Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

ЖЕРНОВА ПОЛІНА ЄВГЕНІЇВНА

УДК 004.032.26

НЕЧІТКА КЛАСТЕРИЗАЦІЯ ПОТОКІВ ДАНИХ ЗА УМОВ НЕВІДОМОЇ КІЛЬКОСТІ КЛАСТЕРІВ

05.13.23 – системи та засоби штучного інтелекту

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2019

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 **Пелешко Дмитро Дмитрович**, Приватний заклад вищої освіти «ІТ СТЕП Університет», проректор з науково-педагогічної роботи; кандидат технічних наук, доцент **Гороховатський Олексій Володимирович**, Харківський національний економічний університет імені Семена Кузнеця, доцент кафедри інформатики та комп'ютерної техніки.

Захист відбудеться «__» _____ 2019 р. о ____ 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 «__» _____ 2019 р.

Учений секретар
спеціалізованої вченої ради Є.І. Литвинова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. На сьогоднішній день для вирішення задач інтелектуального аналізу даних, насамперед кластеризації даних, що є невід'ємною частиною проблеми Data Mining, існує безліч підходів та методів, які засновані на апараті обчислювального інтелекту. Розроблено математичний апарат, який дозволяє вирішувати задачі кластеризації даних у різних сферах, таких як медицина, наука, техніка та інші. Але більшість відомих методів за своєю суттю є чіткими процедурами, тобто кластери є лінійно роздільними, а інформація обробляється в пакетному режимі. Зараз на перший план виходять задачі, які пов'язані з Dynamic Data Mining, Data Stream Mining та Big Data, коли дані надходять у вигляді потоку, а кластери мають довільну форму та перетинаються у просторі ознак, виникає потреба розробити методи нечіткої кластеризації для обробки потоків даних. Така задача може бути вирішена за допомогою нечітких методів та ядерного підходу згідно з гіпотезою Кавера: якщо задача лінійно нероздільна у вихідному просторі, вона може бути лінійно роздільною у просторі вищої розмірності. Також однією з проблем кластеризації даних є те, що заздалегідь невідома кількість кластерів, на яку будуть поділені вхідні дані, які надходять на обробку у вигляді потоку.

Таким чином, на сьогоднішній день є актуальною задача розробки нових методів нечіткої кластеризації потоків даних високої розмірності, призначених для обробки даних в онлайн режимі, коли кількість кластерів невідома заздалегідь та вони мають довільну форму та перетинаються у просторі ознак.

Великий внесок у розвиток напрямку динамічного інтелектуального аналізу даних на основі штучних нейронних мереж внесли У. Маккалох, Ф. Розенблатт, О.Г. Івахненко, Б. Уідроу, Т. Кохонен, Г. Голланд, Дж. Бездек, Е. Мамдані, Т. Суґено, Б. Коско, Дж. Мендель, В. Педрич, Є.В. Бодянський, В.С. Степашко, Д.Д. Пелешко, О.І. Михальов, Ю.П. Кондратенко, О.Г. Руденко, В.І. Литвиненко, А.А. Тунік, Н.М. Куссуль, Ю.П. Зайченко та інші.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана в рамках держбюджетних НДР: «Динамічний інтелектуальний аналіз послідовностей нечіткої інформації за умов суттєвої невизначеності на основі гібридних систем обчислювального інтелекту» (№ДР 0116U002539); «Глибинні гібридні системи обчислювального інтелекту для аналізу потоків даних та їх швидке навчання» (№ДР 0119U001403) В рамках зазначених НДР здобувачем розроблені методи на основі ансамблю нечіткої кластеризації, які призначені для обробки даних в онлайн режимі, коли дані надходять на обробку послідовно, одні за одними, а кластери можуть перетинатися у просторі ознак та мати довільну форму.

Мета і задачі дослідження. Розробка методів нечіткої кластеризації потоків даних високої розмірності з використанням ансамблевого підходу, коли кількість та форма кластерів заздалегідь невідома.

Відповідно до поставленої мети у дисертаційній роботі необхідно вирішити такі завдання:

– провести аналіз існуючих методів та підходів для кластеризації потоків даних;

– розробити метод для кластеризації потоків даних у випадку невідомої кількості кластерів;

– розробити архітектуру ансамблю нейронних мереж для кластеризації потоків даних;

– розробити ансамбль нейронних мереж з використанням ядерних функцій для вирішення задачі за умов лінійної нероздільності класів;

– розробити ансамбль нейро-фаззі систем для кластеризації потоку даних у припущенні, що кількість та форма кластерів невідомі заздалегідь;

– розробити ансамбль нейро-фаззі мереж на основі імовірнісно-можливісного підходу для кластеризації потоків даних;

– провести експериментальні дослідження розроблених методів, вирішити за їх допомогою практичні задачі нечіткої кластеризації потоків даних високої розмірності.

Об'єкт дослідження – процес обробки даних високої розмірності, які надходять на обробку спостереження за спостереженням за умов невизначеної кількості та форми кластерів.

Предмет дослідження – методи нечіткої кластеризації на основі ансамблевого підходу у задачах, коли дані обробляються послідовно за умов невизначеної кількості та форми кластерів.

Методи дослідження – базуються на теорії обчислювального інтелекту, а саме на методах теорії штучних нейронних мереж та теорії нечіткої логіки для побудови ансамблю нечітких нейро-фаззі мереж Т. Кохонена, що дозволяє провести нечітку кластеризацію потоків даних; теорії оптимізації для синтезу методів нечіткої кластеризації та методів самонавчання нейро-фаззі мереж. Імітаційне моделювання використовується для перевірки якості роботи ансамблю нейро-фаззі кластеризаційних мереж Т. Кохонена для потоку даних високої розмірності.

Наукова новизна одержаних результатів. До нових, одержаних особисто автором, належать такі результати:

1. Вперше запропоновано ансамбль самоорганізовних карт Т. Кохонена, який характеризується використанням онлайн методу К-середніх, що дозволяє кластеризувати дані за умов апріорі невідомої кількості класів.

2. Вперше запропоновано ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена, який характеризується використанням модифікованого онлайн методу нечітких С-середніх, коли апріорі невідома кількість та форма кластерів, що дозволяє кластеризувати потоки даних за умов лінійної нероздільності класів, які довільним чином перетинаються у просторі ознак.

3. Удосконалено ансамбль ядерних самоорганізовних карт Т. Кохонена, який відрізняється від аналогів введенням додаткового ядерного шару для підвищення розмірності вхідного простору, що дозволяє кластеризувати потоки даних за умов, коли кластери є лінійно нероздільними.

4. Удосконалено ансамбль самоорганізовних нечітких карт Т. Кохонена, який відрізняється від аналогів одночасним використанням процедури імовірнісної та можливісної кластеризації потоків даних, що дозволяє підвищити рівень якості кластеризації потоків даних.

Практичне значення одержаних результатів. Розроблені у роботі методи кластеризації даних на основі ансамблевого підходу та нейро-фаззі систем обчислювального інтелекту призначені для онлайн обробки потоку даних в умовах невизначеності щодо кількості та форми кластерів. Отриманий підхід є достатньо простим з обчислювальної точки зору та дозволяє вирішувати задачі інтелектуального

аналізу даних та динамічного аналізу потоків даних. Використання методів кластеризації на основі ансамблевого підходу дозволяє підвищити ефективність вирішення задач обробки медичних даних. Ансамбль нейро-фаззі кластеризаційних мереж дозволяє підвищити точність аналізу потоків даних. Ці методи дозволили встановити закономірності формування відповідної реакції організму на об'єднаний вплив екологічних чинників; використання запропонованих підходів допомогло визначити значення біологічних ефектів сполученої дії електромагнітного випромінювання та позитивних низьких температур при аналізі результатів у рамках науково-дослідної роботи, що фінансувалася Міністерством охорони здоров'я України, «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища» (КПКВК 2301020, № 0113U002536) (акт впровадження від 11.01.2019).

Також основні результати дисертаційної роботи використовуються у навчальному процесі Харківського національного університету радіоелектроніки на кафедрі системотехніки в курсі «Нейросистеми та генетичні алгоритми» (акт впровадження від 14.02.2019).

Особистий внесок здобувача. Усі наукові результати дисертаційної роботи, що виносяться на захист, отримані автором самостійно. У роботах, опублікованих у співавторстві, автору належать такі результати: [1] – ансамбль самоорганізовних карт Т. Кохонена для кластеризації потоків даних, коли кількість кластерів апріорно невідома; [2] – архітектура ядерної кластерувальної мережі з використанням самоорганізовних карт Т. Кохонена; [3] – онлайн модифікація методу Х-середніх з використанням ансамблевого підходу для кластеризації потоків даних; [4] – ядерний ансамбль самоорганізовних карт Т. Кохонена для кластеризації потоків даних, коли кількість та форма кластерів заздалегідь невідомі; [5] – ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних з використанням імовірнісного підходу; [6] – ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних з використанням імовірнісно-можливісного підходу; [7] – адаптивна модифікація методу Х-середніх для кластеризації потоків даних; [8] – ансамбль нейронних мереж та його навчання для кластеризації потоків даних; [9] – нечітка імовірнісна нейронна мережа для кластеризації даних; [10] – ансамбль ядерних самоорганізовних карт Т. Кохонена для кластеризації потоків даних у ситуаціях, коли кластери є лінійно не роздільними; [11] – нечітка кластеризація потоків даних на основі методу С-середніх; [13] – ансамбль самоорганізовних карт Т. Кохонена для кластеризації потоків даних на основі імовірнісно-можливісного підходу; [14] – ансамбль нейро-фаззі мереж Т. Кохонена для кластеризації потоків даних.

Апробація результатів дисертації. Основні положення та результати дисертаційної роботи були представлені, доповідалися й обговорювалися на міжнародних і всеукраїнських наукових конференціях і семінарах, зокрема на: VI Міжнародній науково-практичній конференції «Інформаційні управляючі системи та технології», Україна, м. Одеса, 2017 р.; Міжнародній науковій конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту», Україна, м. Залізний порт, 2018 р.; Міжнародній конференції «The 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)», Україна, м. Львів, 2018 р.; Міжнародній науково-практичній конференції "Інформаційні технології та системи", Україна, м. Харків, 2019 р.; XXXIII Міжнародному молодіжному форумі

«Радіоелектроніка та молодь у XXI столітті», Україна, м. Харків, 2019 р.; Міжнародному науковому симпозіумі "Інтелектуальні рішення", Україна, м. Ужгород, 2019 р.

Публікації. Основні положення дисертаційної роботи опубліковані в 14 наукових працях: у тому числі 6 статтях у періодичних фахових виданнях України з технічних наук, що включені до переліку МОН України (1 стаття в журналі, що входить до міжнародних наукометричних баз), 8 публікацій у збірниках міжнародних наукових конференцій.

Структура та обсяг дисертації. Дисертація складається зі вступу; п’яти розділів; висновків, що містять основні результати; списку використаних джерел і додатку. Загальний обсяг роботи складає 140 сторінок тексту, що містить 2 анотації на 10 сторінках, 38 рисунків, 9 таблиць, список використаних джерел з 118 найменувань на 12 сторінках, 2 додатки на 6 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність обраної теми дисертації; сформульовано мету та задачі дослідження; визначено об’єкт, предмет і методи досліджень; визначено наукову новизну та практичну значущість отриманих результатів, а також особистий внесок авторки в роботах, виконаних у співавторстві; наведено відомості про апробацію результатів дисертації та кількість публікацій за темою дисертаційної роботи.

У **першому розділі** виконано огляд стану проблеми динамічного аналізу даних, а саме: послідовної кластеризації потоків даних у ситуації, коли дані надходять на обробку спостереження за спостереженням. Описано основні методи та підходи для кластеризації даних, серед яких присутні методи для нечіткої кластеризації. Розглянуто різні нейронні мережі для кластеризації даних, основна увага приділена самоорганізовним картам Т. Кохонена та методу нечітких С-середніх.

Основною проблемою методів для кластеризації даних є те, що кількість кластерів апріорно невідома заздалегідь. Зважаючи на це, припускається, що ефективним способом кластеризації даних у припущенні щодо невідомої кількості кластерів та їх форми, коли дані надходять спостереження за спостереженням, є обробка даних ансамблем кластерувальних мереж, кожна з яких налаштовується на свою кількість кластерів і обробляє дані в онлайн режимі.

Внаслідок проведеного аналізу визначено задачі дисертаційного дослідження, що полягають у розробці методу кластеризації потоків даних, який засновано на ансамблевому підході з використанням самоорганізовних карт Т. Кохонена та модифікованого методу С-середніх, що дозволяє обробляти дані, коли невідома кількість та форма кластерів.

Другий розділ присвячено дослідженню та розробці методів кластеризації, коли дані надходять на вхід системи спостереження за спостереженням, а кількість кластерів апріорно невідома.

Вперше запропоновано ансамбль самоорганізовних карт Т. Кохонена для кластеризації потоків даних при невідомій кількості класів, що дозволяє забезпечити оптимальну якість кластеризації. У ситуаціях, коли необхідна обробка інформації в online режимі по мірі послідовного надходження на вхід системи нових даних, на перший план виходять нейронні мережі. Завдяки цьому використання самоорганізовних карт Т. Кохонена є доцільним, а ансамбль карт Т. Кохонена SOM^m дозволяє

кластеризувати дані у ситуації, коли кількість кластерів невідома, оскільки кожна з мереж обробляє дані паралельно, кожна з цих мереж налаштована на свою кількість кластерів $m = 2,3,...,M$. Таким чином, перша кластерувальна мережа ансамблю працює в припущенні $m = 2$, тобто в шарі Кохонена містить усього два нейрони з синаптичними вагами-центроїдами w_1^2 і w_2^2. Другий елемент ансамблю містить три нейрони з векторами синаптичних ваг w_1^3, w_2^3, w_3^3 і, нарешті, остання SOM^M ансамблю працює в припущенні, що число можливих кластерів дорівнює M, тобто містить M нейронів – адаптивних лінійних асоціаторів.

Для навчання кожної з окремих SOM^m можуть бути використані як стандартні кохоненівські WTA- і WTM-правила самонавчання, так і їх модифікації з використанням функції сусідства спеціального виду.

Розглянемо процес самонавчання m-ї мережі Кохонена SOM^m, що містить m нейронів з синаптичними вагами $\left\{w_1^m, w_2^m, ...w_m^m\right\} \subset R^n$. В основі алгоритму налаштування синаптичних ваг полягає принцип конкурентного самонавчання, який реалізується в три основних етапи (конкуренція, кооперація, синаптична адаптація) і починається з аналізу вхідного вектора-образу $x(k)$, що надходить з рецепторного (нульового) шару на всі нейрони шару Кохонена. Для кожного з нейронів обчислюється відстань

$$D\left(x(k), w_j^m(k-1)\right) = \left\|x(k) - w_j^m(k-1)\right\|,$$

де $j = 1,2,...m$, при цьому, якщо вхідні сигнали попередньо пронормовані за допомогою перетворення

$$x(k) = \frac{x(k)}{\left\|x(k)\right\|}$$

так, що $\left\|x(k)\right\| = 1$, а в якості відстані використовується евклідова метрика, то мірою схожості (подібності) векторів $x(k), w_j^m(k-1)$ може служити скалярний добуток

$$sim\left(x(k), w_j^m(k-1)\right) = x^T(k) w_j^m(k-1) = \cos\left(x(k), w_j^m(k-1)\right).$$

Далі визначається нейрон-переможець «найближчий» до вхідного образу такий, що

$$sim\left(x(k), w^{m*}(k-1)\right) = \max_j sim\left(x(k), w_j^m(k-1)\right), \quad (1)$$

після чого, опускаючи тимчасово процес кооперації, можна уточнити синаптичні ваги переможця за допомогою рекурентного співвідношення

$$w_j^m(k) = \begin{cases} w_j^m(k-1) + \eta(k) \times \\ \times\left(x(k) - w_j^m(k-1)\right), \text{ якщо } w_j^m(k-1) = w^{m*}(k-1), \\ w_j^m(k-1) \text{ у іншому випадку.} \end{cases} \quad (2)$$

Таким чином, процедура реалізує правило «переможець отримує все» (WTA), при цьому вектор синаптичних ваг переможця $w^{m*}(k-1)$ «підтягується» до вхідного образу на відстань, що визначається величиною кроку

$$0<\eta(k)<1.$$

Використання функції сусідства призводить до алгоритму самонавчання

$$w_l^m(k)=w_l^m(k-1)+\eta(k)\varphi(j,l)\left(x(k)-w_l^m(k-1)\right)\forall l=1,2,...,m,$$

що реалізує правило «переможець отримує більше» (WTM), при цьому при $l=j$ цей алгоритм збігається зі співвідношенням (2).

В принципі, можна взагалі відмовитися від етапу конкуренції та визначення переможця як такого. При цьому в ролі переможця в такому випадку виступає сам вхідний вектор-образ, як функція сусідства використовується міра схожості (1).

При цьому алгоритм самонавчання *m*-го елемента ансамблю набуває вигляду

$$w_l^m(k)=w_l^m(k-1)+\eta(k)\left[\cos\left(x(k),w_l^m(k-1)\right)\right]_+\left(x(k)-w_l^m(k-1)\right)=$$

$$=w_l^m(k-1)+\eta(k)\left[x^T(k)w_l^m(k-1)\right]_+\left(x(k)-w_l^m(k-1)\right)=$$

$$=w_l^m(k-1)+\eta(k)\left[y_l^m(k)\right]_+\left(x(k)-w_l^m(k-1)\right),$$

де $\left[y_l^m(k)\right]_+=\max\left\{y_l^m(k),0\right\}$ – невід'ємне значення *l*-го вихідного сигналу *m*-ої карти Кохонена ансамблю.

У процесі роботи ансамблю постійно проводиться оцінка якості кластеризації за допомогою критерію Цаліньського-Харабаша або в його стандартній формі, або за допомогою його online модифікації. При цьому критерій в загальному вигляді має форму

$$CH(m)=\frac{1}{m-1}TrS_B^m\left(\frac{1}{N-m}TrS_w^m\right)^{-1},$$

де $S_B^m=\frac{1}{N}\sum_{j=1}^{m}N_j^m\left(w_j^m-\bar{w}^m\right)\left(w_j-\bar{w}^m\right)^T$ – матриця міжкластерної відстані для *m* кластерів;

$\bar{w}^m=\frac{1}{N}\sum_{j=1}^{m}N_j^m w_j^m$ – центр ваги масиву даних X;

N_j^m – кількість спостережень, що відносяться до *j*-го кластеру, $j=1,2,...m$;

$S_w^m=\frac{1}{N}\sum_{j=1}^{m}\sum_{k=1}^{N}u_j(k)\left(x(k)-w_j^m\right)\left(x(k)-w_j^m\right)^T$ – матриця розсіяння *m*-го кластеру;

$u_j=\begin{cases}1, \text{ если } x(k) \text{ належить j-му кластеру,}\\ 0-\text{у іншому випадку}\end{cases}$ – чітка функція належності *k*-го спостереження *j*-му кластеру.

При аналізі даних, що надходять на обробку в online режимі, розрахунок критерію доцільно організувати на ковзному вікні розмірності *s* ($s=1,2,...N$), при цьому в поточний момент часу *k* $CH(m)$ можна записати як

$$CH(m,k)=\frac{\frac{1}{m-1}\sum_{j=1}^{m}\sum_{\tau=k-s+1}^{k}N_j^m(\tau)\left\|w_j^m(\tau)-\bar{w}^m(\tau)\right\|^2}{\frac{1}{N-m}\sum_{j=1}^{m}\sum_{\tau=k-s+1}^{k}u_j(\tau)\left\|x(\tau)-w_j^m(\tau)\right\|^2},$$

де

$$\bar{w}^m(\tau)=\frac{1}{s}\sum_{\tau=k-s+1}^{k}x(\tau).$$

За оптимальну кількість кластерів у вибірці m^* приймається m, що забезпечує максимум значенню $CH(m)$, тобто

$$CH(m^*)=\max_m\{CH(2),CH(3),...,CH(M)\}.$$

Запропонована процедура online кластеризації її основі ансамблю нейронних мереж Т. Кохонена є за суттю адаптивною модифікацією методу X-середніх, що орієнтована на обробку потоків даних, досить проста в чисельній реалізації і дозволяє вирішити задачу чіткої кластеризації в умовах апріорно невідомої або змінної кількості кластерів.

У **третьому розділі** описано ансамбль ядерних самоорганізовних карт Т. Кохонена для кластеризації потоків даних у випадку, коли класи є лінійно не роздільними та їх кількість апріорно невідома. Архітектура ансамблю ядерних кластерувальних нейронних мереж містить п'ять шарів обробки інформації.

Вихідна інформація, яка підлягає кластеризації, подається на нульовий (вхідний) шар системи у вигляді послідовності $x(1),x(2),...,x(k),...,x(N),...$, звідки надходить на перший прихований шар (RL) радіально-базисних функцій, утворений R-нейронами. Саме в цьому шарі відбувається підвищення розмірності вхідного простору за допомогою системи ядерних функцій $\varphi_1(x),\varphi_2(x),...,\varphi_l(x),...,\varphi_h(x)$, $h>n$, в якості яких використовуються або традиційні гаусіани, або інші дзвоноподібна функції.

Другий прихований шар NL реалізує елементарну операцію нормалізації сигналу $\varphi(x(k))$ виду $\tilde{\varphi}(x(k))=\frac{\varphi(x(k))}{\left\|\varphi(x(k))\right\|}$ необхідну для ефективної роботи третього прихованого шару SL, утвореного (M-1) самоорганізовними картами Кохонена SOM^m, кожна з яких працює в припущенні, що в оброблюваній виборці даних міститься m класів.

Якість кластеризації, що забезпечується кожною SOM^m, оцінюється за допомогою того чи іншого індексу валідації в четвертому прихованому шарі VL, де обчислюються відповідні індекси $VI^2,VI^3,...,VI^m,...,VI^M$ для кожного з можливих $m=2,3,...,M$.

Процес самонавчання цієї системи реалізується на рівні першого шару RL, де налаштовуються центри w_l, $l=1,2,...,h$ ядерних функцій $\varphi_l(x)$, і третього прихованого шару SL, де уточнюються синаптичні ваги w_j^m, $m=2,3,...,M$, $j=1,2...,m$ кожної нейронної мережі SOM^m ансамблю.

Розглянемо процес налаштування центрів ядерних функцій, що складається з послідовності таких кроків:

Крок 0: задати порогове значення Δ, що визначає рівень нерозрізненості двох сусідніх ядерних функцій, максимально можливу кількість цих функцій h і параметр рецепторного поля γ_φ.

Крок 1: при подаванні на вхід системи першого вектора-спостереження $x(1)$ формується перший центр w_1 і перша радіально-базисна функція

$$\varphi_1(x) = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - w_1\|^2},$$

де $w_1 = x(1)$.

Крок 2: при подаванні на вхід системи другого спостереження $x(2)$ перевіряється нерівність

$$\|x(2) - w_1\| \le \Delta$$

і якщо вона виконується, то $x(2)$ не формує новий центр, якщо ж виконується умова

$$\Delta < \|x(2) - w_1\| \le 2\Delta, \tag{3}$$

то w_1 коригується відповідно до правила самонавчання Т. Кохонена «Переможець отримує все:

$$w_1(2) = w_1(1) + \eta(2)(x(2) - w_1(1)), \tag{4}$$

де $w_1(1) = x(1)$, $0 < \eta(2) < 1$- параметр кроку навчання.

Якщо ж виконується умова

$$2\Delta < \|x(2) - w_1\|,$$

то формується нова ядерна функція

$$\varphi_2(x) = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - w_2\|^2} = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - x(2)\|^2}.$$

Цей процес реалізується при надходженні кожного нового спостереження $x(k)$. Якщо ж на кроці N буде сформовано h радіально-базисних функцій, то в подальшому їх кількість не збільшується, а уточнення вже сформованих центрів w_l, $l = 1,2,...,h$ може проводитися тільки згідно з умовою (3) і правилом самонавчання (4).

У четвертому прихованому шарі системи проводиться оцінка якості кластеризації за допомогою критерій Девіса-Булдена:

$$DB(m) = \sum_{j=1}^{m} \max_{\substack{1 \le q \le m \\ q \ne j}} \frac{s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k)) - s(w_q^m(k), u_q(k), \tilde{\varphi}(x(k)))}{D(w_j^m(k), w_q^m(k))},$$

де $D(w_j^m(k), w_q^m(k))$ – відстань між центроїдами;

$s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k)))$ – характеристики внутрішньокластерного розсіювання для j-го кластеру;

$u_j(k)$ – чітка функція належності вектора $\tilde{\varphi}(x(k))$ до j-го кластеру.

За оптимальну кількість кластерів m^* обирається значення, що забезпечує мінімум $DB(m)$, тобто

$$DB(m^*) = \min_m \{DB(2), DB(3), ..., DB(M)\},$$

який розраховується у вихідному шарі.

При обробці нестаціонарних даних, що надходять в online режимі, індекс $DB(m)$ доцільно модифікувати для роботи в режимі «ковзного вікна» розмірності $1<s<N$. При цьому модифікації піддаються тільки характеристики міжкластерної відстані, які розраховуються на «ковзному вікні» за допомогою виразу

$$\mathrm{s}(w_j^m(k),\mathrm{u}_j(k),\tilde{\varphi}(x(k)),\mathrm{s})=\left(\frac{\sum\limits_{\tau=k-s+1}^{k}u_j(\tau)\left\|\tilde{\varphi}(x(\tau))-w_j^m(k)\right\|^2}{\sum\limits_{\tau=k-s+1}^{k}u_j(\tau)}\right)^{\frac{1}{2}},$$

при цьому передбачається, що обсяг вибірки N необмежений, а зростає з плином часу $k=1,2,...,\mathrm{N},\mathrm{N}+1,...$

Запропонований ансамбль ядерних самоорганізовних карт Т. Кохонена для кластеризації потоків даних завдяки ядерному шару має можливість обробляти дані, коли кластери утворюють довільну форму. Введений ансамбль нейронних мереж простий у реалізації та дозволяє вирішувати досить широкий клас задач динамічного інтелектуального аналізу даних та інтелектуального аналізу потоків даних.

У **четвертому розділі** розглядаються ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних у припущенні про невідому кількість кластерів та їх форму.

Ситуація істотно ускладнюється, якщо кластери, які формуються, перетинаються у просторі ознак. Такі завдання вирішуються за допомогою методів нечіткої кластеризації, найбільш популярним з яких є алгоритм нечітких С-середніх (FCM). Для роботи в online режимі з успіхом можуть бути використані нечіткі карти Кохонена для кластеризації.

При цьому необхідно пам'ятати, що ефективність процедур нечіткої кластеризації обмежується так званим ефектом концентрації норм, коли результати виявляються незадовільними при високих розмірностях простору ознак.

У зв'язку з цим є доцільною розробка online методу нечіткої кластеризації даних високої розмірності на основі ансамблів для кластеризації в умовах невідомої кількості класів у потоці оброблюваної інформації.

У класі процедур нечіткої кластеризації з математичної точки зору найбільш коректними є алгоритми, які засновані на цільових функціях та вирішують задачу їх оптимізації за наявності тих або інших обмежень. Тут найбільш популярним є імовірнісний алгоритм нечіткої кластеризації, заснований на оптимізації цільової функції

$$E\left(u_j(k),w_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}(k)\left\|x(k)-w_j\right\|^2=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}(k)\sum_{i=1}^{n}\left(x_i(k)-w_{ji}\right)^2 \qquad (5)$$

за наявності обмежень

$$\sum_{j=1}^{m}u_j(k)=1,\ 0\le\sum_{k=1}^{N}u_j(k)\le N. \qquad (6)$$

$$0\le\sum_{k=1}^{N}u_j(k)\le N. \qquad (7)$$

Тут $u_j(k) \in [0,1]$ – рівень нечіткої належності вектора спостережень $x(k)$ до j-го кластеру, w_j – центроїд-ваги j-го кластеру, β – фаззіфікатор, що визначає розмитість границь між кластерами.

Вирішення задачі оптимізації (5) за наявності обмежень (6), (7) за допомогою невизначених множників Лагранжа приводить до результату

$$\begin{cases} u_j(k) = \dfrac{\left(\left\|x(k) - w_j\right\|^2\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(\left\|x(k) - w_l\right\|^2\right)^{\frac{1}{1-\beta}}}, \\ w_j = \dfrac{\sum\limits_{k=1}^{N} u_j^{\beta}(k)\, x(k)}{\sum\limits_{k=1}^{N} u_j^{\beta}(k)}, \end{cases} \tag{8}$$

який при $\beta = 2$ повністю збігається з FCM Дж. Бездека.

Імовірнісний алгоритм нечіткого кластеризації (8) набув широкого поширення в Data Mining, однак, втрачає свою ефективність в задачах обробки даних високої розмірності через те, що виникає ефект концентрації норм. Для подолання цього недоліку було запропоновано використовувати, так званий поліноміальний фаззіфікатор, що веде до процедури відомої як нечіткий метод С-середніх з поліноміальним фаззіфікатором (PFCM).

Для вирішення завдань кластеризації даних високої розмірності була запропонована модифікація FCM зі зважуванням кожної з ознак $x_i(k)$, що утворюють вектор-образ $x(k) \in R^n$, $i = 1,2,...,n$.

Об'єднуючи ці два підходи, можна ввести до розгляду цільову функцію нечіткої кластеризації виду

$$\begin{aligned} E\left(u_j(k), w_j, \alpha, \gamma_{ji}\right) &= \sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)\left\|x(k) - w_j\right\|_{\Gamma_j^2}^2 = \\ &= \sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)\sum_{i=1}^{n}\gamma_{ji}^2\left(x_i(k) - w_{ji}\right)^2 \end{aligned} \tag{9}$$

з обмеженнями (6), (7) та

$$\sum_{i=1}^{n}\gamma_{ji} = Tr\Gamma_j = 1 \quad \forall\, j = 1,2,...m. \tag{10}$$

Оптимізація цільової функції (9) при обмеженнях (6), (7) та (10) за допомогою невизначених множників Лагранжа приводить до результату

$$\begin{cases} u_j(k) = \dfrac{\alpha-1}{2\alpha} + \dfrac{1-m\dfrac{\alpha-1}{2\alpha}}{\sum\limits_{l=1}^{m}\dfrac{\left\|x(k)-w_j\right\|^2_{\Gamma_l^2}}{\left\|x(k)-w_l\right\|^2_{\Gamma_l^2}}}, \\ \gamma_{ji} = \left(\sum\limits_{h=1}^{n}\left(\dfrac{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\right)\left(x_i(k)-w_{ji}\right)^2}{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\right)\left(x_h(k)-w_{ji}\right)^2}\right)\right)^{-1}, \\ w_{ji} = \dfrac{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\right)\gamma_{ji}^2 x_i(k)}{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\right)\gamma_{ji}^2}. \end{cases} \quad (11)$$

що є узагальненням (8) і збігається з ним при $\alpha = 1, \gamma_{ji} = m^{-1}$.

Останнє співвідношення (11) для розрахунку центроїдів кластерів може бути переписано у рекурентній формі

$$w_j(k) = w_j(k-1) + \eta(k)\left(\alpha u_j^2(k-1)+(1-\alpha)u_j(k-1)\right)\times \\ \times\Gamma_j^2(k-1)\left(x(k)-w_j(k-1)\right). \quad (12)$$

Для вирішення завдання кластеризації в умовах, коли кількість кластерів невідома, пропонується використовувати ансамбль кластерувальних нейро-фаззі мереж Кохонена, архітектура якого наведена на рисунку 1. Цей ансамбль містить $(M-1)_q$ $FSOM_p^{[m]}$, де індекс [m] означає кількість кластерів, на яку ця мережа розбиває оброблювану вибірку – тобто кількість нейронів в шарі Кохонена KL, а p – індекс конкретного фаззіфікатора, що приймає q значень. Всі $FSOM_p^{[m]}$ навчаються за допомогою однотипних процедур (11), (12), які відрізняються одна від одної тільки значеннями m та α.

У блоках $MEXB_p^{[m]}$ оцінюється якість кластеризації, що забезпечується конкретною FSOM, а вихідний шар ансамблю DM з $(M-1)_q$ результатів попередніх шарів виділяє найкращий, тобто кількість кластерів m^* в оброблюваних даних, центроїди сформованих кластерів $w_1^*, w_2^*, ..., w_{m^*}^*$ і рівні належності кожного спостереження $u_1^*(k), u_2^*(k), ..., u_{m^*}^*(k)$ до відповідного кластеру.

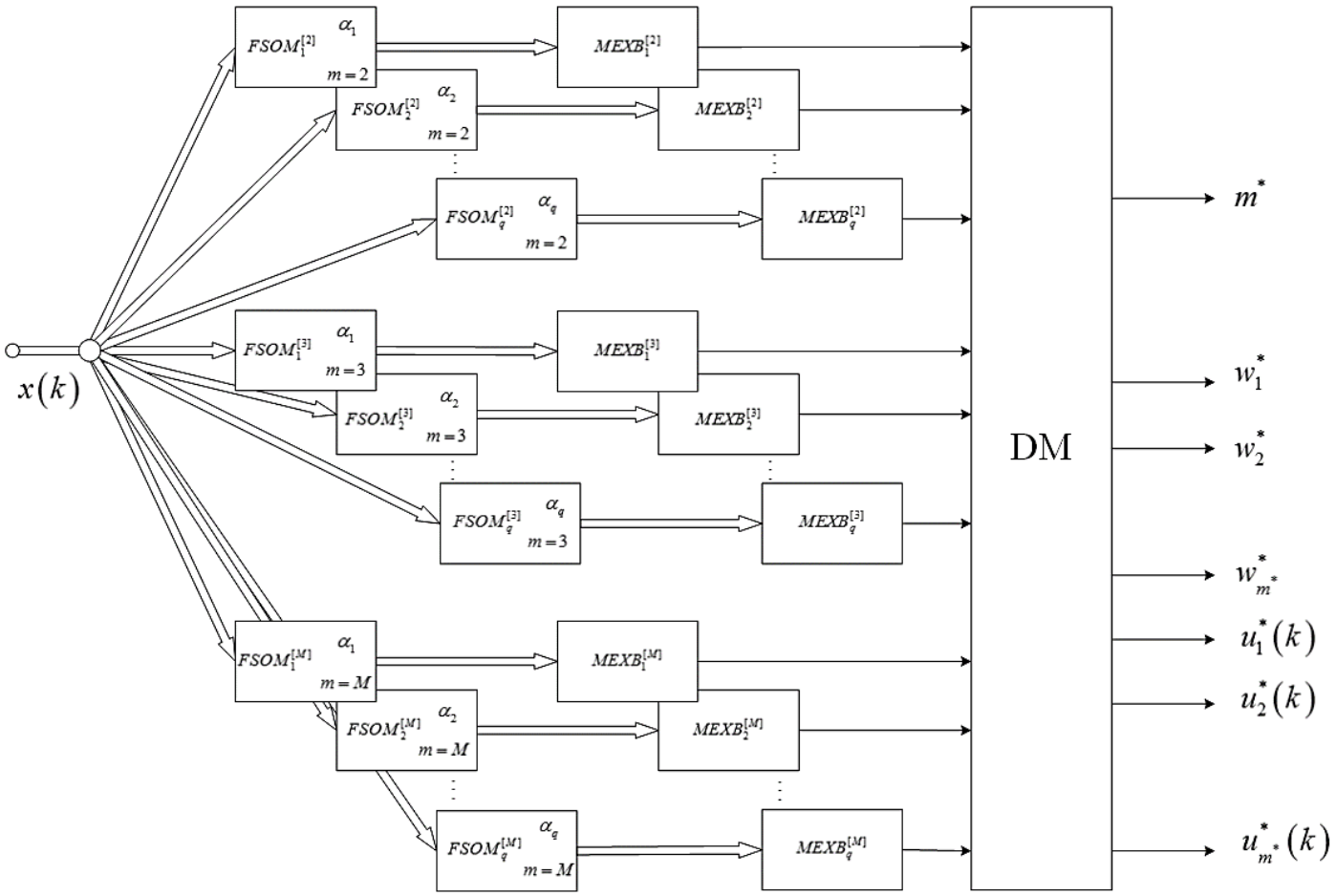

Рисунок 1 – Архітектура нечіткого ансамблю для кластеризації потоку даних

Для оцінки якості кластеризації кожним з елементів ансамблю може бути використаний будь-який з індексів нечіткої кластеризації, де одним з найбільш популярних є індекс Ксі-Бені, який для FCM-процедури в разі m кластерів може бути записаний у формі

$$XB^{[m]} = \frac{\sum_{k=1}^{N}\sum_{j=1}^{m} u_j^2(k)\left\|x(k)-w_j\right\|^2}{\dfrac{N}{\min\limits_{l\neq j}\left\|w_j-w_l\right\|^2}} = \frac{NXB^{[m]}}{DXB^{[m]}}.$$

Для послідовної обробки можна ввести online версію XB-індексу у вигляді

$$XB^{[m]}(k) = \frac{NXB^{[m]}(k)}{DXB^{[m]}(k)} = \frac{NXB^{[m]}(k-1)}{\min\limits_{l\neq j}\left\|w_j(k)-w_l(k)\right\|^2} + \frac{\dfrac{1}{k}\left(\sum\limits_{j=1}^{m} u_j^2(k)\left\|x(k)-w_j(k)\right\|^2 - NXB^{[m]}(k-1)\right)}{\min\limits_{l\neq j}\left\|w_j(k)-w_l(k)\right\|^2}. \quad (13)$$

За аналогією можна ввести модифікацію розширеного індексу Ксі-Бені (extended Xie-Beni index) для цільової функції для online версії

$$MEXB_p^{[m]}(k)=\frac{NMEXB_p^{[m]}(k)}{DMEXB_p^{[m]}(k)}=\frac{NMEXB_p^{[m]}(k-1)}{\min\limits_{l\neq j}\left\|w_{pj}^{[m]}(k)-w_{pl}^{[m]}(k)\right\|^2}+$$

$$+\left(\frac{1}{k}\left(\sum_{j=1}^{m}\left(\alpha_p\left(u_{pj}^{[m]}(k)\right)^2+(1-\alpha)u_{pj}^{[m]}(k)\right)\times\right.\right.$$

$$\left.\left.\times\left\|x(k)-w_{pj}^{[m]}(k)\right\|^2_{\left(\Gamma_{pj}^{[m]}(k)\right)^2}-NMEXB_p^{[m]}(k-1)\right)\right)\left(\min_{l\neq j}\left\|w_{pj}^{[m]}(k)-w_{pl}^{[m]}(k)\right\|^2\right)^{-1}.$$

У процесі обробки даних блок DM знаходить $FSOM_p^{[m^*]}$ з найкращим значенням $MEXB_p^{[m^*]}$ і результати роботи саме цієї нейро-фаззі мережі визначають кінцевий результат кластеризації.

Таким чином, запропоновано ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних великої розмірності, що послідовно надходять на вхід системи спостереження за спостереженням, який дозволяє в процесі самонавчання налаштовувати не тільки свої параметри, а й архітектуру в on-line режимі та вирішувати завдання кластеризації потоку даних за умови апріорно невідомої форми та кількості кластерів.

У **п'ятому розділі** описано ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена з використанням імовірнісно-можливісного підходу для кластеризації потоків даних, коли кількість та форма кластерів заздалегідь невідома.

В імовірнісному алгоритмі нечіткої кластеризації виконується оптимізація цільової функції (5) з обмеженнями (6), (7), яка з використанням невизначених множників Лагранжа набуває вигляду (8), та при $\beta=2$ повністю збігається з FCM Дж. Бездека.

Для можливісного підходу до кластеризації критерій, що мінімізується, має вигляд

$$E\left(u_j(k),w_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}(k)\left\|x(k)-w_j\right\|^2+\sum_{j=1}^{m}\mu_j\sum_{k=1}^{N}\left(1-u_j(k)\right)^{\beta},\qquad(14)$$

де скалярний параметр $\mu_j>0$ визначає відстань, на якій рівень належності приймає значення 0.5, тобто якщо $\left\|x(k)-w_j\right\|^2=\mu_j$, то $u_j(k)=0.5$.

Мінімізація критерію (14) за змінними $u_j(k)$, w_j, μ_j призводить до системи рівнянь:

$$\begin{cases}\dfrac{\partial E(u_j(k),w_j)}{\partial u_j(k)}=0,\\[2ex]\dfrac{\partial E(u_j(k),w_j)}{\partial \mu(k)}=0,\\[2ex]\nabla_{w_j}E(u_j(k),w_j)=\vec{0},\end{cases}\qquad(15)$$

а розв’язання перших двох рівнянь дає результат:

$$u_j^{pos}(k)=\left(1+\left(\frac{\left\|x(k)-w_j\right\|^2}{\mu_j}\right)^{\frac{1}{\beta-1}}\right)^{-1},$$

$$\mu_j=\frac{\sum_{k=1}^{N}u_j^{\beta}(k)\left\|x(k)-w_j\right\|^2}{\sum_{k=1}^{N}u_j^{\beta}(k)}.$$

Розв’язання третього рівняння системи (15) для евклідової норми має вигляд:

$$w_j^{pos}=\frac{\sum_{k=1}^{N}u_j^{\beta}(k)x(k)}{\sum_{k=1}^{N}u_j^{\beta}(k)}.$$

Аналіз (15) показує, що для обчислення рівнів належності $u_j(k)$ замість звичайної функції Лагранжа можна використовувати її локальну модифікацію

$$L_k\left(u_j(k),w_j,\lambda(k)\right)=\sum_{j=1}^{m}u_j^{\beta}(k)\left\|x(k)-w_j\right\|^2+\lambda(k)\left(\sum_{j=1}^{m}u_j(k)-1\right), \tag{16}$$

при чому оптимізація виразу (16) за допомогою процедури Ерроу–Гурвіца–Удзави веде до процедури:

$$u_j^{pr}(k)=\frac{\left(\left\|x(k)-w_j(k)\right\|^2\right)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}\left(\left\|x(k)-w_l(k)\right\|^2\right)^{\frac{1}{1-\beta}}}, \tag{17}$$

$$\begin{aligned}w_j^{pr}(k+1)&=w_j^{pr}(k)-\eta(k)\nabla_{w_j}L_k(u_j(k),w_j^{pr}(k),\lambda(k))=\\&=w_j^{pr}-\eta(k)u_j^{\beta}(k)\left(x(k+1)-w_j^{pr}(k)\right),\end{aligned} \tag{18}$$

де $\eta(k)$ – параметр кроку навчання; $w_j^{pr}(k)$ – прототипи j-го кластера, що обчислюються на вибірці з k спостережень.

Процедура (17)–(18) для $\beta=2$ збігається з градієнтною процедурою кластеризації Парка–Деггера

$$u_j^{pr}(k)=\frac{\left\|x(k)-w_j(k)\right\|^{-2}}{\sum_{l=1}^{m}\left\|x(k)-w_j(k)\right\|^{-2}},$$

$$w_j^{pr}(k+1)=w_j^{pr}(k)+\eta(k)u_j^2(k)\left(x(k+1)-w_j^{pr}(k)\right). \tag{19}$$

Нескладно бачити, що процедура (19) є не що інше, як WTM-правило самонавчання Т. Кохонена з функцією сусідства $u_j^2(k)$.

У межах можливісного підходу локальний критерій набуває вигляду

$$E_k\left(u_j(k), w_j\right) = \sum_{j=1}^{m} u_j^{\beta}(k)\left\|x(k) - w_j\right\|^2 + \sum_{j=1}^{m} \mu_j\left(1 - u_j(k)\right)^{\beta},$$

а результат його оптимізації записується як

$$u_j^{pos}(k) = \left(1 + \left(\frac{\left\|x(k) - w_j(k)\right\|^2}{\mu_j(k)}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \tag{20}$$

$$w_j^{pos}(k+1) = w_j^{pos}(k) - \eta(k) u_j^{\beta}(k)\left(x(k+1) - w_j^{pos}(k)\right), \tag{21}$$

$$\mu_j(k+1) = \frac{\sum_{p=1}^{k} u_j^{\beta}(p)\left\|x(p) - w_j(k+1)\right\|^2}{\sum_{p=1}^{k} u_j^{\beta}(p)}. \tag{22}$$

У квадратичному випадку (при $\beta = 2$) процедура (20)–(22) перетворюється на достатньо просту конструкцію

$$u_j^{pos}(k) = \frac{\mu_j(k)}{\mu_j(k) + \left\|x(k) - w_j(k)\right\|^2},$$

$$w_j^{pos}(k+1) = w_j^{pos}(k) + \eta(k) u_j^{2}(k)\left(x(k+1) - w_j^{pos}(k)\right),$$

$$\mu_j(k+1) = \frac{\sum_{p=1}^{k} u_j^{2}(p)\left\|x(p) - w_j(k+1)\right\|^2}{\sum_{p=1}^{k} u_j^{2}(p)}.$$

Паралельне застосування адаптивних імовірнісної і можливісної процедур призводить до об'єднаної процедури нечіткої кластеризації

$$\begin{cases} w_j^{pr}(k) = w_j^{pos}(k-1) - \eta(k) u_j^{pos\beta}(k-1)\left(x(k+1) - w_j^{pos}(k)\right), \\ u_j^{pr}(k) = \left(\left\|x(k) - w_j^{pr}(k)\right\|^2\right)^{\frac{1}{1-\beta}}\left(\sum_{l=1}^{m}\left(\left\|x(k) - w_j^{pr}(k)\right\|^2\right)^{\frac{1}{1-\beta}}\right)^{-1}, \\ w_j^{pos}(k) = w_j^{pr}(k-1) - \eta(k) u_j^{pr\beta}(k)\left(x(k+1) - w_j^{pr}(k)\right), \\ \mu_j(k) = \left(\sum_{p=1}^{k} u_j^{pr\beta}(p)\left\|x(k) - w_j^{pos}(k)\right\|^2\right)\left(\sum_{p=1}^{k} u_j^{pr\beta}\right)^{-1}, \\ u_j^{pos}(k) = \left(1 + \left(\frac{\left\|x(k) - w_j^{pos}(k)\right\|^2}{\mu_j(k)}\right)\right)^{-1}, \; j = 1, 2, \ldots, m. \end{cases} \tag{23}$$

Ознакою правильного знаходження прототипів (а отже і коректної кластеризації), з використанням процедури (23) є виконання нерівності

$$\sum_{l=1}^{m}\left\|w_j^{pr}(k)-w_l^{pos}(k)\right\|\le\varepsilon,$$

де ε визначає прийнятну точність кластеризації.

Таким чином, процедура (23) може працювати як в пакетному режимі для обробки заданої вибірки, так і в послідовному режимі, де кількість спостережень визначається дискретним часом $k=1,2,\ldots,N,N+1,\ldots$ В останньому випадку за допомогою цієї процедури послідовно оброблюються спостереження, що надходять на опрацювання. Отже, у випадку нестаціонарних даних рівні належності та прототипи кластерів перебудовуються відповідно до нових даних.

Для оцінки якості кластерування кожним з елементів ансамблю може бути використаний будь-який з індексів нечіткого кластерування, де одним з найбільш популярних є індекс Ксі-Бені, який розраховується відповідно до рівняння (13).

Запропоновані архітектура та алгоритм самонавчання нейро-фаззі системи, призначеної для вирішення завдання online кластеризації потоку даних в умовах, коли кластери, які формуються, перекриваються та їх кількість заздалегідь невідома. Запропонована система є ансамблем нейро-фаззі самоорганізовних карт Т. Кохонена, кожна з яких відрізняється від інших кількістю нейронів. Налаштування кожного з членів ансамблю відбувається за допомогою модифікованого WTM-правила самонавчання, при цьому в процесі налаштування проводиться автоматичне зважування усіх компонент оброблюваних векторів. Запропонований підхід є узагальненням низки відомих процедур нечіткої імовірнісної та можливісної кластеризації та може бути використаний для вирішення задач аналізу потоків даних.

У **шостому розділі** описаний процес імітаційного моделювання запропонованих у роботі ансамблів самоорганізовних карт Т. Кохонена для кластеризації потоків даних, коли дані надходять на обробку спостереження за спостереженням. Також використовувалися ансамблі нейронних мереж та нейро-фаззі систем з ядерними функціями для кластеризації даних коли форма кластерів заздалегідь невідома.

Була вирішена задача кластеризації на основі тестових вибірок з UCI-репозиторію за допомогою розробленого ансамблю самоорганізовних карт Т. Кохонена для кластеризації потоків даних, коли кількість класів апріорно невідома. Проведене імітаційне моделювання вирішення задачі нечіткої кластеризації даних на основі вибірки, отриманої у рамках науково-дослідної роботи, що фінансувалася Міністерством охорони здоров'я України «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища», за допомогою запропонованого ансамблю самоорганізовних карт Т. Кохонена для кластеризації потоків даних. Проведено імітаційне моделювання кластеризації даних на основі ансамблю нейро-фаззі самоорганізовних карт Т. Кохонена з використанням імовірнісно-можливісного підходу, що дозволяє обробляти дані, коли апріорно невідома кількість та форма класів.

У **висновках** сформульовано наукові та практичні результати, що одержані у дисертаційній роботі.

У **додатку** наведено акти про впровадження результатів дисертаційної роботи у науково-дослідну роботу, що фінансувалася Міністерством охорони здоров'я України, «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища» (КПКВК 2301020, № 0113U002536) та навчальний процес.

ВИСНОВКИ

У дисертаційній роботі представлені результати, які є відповідно до поставленої мети рішенням актуального завдання обробки багатовимірних масивів даних в умовах невизначеності за допомогою ансамблю нечітких методів ядерної кластеризації на основі ядерних функцій. Проведені дослідження дозволили зробити такі висновки.

1. Розроблено ансамбль самоорганізовних карт Т. Кохонена для кластеризації даних за умов апріорі невідомої кількості класів з використанням онлайн модифікованого методу К-середніх;
2. Розроблено ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних за умов, коли класи є лінійно нероздільними та довільним чином перетинаються у просторі ознак, що базується на використанні онлайн методу нечітких С-середніх за умов, коли кількість кластерів апріорі невідома;
3. Розроблено ансамбль ядерних самоорганізовних карт Т. Кохонена для кластеризації потоків даних за умов, коли кластери є лінійно нероздільними, що характеризується введенням додаткового ядерного шару для підвищення розмірності вхідного простору;
4. Розроблено ансамбль самоорганізовних нечітких карт Т. Кохонена, що одночасно реалізує процедури імовірнісної та можливісної кластеризації потоків даних;
5. Розроблено ансамбль нейро-фаззі мереж на основі імовірнісно-можливісного підходу для кластеризації потоків даних;
6. Проведено імітаційне моделювання розроблених методів та моделей та вирішення практичних задач нечіткої кластеризації потоків даних високої розмірності з використанням даних які було взято з UCI-репозиторію;
7. Було проведено апробацію запропонованих методів на реальних даних для кластеризації даних високої розмірності, які були отримані у ході дослідження щурів у рамках НДР бюджетного фінансування «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища».
8. Вирішено практичне завдання кластеризації даних. У порівнянні з іншими методами кластеризації розроблені методи показали простоту реалізації з точки зору математичного апарату та підвищення точності кластеризації потоків даних.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Список публікацій здобувача, в яких опубліковані основні наукові результати дисертації:

1. P. Zhernova, A. Deyneko, Z. Deyneko, I. Pliss and V. Ahafonov, "Data Stream Clustering in Conditions of an Unknown Amount of Classes," In: Hu Z., Petoukhov S., Dychka I., He M. (eds) *Advances in Computer Science for Engineering and Education. ICCSEEA 2018. Advances in Intelligent Systems and Computing*, vol 754. Springer, Cham, pp. 410-419, 2019. (Входить до міжнародних наукометричних баз SCOPUS.)
2. Є. Бодянський, А. Дейнеко, П. Жернова, О. Золотухін та Я. Хаустова, «Послідовне ядерне нечітке кластерування великих масивів даних на основі гібридної системи обчислювального інтелекту,» *Вісник Національного університету "Львівська політехніка". Інформаційні системи та мережі,* № 829, pp. 20-24, 2017. (Входить до міжнародної наукометричної бази Google Scholar)

3. Є. Бодянський, А. Дейнеко, П. Жернова та В. Репін, «Онлайн модифікація методу Х-середніх на основі ансамблю самоорганізовних мап Т. Когонена,» *Збірник наукових праць «Розвиток транспорту»,* № 1, pp. 96-107, 2017.

4. П. Жернова та Є. Бодянський, «Ядерна нечітка кластеризація потоків даних на основі ансамблю нейронних мереж,» *Сучасний стан наукових досліджень та технологій в промисловості,* № 4(6), pp. 42-49, 2018. (Входить до міжнародної наукометричної бази Index Copernicus International)

5. Y. Bodyanskiy, I. Perova and P. Zhernova, "Online fuzzy clustering of high - dimensional data based on ensembles in data stream mining tasks," *Innovative Technologles & Scientific Solutions for Industries*, no. 1(7), pp. 16-24, 2019.

6. П. Жернова та Є. Бодянський, «Нечітка імовірнісно-можливісна послідовна кластеризація даних на основі ансамблевого підходу,» *Науково-технічний журнал «Прикладна радіоелектроніка»,* № 1,2, pp. 40-45, 2019. (Входить до міжнародної наукометричної бази Index Copernicus International)

Результати, які засвідчують апробацію матеріалів дисертації:

7. Е. Бодянский, А. Дейнеко, П. Жернова и В. Репин, «Адаптивная модификация метода Х-средних на основе ансамбля кластеризующих нейронных сетей Т. Кохонена,» в *Матеріали VI Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології»*, Одесса, 2017.

8. Е. Бодянский, П. Жернова и А. Дейнеко, «Кластеризующий ансамбль нейронных сетей и его обучение в условиях неизвестного количества классов,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018.

9. А. Дейнеко, П. Жернова, І. Плісс та О. Чала, «Модифікована нечітка ймовірнісна нейронна мережа,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018.

10. P. Zhernova, A. Deyneko, Y. Bodyanskiy and V. Riepin, "Adaptive kernel data streams clustering based on neural networks ensembles in conditions of uncertainty about amount and shapes of clusters," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018. (Входить до міжнародної науково-метричної бази SCOPUS)

11. Deineko, P. Zhernova, B. Gordon, O. Zayika, I. Pliss and N. Pabyrivska, "Data stream online clustering based on fuzzy expectation-maximization approaching formation on submission," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018. (Входить до міжнародної науково-метричної бази SCOPUS)

12. П. Жернова, «Вероятностно-возможностный подход для кластеризации потоков данных на основе ансамблей нейронных сетей,» в *Материалы международной научно-практической конференции «Информационные технологии и системы»*, Харьков, 2019.

13. П. Жернова та А. Лобинцев, «Кластеризація даних високої розмірності з використанням можливісного підходу,» в *Матеріали 23-го Міжнародного молодіжного форуму «Радіоелектроніка та молодь в 21 столітті»*, Харьков, 2019.

14. П. Жернова та Є. Бодянський, «Нейро-фаззі мережа та її навчання для кластеризації потоків даних високої розмірності,» в *Матеріали V міжнародної науково-практичної конференції «Обчислювальний інтелект (результати, проблеми, перспективи)»*, Ужгород, 2019.

АНОТАЦІЯ

Жернова П.Є. Нечітка кластеризація потоків даних за умови невідомої кількості кластерів. – 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 Міністерства освіти і науки України, Харків, 2019.

У дисертаційній роботі запропоновано ансамбль самоорганізовних карт Т. Кохонена, який базується на використанні онлайн методу К-середніх. Такий підхід дозволяє обробляти інформацію, яка надходить на вхід системи спостереження за спостереженням. На відміну від існуючих методів кластеризації використання ансамблевого підходу дозволяє обійти проблему, коли кількість класів заздалегідь невідома, оскільки кожна з мереж Кохонена налаштована на свою кількість кластерів. Вдосконалено метод, заснований на ансамблевому підході, з використанням ядерних самоорганізовних карт Т. Кохонена, що дозволило завдяки додатковому прихованому ядерному шару нейромережі підвищити розмірність вхідного простору, що дає змогу кластеризувати дані, які є лінійно нероздільними. Розроблено ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних, який за допомогою використання вдосконаленого методу С-середніх та додаткового ядерного шару здатний обробляти інформацію, що є лінійно нероздільною, а також обробляти кластери довільної форми. Саме це дозволяє обробляти дані високої розмірності та уникнути ефекту концентрації норм. Вдосконалено ансамбль самоорганізовних карт Т. Кохонена для кластеризації потоків даних високої розмірності, який обробляє інформацію, що надходить на вхід системи з використанням двох підходів: імовірнісного та можливісного.

Ключові слова: ансамбль нейронних мереж, самоорганізовна карта Т. Кохонена, метод К-середніх, нечіткий метод С-середніх, нейро-фаззі мережа, індекс валідації.

АННОТАЦИЯ

Жернова П.Е. Нечеткая кластеризация потоков данных при условии неизвестного количества кластеров. – 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 Министерства образования и науки Украины, Харьков, 2019.

Диссертационная работа посвящена решению актуальной научной задачи разработки новых методов для кластеризации потоков данных в условиях априорной неопределенности о количестве и форме кластеров.

Впервые предложен ансамбль самоорганизующихся карт Т. Кохонена для кластеризации данных в условиях априори неизвестного количества классов с использованием онлайн метода К-средних.

Впервые предложен ансамбль нейро-фаззи самоорганизующихся карт Т. Кохонена для кластеризации потоков данных в условиях, когда классы являются линейно неразделимыми и произвольным образом пересекаются в пространстве

признаков, основанный на использовании онлайн метода нечетких С-средних в условиях, когда количество кластеров априори неизвестно.

Усовершенствован ансамбль ядерных самоорганизующихся карт Т. Кохонена для кластеризации потоков данных в условиях, когда кластеры являются линейно неразделимыми, характеризующийся введением дополнительного ядерного слоя для повышения размерности входного пространства.

Усовершенствован ансамбль самоорганизующихся нечетких карт Т. Кохонена, одновременно реализующий процедуры возможностной и вероятностной кластеризации потоков данных.

Проведено численное моделирование и решение ряда задач, на основе которого показана эффективность использования предложенных методов кластеризации с использованием ансамблей нейронных сетей и нейро-фаззи систем на основе самоорганизующихся карт Т. Кохонена. Они могут быть использованы в любой области, где данные представлены в числовой форме.

Результаты исследований внедрены в научно-исследовательскую работу, финансируемую Министерством здравоохранения Украины «Установить механизмы адаптации к совмещенному действию химических и физических факторов окружающей среды», г. Харьков.

Ключевые слова: ансамбль нейронных сетей, самоорганизующаяся карта Т. Кохонена, метод К-средних, нечеткий метод С-средних, нейро-фаззи сеть, индекс валидации.

ABSTRACT

Zhernova P.Ye. Data streams fuzzy clustering in conditions of unknown number of clusters. – Manuscript.

A thesis for the candidate degree in technical sciences in the specialty 05.13.23 – systems and methods of artificial intelligence. – Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2019.

In the thesis proposed the ensemble of T. Kohonen's self-organizing maps, which is based on using the online method of K-means. This approach allows processing the information in online mode that is fed to the input of the system. Unlike existing clustering methods, the use of the ensemble approach allows to bypass the problem when the number of classes is unknown in advance because each of the Kohonen's networks is configured for its own number of clusters. The method based on the ensemble approach, using the neural self-organizing T. Kohonen's maps, has been improved, which allowed using additional hidden layer of the neural network to increase the dimension of the input space. A neuro-fuzzy ensemble of T. Kohonen's self-organizing maps has been developed for data stream clustering, when the use of the improved C-means and additional neural layer method is able to process information that is linear inseparable, as well as to process clusters of arbitrary form. This allows processing data of high dimensionality and avoiding the concentration of norms effect. The ensemble of self-organizing maps has been improved for the clustering of high-dimensional data streams, which processes information entering the system using several approaches: probabilistic and possibilistic.

Key words: ensemble of neural networks, T. Kohonen's self-organizing maps, K-means method, fuzzy C-means method, neuro-fuzzy network, validation index.