Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

САМІТОВА ВІКТОРІЯ ОЛЕКСАНДРІВНА

УДК 004.032.26

КЛАСИФІКАЦІЯ ТА КЛАСТЕРИЗАЦІЯ ДАНИХ, ЩО ЗАДАНІ В НЕЧИСЛОВИХ ШКАЛАХ

05.13.23 – системи та засоби штучного інтелекту

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2017

Дисертацією є рукопис.

Робота виконана в Харківському національному університеті радіоелектроніки Міністерства освіти і науки України.

Науковий керівник – доктор технічних наук, професор
Бодянський Євгеній Володимирович,
Харківський національний університет радіоелектроніки, професор кафедри штучного інтелекту.

Офіційні опоненти: доктор технічних наук, професор
Литвиненко Володимир Іванович,
Херсонський національний технічний університет, МОН України, завідувач кафедри інформатики та комп'ютерних наук;

доктор технічних наук, професор
Субботін Сергій Олександрович,
Запорізький національний технічний університет МОН України, завідувач кафедри програмних засобів.

Захист відбудеться « 22 » березня 2017 р. о 15.30 годині на засіданні спеціалізованої вченої ради Д 64.052.01 Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Науки, 14.

З дисертацією можна ознайомитись у бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, пр. Науки, 14.

Автореферат розісланий « 06» лютого 2017 р.

Учений секретар
спеціалізованої вченої ради,
д.т.н., проф. О.А. Винокурова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Актуальними методами інтелектуального аналізу даних є кластеризація та класифікація. Часто у таких галузях, як медицина, соціологія, освіта дані можуть бути представлені у нечислових шкалах. Для вирішення задач кластеризації та класифікації даних у нечислових шкалах найбільш популярним є підхід, що базується на заміні лінгвістичних характеристик їх рангами. Однак у більшості випадків цей прийом виявляється некоректним, оскільки припускає, що відстані між сусідніми рангами є однаковими, що не завжди відповідає дійсності. Крім того, традиційні методи і підходи не здатні працювати у послідовному режимі опрацювання даних в умовах кластерів, що перетинаються. Тому на сьогоднішній день актуальною є розробка методів нечіткої кластеризації та класифікації даних, заданих у нечислових шкалах, що дозволяють подолати обмеження існуючих підходів.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана в рамках НДР «Нейро-фаззі системи для поточної кластеризації та класифікації послідовностей даних за умов їх викривленості відсутніми та аномальними спостереженнями» (№ ДР 0113U000361) та «Динамічний інтелектуальний аналіз послідовностей нечіткої інформації за умов суттєвої невизначеності на основі гібридних систем обчислювального інтелекту» (№ДР 0116U002539), які виконувалися згідно наказів Міністерства освіти і науки України за результатами конкурсного відбору проектів наукових досліджень. В рамках зазначених НДР здобувачкою як виконавцем розроблено: методи нечіткої кластеризації даних, що задані у порядковій шкалі; методи нечіткої кластеризації даних, що задані у категоріальній шкалі; архітектура гібридної системи обчислювального інтелекту для класифікації порядкових даних.

Мета та задачі дослідження. Метою дослідження є розробка методів нечіткої кластеризації та класифікації даних, що задані у нечислових шкалах.

Згідно поставленої мети необхідно вирішити такі наукові задачі:

- проаналізувати існуючі методи та підходи інтелектуальної обробки даних, що представлені у нечислових шкалах;
- синтезувати адаптивні та робастні методи нечіткої кластеризації порядкових даних;
- синтезувати метод можливісної нечіткої кластеризації даних, що задані у категоріальній шкалі;
- синтезувати архітектуру гібридної системи обчислювального інтелекту для класифікації порядкових даних та метод її навчання;
- за допомогою розроблених методів вирішити низку практичних задач інтелектуального аналізу даних.

Об'єктом дослідження є процес нечіткої кластеризації та класифікації даних, що задані у нечислових шкалах.

Предметом дослідження є методи нечіткої кластеризації та класифікації даних, що задані у нечислових шкалах, з використанням гібридних систем обчислювального інтелекту.

Методи дослідження. Основні результати роботи отримані за допомогою теорії оптимізації і статистичного аналізу, що дала можливість знаходити приховані

закономірності в інформації, теорії нечіткої кластеризації, що дозволила синтезувати адаптивні та робастні методі кластеризації даних у нечислових шкалах в умовах кластерів, що перетинаються, теорії штучних нейронних мереж, що дозволила синтезувати архітектуру нейро-фаззі системи для класифікації порядкових даних та методів її навчання, а також імітаційного моделювання, що дозволило визначити ефективність розроблених методів та архітектур.

Наукова новизна отриманих результатів.

1. Вперше запропоновано метод нечіткої кластеризації порядкових даних на основі частотних прототипів та функцій належності, що дало можливість обробляти порядкові дані, які не підпорядковуються нормальному розподілу.
2. Вперше запропоновано метод нечіткої кластеризації порядкових даних на основі порядково-цифрового відображення, що дало можливість обробляти порядкові дані у послідовному режимі.
3. Удосконалено метод нечіткої кластеризації порядкових даних шляхом сумісного використання функцій належності і функцій правдоподібності, що дозолило підвищити точність кластеризації порядкових даних.
4. Удосконалено метод можливісної нечіткої кластеризації масивів категоріальних даних шляхом сумісного використання частотних прототипів і мір несхожості, що дозволило подолати недоліки класичних методів та підвищити точність кластеризації даних.
5. Удосконалено нейро-фаззі систему на основі нео-фаззі нейрону шляхом використання додаткового вихідного шару, що дозволило обробляти дані, задані у порядковій шкалі.
6. Отримали подальший розвиток методи адаптивної робастної нечіткої кластеризації порядкових даних шляхом використання критерію спеціального виду (міри схожості), що дозволило обробляти порядкові дані, які містять викиди, у послідовному режимі.

Практичне значення отриманих результатів. Запропоновані в роботі методи нечіткої кластеризації та класифікації даних, що задані у нечислових шкалах, реалізовані у вигляді програмних засобів, що дозволило автоматизувати процес обробки даних у порядковій та категоріальній шкалах для вирішення задач інтелектуального аналізу даних. Розроблені методи та архітектури дозволяють опрацьовувати дані у послідовному режимі і підвищують точність їх кластеризації. Імітаційне моделювання отриманих теоретичних результатів довело перевагу над існуючими методами.

Результати дисертаційної роботи були впроваджені у ТОВ «Південелектропроект», м. Харків, для розв’язання задачі аналізу клієнтської бази даних підприємства (акт впровадження від 25.05.2016), а також у ТОВ НВП «Мідіел», м. Харків, для розв’язання задачі діагностування електрообладнання (акт впровадження від 20.05.2016).

Особистий внесок здобувача. Усі положення, що виносяться на захист, основні результати теоретичних та експериментальних досліджень отримані здобувачкою особисто. Внесок авторки в публікаціях, написаних у співавторстві такий: [1] – розроблено метод нечіткої кластеризації порядкових даних на основі частотних прототипів та функцій належності, [2] – запропоновано метод нечіткої кластеризації даних, що задані у порядковій шкалі, на основі сумісного використання функцій належності та правдоподібності, [4] – розроблена архітектура нейро-фаззі

системи на основі подвійного нео-фаззі нейрона для обробки порядкових даних, [5] – запропоновано адаптивні методи робастної нечіткої кластеризації порядкових даних на основі мір схожості, [6] – запропоновано можливісний метод нечіткої кластеризації масивів категоріальних даних з використанням частотних прототипів та мір несхожості.

Робота [3] опублікована без співавторів.

Апробація результатів дисертації. Основні результати дисертаційної роботи були представлені та обговорені на міжнародних наукових конференціях: 12-му, 19-му, 20-му Міжнародних молодіжних форумах «Радіоелектроніка та молодь в XXI столітті» (Харків, 2007-2016 рр.); міжнародній науково-технічній конференції «Системний аналіз та інформаційні технології» (Київ, 2007 р.); міжнародній науково-технічній конференції «Поліграфічні, мультимедійні та web-технології» (Харків, 2016 р.); XII міжнародній науковій конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту ISDMCI'2016» (Залізний порт, 2016 р.).

Публікації. Основні положення дисертаційної роботи опубліковані в 12 наукових роботах: 6 статтях у періодичних фахових виданнях з технічних наук, що входять до міжнародних наукометричних баз (у тому числі 5 статей у виданнях, що включено до переліків МОН України, 1 статтю видано за кордоном), 6 публікаціях у працях міжнародних наукових конференцій та форумів.

Структура та обсяг дисертації. Дисертація складається із вступу, п'яти розділів, висновків, що містять основні результати, списку використаних джерел і додатку. Загальний обсяг дисертації складає 139 сторінки (з них 118 – основного тексту), містить 30 рисунків, 5 таблиць, список використаних джерел, що включає 134 найменування та займає 13 сторінок, 1 додаток на 4 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету і задачі дослідження, наукову новизну і практичне значення одержаних результатів. Наведено відомості про впровадження результатів роботи, апробацію, особистий внесок здобувачки.

У **першому розділі** виконано огляд стану проблеми класифікації та кластеризації даних, що задані у нечислових шкалах, і розглянуто існуючі підходи до їх вирішення. Розглянуто основні принципи нечіткої логіки та систем нечіткого висновування. Проведено аналіз відомих архітектур нейро-фаззі систем, що дістали найбільшого поширення в задачах інтелектуального аналізу даних. Показано, що існуючі архітектури та методи мають свої недоліки і переваги та містять обмеження при розв'язанні задач, де спостереження надходять на опрацювання у послідовному режимі за умов дефіциту апріорної та поточної інформації, а також у випадку, коли дані не підпорядковуються нормальному розподілу.

На основі проведеного аналізу визначено мету та задачі дослідження, що полягають у розробці архітектур та методів нечіткої класифікації та кластеризації даних, що задані у нечислових шкалах.

Другий розділ присвячено розробці методів нечіткої кластеризації порядкових даних. Класичні методи кластеризації таких даних пропонують заміняти лінгвістичні характеристики числами, однак такий підхід припускає рівність відстаней між

сусідніми числовими рангами, що не завжди відповідає дійсності. Більш природним виглядає підхід, в основі якого лежить фаззіфікація порядкових даних і подальше використання методів нечіткої кластеризації.

Запропоновано метод, в основі якого лежить використання частотних прототипів та функцій належності. Розглянемо метод фаззіфікації послідовності лінгвістичних змінних на прикладі одновимірної вибірки $x(1), x(2), ..., x(N)$, де кожному із спостережень $x(k)$ може бути приписаний один із рангів l $l = 1, 2, ...m$.

Розрахуємо відносні частоти появи l-го рангу

$$f_l = \frac{N_l}{N} \tag{1}$$

и накопичені частоти

$$F_1 = \frac{f_1}{2}, \quad F_l = \frac{f_l}{2} + \sum_{s=1}^{l-1} f_s, l = 2, 3, ..., m, \tag{2}$$

при цьому виконується умова

$$\sum_{l=1}^{m} f_l = 1.$$

На основі накопичених частот формуються центри функцій належності $\mu_l(x)$ за допомогою рекурентного співвідношення

$$c_1 = 0,5F_1, \; c_l = c_{l-1} + 0,5(\mathrm{F}_{l-1} + F_l), \; l = 2, 3, ...m, \tag{3}$$

а самі функції належності задаються у формі

$$\begin{cases} \mu_1(x) = 1, \; x \in [0, c_1], \\ \mu_l(x) = \begin{cases} \dfrac{x - c_{l-1}}{c_l - c_{l-1}}, \; x \in [c_{l-1}, c_l], \\ \dfrac{c_{l+1} - x}{c_{l+1} - c_l}, \; x \in [c_l, c_{l+1}], \\ 0, \; x \notin [c_{l-1}, c_{l+1}], \end{cases} \\ \mu_m(x) = 1, \; x \in [c_m, 1]. \end{cases} \tag{4}$$

Такий спосіб задання функцій належності автоматично забезпечує розбиття Руспіні, тобто виконання умови

$$\sum_{l=1}^{m} \mu_l(x) = 1.$$

Розглянемо дві сусідні функції належності $\mu_l(x)$ та $\mu_{l+1}(x)$. Використовуючи поняття α - розрізу у вигляді $A_\alpha = \{x \in X : \mu(x) \geq \alpha\}$, можна ввести області впливу двох сусідніх рангів у формі

$$\begin{cases} A_l^R = \{x \in [c_l, c_l + 0{,}5\, f_l] : \mu_l(x) \geq \alpha_l^R = 1 - 0{,}5 \dfrac{f_l}{c_{l+1} - c_l}, \\ A_{l+1}^L = \{x \in [c_{l+1} - 0{,}5\, f_{l+1}, c_{l+1}] : \mu_{l+1}(x) \geq \alpha_{l+1}^L = 1 - 0{,}5 \dfrac{f_{l+1}}{c_{l+1} - c_l}, \end{cases} \quad (5)$$

де R та L означають праву та ліву сторони сусідніх функцій належності.

У процесі фаззіфікації багатовимірної вибірки даних формується nm функцій належності з центрами c_{jl}.

Процес кластеризації багатовимірної вибірки даних наведено на рис. 1. Після обчислення відстаней між $x(k)$ та усіма центроїдами c_i $d(x(k), c_i) = \| x(k) - c_i \|$, можна визначити рівень належності $u_i(k)$ вектора $x(k)$ i-му кластеру відповідно з FCM-методом у вигляді

$$u_i(k) = \frac{\| x(k) - c_i \|^{-2}}{\sum_{t=1}^{m} \| x(k) - c_t \|^{-2}} = \frac{d^{-2}\left(x(k), c_i\right)}{\sum_{t=1}^{m} d^{-2}\left(x(k), c_t\right)}. \quad (6)$$

Обмеженням такого підходу є те, що кожне спостереження «розмивається» по усім кластерам, що у порядковій шкалі веде до втрати фізичного змісту. У зв'язку з цим, доцільно після розрахунку усіх відстаней $d\left(x(k), c_i\right)$ провести їх ранжування за збільшенням і вибрати дві найменші відстані. Далі скористаємося виразом (6), беручи до уваги лише такі відстані.

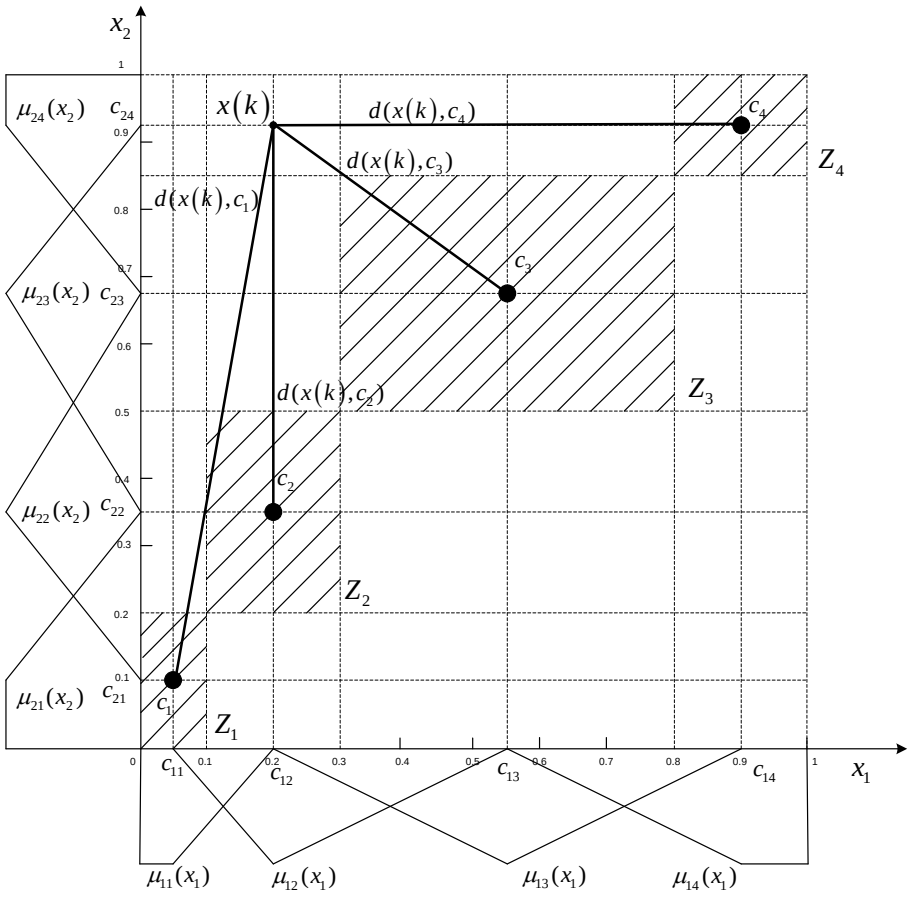

Рисунок 1 – Нечітка кластеризація рангових змінних

В основі класифікації за прототипами лежить порівняння їх схожості (similarity). Схожість між двома спостереженнями одного типу являє собою агрегацію схожостей між параметрами цих двох спостережень.

Вихідною інформацією для вирішення задачі нечіткої класифікації є вибірка спостережень, що сформована з N n-вимірних векторів ознак $X = \{x(1), x(2), ..., x(k), ..., x(N)\}$, де $k = 1, ..., N$. Нехай задана порядкова ознака і $L = \{l_1, ..., l_m\}$ - це множина можливих значень даної ознаки, яка задовольняє умові $l_1 < l_2 < ... < l_m$.

Для кожного значення l_s нехай існує підмножина об'єктів $X_s \subseteq X$, що включає в себе l_s. Схожість між l_s і l_t визначається як середнє схожості між об'єктами в X_s та X_t відповідно

$$sim(U, l_s, l_t) = sim(U, X_s, X_t), \forall s = 1...m; t = 1...m, s \neq t.$$

Схожість між двома підмножинами X, що не перетинаються, визначається наступним чином:

$$\forall A, B \subseteq X \ э \ A \cap B \neq \varnothing,$$

$$sim(U, A, B) = \frac{\sum_{x \in A; y \in B} sim(U, x, y)}{|A||B|},$$

де $sim(U, x, y)$ – це схожість між $x \in X$ і $y \in X$.

Схожість між двома спостереженнями x та y в X визначається контекстно-орієнтовною близькістю між x та y

$$sim(U, x, y) = prox(U, x, y),$$

а контекстно-орієнтовна близькість між $x(j)$ та $x(k)$ у відношенні U, визначається наступним чином:

$$prox\big(U, x(j), x(k)\big) = \sum_{i=1}^{r} \min\big(u_i(j), u_i(k)\big).$$

Введемо для зручності дві додаткові характеристики l_0 та l_{m+1} такі, що $l_0 < l_1$, а $l_m < l_{m+1}$. Тоді порядково-чисельне відображення g у даній множині об'єктів X описується виразом

$$\begin{cases} g(l_0) = 0, \\ g(l_l) = \sum_{s=1}^{l} \dfrac{1 - sim(U, l_{s-1}, l_s)}{\sum_{t=1}^{m+1} \big(1 - sim(U, l_{t-1}, l_t)\big)}, \forall l = 1...m, \\ g(l_{m+1}) = 1, \end{cases} \quad (7)$$

де

$$\begin{cases} sim(U, l_0, l_1) = sim(U, X_1, X - X_1), \\ sim(U, l_{l-1}, l_l) = sim(U, X_s, X_t), \forall l = 2...m, \\ sim(U, l_m, l_{m+1}) = sim(U, X_m, X - X_m). \end{cases}$$

Задача кластеризації числових характеристик вирішується шляхом мінімізації цільової функції

$$E(u_i, c_i) = \sum_{k=1}^{N} \sum_{i=1}^{r} u_i^{\beta}(k) d^2\left(x(k), c_i\right), \tag{8}$$

за обмежень

$$\begin{cases} u_i(k) \geq 0, \forall i = 1,2,...,r; \forall k = 1,2,...,N, \\ \sum_{i=1}^{r} u_i(k) = 1, \forall k = 1,2,...,N, \\ \sum_{k=1}^{N} u_i(k) > 0, \forall i = 1,2,...,r, \end{cases} \tag{9}$$

де N – обсяг вибірки даних, r – кількість кластерів, $k = 1,2,...,N$ – номер спостереження, c_i – прототип (центроїд) i-го кластеру, $d\left(x(k), \mathrm{c}_i\right)$ – відстань між прототипом i-го кластеру і k-м спостереженням, $u_i(k) \in [0,1]$ – рівень належності вектора $x(k)$ до i-го кластеру, β – невід'ємний параметр фаззіфікації (фаззіфікатор), що визначає розмитість границь між кластерами.

Задача пошуку сідлової точки функції Лагранжа може бути зведена до задачі пошуку сідлової точки її локальної модифікації. У зв'язку з цим введемо локальну модифікацію лагранжіана

$$L_S\left(u_i(k), c_i, \lambda(k)\right) = \sum_{i=1}^{r} u_i^{\beta}(k) d^2\left(x(k), c_i\right) + \lambda(k)\left(\sum_{i=1}^{r} u_i(k) - 1\right), \tag{10}$$

де $\lambda(k)$– невизначений множник Лагранжа.

Використовуючи далі метод оптимізації Ерроу-Гурвіца-Удзави, отримуємо рекурентну процедуру

$$\begin{cases} u_i(k) = \dfrac{\left(d^2\left(x(k), c_i(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum_{t=1}^{r}\left(d^2\left(x(k), c_t(k)\right)\right)^{\frac{1}{1-\beta}}}, \\ c_i(k+1) = c_i(k) - \eta(k) \nabla_{c_i} L(k)\left(u_i(k), \mathrm{c}_i(k), \lambda(k)\right) = \\ = c_i(k) - \eta(k) u_i^{\beta}(k) d\left(x(k+1), c_i(k)\right) \nabla_{c_i} d\left(x(k+1), c_i(k)\right). \end{cases} \tag{11}$$

Процедура (11) за структурою близька до методу нечіткого конкурентного навчання Чанга-Лі, а у випадку, коли β=2 – до градієнтного методу нечіткої кластеризації Парка-Деггера

$$\begin{cases} u_i(k) = \dfrac{\| x(k) - c_i(k) \|^{-2}}{\sum\limits_{t=1}^{r} \| x(k) - c_t(k) \|^{-2}}, \\ c_i(k+1) = c_i(k) + \eta(k) u_i^2(k) \big(x(k+1) - c_i(k) \big). \end{cases} \tag{12}$$

Використовуючи порядково-чисельне відображення за допомогою рекурентної нечіткої кластеризації, отримуємо метод, що здатний обробляти порядкові дані в онлайн режимі.

Ідея наступного методу нечіткої кластеризації порядкових даних полягає у використанні правдоподібності спостережень замість квадрата евклідової відстані у методі нечітких с-середніх.

Задача вирішується шляхом максимізації цільової функції

$$Q = \sum_{k=1}^{N} \sum_{i=1}^{r} u_i^{\beta}(k) L_i(k), \tag{13}$$

або мінімізації

$$Q = \sum_{k=1}^{N} \sum_{i=1}^{r} u_i^{\beta}(k) U_i(k), \tag{14}$$

де $L_i(k)$ - правдоподібність належності k - го спостереження до i - го кластера, $U_i(k)$ – логарифм несхожості k -го спостереження з i -м кластером.

Правдоподібність $L_i(k)$ в (13) розраховується згідно з формулою

$$L_i(k) = \prod_{j=1}^{n} p_{ij}(k), \tag{15}$$

де $p_{ij}(k)$ – умовна ймовірність появи визначеного значення j - ої характеристики k -го спостереження у i - ому кластері.

При цьому, оскільки $p_{ij}(k)$ залежить від частоти зустрічаємості визначеного значення характеристики у виборці, а ознаки розташовані згідно порядку, то можна сказати, що

$$p_{ij}(k) = \mu_{ij}(k). \tag{16}$$

Логарифм несхожості у (14) визначається наступним чином:

$$U_i(k) = -\ln L_i(k), \tag{17}$$

а цільову функцію (14) можна переписати у вигляді

$$Q = \sum_{k=1}^{N} \sum_{i=1}^{r} u_i^{\beta}(k) U_i(k) = \sum_{k=1}^{N} \sum_{r=1}^{i} u_i^{\beta}(k) \left(-\ln \prod_{j=1}^{n} p_{ij}(k) \right) = -\sum_{k=1}^{N} \sum_{i=1}^{r} u_i^{\beta}(k) \sum_{j=1}^{n} \ln p_{ij}(k).$$

Для обчислення $u_i(k)$ використаємо формулу за типом нечітких c - середніх

$$u_{t,j} = \frac{1}{\sum\limits_{i=1}^{r} \left(\dfrac{\mu_t(k)}{\mu_i(k)} \right)^{\frac{1}{\beta-1}}}, \forall t = 1,...,r; \forall k = 1,...,N. \tag{18}$$

Для фаззіфікації порядкових характеристик використаємо розглянутий вище частотний підхід. На основі відносних частот (1) розраховуються усереднені частоти зустрічаємості характеристик у вибірці за допомогою рекурентного виразу

$$F_1 = 0.5 f_1, \quad F_l = F_{l-1} + 0.5(f_{l-1} + f_l), \forall l = 2,...,m. \tag{19}$$

Припускаючи, що рівень належності до кластерів $u_i(k), \forall i = 1,...,r; \forall k = 1,...,N$ відомий, розраховуються моди для кожної характеристики по кожному із кластерів $x_{ij}^*, \forall i = 1,...,r; \forall j = 1,...,n$ і будуються асиметричні функції належності відповідно до наступних умов:

1. Якщо $x_{ij}^* > 0.5$, то функція належності описується формулою

$$\mu_{ij}(k) = \begin{cases} \dfrac{x(k)}{x_{ij}^*}, x \in [0, x_{ij}^*], \\ \dfrac{2x_{ij}^* - x(k)}{x_{ij}^*}, x \notin [0, x_{ij}^*]. \end{cases} \tag{20}$$

2. Якщо $x_{ij}^* < 0.5$, то функція належності описується формулою

$$\mu_{ij}(k) = \begin{cases} \dfrac{1 - x(k)}{1 - x_{ij}^*}, x \in [x_{ij}^*, 1], \\ \dfrac{x(k) - 2x_{ij}^* + 1}{1 - x_{ij}^*}, x \notin [x_{ij}^*, 1]. \end{cases} \tag{21}$$

3. Якщо $x_{ij}^* = 0.5$, то функція належності описується формулою

$$\mu_{ij}(k) = \begin{cases} \dfrac{x(k)}{x_{ij}^*}, x \in [0, x_{ij}^*], \\ \dfrac{1 - x(k)}{1 - x_{ij}^*}, x \in [x_{ij}^*, 1]. \end{cases} \tag{22}$$

При вирішенні реальних задач часто данні можуть містити викиди. Класичні методи кластеризації у цьому випадку показують значні зміщення прототипів кластерів та їх радіусів. Вирішенням цієї проблеми стало використання робастних методів кластеризації.

Для обробки таких даних було розроблено адаптивні методи кластеризації, в яких замість метрики у цільовій функції (1) використовується критерій близькості, оскільки він убуває повільніше, ніж квадрат евклідової відстані.

Більш зручним видається використання так названих «мір схожості» замість цільових функцій, до яких висуваються дещо м'якіші умови, ніж до метрик

$$\begin{cases} S\left(\tilde{x}(k), \tilde{x}(p)\right) \geq 0, \\ S\left(\tilde{x}(k), \tilde{x}(p)\right) = S\left(\tilde{x}(p), \tilde{x}(k)\right), \\ S\left(\tilde{x}(k), \tilde{x}(k)\right) = 1 \geq S\left(\tilde{x}(k), \tilde{x}(p)\right). \end{cases}$$

Рисунок 2 ілюструє використання традиційного гаусіана як міри схожості з різними параметрами ширини $\sigma^2 < 1$ у вигляді

$$S\left(\tilde{x}(k), c_i\right) = e^{-\frac{\left\|\tilde{x}(k)-c_i\right\|^2}{2\sigma^2}} = e^{-\frac{d^2\left(\tilde{x}(k), c_i\right)}{2\sigma^2}}. \tag{23}$$

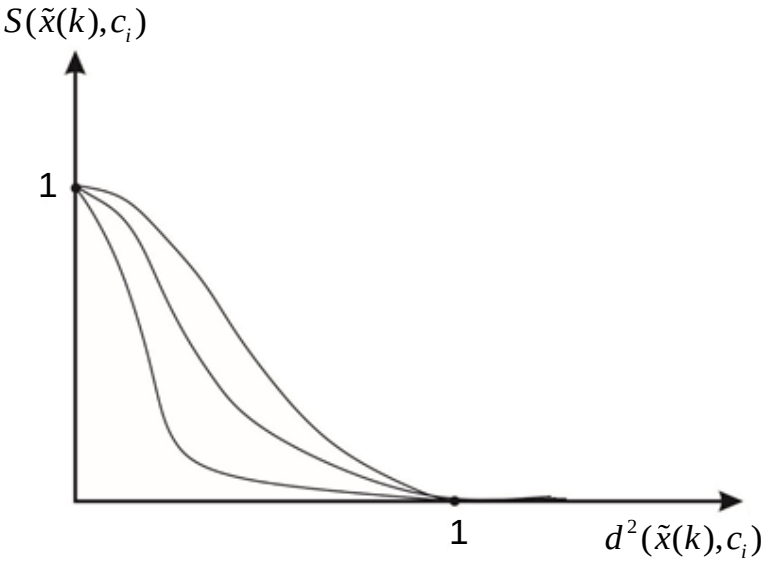

Рисунок 2 – Гаусіан як міра схожості

Цільова функція на основі міри схожості має вигляд

$$E_s\left(u_i(k), c_i\right) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k) S\left(\tilde{x}(k), c_i\right) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k) e^{-\frac{\left\|\tilde{x}(k)-c_i\right\|^2}{2\sigma^2}}. \tag{24}$$

Вводячи функцію Лагранжа

$$L_S\left(u_i(k), c_i, \lambda_i\right) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k) e^{-\frac{\left\|\tilde{x}(k)-c_i\right\|^2}{2\sigma^2}} + \sum_{k=1}^{N}\lambda(k)\left(\sum_{i=1}^{r} u_i(k) - 1\right), \tag{25}$$

та розв’язавши систему рівнянь Каруша-Куна-Таккера, отримуємо шукане рішення у вигляді

$$\begin{cases} u_i(k) = \dfrac{S\left(\tilde{x}(k), c_i\right)^{\frac{1}{\beta-1}}}{\sum\limits_{l=1}^{r} S\left(\tilde{x}(k), c_l\right)^{\frac{1}{\beta-1}}}, \\ \lambda(k) = -\left(\sum\limits_{l=1}^{r} \beta S\left(\tilde{x}(k), c_l\right)^{\frac{1}{\beta-1}}\right)^{\beta-1}, \\ \nabla_{c_i} L_S\left(u_i(k), c_i, \lambda(k)\right) = \sum\limits_{k=1}^{N} w_i^{\beta}(k) e^{-\frac{\left\|\tilde{x}(k)-c_i\right\|^2}{2\sigma^2}} \cdot \dfrac{\tilde{x}(k) - c_i}{\sigma^2} = \vec{0}. \end{cases} \tag{26}$$

Використовуючи процедуру оптимізації Ерроу-Гурвіца-Удзави і вважаючи значення фаззіфікатора $\beta = 2$, отримуємо робастний варіант нечітких c-середніх на основі міри схожості

$$\begin{cases} u_i(k+1) = \dfrac{S\left(\tilde{x}(k+1), c_i(k)\right)}{\sum\limits_{l=1}^{r} S\left(\tilde{x}(k+1), c_l(k)\right)}, \\ c_i(k+1) = c_i(k) + \eta(k+1) u_i^2(k+1) e^{-\frac{\left\|\tilde{x}(k+1) - c_i(k)\right\|^2}{2\sigma^2}} \cdot \dfrac{\tilde{x}(k+1) - c_i(k)}{\sigma^2}. \end{cases} \quad (27)$$

Аналогічним чином може бути синтезовано метод можливісної нечіткої кластеризації за критерієм

$$E_s\left(u_i(k), c_i, \mu_i\right) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k) S\left(\tilde{x}(k), c_i\right) + \sum_{i=1}^{r} \mu_i \left(1 - u_i(k)\right)^{\beta}, \quad (28)$$

де параметр $\mu_i \geq 0$ визначає відстань, на якій рівень належності приймає значення 0,5.

Вирішуючи задачу максимізації (28), отримуємо

$$\begin{cases} u_i(k+1) = \left(1 + \left(\dfrac{S^{-1}\left(\tilde{x}(k+1), c_i(k)\right)}{\mu_i(k)}\right)\right)^{-1}, \\ c_i(k+1) = c_i(k) + \eta(k+1) u_i^{\beta}(k+1) e^{-\frac{\left\|\tilde{x}(k+1) - c_i(k)\right\|^2}{2\sigma^2}} \cdot \dfrac{\tilde{x}(k+1) - c_i(k)}{\sigma^2}, \\ \mu_i(k+1) = \dfrac{\sum\limits_{p=1}^{k+1} u_i^{\beta}(p) S^{-1}\left(\tilde{x}(p), c_i(k+1)\right)}{\sum\limits_{p=1}^{k+1} u_i^{\beta}(p)}. \end{cases} \quad (29)$$

У **третьому розділі** запропоновано методи нечіткої кластеризації масивів категоріальних даних. У багатьох практичних задачах в таких галузях, як Web Mining, Text Mining, Medical Data Mining тощо, часто виникає ситуація, коли ознаки $x_j(k)$ задані не в числовій, а у категоріальній (номінальній) шкалі, при цьому кожна така ознака може приймати скінченне значення «імен» $x_j^l(k)$, де $j = 1,2,...,n$; $l = 1,2,...,\mathrm{m}_j$; $k = 1,2,...,N$.

Прикладом таких даних можуть слугувати покупки у супермаркеті. У такому випадку дані будуть мати вигляд: «молоко», «хліб», «масло», «вино», тощо. Оскільки номенклатура товарів в супермаркете дуже різноманітна, то і вибірки подібних даних відрізняються великими розмірами.

Категоріальні дані можуть бути трансформовані у бінарну шкалу, однак, при цьому різко збільшується розмірність простору ознак, що веде до виникнення ефекту «прокльону розмірності».

Для подолання проблеми «прокльону розмірності» було запропоновано використовувати замість традиційної евклідової відстані несхожість між векторами-образами, а замість прототипів кластерів – моди окремих ознак.

При цьому несхожість між двома векторами $x(k)$ і $x(q)$ може бути описана за допомогою формули

$$d(x(k),x(q))=\sum_{j=1}^{n}\delta(x_j(k),x_j(q)), \tag{30}$$

де

$$\delta(x_j(k),x_j(q))=\begin{cases}0, if\ x_j(k)=x_j(q),\\ 1, if\ x_j(k)\neq x_j(q),\end{cases}$$

при цьому, якщо $x(k)=x(q)$, то $d(x(k),x(q))=0$, а при повній незбіжності компонент цих векторів $d(x(k),x(q))=n$, тобто $0\le d(x(k),x(q))\le n$.

Недоліком такого підходу є те, що мода кожного кластеру не єдина, що не дозволяє отримати стійкий результат.

Для подолання цього недоліку пропонується в якості прототипів використовувати не звичайні моди, а так звані «представники», які здатні враховувати частоти появи окремих ознак.

Нехай i-й кластер містить N_i спостережень $x(k)$ таких, що $Cl_i=\{x(1),x(2),...,x(N_i)\}\subset R^n$, $\sum_{i=1}^{r}N_i=N$. При цьому вектор-прототип цього кластеру може бути представлений у вигляді $c_i=(c_{i1},c_{i2},...,c_{in})^T$, а для кожної компоненти c_{ij} може бути розрахована частота появи відповідного значення ознаки в кластері у вигляді

$$f_{ij}^{l}=\frac{N_{ij}^{l}}{N_i}, l=1,2,...,m_j, \tag{31}$$

де N_{ij}^{l} - кількість появи ознаки x_j^l в Cl_i.

Тоді в якості міри несхожості між прототипом c_i і спостереженням $x(k)$ замість (30) використовується оцінка

$$d\left(x(k),c_{ij}\right)=\sum_{j=1}^{n}\sum_{l_j=1}^{m_j}f_{ij}^{l}\delta\left(x(k),c_{ij}\right). \tag{32}$$

Зрозуміло, що (32) також лежить в інтервалі $0\le d\left(x(k),c_i\right)\le n$.

Задача формування кластера вирішується шляхом мінімізації (8) за обмежень (9). Використання стандартних технік у випадку категоріальних даних приводить до результату

$$\begin{cases}c_i=\dfrac{\sum_{k=1}^{N}u_i^{\beta}(k)x(k)}{\sum_{k=1}^{N}u_i^{\beta}(k)},\\ u_i(k)=\dfrac{d^{\frac{1}{1-\beta}}\left(x(k),c_i\right)}{\sum_{t=1}^{r}d^{\frac{1}{1-\beta}}\left(x(k),c_t\right)}.\end{cases} \tag{33}$$

Для розрахунку мод-прототипів вектор $x(k)$ приписується до кластера Cl_i, для якого

$$u_i(k) > u_t(k), \forall t = 1,2,...,c; t \neq i. \tag{34}$$

Незважаючи на ефективність і широке використання методу FCM, йому властивий недолік, який можна проілюструвати простим прикладом.

Нехай сформовано два кластера з прототипами c_1 і c_2, і нехай на опрацювання надійшло спостереження $x(k)$, що не належить жодному з кластерів, однак відповідно до міри несхожості (32) рівновіддалене від обох прототипів. Тоді, в силу першого обмеження (9) це спостереження з однаковими рівнями належності буде приписане до обох кластерів відповідно до оцінки (33).

Цього недоліку позбавлений метод можливісних нечітких с-середніх (PCM), в основі якого лежить цільова функція

$$E\left(u_i(k), c_i\right) = \sum_{k=1}^{N}\sum_{i=1}^{r} u_i^{\beta}(k)\left\|x(k) - c_i\right\|^2 + \sum_{i=1}^{r}\mu_i \sum_{k=1}^{N}\left(1 - u_i(k)\right)^{\beta}. \tag{35}$$

Мінімізація (35) по c_i, $u_i(k)$ і μ_i веде до результату

$$\begin{cases} c_i = \dfrac{\sum\limits_{k=1}^{N} u_i^{\beta}(k)x(k)}{\sum\limits_{k=1}^{N} u_i^{\beta}(k)}, \\ u_i(k) = \dfrac{1}{1 + \left(\dfrac{\left\|x(k) - c_i\right\|^2}{\mu_i}\right)^{\frac{1}{\beta-1}}}, \\ \mu_i = \left(\sum\limits_{k=1}^{N} u_i^{\beta}(k)\right)^{-1}\left(\sum\limits_{k=1}^{N} u_i^{\beta}(k)\left\|x(k) - c_i\right\|^2\right), \end{cases} \tag{36}$$

який у випадку категоріальних змінних набуває вигляд

$$\begin{cases} u_i(k) = \left(1 + \dfrac{d\left(x(k), c_i\right)}{\mu_i}\right)^{\frac{1}{1-\beta}}, \\ \mu_i = \dfrac{\sum\limits_{k=1}^{N} u_i^{\beta}(k) d\left(x(k), c_i\right)}{\sum\limits_{k=1}^{N} u_i^{\beta}(k)}. \end{cases} \tag{37}$$

Сам же процес можливісної нечіткої кластеризації відбувається у формі послідовності кроків аналогічних методу нечітких c-середніх.

Четвертий розділ присвячено розробці архітектури гібридної системи обчислювального інтелекту для класифікації порядкових даних та методу її навчання. Нео-фаззі нейрон – це нелінійна нейро-фаззі система, що навчається з «вчителем». Структурними елементами нео-фаззі нейрона є нелінійні синапси NS_i, які реалізують

базові правила нечіткого виведення у вигляді

$$if\ \ x_i\ \ is\ \ X_i\ \ then\ \ f_i(x_i)=\sum_{j=1}^{m_i}\mu_{ij}(x_i)w_{ij}, i=1...n.$$

Розглянувши нелінійний синапс нео-фаззі нейрону з позицій нечіткої логіки, можна побачити, що він є дуже схожим на шар фаззіфікації таких нейро-фаззі систем, як мережі Такаґі-Суґено-Канґа, Дженґа, Ванґа-Менделя і фактично реалізує нечітке висновування Такаґі-Суґено нульового порядку. Для покращення апроксимуючих властивостей нео-фаззі нейрону запропоновано конструкцію, що отримала назву подвійний нео-фаззі нейрон, архітектура якого наведена на рис. 3.

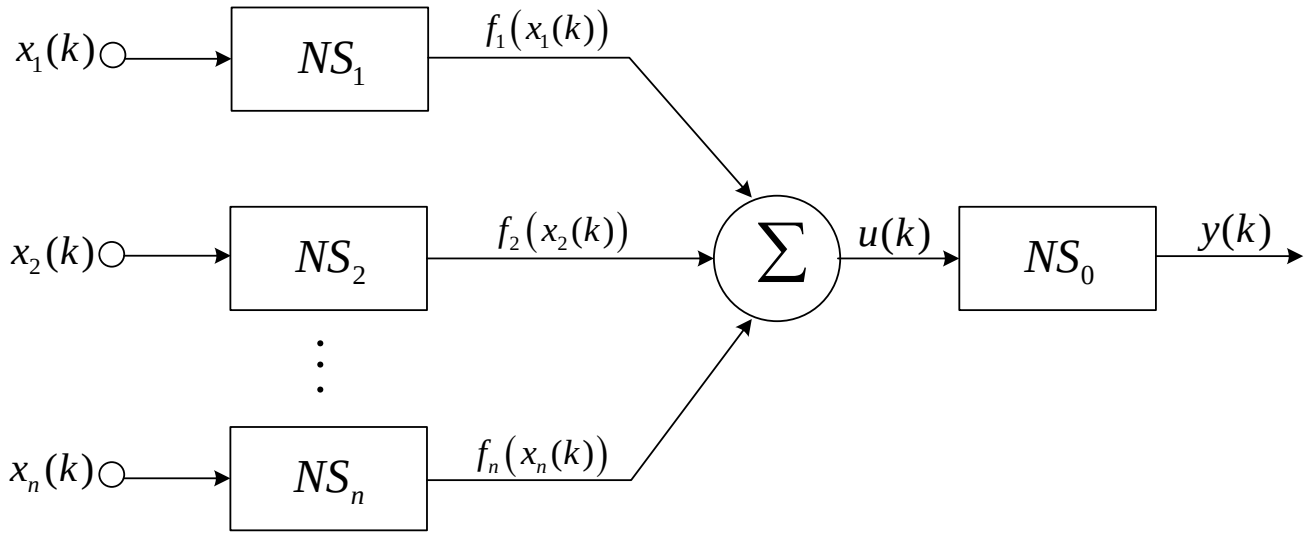

Рисунок 3 – Архітектура подвійного нео-фаззі нейрона

Подвійний нео-фаззі нейрон містить два шари: перший шар, який складається з n нелінійних синапсів NS_i з m_i функціями належності та синаптичними вагами w_{li} кожний, і вихідний шар, який складається з нелінійного синапсу NS_0 з m_0 функціями належності μ_{l0}, $l=1,2,...,m_0$ та синаптичними вагами w_{l0}.

Вихідною інформацією для вирішення задачі нечіткої класифікації є вибірка спостережень, сформована з N n-вимірних векторів ознак $X=\{x(1),x(2),...,x(k),...,x(N)\}$ (тут $x(k)=\{x_i^{r_i}(k)\}$, $i=1,2,...,n$, $r_i=1,2,...,m_i$ - ранг конкретного значення лінгвістичної змінної по i-й координаті n-вимірного простору для k-го спостереження) та вибірка навчальних сигналів $D=\{d(1),d(2),...,d(k),...,d(N)\}$, де $d(k)=d^{r_0}(k)$, $r_0=1,2,...,m_0$ - ранг значення навчального сигналу у вибірці D.

Для фаззіфікації вхідних даних, що задані у порядковій шкалі, використаємо частотний метод, розглянутий у другому розділі. Після обчислення відносних частот характеристик у вибірці (1), формуються несиметричні нерівномірно розташовані функції належності μ_{ij}, μ_{l0} з центрами, які обчислюються за рекурентним відношенням

$$c_1=0.5f_1,\ c_l=c_{l-1}+0.5(f_{l-1}+f_l), \forall l=2,...,m. \qquad (38)$$

При надходженні до подвійного нео-фаззі нейрона вектора-спостереження $x(k)$ на виході отримуємо сигнал

$$y(k) = f_0\left(\sum_{i=1}^{n} f_i\left(x_i(k)\right)\right) = \sum_{l=1}^{m_0} \mu_{l0}\left(u(k)\right) w_{l0} = \sum_{l=1}^{m_0} \mu_{l0}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i} \mu_{ij}\left(x_i(k)\right) w_{ij}\right) w_{l0}. \qquad (39)$$

В якості функцій належності використовуються традиційні трикутні та трапецеїдальні структури, які задовольняють умові одиничного розбиття

$$\sum_{j=1}^{m_i} \mu_{ij}\left(x_i(k)\right) = 1, i = 1,2,...,n,$$

що робить непотрібним введення шару нормалізації, зазвичай присутнього в нейро-фаззі системах.

Такий вибір функцій належності гарантує, що на кожному кроці навчання активуються лише дві сусідні функції належності. Позначаючи їх μ_{ip}, $\mu_{i,p+1}$, $i = 0,1,...,n$ можна записати вихідний сигнал першого шару наступним чином:

$$\begin{aligned} f_i\left(x_i(k)\right) &= \sum_{j=1}^{m_i} \mu_{ij}\left(x_i(k)\right) w_{ij}(k) = \mu_{ip}\left(x_i(k)\right) w_{ip}(k) + \mu_{i,p+1}\left(x_i(k)\right) w_{i,p+1}(k) = \\ &= \frac{c_{i,p+1} - x_i(k)}{c_{i,p+1} - c_{ip}} w_{ip}(k) + \frac{x_i(k) - c_{ip}}{c_{i,p+1} - c_{ip}} w_{i,p+1}(k) = a_i(k) x_i(k) + b_i(k), \end{aligned} \qquad (40)$$

де

$$a_i(k) = \frac{w_{i,p+1}(k) - w_{ip}(k)}{c_{i,p+1} - c_{ip}}, \quad b_i(k) = \frac{c_{i,p+1} w_{ip}(k) - c_{ip} w_{i,p+1}(k)}{c_{i,p+1} - c_{ip}},$$

та

$$u(k) = \sum_{i=1}^{n} a_i(k) x_i(k) + b_i(k). \qquad (41)$$

При цьому вихідний сигнал системи описується за допомогою формули

$$\begin{aligned} y(k) &= \sum_{l=1}^{m_0} \mu_{l0}\left(u(k)\right) w_{l0}(k) = \mu_{p0}\left(u(k)\right) w_{po}(k) + \mu_{p+1,0}\left(u(k)\right) w_{p+1,0}(k) = \\ &= \frac{c_{p+1,0} - u(k)}{c_{p+1,0} - c_{p0}} w_{p0}(k) + \frac{u(k) - c_{p0}}{c_{p+1,0} - c_{p0}} w_{p+1,0}(k) = a_0(k) u(k) + b_0(k), \end{aligned} \qquad (42)$$

де

$$a_0(k) = \frac{w_{p+1,0}(k) - w_{p0}(k)}{c_{p+1,0} - c_{p0}}, \; b_0(k) = \frac{c_{p+1,0} w_{p0}(k) - c_{p0} w_{p+1,0}(k)}{c_{p+1,0} - c_{p0}}.$$

Таким чином, подвійний нео-фаззі нейрон забезпечує кусочно-лінійну апроксимацію деякої нелінійної розділяючої функції у вигляді

$$y(k)=a_0(k)\left(\sum_{i=1}^{n}a_i(k)x_i(k)+b_i(k)\right)+b_0(k). \tag{43}$$

Для навчання подвійного нео-фаззі нейрону використовується градієнтна процедура мінімізації зі змінним кроком пошуку $\eta_i(k)$.

Для налаштування вихідного синапсу NS_0 використовується процедура

$$\begin{cases} w_{l0}(k+1)=w_{l0}(k)+\eta_0(k)e(k)\mu_{l0}\big(u(k)\big),\ l=p,\ p+1,\\ w_{l0}(k+1)=w_{l0}(k),\ \forall l\neq p\neq p+1. \end{cases} \tag{44}$$

Таким чином, на кожній ітерації проходить налаштування вагових коефіцієнтів, відповідно активованим функціям належності μ_{p0} та $\mu_{p+1,0}$. Для збільшення швидкості збіжності та введення додаткових фільтруючих властивостей доцільно використати процедуру вигляду

$$\begin{cases} w_{l0}(k+1)=w_{l0}(k)+\eta_0^{-1}(k)e(k)\mu_{l0}\big(u(k)\big),\ l=p,\ p+1,\\ \eta_0(k+1)=\alpha\eta_0(k)+\mu_{p0}^2\big(u(k+1)\big)+\mu_{p+1,0}^2\big(u(k+1)\big),\\ w_{l0}(k+1)=w_{l0}(k),\ \forall l\neq p\neq p+1,\\ 0\le\alpha\le 1. \end{cases} \tag{45}$$

Для налаштування вагових коефіцієнтів першого шару запишемо критерій навчання у вигляді

$$E(k)=\frac{1}{2}\big(d(k)-f_0\big(u(k)\big)\big)^2=\frac{1}{2}\left(d(k)-f_0\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\mu_{ij}\big(x(k)\big)w_{ij}\right)\right) \tag{46}$$

та візьмемо похідну

$$\frac{\partial E(k)}{\partial w_{ij}}=-e(k)\frac{\partial f_0\big(u(k)\big)}{\partial u(k)}\cdot\frac{\partial u(k)}{\partial w_{ij}}=-e(k)a_0(k)\frac{\partial u(k)}{\partial w_{ij}}.$$

Тоді градієнтна процедура мінімізації (46) має вигляд

$$\begin{cases} w_{ij}(k+1)=w_{ij}(k)+\eta_i(k)e(k)a_0(k)\mu_{ij}\big(x_i(k)\big),\ j=p,\ p+1;\ i=1,2,...,n,\\ w_{ij}(k+1)=w_{ij}(k),\ \forall j\neq p\neq p+1. \end{cases} \tag{47}$$

Вводячи позначення

$$\mu_{ij0}\big(x_i(k)\big)=a_0(k)\mu_{ij}\big(x_i(k)\big),$$

можна записати простий та ефективний метод навчання нелінійних синапсів першого шару

$$\begin{cases} w_{ij}(k+1)=w_{ij}(k)+\eta_i^{-1}(k)e(k)\mu_{ij0}\big(x_i(k)\big),\ j=p,\ p+1;\ i=1,2,...,n,\\ \eta_i(k+1)=\alpha\eta_i(k)+\mu_{ip0}^2\big(x_i(k+1)\big)+\mu_{i,p+10}^2\big(x_i(k+1)\big),\\ w_{ij}(k+1)=w_{ij}(k),\ \forall j\neq p\neq p+1,\\ 0\le\alpha\le 1, \end{cases} \tag{48}$$

який по структурі збігається з процедурою (45).

У **п'ятому розділі** викладено результати проведених експериментальних досліджень та їх використання для розв'язання практичних задач інтелектуального аналізу даних. Виконано програмну реалізацію запропонованого методу нечіткої кластеризації порядкових даних на основі частотних прототипів та функцій належності, а також методу нечіткої кластеризації на основі функцій належності і правдоподібності. Продемонстровано, що представлені методи мають більшу точність порівняно з традиційними аналогами. Досліджено залежність точності кластеризації даних, що мають викиди, від обраного методу. Доведено стійкість запропонованого методу робастної нечіткої кластеризації порядкових даних до викидів у порівнянні з традиційними методами. Проведено імітаційне моделювання запропонованого методу адаптивної нечіткої кластеризації порядкових даних на основі порядково-чисельного відображення. Проведено порівняння запропонованого методу з пакетними аналогами і показано, що адаптивний метод зберігає монотонність збільшення точності кластеризації з кожним наступним спостереженням. Проведено імітаційне моделювання низки методів нечіткої кластеризації масивів категоріальних даних з порівнянням точності кластеризації даних. Змодельовано архітектуру подвійного нео-фаззі нейрону для класифікації порядкових даних. Показано, що запропонована система має високу точність класифікації порядкових даних. Розв'язано ряд практичних задач.

У **висновках** сформульовано наукові та практичні результати, що їх одержано у дисертаційній роботі. У **додатку** наведено акти про впровадження результатів дослідження.

ВИСНОВКИ

У дисертаційній роботі представлено результати, що відповідають меті дослідження, а саме – розробці методів нечіткої кластеризації та класифікації даних, що задані у нечислових шкалах. Проведені дослідження дозволили зробити такі висновки:

1. Виконано огляд стану проблеми інтелектуального аналізу даних, що задані у нечислових шкалах. Виявлено недоліки існуючих методів: неможливість функціонувати в онлайн режимі та обмеженість процедур трансформації лінгвістичних характеристик у чисельну шкалу.

2. Синтезовано адаптивні та робастні методи нечіткої кластеризації порядкових даних: запропоновано метод нечіткої кластеризації порядкових даних на основі частотних прототипів та функцій належності, що дало можливість обробляти порядкові дані, які не підпорядковуються нормальному розподілу; запропоновано метод нечіткої кластеризації на основі порядково-чисельного відображення для інтелектуального опрацювання порядкових даних, що дало можливість обробляти дані у послідовному режимі; удосконалено метод нечіткої кластеризації порядкових даних шляхом сумісного використання функцій належності і функцій правдоподібності, що дозволило підвищити точність кластеризації порядкових даних; отримали подальший розвиток методи адаптивної робастної нечіткої кластеризації порядкових даних шляхом використання критерія міри схожості, що дозволило обробляти порядкові дані, які мають викиди, у послідовному режимі.

3. Удосконалено метод можливісної нечіткої кластеризації категоріальних даних шляхом використання частотних прототипів та мір несхожості, що дозволило подолати недоліки класичних методів та підвищити точність кластеризації.

4. Синтезовано нейро-фаззі систему на основі нео-фаззі нейрону шляхом використання додаткового вихідного шару, що дозволило вирішити задачу класифікації даних, заданих у порядковій шкалі.

5. Розв'язано практичну задачу аналізу клієнтської бази даних ТОВ «Південелектропроект» за допомогою запропонованого методу можливісної нечіткої кластеризації масивів категоріальних даних.

6. Розв'язано практичну задачу автоматичної обробки термограм з метою діагностики електрообладнання у ТОВ НВП «Мідіел» за допомогою запропонованого методу нечіткої кластеризації порядкових даних на основі сумісного використання функцій належності і правдоподібності.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Бодянский, Е.В. Нечеткая кластеризация данных, заданных в порядковой шкале / Е.В. Бодянский, В.А. Опанасенко (В.А Самитова), А.Н. Слипченко // Системы обработки информации. – 2007. – 4(62). – С. 5 - 9.

2. Бодянский, Е.В. Нечеткая кластеризация данных в порядковой шкале на основе совместного использования функций принадлежности и правдоподобия / Е.В. Бодянский, В.А. Самитова // Сборник научных работ ХУПС. – 2010. – 3(25). – С. 91 - 95.

3. Самитова, В.А. Отображение порядковых характеристик в цифровую шкалу на основе нечеткой кластеризации / В.А. Самитова // Системы обработки информации. – 2015. – 7(132). – С. 107 - 110.

4. Бодянский, Е.В. Нечеткая классификация данных в ранговой шкале на основе двойного нео-фаззи нейрона / Е.В. Бодянский, В.А. Самитова // Восточно-Европейский журнал передовых технологий. – 2008. – 4/2 (34). – С. 4 - 7.

5. Bodyanskiy, Ye. Robust fuzzy data clustering in an ordinal scale based on a similarity measure / Ye. Bodyanskiy, O. Tyshchenko, V. Samitova // International Journal of Reseach In Engineering And Science (IJRES). – 2014. – 2(4). – P. 21-25. (Входить до міжнародних наукометричних баз JOUR Informatics, NEW JOUR, Index Copernicus).

6. Бодянский, Е.В. Возможностная нечеткая кластеризация массивов категориальных данных с использованием частотных прототипов и мер несходства / Е.В. Бодянский, В.А. Самитова // Бионика интеллекта. – 2016. – 1(82). – С. 72 - 75.

7. Опанасенко, В.А. (Самитова В. А.) Алгоритм нечеткой кластеризации данных, представленных порядковыми атрибутами / В.А. Опанасенко (В.А Самитова), А.Н. Слипченко // Системный анализ и информационные технологии: IX международная научно-техническая конференция, 15-19 апреля 2007г.: тез. докл. – Киев. – 2007. – С. 126.

8. Самитова, В.А. Нечеткая кластеризация порядковых данных с помощью двойного нео-фаззи нейрона / В.А. Самитова // Радиоэлектроника и молодежь в XXI веке: 12-й Международный молодежный форум, 1 – 3 апреля 2008 г.: мат. конф. – Харьков. – 2008. – С. 144.

9. Самитова, В.А. Нечеткая робастная кластеризация порядковых данных на основе мер схожести / В.А. Самитова // Радиоэлектроника и молодежь в XXI веке:

19-й Международный молодежный форум, 20 – 22 апреля 2015 г.: мат. конф. – Харьков. – 2015. – С. 58 - 59.

10. Самитова, В.А. Нечеткая кластеризация порядковых данных на основе функций принадлежности и функций правдоподобия / В.А. Самитова // Радиоэлектроника и молодежь в XXI веке: XX Юбилейный Международный молодежный форум, 19 – 21 мая 2016 г.: мат. конф. – Харьков. – 2016. – С. 47 - 48.

11. Бодянский, Е.В. Возможностная нечеткая кластеризация категориальных данных на основе частотных прототипов и мер несходства / Е.В. Бодянский, В.А. Самитова // Полиграфические, мультимедийные и web- технологии: I Международная научно-техническая конференция, 16 – 20 мая 2016 г.: мат. конф. – Харьков. – 2016. – С. 39 - 40.

12. Бодянский, Е.В. Рекуррентная нечеткая кластеризация данных на основе отображения порядковых характеристик в цифровую шкалу / Е.В. Бодянский, В.А. Самитова // Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта (ISDMCI'2016): XII международная научная конференция, 24 – 28 мая 2016 г.: мат. конф. – Железный порт. – 2016. – С. 258 - 260.

АНОТАЦІЯ

Самітова В.О. Класифікація та кластеризація даних, що задані в нечислових шкалах. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту. – Харківський національний університет радіоелектроніки, Міністерство освіти і науки України, Харків, 2016.

Метою дисертаційної роботи є розробка методів нечіткої кластеризації та класифікації даних, що задані у нечислових шкалах.

Виконано огляд стану проблеми інтелектуального аналізу даних, що задані у нечислових шкалах. Виявлено недоліки існуючих методів: неможливість функціонувати в онлайн режимі та обмеженість процедур трансформації лінгвістичних характеристик у чисельну шкалу. Запропоновано метод нечіткої кластеризації порядкових даних на основі частотних прототипів та функцій належності, що дало можливість обробляти порядкові дані, які не підпорядковуються нормальному розподілу. Запропоновано метод нечіткої кластеризації на основі порядково-чисельного відображення для інтелектуального опрацювання порядкових даних, що надходять у послідовному режимі. Вдосконалено метод нечіткої кластеризації порядкових даних шляхом сумісного використання функцій належності і функцій правдоподібності, що дозолило підвищити точність кластеризації порядкових даних. Розроблено методи адаптивної робастної нечіткої кластеризації порядкових даних шляхом використання критерію спеціального виду (міри схожості), що дозволило обробляти порядкові дані, які мають викиди, у послідовному режимі. Удосконалено метод можливісної нечіткої кластеризації категоріальних даних шляхом використання частотних прототипів та мір несхожості, що дозволило подолати недоліки класичних методів та підвищити точність кластеризації даних. Вдосконалено архітектуру нейро-фаззі системи на основі нео-фаззі нейрону для класифікації порядкових даних шляхом використання додаткового вихідного шару, що дозволило покращити апроксимуючі властивості гібридної системи

обчислювального інтелекту при роботі з порядковими даними. Розв’язано практичну задачу аналізу клієнтської бази даних ТОВ «Південелектропроект», а також задачу автоматичної обробки термограм з метою діагностування електрообладнання у ТОВ НВП «Мідіел».

Ключові слова: інтелектуальний аналіз даних, порядкові дані, категоріальні дані, класифікація, нечітка кластеризація, штучні нейронні мережі, подвійний нео-фаззі нейрон.

АННОТАЦИЯ

Самитова В.А. Классификация и кластеризация данных, заданных в нечисловых шкалах. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.23 – системы и средства искусственного интеллекта. – Харьковский национальный университет радиоэлектроники, Министерство образования и науки Украины, Харьков, 2016.

Целью диссертационной работы является разработка методов нечеткой кластеризации и классификации данных, заданных в нечисловых шкалах.

Выполнен обзор состояния проблемы интеллектуального анализа данных, заданных в нечисловых шкалах. Выявлены недостатки существующих методов: невозможность функционировать в онлайн режиме и ограниченность процедур трансформации лингвистических характеристик в цифровую шкалу. Впервые предложен метод нечеткой кластеризации порядковых данных на основе частотных прототипов и функций принадлежности, что позволило обрабатывать данные, не подчиняющиеся нормальному распределению. Впервые предложен адаптивный метод нечеткой кластеризации порядковых данных на основе порядково-цифрового отображения, что позволило обрабатывать данные, поступающие в последовательном режиме. Усовершенствован метод нечеткой кластеризации данных, заданных в порядковой шкале путем совместного использования функции принадлежности и функции правдоподобия, что позволило повысить точность кластеризации порядковых данных. Разработаны методы адаптивной робастной нечеткой кластеризации порядковых данных путём введения критерия специального вида (меры схожести), что позволило обрабатывать порядковые данные, содержащие выбросы, в последовательном режиме. Усовершенствован метод возможностной нечеткой кластеризации массивов категориальных данных путем совместного использования частотных прототипов и мер несходства, что позволило преодолеть недостатки классических методов и повысить точность кластеризации данных. Улучшена нейро-фаззи система на основе нео-фаззи нейрона для классификации порядковых данных путем введения дополнительного выходного слоя, что позволило улучшить аппроксимирующие свойства гибридной системы вычислительного интеллекта при работе с порядковыми данными. Решена практическая задача анализа клиентской базы ООО «Южэлектропроект», а также задача автоматической обработки термограмм при диагностике электрооборудования для ООО НПФ "Мидиэл".

Ключевые слова: интеллектуальный анализ данных, порядковые данные, категориальные данные, классификация, нечеткая кластеризация, искусственные нейронные сети, двойной нео-фаззи нейрон.

ABSTRACT

Samitova V.A. Non-numerical data classification and clusterization. – As the Manuscript.

The thesis for the candidate degree in engineering science in the specialty 05.13.23 – systems and means of artificial intelligence. – Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2016.

The thesis is devoted to research and development of fuzzy clusterization and classification methods for non-numerical data.

The fuzzy clustering method for ordinal data without normal distribution based on frequency prototypes and membership functions is proposed. The adaptive fuzzy clustering method for ordinal data using ordinal-numerical mapping is proposed. This method specifically designed to process ordinal data in an online mode. The fuzzy clustering method for ordinal data based on membership functions and likelihood is developed. The proposed method is proven to have higher clustering accuracy comparatively with traditional methods. The thesis proposes a number of robust fuzzy clustering methods based on similarity measure for ordinal data. These methods are resistant to noise and permit to process ordinal data in an online mode. The fuzzy clustering methods for categorical data based on frequency prototypes and dissimilarity measure is developed. The proposed method solves traditional methods` weaknesses and has high clustering accuracy. The double neo-fuzzy neuron and its learning algorithm are proposed. The proposed architecture is proven to have enhanced approximating capabilities as well as high operating speed. An experimental study of the properties and characteristics of the developed methods is carried out and recommendations on their use in solving practical tasks are proposed.

Keywords: data mining, ordinal data, categorical data, classification, fuzzy clustering, artificial neural networks, double neo-fuzzy neuron.