна простота, що є важливим фактором у роботі з нечіткими множинами другого та вищих порядків.

За останнє десятиліття подібну адаптацію математичного апарату нечітких множин було виконано також для інших підходів, в результаті чого з'явились методи IT2FDEA для Data Envelopment Analysis (DEA), IT2FELECTRE на базі методу ELECTRE (Interval Type-2 Fuzzy ELimination Et Choix Traduisant la REalité), IT2FENTROPY, IT2FLINMAP на основі методу LINMAP, Linear Programming Technique for Multidimensional Analysis of Preference, IT2FVIKOR як розширення методу VIKOR (VIseKriterijumska Optimizacija I Kompromisno Resenje, метод багатокритеріальної оптимізації та компромісного розв'язку) та ін.

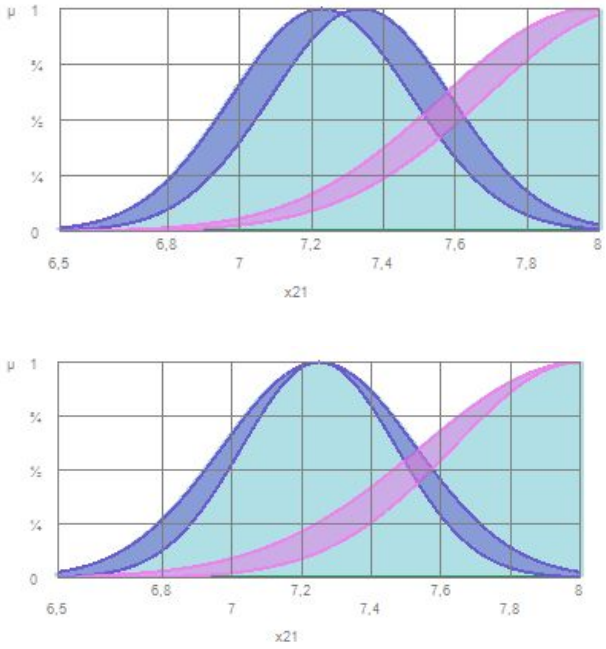
В усіх розглянутих підходах до формування нечітких множин термів лінгвістичних змінних так чи інакше залучається експерт, що зумовлює використання простих форм функцій належності – трикутних або трапецієвидних. Вимога подання цих функцій відразу в інтервальній формі ще більше ускладнює роботу експерта, який, в загальному випадку, не обов'язково має необхідне для цього математичне підгрунтя. В той самий час, у порівнянні з многокутними гладкі функції належності, зокрема гаусова, мають ряд переваг, таких як ненульове значення в будь-якій точці області визначення та більш інтуїтивний характер (нормальний розподіл – найбільш поширений в природі). Крім того, гладкі функції належності неперервно диференційовні. У деяких роботах висловлюється думка про те,

що відсутність неперервно диференційовної функції належності ускладнює процес навчання нечітких моделей. Гаусова ж функція, будучи неперервно та нескінченно диференційовною, дає можливість проведення теоретичного аналізу нечітких систем.

У випадку гаусових функцій можливі два типи первинних функцій належності інтервальних нечітких множин типу 2:

– функції з невизначеним центром та сталим відхиленням (рис. 2.1, а);

– функції зі сталим центром та невизначеним відхиленням (рис. 2.1, б).



а) б)

Рисунок 2.1 – Первинні функції належності

В межах першого підходу до синтезу нечітких моделей та побудови нечітких баз знань, коли архітектура нечіткої моделі встановлюється

дослідником, інтервальна форма функції належності отримується емпіричним та як правило дещо штучним шляхом, наприклад приймаючи значення верхньої та нижньої функції належності на рівні 1 та 0.8 відповідно.

Існує також механізм побудови інтервальних нечітких функцій належності типу 2 з функцій належності типу 1 в межах третього підходу, що передбачає використання експериментальних даних для побудови функцій належності. Відповідно до нього, виконується зміна відповідного параметра незначними інкрементами доти, поки вихід нечіткої логічної системи типу 1 знаходиться в межах одного терму:

$$\forall x_i \in X \; F(x_i, P^{(k)}) = F(x_i, P^{(1)}),$$

де $P^{(1)}$ – початкові параметри функцій належності,

$$P^{(k)} = \left\{ \mu_1 \cdot k, ..., \mu_p \cdot k \right\}, \quad k = k \pm 0,001$$

$F(x_i, P^{(k)})$ – вихід системи без дефазифікації – номер терма з максимальним покриттям результуючою функцією належності. Форма інтервальної функції, отриманої таким чином, має прямий зв’язок із апроксимаційними властивостями моделі. Окрім власне зони невизначеності, яка застерігає від прийняття хибного точкового рішення, вона також являє собою змістовну характеристику ступеня невизначеності, асоційованої з об’єктом дослідження.

2.3 Побудова нечітких логічних систем типу 2

Нечітку логічну систему можна модифікувати для роботи з нечіткими множинами типу 2. Як правило, це робиться для того, щоб забезпечити

можливість роботи в умовах невизначеності, а саме тоді, коли вхідний вектор містить пропуски даних.

Інтервальна модель в цілому складається зі структурних блоків, аналогічних до моделі типу 1 (рис. 2.2). Виходом системи є інтервал значень вихідної змінної. Перетворимо функції належності типу 1, отримані після генетичної оптимізації, на інтервальні функції належності типу 2 з невизначеним середньоквадратичним відхиленням (рис. 2.3).

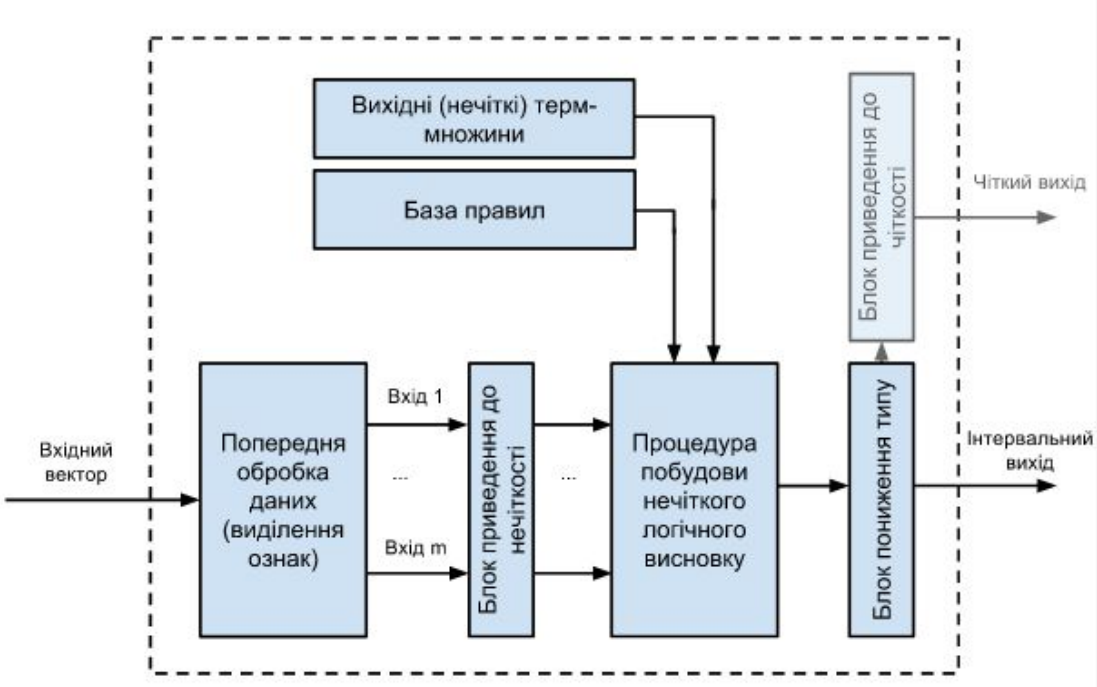


Рисунок 2.2 – Нечітка логічна система класифікації даних типу 2

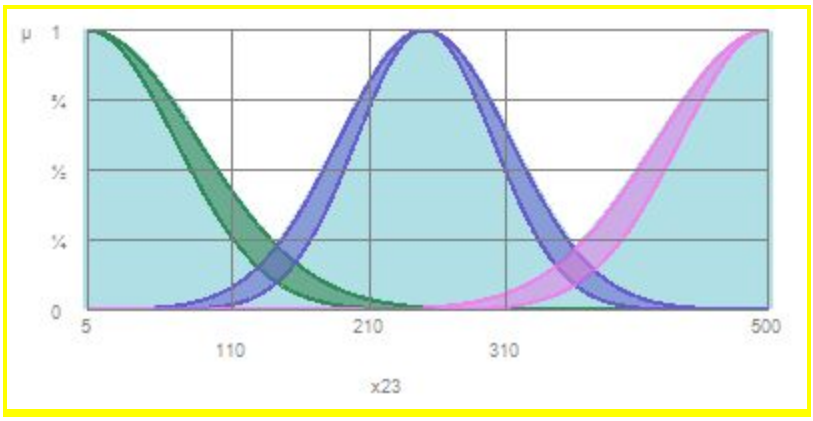


Рисунок 2.3 – Приклад інтервальної функції належності з невизначеним середньоквадратичним відхиленням

Інтервальний нечіткий логічний висновок будується за алгоритмом Карніка-Менделя. Вихідна змінна приймає значення $[y_l;y_r] \in D(Y)$. Ширина інтервалу характеризує ступінь невизначеності, пов'язаної з прийнятим рішенням.

2.4 Нечіткий логічний висновок в системах на основі інтервальних нечітких множин типу 2 (алгоритм Карніка-Менделя)

У моделях на основі нечітких множин типу 2 правила бази знань в загальному вигляді мають формат

$$R^i : IF\ x_1 \in A^i_{x_1} \wedge x_2 \in A^i_{x_2} \wedge \ldots \wedge x_m \in A^i_{x_m}\ THEN\ y \in L^k_y \in \left\{ L_1,\ldots, L_p \right\},$$

де x_i – вхідні змінні, y – вихідна змінна, $L_y \in \left\{ L_1,\ldots, L_p \right\}$ – терм-множини вихідної змінної.

Інтервальні ступені належності кожного правила розраховуються як мінімум усіх антецедентів:

$$\mu_{R_i} = \left[\min(\underline{\mu}^{(2)A_j^{(i)}}_{j}(x_j^*)) \; ; \min(\overline{\mu}^{(2)A_j^{(i)}}_{j}(x_j^*)) \right].$$

Для знаходження лівої та правої границь інтервалу вихідної змінної *[yl;yr]* на основі розрахованих ступенів належності правил та інтервальних значень консеквентів правил будується вихідна нечітка множина типу 2. Вихідний інтервал значень отримуємо за допомогою процедури пониження порядку нечіткої множини. Для правої границі інтервалу:

1. Обчислити

$$f_r^i = \left(\underline{\mu}_i + \overline{\mu}_i\right) \Big/ 2 \; ; \; y_r = \frac{\sum_{i=1}^{M} f_r^i w_r^i}{\sum_{i=1}^{M} f_r^i}; \; y_r' = y_r.$$

2. Знайти R (R=1...M – 1) таке, що $w_r^R \le y_r' \le w_r^{R+1}$.

3. $$y_r = \frac{\sum_{i=1}^{R} \underline{f}^i w_r^i + \sum_{i=R+1}^{M} \overline{f}^i w_r^i}{\sum_{i=1}^{R} \underline{f}^i + \sum_{i=R+1}^{M} \overline{f}^i}; \; y_r'' = y_r.$$

4. Якщо $y_r'' \ne y_r$, перейти на крок 5, інакше $y_r = y_r''$ та перейти на крок 6.

5. $y_r' = y_r''$, перейти на крок 2.

Для лівої границі інтервалу:

1. Обчислити

$$f_l^i = \left(\underline{\mu}_i + \overline{\mu}_i\right) \Big/ 2 \; ; \; y_l = \frac{\sum_{i=1}^{M} f_l^i w_l^i}{\sum_{i=1}^{M} f_l^i}; \; y_l' = y_l.$$

2. Знайти L (L=1...M – 1) таке, що $w_l^L \le y_l' \le w_l^{L+1}$.

3. $$y_l = \frac{\sum_{i=1}^{L} \overline{f}^i w_l^i + \sum_{i=L+1}^{M} \underline{f}^i w_l^i}{\sum_{i=1}^{L} \overline{f}^i + \sum_{i=L+1}^{M} \underline{f}^i}; \; y_l'' = y_l.$$

4. Якщо $y_l'' \neq y_l$, перейти на крок 5, інакше $y_l = y_l''$ та перейти на крок 6.

5. $y_l' = y_l''$, повернутися на крок 2.

Ширина отриманого інтервалу $[y_l; y_r]$ характеризує ступінь невизначеності, пов'язаної з прийнятим рішенням. У випадках, коли специфіка прикладної задачі вимагає подання результату у вигляді єдиного числа, за вихідне значення приймається середнє арифметичне границь інтервалу $[y_l; y_r]$: $y = (y_l + y_r) / 2$.

2.5 Інтервальні нечіткі множини в задачі кластеризації

Розв’язки задачі кластерного аналізу, отримані за допомогою методу можливісної кластеризації PCM, більше відповідають дійсності та інтуїтивному уявленню про природу кластерів. Не зважаючи на таке вдосконалення, одна проблема залишається спільною для FCM та PCM:

обидва методи в усіх обчисленнях спираються на параметр *m*, що задає рівень нечіткості кластерів.

Випадок *m = 1* відповідає чіткій кластеризації. Зі зростанням *m* ступені належності всіх без винятку точок до всіх кластерів наближаються до *0,5*, як показано на рис. 2.4 (для випадку двох кластерів). Кожна крива зображає зміну ступеня належності точки до одного з кластерів.

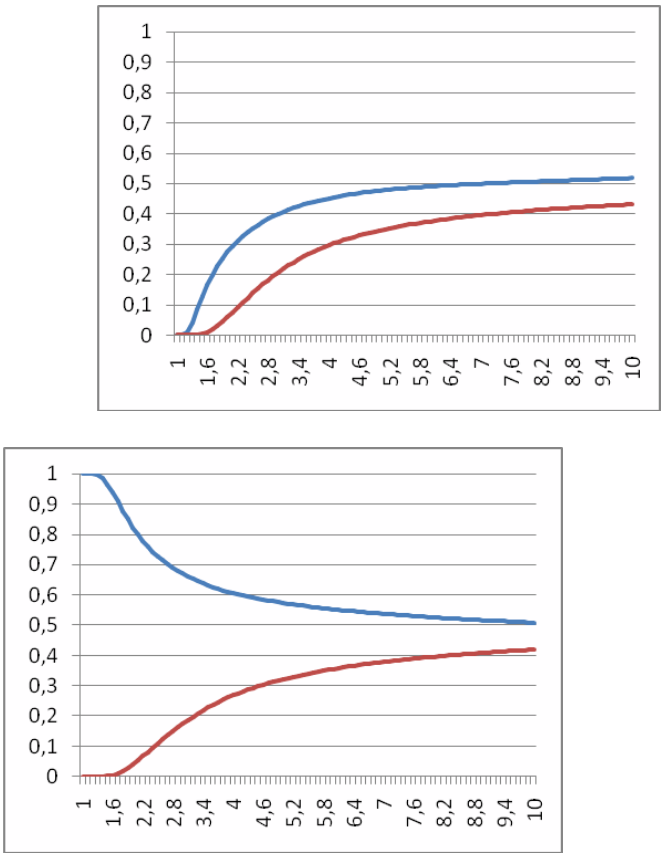


Рисунок 2.4 – Зміна ступенів належності точки до кластерів при зміні рівня нечіткості:

а) точка нерепрезентативна;

б) точка репрезентативна.

На рис. 2.4 видно, що в усіх випадках m змінюється монотонно та не утворює локальних екстремумів. Тому закономірно, що строго обґрунтованих механізмів визначення m не існує. Існує ряд критеріїв якості (validity indices) кластеризації, пряме призначення яких – кількісна оцінка якості результатів кластеризації як поєднання умов подібності об'єктів у межах одного кластера та відмінності об'єктів із різних кластерів. Найбільш традиційне їх застосування – визначення оптимального числа кластерів, але можливо використовувати їх і для оцінки впливу інших параметрів. Наведемо найпоширеніші критерії якості.

Індекс розбиття (Partition Index):

$$SC(c,m) = \sum_{i=1}^{c} \frac{\sum_{k=1}^{N} \left(\mu_{i,k}\right)^m \left\| x_k - v_i \right\|^2}{\sum_{k=1}^{N} \mu_{i,k} \sum_{j=1}^{c} \left\| v_j - v_i \right\|^2}$$

де μ_{ij} – ступінь належності точки j до кластера i;

v_j - центр j-го кластеру;

m – рівень нечіткості;

c – кількість кластерів;

N – кількість точок.

Критерії Хіе-Бені та Квона беруть до уваги геометричні властивості кластерів, а не лише відстані об'єктів до їхніх центрів.

Критерій Квона:

$$K(c,m)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^m\left\|x_j-v_i\right\|^2+\frac{1}{c}\sum_{i=1}^{c}\left\|v_i-\bar{v}\right\|^2}{\min_{i\neq j}\left\|v_i-v_j\right\|^2}$$

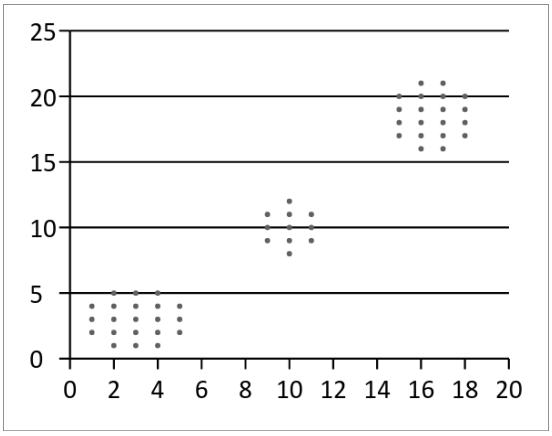
де $\bar{v}$ - середнє значення центрів кластерів.

Критерій Хіе-Бені:

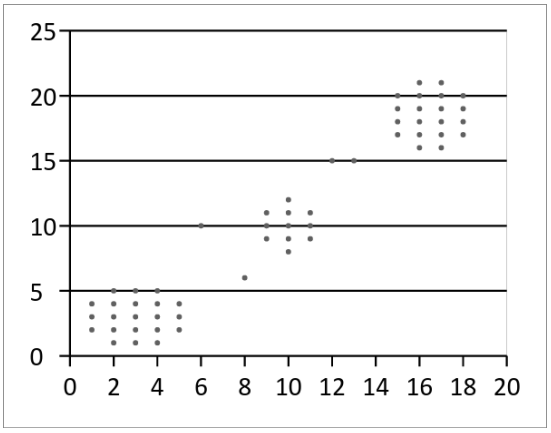
$$XB(c,m)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^m\left\|x_j-v_i\right\|^2}{N\min_{i,j}\left\|v_i-v_j\right\|^2}$$

Менші значення наведених критеріїв відповідають кращим варіантам розбиття. Число кластерів вважатимемо вихідним параметром, заданим заздалегідь. Критерії ж якості пропонується використовувати для визначення оптимального значення рівня нечіткості. Дослідимо характер зміни їхніх значень залежно від зміни рівня нечіткості *m* на прикладі тестових наборів даних, зображених на рис. 2.5. Випадок 2.5 (б) відрізняється від 2.5 (а) наявністю аномальних об’єктів (шуму).

Розв’яжемо поставлену задачу методом PCM при різних значеннях параметру *m* та обчислимо значення критеріїв якості за наведеними вище співвідношеннями. Для початкової ініціалізації центрів кластерів та ступенів належності будемо використовувати метод FCM Дж. Беждека.



а)



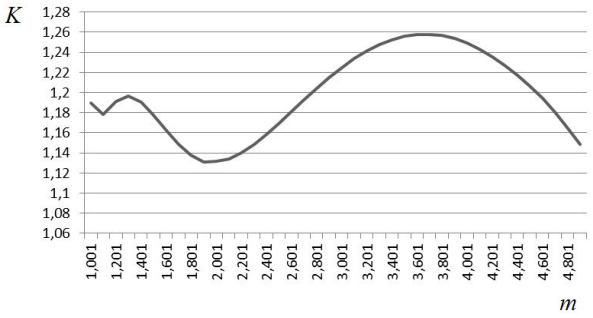
б)

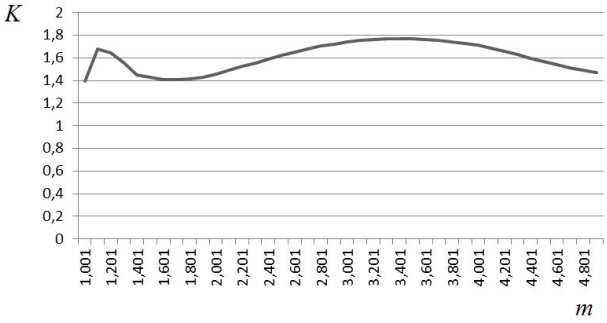
Рисунок 2.5 – Тестові набори даних:

а) ідеальний (без аномальних спостережень)

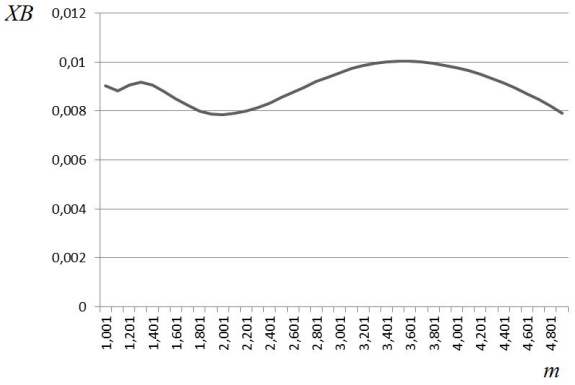
б) зашумлений (з аномальними спостереженнями)

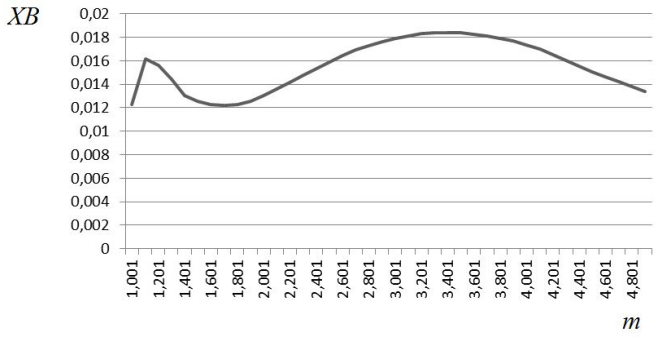
Залежності значення відповідного критерію від значення рівня нечіткості m наведено на рис. 2.6.


K
1,28
1,26
1,24
1,22
1,2
1,18
1,16
1,14
1,12
1,1
1,08
1,06
1,001
1,201
1,401
1,601
1,801
2,001
2,201
2,401
2,601
2,801
3,001
3,201
3,401
3,601
3,801
4,001
4,201
4,401
4,601
4,801
m

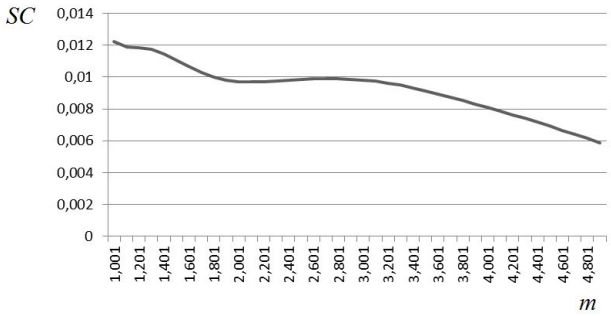

K
2
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
1,001
1,201
1,401
1,601
1,801
2,001
2,201
2,401
2,601
2,801
3,001
3,201
3,401
3,601
3,801
4,001
4,201
4,401
4,601
4,801
m

а) б)

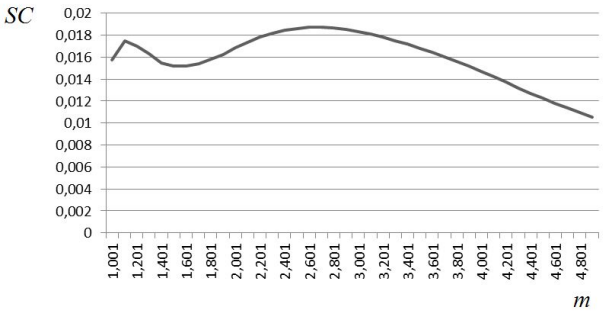

XB
0,012
0,01
0,008
0,006
0,004
0,002
0
1,001
1,201
1,401
1,601
1,801
2,001
2,201
2,401
2,601
2,801
3,001
3,201
3,401
3,601
3,801
4,001
4,201
4,401
4,601
4,801
m


XB
0,02
0,018
0,016
0,014
0,012
0,01
0,008
0,006
0,004
0,002
0
1,001
1,201
1,401
1,601
1,801
2,001
2,201
2,401
2,601
2,801
3,001
3,201
3,401
3,601
3,801
4,001
4,201
4,401
4,601
4,801
m

в) г)



д)



е)

Рисунок 2.6 – Характер зміни значень критеріїв якості залежно від значення рівня нечіткості:

а) критерій Квона, набір 1;

б) критерій Квона, набір 2;

в) критерій Хіе-Бені, набір 1;

г) критерій Хіе-Бені, набір 2;

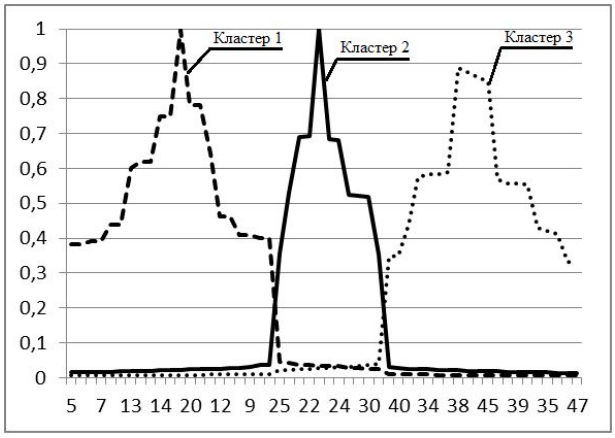
д) індекс розбиття, набір 1;

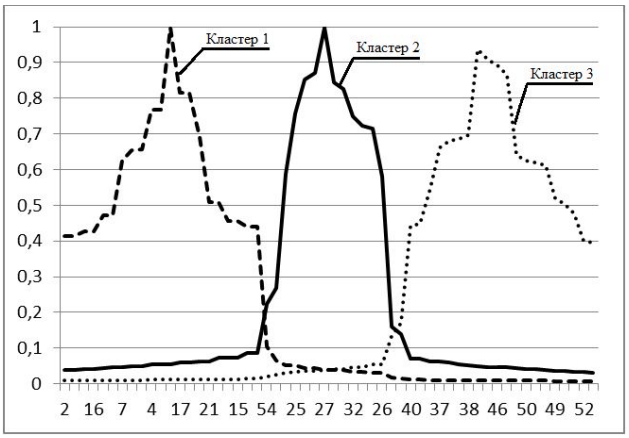
е) індекс розбиття, набір 2.

Випадки (а), (в) і (д) відповідно демонструють характер зміни значень критеріїв Квона, Хіе-Бені та індексу розбиття на наборі даних 2.5 (а), в якому відсутні аномальні спостереження. Випадки (б), (г) та (е) відповідають зашумленому набору даних 2.5 (б).

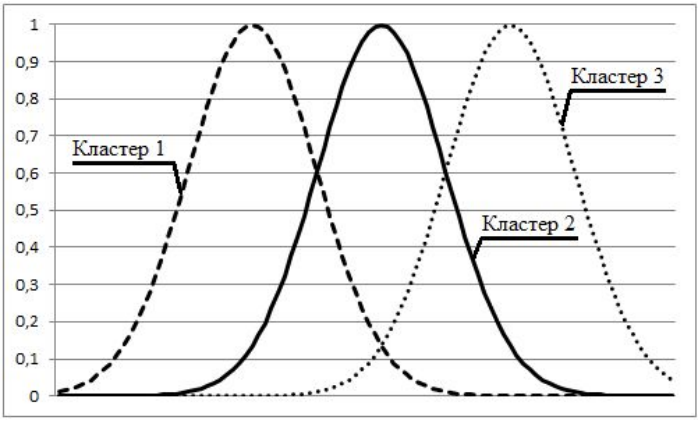
Якщо говорити про самі результати кластеризації (ступені належності точок до кластерів), то кожен із наведених критеріїв має глобальний або локальний мінімум, що відповідає рекомендованому значенню m. Прийнявши це значення за оптимальне, можна побудувати розв’язок, подібний до зображеного на рис. 2.7.

На рис. 2.7 (а) наведено розв’язок задачі кластеризації за значення $m = 2$, що відповідає мінімуму кривої зміни індексу Хіе-Бені на наборі даних 2.5 (а). Результати кластеризації представлено у вигляді нечітких множин: для кожного об’єкта на осі абсцис по осі ординат відкладено його ступені належності до кожного з кластерів. Набору даних 2.5 (б) відповідає розв’язок 2.7 (б). На рис. 2.7 (в) зображено класичні нечіткі множини типу 1.





а) б)



в)

Рисунок 2.7 – Результати кластеризації як нечіткі множини типу 1.

Отже, прийнявши значення рівня нечіткості $m = 2$, ми отримали деякий розв'язок задачі кластеризації для заданого набору даних. Але, проаналізувавши криві 2.6 а – е, можна помітити, що вони мають мінімуми за різних значень m. Так, окрім уже згаданого $m = 2$ у випадку 2.6 (в), маємо результати $m = 1,9$ для 2.6 (а), $m = 2,1$ для 2.6 (д), $m = 1,7$ для 2.6 (б) та (г), $m = 1,6$ для 2.6 (е). Таким чином, залежно від способу розв'язання задачі (критерію якості) та якості вхідних даних (наявності чи відсутності аномалій) нами отримано п'ять варіантів розв'язку, кожен із яких претендує на оптимальність. Шостим претендентом можна вважати значення $m = 1,5$, рекомендоване для даної задачі. Остаточно визначити, який із розв'язків-претендентів є вірним, і чи є серед них вірний розв'язок, не видається можливим. Тому для того, щоб убезпечити себе від помилкового результату, пов'язаного з неправильним вибором значення m, доцільно використовувати інтервальні ступені належності.

Рівень нечіткості m, як правило, задається емпірично дослідником, при цьому доводиться повністю покладатися на це заздалегідь задане значення без жодних гарантій його правильності. З цим пов'язана невизначеність, яку неможливо врахувати, коли отримане значення міри належності точки до кластера являє собою єдине число. Тому для того, щоб убезпечити себе від помилкового результату, пов'язаного з неправильним вибором значення m, доцільно використовувати інтервальні функції належності типу 2. Такий підхід найчастіше застосовується тоді, коли точний характер розподілу ступенів належності другого типу в області між границями інтервалу невідомий. Саме такий випадок являє собою задача кластеризації: невідомо, чи піддається виділенню та математичному опису закономірність, за якою розподілені ступені належності другого типу, та чи має дослідження цієї закономірності практичний сенс. З іншого боку, інформація про верхню та нижню функції належності, що описують кожен кластер залежно від значення параметру m, має виняткову цінність, оскільки інтервал (його ширина та розташування відносно нуля та одиниці) несе значно більше інформації про міру належності точки до кластера, ніж єдине число. Наприклад, ширина інтервалу може свідчити про ступінь достовірності отриманого розв'язку. Тому пропонується модифікувати алгоритм можливісної кластеризації PCM для роботи з інтервальними ступенями належності. Цим буде досягнуто повне врахування невизначеності, пов'язаної з різними можливими значеннями рівня нечіткості, для подальшого аналізу результатів кластеризації.

Адаптуємо метод для роботи з інтервальними ступенями належності точок до кластерів. В основі пропонованого методу лежить алгоритм

кластеризації PCM. Окрім нетрадиційного трактування ступенів належності та стійкості до шуму він володіє ще однією властивістю. Йдеться про те, що, оскільки міри належності однієї й тієї самої точки до різних кластерів незалежні одна від одної, ступінь належності точки до одного з них можна змінити без обов'язкової процедури перерахунку ступенів її належності до всіх інших кластерів. Дана властивість є надзвичайно корисною, оскільки вона дає змогу «розтягти» ступінь належності точки до кластера з чіткого значення в інтервал, і це не ставить під загрозу виконання обмеження на суму значень ступенів належності точки до всіх наявних кластерів.

Отже, виходячи з визначення ступеня належності як міри типовості заданої точки для відповідного кластеру, знайдемо такі значення невідомих параметрів, які ведуть до мінімуму функціоналу (1.2). Враховуючи властивості рівня нечіткості m та його вплив на результати кластерного аналізу, необхідно подати ступені належності у вигляді інтервалів, ліва та права границі яких лежать у межах $[0, 1]$.

Для визначення меж інтервалу розтягу ступеня належності скористаємося критеріями кластеризації. Перший локальний мінімум критеріїв K, SC та XB відповідає оптимальному значенню m, тому за межі інтервалу зміни рівня нечіткості приймемо точки перегину кривих 2.6 (а) та (б) по обидва боки від цієї точки. Тоді ступінь належності точки до кожного з кластерів буде лежати в межах $[\mu_{ijL}, \mu_{ijU}]$, де границями інтервалу є відповідні значення μ_{ij} у згаданих точках перегину.

Для початкової ініціалізації центрів кластерів використаємо звичайний метод FCM. Він збігається за лічені ітерації, тому якнайкраще підходить для цього завдання, адже воно вимагає грубого наближеного розв'язку.

Виходячи з усього сказаного вище, можна сформулювати покроковий алгоритм розв'язання задачі кластерного аналізу.

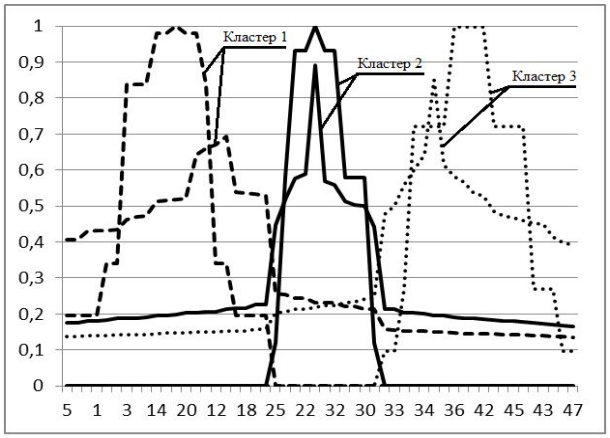
1. Глобальні значення критеріїв якості ініціалізувати максимально можливим значенням.
2. Визначити приблизні місця розташування центрів кластерів за допомогою алгоритму FCM.
3. Оцінити значення η для результату роботи FCM.
4. Сформувати матрицю D як матрицю Евклідових відстаней від кожної точки з вихідної множини до центру кожного з кластерів.
5. Задати початкове значення рівня нечіткості $m = 1$.
6. Розрахувати початкове значення локальних критеріїв якості.
7. Розрахувати функцію належності для кожної з пар (точка, кластер), користуючись відповідним співвідношенням із методу PCM.
8. Перерахувати положення центрів кластерів за формулою, спільною для обох методів.
9. Перерахувати матрицю відстаней D.
10. Розрахувати цільову функцію PCM при заданих значеннях ступенів належності, координат центрів кластерів, елементів матриці D та вектора η.
11. Якщо розраховане значення цільової функції менше за отримане на попередній ітерації, повернутись до кроку 8.
12. Розрахувати значення локальних критеріїв якості при заданому m. Знаючи поточне значення критерію та його значення з попередньої ітерації, визначити, чи є попереднє значення локальним екстремумом. Якщо так, зберегти попереднє значення m, характер

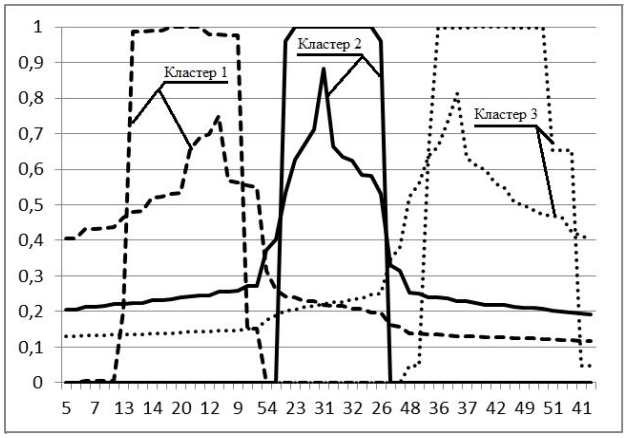
екстремуму відповідного критерію в цій точці та побудоване розбиття (центри кластерів і ступені належності).

13. Перерахувати значення глобальних критеріїв якості.
14. Серед усіх проміжних результатів вибрати три значення *m,* за яких глобальні критерії досягають мінімального значення. Для кожного з них серед збережених на кроці 13 екстремальних точок вибрати найближчі справа та зліва. Побудувати інтервали зміни значень рівня нечіткості за кожним критерієм: $\tilde{m}_K = [m_{Kl}; m_{Kr}]$, $\tilde{m}_{XB} = [m_{XBl}; m_{XBr}]$ та $\tilde{m}_{SC} = [m_{SCl}; m_{SCr}]$.
15. Об'єднати інтервали рівня нечіткості за правилом

$$\tilde{m} = \tilde{m}_K \cup \tilde{m}_{XB} \cup \tilde{m}_{SC} = [\min(m_{Kl}, m_{XBl}, m_{SCl}); \max(m_{Kr}, m_{XBr}, m_{SCr})] \quad (3.6)$$

Інтервали зміни рівня нечіткості, отримані після кроку 16 для кожного з критеріїв якості, можна використовувати напряму для розрахунку ступенів належності. Тоді отримаємо кластери у вигляді інтервальних нечітких множин, як показано на рис. 2.8 для критерію Квона. Випадок (а) відповідає набору 2.5 (а), випадок (б) – набору 2.5 (б).





а) б)

Рисунок 2.8 – Результати інтервальної нечіткої кластеризації на основі критерію Квона

Ширина інтервалу рівня нечіткості відображається також на значення ступенів належності. Хоча в результатах у цілому прослідковується тенденція до утворення компактних кластерів, деякі об’єкти мають значну невизначеність у ступенях належності. Серед таких об’єктів, наприклад, точка 20 ($\mu_{20,2}$ = *[0,52; 0,98]*). Ця невизначеність є природною особливістю, характерною для будь-якої емпіричної оцінки.

Аномальні спостереження, присутні в наборі 2.5 (б), також вносять невизначеність у результати кластеризації. Але найбільшим недоліком цього методу є те, що сам критерій якості, що використовується, вносить певну невизначеність у результати та генерує ступені належності з широким інтервалом навіть для «чистого» набору даних, апріорі вільного від аномальних спостережень. Про це свідчать розбіжності оцінок, отриманих за допомогою різних критеріїв.

Для того, щоб зменшити вплив внутрішніх особливостей одного конкретного критерію якості на кінцевий результат, до формування інтервалу залучаються інші критерії. В постановці задачі кластерного аналізу як задачі навчання без учителя єдино вірний розв'язок завжди невідомий, тому для гарантованого попадання вірного розв'язку в остаточний інтервал необхідно виконувати об'єднання інтервалів, отриманих за критеріями Квона, Хіе-Бені та індексом розбиття.

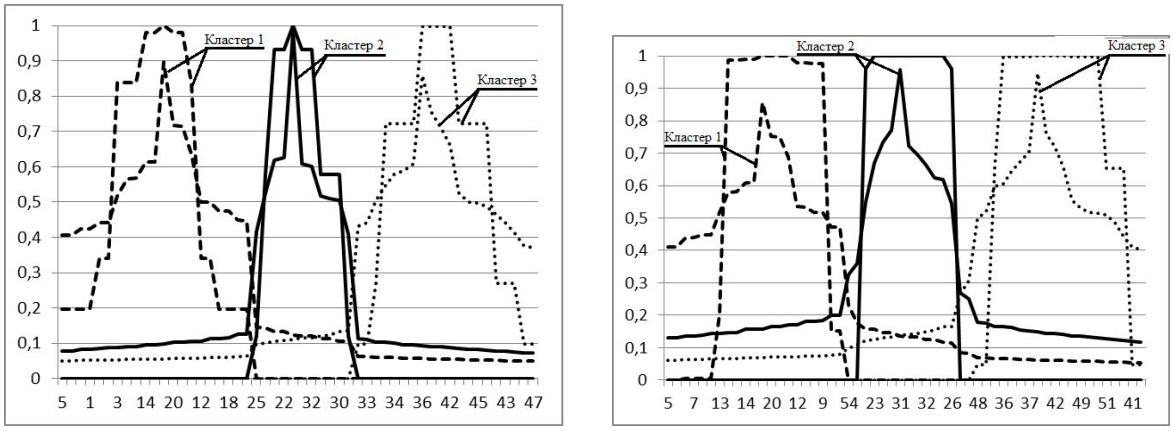
Так, у наведеному тестовому прикладі для набору 2.5 (а)

$$\tilde{m} = \tilde{m}_K \cup \tilde{m}_{XB} \cup \tilde{m}_{SC} = [1{,}3; 3{,}7] \cup [1{,}3; 3{,}5] \cup [1{,}3; 2{,}7] = [1{,}3; 3{,}7]$$

Для набору 2.5 (б):

$$\tilde{m}_б = \tilde{m}_{K_б} \cup \tilde{m}_{XB_б} \cup \tilde{m}_{SC_б} = [1{,}1; 3{,}4] \cup [1{,}1; 3{,}4] \cup [1{,}1; 2{,}7] = [1{,}1; 3{,}4].$$

Інтервальні нечіткі кластери, отримані за таких значень рівня нечіткості, наведено на рис. 2.9.



а) б)

Рисунок 2.9 – Результати інтервальної нечіткої кластеризації на основі комбінації критеріїв Квона, Хіе-Бені та індексу розбиття.

Таким чином, скомбінувавши три критерії якості та визначивши область об’єднання інтервалів, отриманих за кожним із них, бачимо, що значення рівня нечіткості для зашумлених даних набору 2.5 (б) знаходяться приблизно в тому ж інтервалі, що й для позбавленого аномалій набору даних 2.5 (а), що ще раз підтверджує незалежність результатів роботи методу від наявності в аналізованому наборі даних шумів та аномалій.

Отже, інтервальні ступені належності дають змогу враховувати та моделювати невизначеності, пов’язані з браком експертних знань, характерним для моделей на основі навчання без учителя, таких як кластерний аналіз.

3 Список рекомендованої літератури

1. Алтунин А. Е., Семухин М. В. Модели и алгоритмы принятия решений в нечетких условиях. Тюмень : Издательство Тюменского государственного университета, 2000. 352 с.
2. Блюмин С. Л., Шуйкова И. А. Модели и методы принятия решений в условиях неопределенности. Липецк : ЛЭГИ, 2001. 138 с.
3. Борисов А. Н., Крумберг О. А., Федоров И. П. Принятие решений на основе нечетких моделей: Примеры использования. Рига : Зинатне, 1990. 184 с.
4. Вятченин Д. А. Нечеткие методы автоматической классификации: Монография. Мн. : УП «Технопринт», 2004. 219 с.
5. Гитис Л. Х. Статистическая классификация и кластерный анализ. – М. : МГГУ, 2003. 57 с.
6. Дубовой В. М., Глонь О. В. Моделювання систем керування в умовах невизначеності: монографія. Вінниця : УНІВЕРСУМ-Вінниця, 2004. 169 с.
7. Дюран Б., Оделл П. Кластерный анализ. – М. : Статистика, 1977. 128 с.
8. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. М. : Мир, 1976. 163 с.
9. Ивахненко А. Г. Моделирование сложных систем по экспериментальным данным. – М. : Радо и связь, 1987. 120 с.
10. Мандель И. Д. Кластерный анализ. М. : Статистика, 1988. 176 с.
11. Минаев Ю. Н. Методы и алгоритмы решения задач идентификации и прогнозирования в условиях неопределенности в нейросетевом логическом базисе. М. : Горячая Линия-Телеком, 2003. 205 с.

12. Мокін Б. І., Мокін В. Б., Мокін О. Б. Математичні методи ідентифікації динамічних систем: навчальний посібник. Вінниця : ВНТУ, 2010. 260 с.
13. Нариньяни А. С. Недоопределенность в системе представления и обработки знаний // Изв. АН СССР. Тех. кибернетика. 1986. № 5. С. 3–28.
14. Олдендерфер М. С., Блешфилд Р. К. Кластерный анализ // Факторный, дискриминантный и кластерный анализ. М. : Финансы и статистика, 1989. С. 139-215.
15. Олизаренко С. А., Перепелица А. В., Капранов В. А. Нечеткие логические системы интервального типа 2. Архитектура и механизм вывода // Системи обробки інформації. 2011. №5 (95). С. 156-164.
16. Ротштейн А. П. Интеллектуальные технологии идентификации: нечеткие множества, генетические алгоритмы, нейронные сети. Винница : Универсум-Винница, 1999. 320 с.
17. Целых А. Н., Тимошенко Р. П. Некоторые теоретико-множественные операции над интервальными нечеткими множествами в моделях искусственного интеллекта // Перспективные информационные технологии и интеллектуальные системы. 2001. № 2. С. 69-76.
18. Wang L. X., Mendel J. M. Generating Fuzzy Rules from Numerical Data, with Applications // Signal and Image Processing Institute, University of Southern California, Department of Electrical Engineering-Systems, 1991. 63 p