Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

ХАУСТОВА ЯНА ВОЛОДИМИРІВНА

УДК 004.032.26

МЕТОДИ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ НА ОСНОВІ ЯДЕРНИХ ФУНКЦІЙ В ЗАДАЧАХ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

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 **Пелешко Дмитро Дмитрович**, Національний університет «Львівська політехніка», МОН України, професор кафедри інформаційних технологій видавничої справи; кандидат технічних наук, доцент **Гороховатський Олексій Володимирович,** Харківський національний економічний університет імені Семена Кузнеця, МОН України, доцент кафедри інформатики та комп'ютерної техніки

Захист відбудеться «01» березня 2017 р. о 13.00 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 «_26_» січня 2017 р.

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

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. На цей час для розв’язання широкого класу задач інтелектуального аналізу даних, і перш за все кластеризації, що є невід'ємною частиною загальної проблеми Data Mining, відома низка методів обчислювального інтелекту на основі нейро-фаззі підходу. Розроблено потужний математичний апарат, що дозволяє розв’язувати задачі кластеризації в різних областях: медицині, науці та техніці, при керуванні виробництвом.

Однак, більшість з відомих методів за своєю суттю є чіткими процедурами, де припускається, що кластери лінійно розділимі, дані надходять у пакетному режимі. Також розроблена низка методів нечіткої кластеризації, де припускається, що кожне окреме спостереження може з різними рівнями належності відноситися до декількох класів. Однак, при цьому кластери, що формуються мають опуклу форму. Разом з тим, зараз на перший план виходять задачі, що пов'язані з Dynamic Data Mining, Data Stream Mining і Big Data, коли дані надходять на обробку у вигляді потоку інформації. Більш того, в ситуаціях, коли кластери можуть перетинатися та мати довільну форму, виникає необхідність розробки методів нечіткої кластеризації в їх рекурентній формі. Ця проблема може бути вирішена на основі синтезу нечітких методів та ядерного підходу, пов'язаного з гіпотезою, сформульованою Кавером, яка стверджує, що якщо задача лінійно нерозділима в вихідному просторі, то вона може бути розв’язана в просторі підвищеної розмірності.

Таким чином, на сьогоднішній день актуальною є наукова задача розробки нових методів нечіткої ядерної кластеризації призначених для обробки даних в on-line режимі, коли дані надходять на обробку послідовно, одне за одним, а кластери можуть перетинатися і мати довільну форму.

Зв’язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана в рамках держбюджетних НДР: «Нейро-фаззі системи для поточної кластеризації та класифікації послідовностей даних за умов їх викривленості відсутніми та аномальними спостереженнями» (№ДР 0113U000361); «Динамічний інтелектуальний аналіз послідовностей нечіткої інформації в умовах істотної невизначеності на основі гібридних систем обчислювального інтелекту» (№ДР 0116U002539). В рамках зазначених НДР здобувачкою в якості виконавця розроблено методи нечіткої ядерної кластеризації, які призначені для обробки даних в on-line режимі, коли дані надходять на обробку послідовно, одне за одним, а кластери можуть перетинатися.

Мета і задачі дослідження. Метою дослідження є розробка методів нечіткої кластеризації на основі ядерних нейронних мереж і нейро-фаззі систем, які налаштовують свою архітектуру в процесі навчання-самонавчання в умовах, коли кластери можуть перетинатися і мати довільну форму. Досягнення поставленої мети здійснюється шляхом вирішення таких задач:

– аналіз існуючих методів та підходів для кластеризації даних різної фізичної природи;

– удосконалення метода ЕМ (очікування-максимізація) кластеризації даних, які послідовно надходять на обробку одне за одним в on-line режимі;

– розробка штучної нейронної мережі, яка об'єднує в собі ідеї ядерних систем і самонавчання, і побудована на основі радіально-базисної нейронної мережі та самоорганізовної мапи;

– розробка архітектури гібридної нейронної мережі на основі узагальненої регресійної нейронної мережі та самоорганізовної мапи Т.Кохонена;

– розробка багатошарової нейро-фаззі системи, що є гібридом системи Ванга-Менделя і нечіткої кластерувальної самоорганізовної мережі;

– імітаційне моделювання розроблених методів і моделей та розв'язання за їхньою допомогою низки практичних задач нечіткої кластеризації даних.

Об'єкт дослідження – процес обробки даних, які надходять одне за одним, в послідовному режимі за допомогою нечітких методів кластеризації на основі ядерних функцій.

Предмет дослідження –методи нечіткої кластеризації на основі ядерних функцій в задачах інтелектуального аналізу даних.

Методи дослідження базуються на теорії обчислювального інтелекту, а саме на методах теорії штучних нейронних мереж і теорії нечіткої логіки для побудови архітектур гібридних нейро-фаззі мереж, що дозволяють проводити нечітку кластеризацію; теорії оптимізації і статистичного аналізу для синтезу ефективних методів нечіткої кластеризації та методів навчання і самоорганізації гібридних нейро-фаззі мереж. Імітаційне моделювання застосовується для перевірки якості кластеризації з використанням синтезованих методів і архітектур.

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

1. Вперше запропоновано ядерні кластерувальні нейронні мережі, які засновані на радіально-базисній нейронній мережі та узагальненій регресійній мережі, що дозволяють обробляти потоки даних різної фізичної природи в послідовному режимі.

2. Вперше запропоновано багатошарову гібридну нейро-фаззі систему обчислювального інтелекту на основі системи Ванга-Менделя і нечіткої кластерувальної самоорганізовної мережі, що дозволяє в процесі самонавчання налаштовувати не тільки свої параметри, але і архітектуру в on-line режимі і вирішувати задачі кластеризації потоку даних за умов апріорно невідомої форми кластерів і рівнів їх перетинання.

3. Удосконалено метод кластеризації ЕМ шляхом використання ядерних функцій спеціального виду, що дозволяє на відміну від стандартного підходу вирішувати задачу кластеризації в умовах перетинних кластерів з розрахунком оцінки належності кожного спостереження до кожного кластеру.

4. Удосконалено штучну нейронну мережу для аналізу головних компонент шляхом введення додаткових шарів ядерних функцій для підвищення розмірності вхідного простору, що дозволило обробляти інформацію, яка міститься в класах довільної форми.

Практичне значення одержаних результатів. Запропоновані в роботі методи нечіткої кластеризації на основі ядерних функцій і нейро-фаззі системи обчислювального інтелекту призначені для on-line кластеризації потоку інформації в умовах невизначеності як про форму класів, так і про рівень їх перетинання. Введені нейро-фаззі системи досить прості в чисельної реалізації і дозволяють вирішувати задачі динамічного інтелектуального аналізу даних (DDM) і інтелектуального аналізу потоків даних (DSM). Використання розроблених методів кластеризації дозволило підвищити ефективність вирішення задач кластеризації і аналізу документації програмного забезпечення, зменшити на 30% час на пошук того або іншого типу документації, а також вирішити задачу автоматичної каталогізації документації програмного забезпечення в системах інтелектуального аналізу даних у ТОВ «Академія СМАРТ».

Також основні результати дисертаційної роботи використовуються у навчальному процесі Харківського національного університету радіоелектроніки на кафедрі штучного інтелекту в курсах «Штучні нейронні мережі: архітектури, навчання та застосування» та «Нейромережеві методи обчислювального інтелекту».

Особистий внесок здобувача. Основні положення і результати дисертаційної роботи одержані здобувачкою особисто. Внесок авторки в публікаціях, опублікованих у співавторстві такий: [1] – архітектура ядерної гібридної штучної нейронної мережі, яка побудована на основі радіально-базисної нейронної мережі та самоорганізовної мапи Кохонена; [2] – архітектура гібридної нейронної мережі для ядерної кластеризації даних; [3] – архітектура для нечіткої ядерної кластеризації даних на основі ЕМ-методу; [4] –архітектура гібридної нейро-фаззі системи обчислювального інтелекту для нечіткої on-line кластеризації потоків інформації в умовах апріорної невизначеності і довільної форми кластерів; [5] – введення в ядерну еволюційну нейронну мережу шар радіально-базисних функцій та шар відновлення вхідного простору; [6] – гібридна архітектура на основі мапи Кохонена і м'якого ЕМ-методу; [7] – в якості функції щільності запропоновано ядра Єпанечнікова; [8] – модель еволюційної нейро-фаззі системи; [9] – архітектура ядерної мапи Кохонена; [10] – модифікація ЕМ-методу для ймовірнісної кластеризації; [11] – шар ядерних функцій активації для підвищення розмірності вихідного простору; [12] – гібридна система для обробки великих масивів даних; [13] – архітектура ядерної гібридної нейронної мережі.

Апробація результатів дисертації. Основні результати дисертаційної роботи були представлені та обговорені на 20-му Міжнародному молодіжному форумі «Радіоелектроніка і молодь в XXI столітті» (Харків, 2016), 8-й Міжнародній школі-семінарі «Теорія прийняття рішень» (Ужгород, 2016), Xth International Scientific and Technical Conference «Computer Science and Information Technologies» (CSIT 2015) (Lviv, 2015), 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP 2016) (Lviv, 2016), Міжнародній

науково-практичній конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (Залізний порт, 2015, 2016).

Публікації. Основні положення дисертаційної роботи опубліковані в 13 наукових роботах: у тому числі 5 статтях, серед яких 4 опубліковано у періодичних виданнях з технічних наук, що включені до переліку МОН України та 1 за кордоном (3 статті входять до міжнародних наукометричних баз), 8 публікацій в матеріалах міжнародних наукових конференцій і форумів (2 доповіді входять до міжнародних наукометричних баз Scopus, Web of Science).

Структура та обсяг дисертації. Дисертація складається зі вступу, п’яти розділів, висновків, що містять основні результати, списку використаних джерел і додатку. Загальний обсяг дисертації складає 165 сторінок (з них 132 – основного тексту), 18 рисунків з них 1 на окремій сторінці, 10 таблиць, список використаних джерел, що включає 154 найменування та займає 15 сторінок, 1 додаток на 3 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 та публікації.

У **першому розділі** проаналізовано стан проблеми кластеризації даних та існуючі підходи до її вирішення. Проведено аналіз основних методів і процедур навчання нечітких систем. виконано огляд існуючих парадигм та методів навчання та самонавчання.

На основі проведеного аналізу сформульовано мету та задачі дослідження, які полягають у створенні методів нечіткої ядерної кластеризації, призначених для обробки даних в on-line режимі, коли дані надходять на обробку послідовно одне за одним, а кластери можуть перетинатися і мати довільну форму, а також розв’язання за їх допомогою тестових і реальних задач.

У **другому розділі запропоновано** методи кластеризації, що враховують можливість взаємного перетинання кластерів, а також дозволяють аналізувати потік даних, які послідовно надходять на обробку в on-line режимі.

Запропоновано м’який ймовірнісний нечіткий метод кластеризації багатовимірних даних, що послідовно надходять на обробку. Цей підхід призначений для вирішення задач Data Stream Mining в умовах перетинних класів в порівняні зі своїми прототипами значно простіше в обчислювальній реалізації та не використовує ніяких ймовірнісних припущень щодо природи оброблювальних даних.

Задача ймовірнісної кластеризації в загальній постановці зводиться до проблеми самонавчання при невідомій кількості областей, при цьому передбачається, що щільність розподілу даних у кожному кластері підпорядковується багатовимірному нормальному (гауссівському) закону

$$p_j(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \Sigma_j}} \exp\left(-\frac{1}{2}(x-c_j)^T \Sigma_j^{-1}(x-c_j)\right), j = 1,2,...m, \quad (1)$$

а сумісна функція розподілу всіх даних описується виразом

$$p(x) = \sum_{j=1}^{m} p_j p_j(x) = \sum_{j=1}^{m} \frac{p_j}{(2\pi)^{\frac{n}{2}}\sqrt{\det \Sigma_j}} \exp\left(-\frac{1}{2}\left(x-c_j\right)^T \Sigma_j^{-1}\left(x-c_j\right)\right), j = 1,2,...m, \quad (2)$$

де $c_j - (n \times 1)$ -вимірний вектор-центроїд $j-$го кластеру, $\sum_j - (n \times n) -$ кореляційна матриця $j-$го кластеру така, що

$$\Sigma_j = \frac{1}{N}\sum_{k=1}^{N}(x(k)-c_j)(x(k)-c_j)^T, \quad (3)$$

$p_j -$апріорні ймовірності (ваги), що задовольняють природній умові

$$\sum_{j=1}^{m} p_j = 1, \quad (4)$$

при цьому припускається, що c_j, Σ_j і $p_j \forall j = 1,2,...,m$ апріорі невідомі і підлягають оцінюванню в процесі кластеризації.

Але в реальних задачах дані дуже рідко підпорядковуються нормальному закону, тому з метою спрощення обчислювальної реалізації методу можна в якості функцій розподілу використовувати багатовимірну конструкцію В. Єпанечнікова, що має вигляд:

$$p_j\left(x\right) = \max\left\{0, 1-\frac{1}{2}(x-c_j)^T \Sigma_j^{-1}(x-c_j)\right\}, j = 1,2,...m.$$

Робота метода ЕМ складається з повторюваної послідовності двох кроків, при цьому на кроці Е (expectation step) виконується оцінювання параметрів спільного розподілу (2), а на кроці М (maximization step) максимізується критерій самонавчання у вигляді логарифмічної функції правдоподібності

$$E(c_j, \Sigma_j, p_j, x(k)) = \sum_{k=1}^{N} \log(\sum_{j=1}^{m} p_j p_j(x(k-1)),$$

для чого можуть бути використані як традиційні градієнтні, так і квазіньютонівські процедури оптимізації.

І, нарешті, для оцінки ймовірності належності спостереження $j-$му кластеру використовується вираз:

$$p_j(x(k)) = \frac{p_j \exp\left(-\frac{1}{2}\left(x(k)-c_j\right)^T \Sigma_j^{-1}\left(x(k)-c_j\right)\right)}{\sum_{l=1}^{m} p_l \exp\left(-\frac{1}{2}\left(x(k)-c_j\right)^T \Sigma_j^{-1}\left(x(k)-c_l\right)\right)}.$$

Поряд з незаперечними перевагами ЕМ-метод має і низку істотних обмежень. У зв'язку з цим доцільною є розробка чисельно простого алгоритму кластеризації на основі метрики Махаланобіса, що враховує можливість взаємного перетинання сформованих кластерів і дозволяє аналізувати потік даних, що послідовно надходять на обробку в on-line режимі.

У ситуації, коли дані надходять на обробку послідовно в on-line режимі, вирішення такого роду задач може бути отримано за допомогою кластерувальної нейронної мережі Т. Кохонена, синаптичні ваги якої, є по суті компонентами векторів-центроїдів, а їх налаштування проводиться за допомогою тих або інших алгоритмів конкурентного самонавчання. При цьому сама процедура налаштування подібно ЕМ-методу складається з послідовності двох етапів: конкуренції (відповідає Е-кроку) і синаптичної адаптації (відповідає М-кроку).

Введена у другому розділі процедура нечіткої ймовірнісної кластеризації є своєрідним гібридом ЕМ-методу, методу K-середніх Махаланобіса, алгоритмів нечіткої кластеризації Бездека і Гата-Геви, а також нечіткої кластерувальної нейронної мережі Кохонена в її адаптивному варіанті, характеризується обчислювальною простотою і дозволяє аналізувати дані, що послідовно надходять на обробку в on-line режимі.

У третьому розділі розглядаються гібридні штучні нейронні мережі, які об'єднують в собі ідеї ядерних систем і самонавчання, і побудовані на основі еволюційних радіально-базисної нейронної мережі, узагальненої регресійної нейронної мережі та самоорганізовної мапи Т. Кохонена. Запропоновані системи дозволяють вирішувати задачі on-line кластеризації в умовах, коли утворені вихідними даними класи мають довільну форму.

Вперше запропоновано ядерні кластерувальні нейронні мережі, які засновані на радіально-базисній нейронній мережі та узагальненій регресійній мережі, що дозволяють обробляти кластери різної форми в послідовному режимі.

Архітектура ядерної самоорганізовної мапи на основі радіально-базисної нейронної мережі зображена на рис. 1

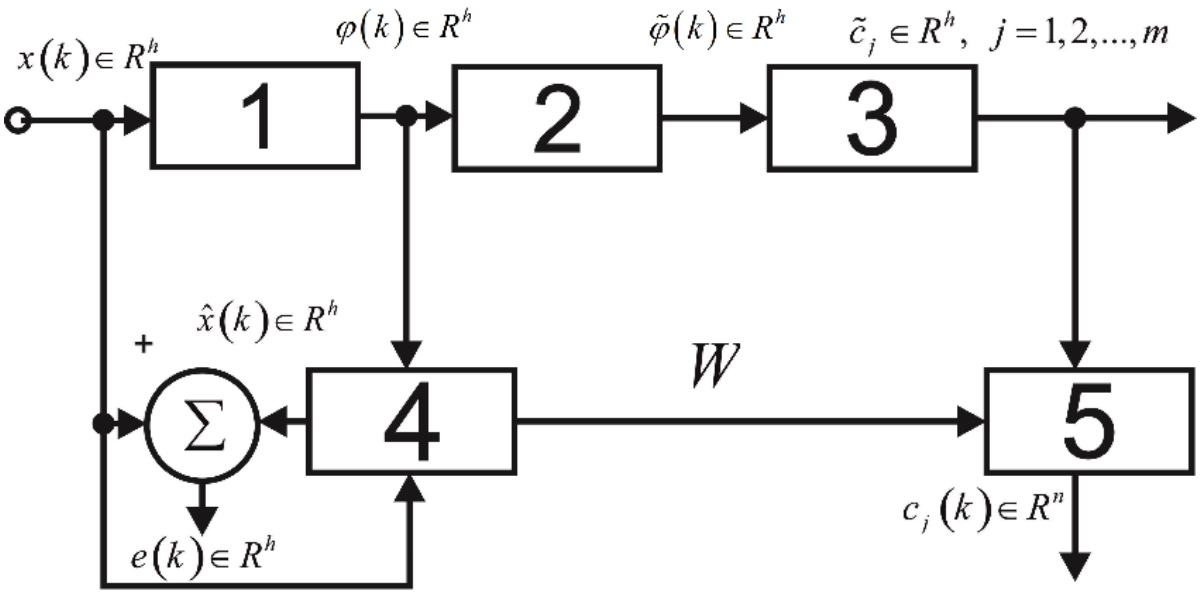

Рисунок 1 – Ядерна самоорганізовна мапа на основі радіально-базисної нейронної мережі

Вектори спостережень $x(k)$ послідовно надходять на перший шар **(1)** радіально-базисних функцій (1), а в якості таких функцій можуть використовуватися традиційні гауссіани, причому в якості їх центрів c_l в найпростішому випадку можуть бути взяті h довільно обраних векторів спостережень $c_l = x(l)$ (концепція «нейрони в точках даних»). Таким чином, при подаванні на вхід системи вектора спостережень $x(k)$ на виході першого шару формується векторний сигнал $\varphi(k) = \left(\varphi_1(k),..., \varphi_l(k),..., \varphi_h(k)\right)^T \in R^h$, де

$$\varphi_l(k) = e^{-\frac{\|x(k)-c_l\|^2}{2\sigma_l^2}}.$$

Другий шар системи **(2)** – шар нормалізації реалізує елементарне перетворення

$$\tilde{\varphi}_l(k) = \frac{\varphi_l(k)}{\|\varphi_l(k)\|}.$$

У третьому шарі **(3)** вирішується задача власне кластеризації, тобто. розбиття послідовності образів $\tilde{\varphi}(1),..., \tilde{\varphi}(k),...$, на m кластерів з находженням в процесі самонавчання прототипів – центроїдів класів $\tilde{c}_j^K \in R^h,\ j=1,2,...,m$.

На виході четвертого шару **(4)** формується сигнал $\hat{x}(k)\in \mathrm{R}^n$, який є оцінкою вхідного сигналу $x(k)$. Якість відновлення оцінюється на основі векторної похибки $e(k) = x(k) - \hat{x}(k)$ за допомогою скалярного критерія.

Результатом навчання четвертого шару є $(n \times h)$ - матриця синаптичних ваг $W(N)$, що отримана на підставі N спостережень.

Ця матриця є вихідною інформацією для п'ятого шару **(5)** відновлення прототипів кластерів у вихідному просторі R^n. При цьому прототипи, сформовані самоорганізовною мапою в h-вимірному просторі, проектуються в вихідний n-вимірний простір.

Таким чином, розглянута система є за суттю об'єднанням двох нейромереж: еволюційної радіально-базисної нейронної мережі (ERBFN) і самоорганізовної мапи Кохонена (SOM), які паралельно налаштовують свої синаптичні ваги в режимі самонавчання, одночасно з цим вирішуючи задачу кластеризації даних, утворюючи класи довільної форми.

Навчання введеної системи може розглядатися як дві відносно незалежні задачі: самонавчання радіально-базисної підсистеми і самонавчання власне самоорганізовної мапи.

Задача шару відновлення вхідного простору полягає у знаходженні $(n \times h)$ -матриці синаптичних ваг $W = \{w_{il}\}$ за вибіркою, що містить N спостережень шляхом мінімізації критерію навчання

$$E^N = \sum_{k=1}^{N} E(k) = \frac{1}{2}\sum_{k=1}^{N} \left\| x(k) - W\varphi(k) \right\|^2 = \frac{1}{2}\sum_{k=1}^{N} \left\| e(k) \right\|^2, \tag{5}$$

а процедура мінімізації цього критерія в рекурентній формі має вигляд

$$\begin{cases} W(k) = W(k-1) + \dfrac{(x(k) - W(k-1)\varphi(k))\varphi^T(k)P(k-1)}{1 + \varphi^T(k)P(k-1)\varphi(k)}, \\ P(k) = P(k-1) - \dfrac{P(k-1)\varphi(k)\varphi^T(k)P(k-1)}{1 + \varphi^T(k)P(k-1)\varphi(k)}. \end{cases} \tag{6}$$

Покращити якість відновлення вхідного простору можна, налаштовуючи не лише синаптичні ваги четвертого шару, а й параметри центрів c_l і ширини σ_l активаційних функцій першого шару. Для цього можна ввести рекурентні градієнтні алгоритми навчання всіх параметрів вигляду:

$$\begin{cases} c_l(k) = c_l(k-1) - \eta_c(k) e_i(k) w_{il}(k-1) \times e^{-\frac{\left\| x(k) - c_l(k-1) \right\|^2}{2\sigma_l^2(k-1)}} \dfrac{x(k) - c_l(k-1)}{\sigma_l^2(k-1)}, \\ \sigma_l^{-2}(k) = \sigma_l^{-2}(k-1) + \eta_\sigma(k) e_i(k) w_{il}(k-1) \times e^{-\frac{\left\| x(k) - c_l(k-1) \right\|^2}{2\sigma_l^2(k-1)}} \dfrac{\left\| x(k) - c_l(k-1) \right\|^2}{2}, \end{cases} \tag{7}$$

де $\eta_c(k), \eta_\sigma(k)$ – скалярні параметри кроку навчання.

Таким чином, співвідношення (7) є процедурою навчання всіх параметрів радіально-базисної нейронної мережі.

Навчання шару 3 – це процедура самонавчання, при цьому процес налаштування починається з ініціалізації синаптичних ваг мережі, в якості яких і виступають початкові значення прототипів $\tilde{c}_j^K(0)$. При цьому ці значення в процесі обробки нормуються подібно вхідним образам, тобто $\left\|\tilde{c}_j^K(k)\right\|=1$.

При подаванні на вхід третього шару сигналу $\tilde{\varphi}(k)$ спочатку обчислюється m відстаней, при цьому якщо в якості відстаней використовується евклідова метрика, то набагато зручніше використовувати міру подібності вигляду

$$SM\left(\tilde{\varphi}(k),\tilde{c}_j^K(k-1)\right)=\tilde{\varphi}^T(k)\tilde{c}_j^K(k-1)=\cos(\tilde{\varphi}(k),\tilde{c}_j^K(k-1))=\cos\theta_j(k). \qquad (8)$$

На підставі (8) визначається нейрон-переможець, «найближчий» до вхідного образу такий, що

$$SM(\tilde{\varphi}(k),\tilde{c}_*^K(k-1))=\max_j SM(\tilde{\varphi}(k),\tilde{c}_j^K(k-1)).$$

Далі налаштовуються ваги нейрона-переможця за допомогою правила самонавчання Кохонена в формі

$$\tilde{c}_j^K(k)=\begin{cases}\dfrac{\tilde{c}_j^K(k-1)+\eta(k)\left(\tilde{\varphi}(k)-\tilde{c}_j^K(k-1)\right)}{\left\|\tilde{c}_j^K(k-1)+\eta(k)\left(\tilde{\varphi}(k)-\tilde{c}_j^K(k-1)\right)\right\|}, \\ \textit{якщо } j-\textit{й нейрон переміг}, \\ \\ \tilde{c}_j^K(k-1) \textit{ в іншому випадку}.\end{cases} \qquad (9)$$

В результаті пред'явлення самоорганізовній мапі N образів $\tilde{\varphi}(k)$ буде отримано m прототипів-центроїдів $\tilde{c}_j^K(N)$, які далі з простору підвищеної розмірності R^h можуть бути спроектовані в початковий простір R^n за допомогою співвідношення:

$$c_j^K(N)=W(N)\tilde{c}_j^K(N)\ \forall\ j=1,2,...,m.$$

На рис. 2 наведена архітектура розглянутої ядерної самоорганізовної мапи на основі узагальненої регресійної нейронної мережі.

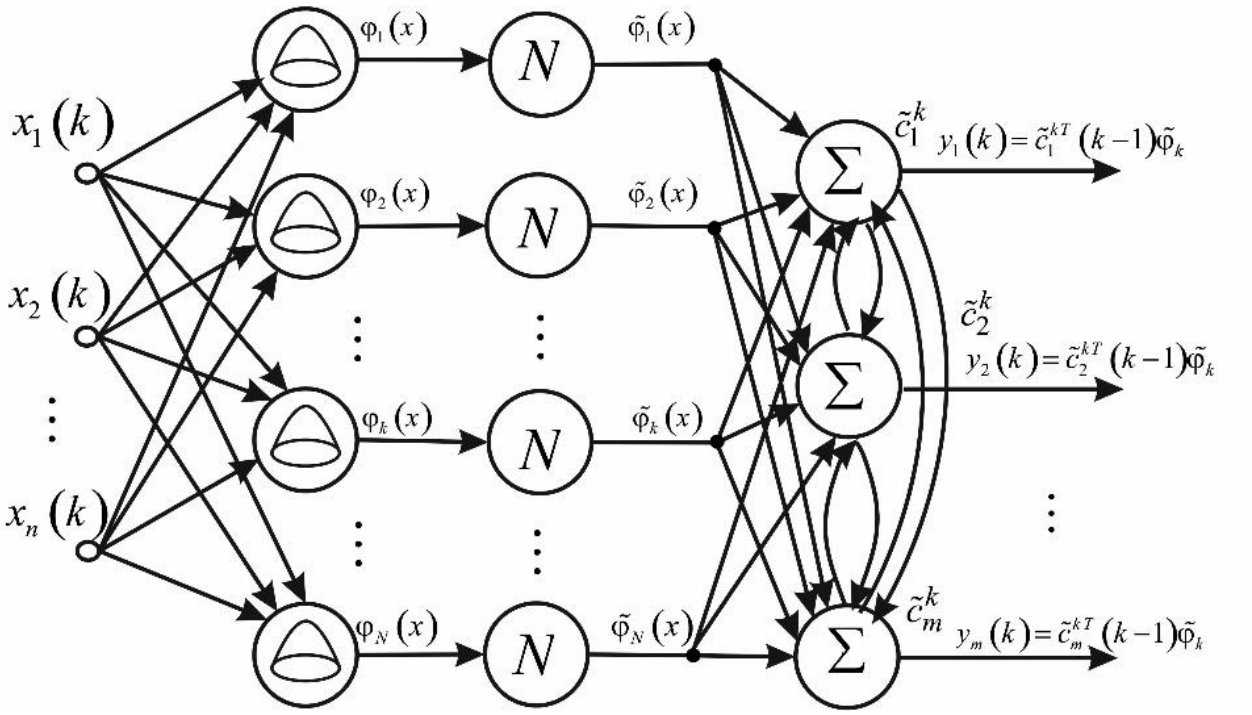

Рисунок 2 – Архітектура ядерної самоорганізовної мапи на основі узагальненої регресійної нейронної мережі

Вихідною інформацією для цієї мережі є вибірка (можливо зростаюча) векторів $x(1), x(2), ..., x(k), ..., x(N), ...; x(k) = (x_1(k), x_2(k) ..., x_i(k), ..., x_N(k))^T \in R^n$, яка має бути поділена на m кластерів довільної форми, при цьому k може бути як номером спостереження, так і моментом поточного часу.

Вектори спостережень $x(k)$ послідовно надходять на перший шар радіально-базисних функцій (R -нейронів), який повністю збігається за структурою з першим шаром (шаром образів) стандартної узагальненої регресійної мережі Д. Шпехта і сформований ядерними дзвонуватими функціями активації $\varphi_1, ..., \varphi_k, ..., \varphi_N$, за допомогою яких здійснюється підвищення розмірності вхідного простору. В якості таких функцій зазвичай використовуються традиційні гауссіани, а налаштування цього шару забезпечується за допомогою «лінивого навчання» на основі концепції «нейрони в точках даних». Таким чином, при подаванні на вхід нейронної мережі деякого некласифікованого образу x, на виходах R - нейронів першого шару з'являються значення

$$\varphi_k(x) = e^{-\frac{\|x - x(k)\|^2}{2\sigma^2}}, k = 1, 2, ..., N,$$

(тут σ^2 - параметр рецепторного поля дзвонуватої функції), а на виході GRNN в цілому – сигнал

$$\hat{y}(x)=\frac{\sum_{k=1}^{N} y(k)\varphi_k(x)}{\sum_{k=1}^{N}\varphi_k(x)}, \qquad (10)$$

де $y(k)$ – зовнішній навчальний сигнал, відповідний образу $x(k)$.

Другий прихований шар розглянутої мережі - шар нормалізації реалізує елементарне перетворення

$$\tilde{\varphi}(x)=\frac{\varphi(x)}{\|\varphi(x)\|},$$

(тут $\varphi(x)=(\varphi_1(x),\varphi_2(x),...,\varphi_N(x))^T$), необхідне для обробки інформації вихідним шаром, який є за суттю кластерувальною нейронною мережею Т.Кохонена, налаштування параметрів якої виконується на основі конкурентного самонавчання. У цьому вихідному шарі вирішується задача розбиття послідовності образів підвищеної розмірності $\tilde{\varphi}_1,...,\tilde{\varphi}_2,...,\tilde{\varphi}_k,...,\tilde{\varphi}_N$ на m кластерів з знаходженням прототипів центроїдів $\tilde{c}_1^{\,K},\tilde{c}_2^{\,K},...,\tilde{c}_m^{\,K}\in R^N$.

Навчання ядерної самоорганізовної мапи на основі узагальненої регресійної нейронної мережі полягає у реалізації послідовності кроків:

Крок 0: задати поріг нерозрізненості векторів центрів активаційних функцій Δ, максимально допустиму кількість нейронів в першому шарі $H \le N$ і параметр ширини рецепторного поля σ^2.

Крок 1: при надходженні спостереження $x(1)$ формується перший центр $c_1 = x(1)$ і власне активаційна функція $\varphi_1(x)=e^{-\frac{\|x-x(1)\|^2}{2\sigma^2}}$.

Крок 2: при надходженні спостереження $x(2)$ перевіряється умова $\|x(2)-c_1\|^2 \le \Delta$, якщо вона виконується, то спостереження $x(2)$ не формує новий центр та автоматично належить тому ж кластеру, що і $x(1)$, якщо виконується умова $\Delta < \|x(2)-c_1\| \le 2\Delta$, то виконується корекція c_1 відповідно WTA-правилу самонавчання Т. Кохонена, якщо ж $2\Delta \le \|x(2)-c_1\|$, то формується друга функція активації.

Крок N: Якщо до моменту k надходження N-го вектору-образу $x(N)$ сформовано $h \le H$ активаційних функцій та виконується умова (2), процес нарощування кількості R-нейронів першого шару закінчується і надалі структура цього шару залишається незмінною.

Оцінити якість функціонування першого шару можна було б за допомогою виразу (10), який для $h = H = N$ має вигляд

$$\hat{x}(k) = \frac{\sum\limits_{k=1}^{N} x(k)\varphi_k(x(k))}{\sum\limits_{k=1}^{N} \varphi_k(x(k))}, \qquad (11)$$

після чого оцінити похибку відновлення вхідних образів:

$$\varepsilon = \frac{1}{N}\sum_{k=1}^{N} \frac{\left\| x(k) - \hat{x}(k) \right\|}{\left\| x(k) \right\|}. \qquad (12)$$

Однак, оскільки в процесі формування першого шару за допомогою розглянутого вище підходу деякі з центрів активаційних функцій не збігаються з векторами спостережень, використання виразу (11) коректного для «класичної» GRNN, в нашому випадку може виявитися неправомірним.

У цій ситуації доцільно модифікувати GRNN, архітектуру якої наведено на рис. 3

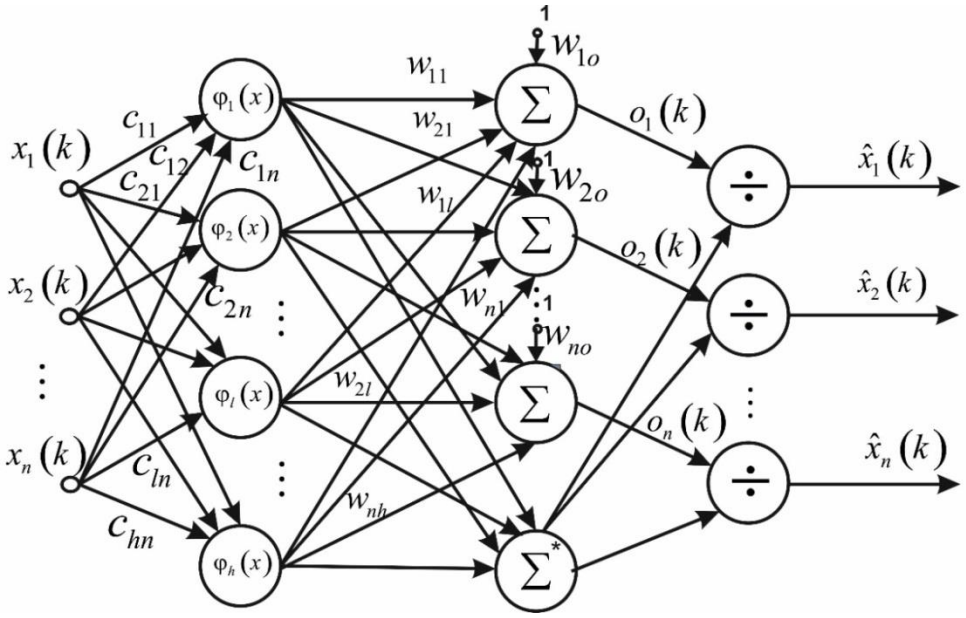

Рисунок 3 – Модифікована узагальнена регресійна нейронна мережа

Ця мережа подібна тришаровому персептрону, що містить три шари обробки інформації, проте в якості активаційних функцій використовує радіально-базисні конструкції в першому прихованому шарі. Другий прихований шар містить $n+1$ вузлів, n з яких є адаптивними лінійними асоціаторами, а $(n+1)-й$ стандартним блоком підсумовування $\sum^*$. Вихідний сигнал мереж і утворений n блоками поділу $\div$.

Другий прихований шар налаштовується аналогічно процесу навчання радіально-базисних нейронних мереж. При цьому на виходах n адаптивних лінійних асоціаторів формуються сигнали

$$o_i(k) = \sum_{l=0}^{h} w_{il}(k)\varphi_l(x(k)),\ \ i = 1,2...,n,$$

а на виході суматорів $\sum^*$ з'являється сума $\sum_{l=0}^{h} \varphi_l(x(k)).$

У вихідному шарі виконується нормування вихідного сигналу по типу нормованої радіально-базисної мережі (NRBFN) так, що

$$\hat{x}_i(k) = \frac{\sum_{l=0}^{h} w_{il}(k)\varphi_l(x(k))}{\sum_{l=0}^{h} \varphi_l(x(k))} = \sum_{l=0}^{h} w_{il}(k)\frac{\varphi_l(x(k))}{\sum_{l=0}^{h} \varphi_l(x(k))} = \tag{13}$$

$$= \sum_{l=0}^{h} w_{il}(k)\varphi_l^*(x(k)),$$

$$\varphi_0(x(k)) = 1,\ \varphi_l^*(x(k)) = \varphi_l(x(k))(\sum_{l=0}^{h} \varphi_l(x(k)))^{-1} p,$$

або

$$\hat{x}(k) = W(k)\varphi^*(k), \tag{14}$$

де

$$\hat{x}_i(k) = (\hat{x}_1(k),...\hat{x}_i(k),...,\hat{x}_n(k))^T,\ \varphi^*(k) = (\varphi_0^*(x(k)),\varphi_1^*(x(k)),...,\varphi_h^*(x(k)))^T,$$

$\mathrm{W}(k) - (n \times (h+1))$ - матриця синаптичних ваг.

Таким чином можна оцінити якість роботи першого шару, використовуючи вираз (12) замість стандартного співвідношення (11).

Отже, запропоновані в цьому розділі системи дозволяють вирішувати задачі on-line кластеризації в умовах, коли утворені вихідними даними класи мають довільну форму. Запропоновані нейронні мережі є простими у реалізації і дозволяють вирішувати досить широкий клас задач динамічного інтелектуального аналізу даних і інтелектуального аналізу потоків даних.

У четвертому розділі запропоновано on-line нейро-фаззі систему для вирішення задач послідовного нечіткого кластерування даних, що дозволяє опрацьовувати вектори спостережень за умов обмеженого числа даних в оброблюваній вибірці, а також метод її самонавчання на основі самоорганізовної мапи Т. Кохонена. Архітектура системи є за своєю суттю гібридом системи

Ванга-Менделя та нечіткої кластерувальної самоорганізовної мережі. Запропонована система в процесі самонавчання налаштовує не лише свої параметри, але й архітектуру в on-line режимі. Для налаштування параметрів функцій належності гібридної нейро-фаззі системи введено метод, що ґрунтується на використанні конкурентного навчання. У процесі навчання гібридна нейро-фаззі система налаштовує синаптичні ваги, центри і параметри ширини функцій належності.

Архітектура системи містить сім шарів обробки інформації. Вектори спостережень $x(k)$ послідовно надходять на нульовий (рецепторний) шар системи, звідки передаються на перший прихований шар, утворений nh (по h на кожний вхід) функціями $\mu_{li}(x_i), l=1,2,...,h; i=1,2,...,n$, що виконує фаззіфікування вхідного простору R^n. Другий прихований шар забезпечує аґреґування рівнів належності, розрахованих у першому шарі, і містить h блоків множення. Третій прихований шар системи – шар нормалізації. Для вирішення власне задачі адаптації призначені п'ятий, шостий і сьомий додаткові шари, утворені nh налаштованими синаптичними вагами, $n+1$ суматорами і n блоками ділення, що вирішують задачі дефаззіфікування.

Таким чином, перший, другий, п'ятий, шостий і сьомий шари системи утворюють по суті багатовихідну нейро-фаззі систему Ванґа-Менделя (TSK-система нульового порядку). Виходом сьомого шару є векторний сигнал $\hat{x}(k) \in R^n$, який є оцінкою вхідного сигналу $x(k)$.

Необхідно також зауважити, що на виходах п'ятого прихованого шару формується nh сигналів

$$w_{il}\prod_{i=1}^{n}\mu_{li}(x_i(k)) = w_{il}\varphi_l(k),$$

шостого $-n+1$ сигналів

$$\sum_{l=1}^{h} w_{il}\prod_{i=1}^{n}\mu_{li}(x_i(k)) = \sum_{l=1}^{h} w_{il}\varphi_l(k),$$

$$\sum_{l=1}^{h}\prod_{i=1}^{n}\mu_{li}(x_i(k)) = \sum_{l=1}^{h}\varphi_l(k)$$

і, нарешті, на виходах сьомого шару – n сигналів

$$\hat{x}_i(k) = \frac{\sum_{l=1}^{h} w_{il}\prod_{i=1}^{n}\mu_{li}(x_i(k))}{\sum_{l=1}^{h}\prod_{i=1}^{n}\mu_{li}(x_i(k))} = \sum_{l=1}^{h} w_{il}\frac{\prod_{i=1}^{n}\mu_{li}(x_i(k))}{\sum_{l=1}^{h}\prod_{i=1}^{n}\mu_{li}(x_i(k))} = \sum_{l=1}^{h} w_{il}\hat{\varphi}_l(x(k)) = w_i^T\hat{\varphi}^h(x(k)),$$

де $\hat{\varphi}_l(x(k)) = \dfrac{\prod_{i=1}^{n} \mu_{li}(x_i(k))}{\sum_{l=1}^{h}\prod_{i=1}^{n} \mu_{li}(x_i(k))}$,

$$w_i = (w_{i1}, w_{i2}, ..., w_{ih})^T, \hat{\varphi}^h(x(k)) = (\hat{\varphi}_1(x(k)), ..., \hat{\varphi}_l(x(k)), ..., \hat{\varphi}_h(x(k)))^T.$$

Вводячи далі у розгляд $(n \times 1)$ -вектор $\hat{x}(k) = (\hat{x}_1(k), ..., \hat{x}_i(k), ..., \hat{x}_n(k))^T$ та $(n \times h)$-матрицю $W = (w_1, w_i, ..., w_n)^T$, остаточно можна записати

$$\hat{x}(k) = W\hat{\varphi}^h(x(k))$$

і

$$e(k) = x(k) - W\hat{\varphi}^h(x(k)).$$

Процес навчання нейро-фаззі системи, що розглядається, зводиться до самонавчання - еволюції першого прихованого шару, навчання з вчителем матриці синаптичних ваг W п'ятого прихованого шару і конкурентного самонавчання нейро-фаззі-мережі Кохонена четвертого вихідного шару.

В основі налаштування першого прихованого шару полягають ідеї еволюційних нейро-фаззі-систем і, перш за все, адаптивний метод навчання еволюційних систем Н. Касабова.

П’ятий розділ присвячено імітаційному моделюванню розроблених методів ядерної кластеризації. Імітаційне моделювання виконувалось як на тестових вибірках, так і на реальних даних. Результати моделювання запропонованих методів порівнювалися зі стандартними методами кластеризації з метою порівняльної оцінки якості вирішення задач, що розглядалися.

Для підтвердження працездатності розробленої ядерної кластерувальної нейронної мережі на основі узагальненої регресійної нейронної мережі була вирішена задача кластеризації на основі тестових вибірок з UCI-сховища.

Ефективність запропонованої багатошарової гібридної нейро-фаззі системи і процедур її самонавчання продемонстрована під час розв'язання задачі нечіткої кластеризації тестових даних на основі двох вибірок: вибірка даних діагностування раку молочної залози (Wisconsin Diagnostic Breast Cancer), яка включає 569 спостережень, розділених на 2 класи, кожне спостереження містить 30 ознак та вибірка даних діагностування діабету (Pima Indians Diabetes), яка включає 768 спостережень, розділених на 2 класи, кожне спостереження містить 8 ознак

Для порівняння ефективності кластеризації запропонованої багатошарової гібридної нейро-фаззі системи були обрані самоорганізовна мапа Т. Кохонена і метод нечітких С-середніх (FCM) з параметром фаззіфікації $\beta = 2$.

Оскільки для кожної вибірки існують мітки вірної класифікації, ефективність кластеризації вимірювалася у відсотках точності кластеризації

щодо еталонної. У кожній клітинці таблиці 1 наведений середній, мінімальний і максимальний результат для серії з 50 експериментів. Результати порівняння з існуючими методами представлені в табл. 1.

Таблиця 1 - Порівняння точності кластеризації на тестових вибірках

Досліджувані методи	Wisconsin Diagnostic Breast Cancer			Pima Indians Diabetes		
	avg	max	min	avg	max	min
Стандартна мапа Т. Кохонена (SOM)	80%	90%	71%	81%	90%	73%
Метод нечітких С-середніх (FCM with $\beta = 2$)	89%	94%	81%	89%	96%	84%
Багатошарова гібридна нейро-фаззі система	93%	96%	89%	92%	95%	91%

Як видно з табл. 1, запропонована гібридна нейро-фаззі система має перевагу в порівнянні з досліджуваними методами. Варто зазначити, що зі збільшенням вибірки похибка кластеризації запропонованої нейро-фаззі системи зменшувалася.

Впроваджений метод кластеризації дозволив зменшити на 30% час на пошук того або іншого типу документації, а також була вирішена задача каталогізації документації програмного забезпечення у ТОВ «Академія СМАРТ».

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

У **додатку** наведено акти про впровадження результатів дослідження в ТОВ «Академія СМАРТ», а також в навчальний процес Харківського національного університету радіоелектроніки.

ВИСНОВКИ

У дисертаційній роботі представлені результати, які відповідно до поставленої мети є розробкою методів нечіткої кластеризації на основі ядерних нейронних мереж і нейро-фаззі систем, які налаштовують свою архітектуру в процесі навчання-самонавчання в умовах, коли кластери можуть перетинатися і мати довільну форму. Проведені дослідження дозволили зробити такі висновки:

1. Удосконалено метод ЕМ-(очікування-максимізація) кластеризації даних, що послідовно надходять на обробку одне за одним в on-line режимі, за допомогою ядерних функцій спеціального виду. Це дозволило на відміну від стандартного підходу кластеризувати данні в умовах перетинних кластерів.

2. Розроблено ядерну кластерувальну нейронну мережу, яка об'єднує в собі ідеї ядерних систем і самонавчання та побудована на основі радіально-базисної нейронної мережі і самоорганізовної мапи Т. Кохонена.

3. Розроблено архітектуру гібридної кластерувальної нейронної мережі на основі узагальненої регресійної нейронної мережі та самоорганізовної мапи Т.Кохонена. Запропонована система дозволяє вирішувати задачі on-line кластеризації в умовах, коли утворені вихідними даними класи мають довільну форму.

4. Розроблено багатошарову нейро-фаззі систему, що є гібридом системи Ванга-Менделя і нечіткої кластерувальної самоорганізовної мережі. Це дозволило в процесі самонавчання налаштовувати як її параметри, так й архітектуру в послідовному режимі і кластеризувати данні в умовах апріорно невідомої форми кластерів і рівнів їх перетинаня.

5. Проведено імітаційне моделювання методів нечіткої кластеризації на основі ядерних нейронних мереж, нейро-фаззі систем і процедур їх самонавчання. Результати моделювання доводять доцільність та ефективність розроблених методів.

6. Синтезовані в роботі методи нечіткої кластеризації даних на основі ядерних функцій підтвердили свою ефективність в системах інтелектуального аналізу даних, які розробляються ТОВ «Академія СМАРТ», а запропоновані інтелектуальні методи кластеризації дозволили впровадити їх до системи каталогізації документів та інформаційного пошуку.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Бодянский Е.В. Ядерная самоорганизующаяся карта на основе радиально-базисной нейронной сети / Е.В. Бодянский, А.А. Дейнеко, Я.В. Куценко (Я.В. Хаустова) // Электротехнические и компьютерные системы. – Одесса. – 2015. – 20 (96). – С. 97-105. (Входить до міжнародної наукометричної бази BASE, Index Copernicus).

2. Бодянский Е.В. Ядерная кластеризация на основе обобщенной регресионной нейронной сети и самоорганизующейся карты Т.Кохонена / Е.В. Бодянский, А.А. Дейнеко, Я.В. Куценко (Я.В. Хаустова) // Інформаційно-керуючі системи на залізничному транспорті. – 2016. – 3 (118). – С. 15-22.

3. Бодянский Е.В. Нечеткая кластеризация потоков данных с помощью EM–алгоритма на основе самообучения по Т. Кохонену / Е.В. Бодянский, А.А. Дейнеко, А.А. Заика, Я.В. Куценко (Я.В. Хаустова) // Прикладная радиоэлектроника. – 2016. – Том 15. – № 1. – С. 80-83.

4. Бодянський Є.В. Послідовне нечітке кластерування на основі нейро-фаззі підходу / Є.В. Бодянский, А.О. Дейнеко, Я.В. Куценко (Я.В. Хаустова) // Радіоелектроніка, інформатика, управління. – Запоріжжя. -2016. – № 3(38) – С. 30-39. (Входить до міжнародних наукометричних баз Web of Science, Index Copernicus, BASE).

5. Deineko A. Neural Network for Kernel Principal Component Analysis / A. Deineko, I. Perova, O. Turuta, Y. Kutsenko (Y. Khaustova), M. Shalamov // International Journal of Computer Science and Mobile Computing. – 2015. – Vol.4 Issue 9. – P. 356-363. (Входить до міжнародних наукометричних баз Index Copernicus, BASE, INSPEC, World Cat, ESJT).

6. Bodyanskiy Ye. Data streams fast EM-fuzzy clustering based on Kohonen`s self-learning / Ye. Bodyanskiy, A. Deineko, Y. Kutsenko (Y. Khaustova), O. Zayika // The 1th IEEE International Conference on Data Stream Mining & Processing (DSMP 2016): proc. of int. conf., Lviv, August 23-27, 2016. – Lviv. – 2016. – P. 309-313. (Входить до міжнародних наукометричних баз Scopus, Web of Science).

7. Дейнеко А.А. Гибридный ЕМ-алгоритм вероятностной кластеризации потоков данных/ А.А. Дейнеко, А.А. Заика, Я.В. Куценко (Я.В. Хаустова), И.П. Плисс // Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта: сб. трудов междунар. научн. конф., Железный порт, 24-28 мая, 2016 г. – Херсон: ХНТУ. – 2016. – С. 274-276.

8. Дейнеко А.О. Нечітке кластерування потоків даних на основі гібридних систем обчислювального інтелекту / А.О. Дейнеко, А.І. Долотов, Я.В. Куценко (Я.В. Хаустова), І.П. Плісс, Д.Р. Чигрин. // Теорія прийняття рішень: праці VIII міжнар. школи-семінару, Ужгород, 26 вересня - 1 жовтня, 2016 р. – Ужгород: УжНУ. – 2016. – С. 108-109.

9. Bodyanskiy Ye. Kernel clustering based on the hybrid neural network in Data Stream Mining tasks / Ye. Bodyanskiy, A. Deineko, Y. Kutsenko (Y. Khaustova) // Intelectual systems for decision making and problems of computational intelligence: proc. of int. conf. Zalizniy Port, May 24-28, 2016. – Kherson. – 2016. – P. 242-244.

10. Заика А.А. ЕМ-алгоритм вероятностной кластеризации / А.А. Заика, Я.В. Куценко (Я.В. Хаустова) // 20-й Международный молодежный форум «Радиоэлектроника и молодежь в 21 веке»: матер. конф., Харьков, 22-24 апреля, 2016 г. – Харьков: ХНУРЭ. – 2016. – Том 6. – С. 19-20.

11. Deineko A. Kernel evolving neural networks for sequential principal component analysis and its adaptive learning algorithm / A. Deineko, Y. Kutsenko (Y. Khaustova), I. Pliss, M. Shalamov // Int. Scientific and Technical Conf. «Computer science and information technologies»: proc. of int. conf., Lviv, September 14-17, 2015. – Lviv: LPNU. – 2015. – P. 107-110. (Входить до міжнародної наукометричної бази Scopus).

12. Ганжа Д.Д. Ядерная самоорганизующаяся карта Т. Кохонена/ Д.Д. Ганжа, Я.В. Куценко (Я.В. Хаустова) // 20-й Международный молодежный форум «Радиоэлектроника и молодежь в 21 веке»: матер. конф., Харьков, 22-24 апреля, 2016 г. – Харьков: ХНУРЭ. – 2016. – Том 6. – С. 15-16.

13. Бодянський Є.В. Кластеризація даних на основі радіально-базисної самоорганізовної мапи / Є.В. Бодянский., А.О. Дейнеко, Я.В. Куценко (Я.В. Хаустова), М.О. Шаламов // Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту: матеріали міжнар. наук. конф., Залізний порт, 25-28 травня, 2015 г. – Херсон: ХНТУ. – 2015. – С. 257-259.

АНОТАЦІЯ

Хаустова Я.В. (Куценко Я.В.) методи нечіткої кластеризації на основі ядерних функцій в задачах інтелектуального аналізу даних. – 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, 2017.

Дисертаційна робота присвячена розв'язанню актуальної наукової задачі розробки нових методів нечіткої кластеризації на основі ядерних нейронних мереж і нейро-фаззі систем, які налаштовують свою архітектуру в процесі навчання-самонавчання в умовах перетинних кластерів довільної форми. Вперше запропоновано ядерні кластерувальні нейронні мережі, які засновані на радіально-базисній нейронній мережі та узагальненій регресійній мережі, що дозволяють обробляти потоки даних різної фізичної природи в послідовному режимі. Вперше запропоновано багатошарову гібридну нейро-фаззі систему обчислювального інтелекту на основі системи Ванга-Менделя і нечіткої кластерувальної самоорганізовної мережі, що дозволяє в процесі самонавчання налаштовувати не тільки свої параметри, але і архітектуру в on-line режимі і вирішувати задачі кластеризації потоку даних за умов апріорно невідомої форми кластерів і рівнів їх перетинання. Удосконалено метод кластеризації EM (expectation-maximization) шляхом використання ядерних функцій спеціального виду, що дозволяє на відміну від стандартного підходу вирішувати задачу кластеризації в умовах перетинних кластерів з розрахунком оцінки належності кожного спостереження до кожного кластеру. Удосконалено штучну нейронну мережу для аналіза головних компонент шляхом введення додаткових шарів ядерних функцій для підвищення розмірності вхідного простору, що дозволило обробляти інформацію, яка міститься в класах довільної форми.

Ключові слова: ядерна нейронна мережа, нейро-фаззі системи, самоорганізовна мапа Кохонена, радіально-базисна нейронна мережа, узагальнена регресійна нейронна мережа, EM-ймовірнісний алгоритм кластеризації, ядерна функція активації.

АННОТАЦИЯ

Хаустова Я.В. (Куценко Я.В.) Методы нечеткой кластеризации на основе ядерных функций в задачах интеллектуального анализа данных. – 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, 2017.

Диссертационная работа посвящена решению актуальной научной задачи разработки новых методов нечеткой кластеризации на основе ядерных нейронных сетей и нейро-фаззи систем, которые настраивают свою архитектуру в процессе обучения-самообучения в условиях перекрывающихся кластеров произвольной формы.

Рассмотрены различные парадигмы обучения искусственных нейронных сетей, основные и наиболее популярные нейронные сети и нейро-фаззи системы, в которых в качестве функций активации-принадлежности используются ядерные конструкции.

Впервые предложено ядерные кластеризующие нейронные сети, на основе радиально-базисной нейронной сети и обобщенной регрессионной сети, которые позволяют обрабатывать потоки данных разной физической природы в последовательном режиме.

Впервые предложена многослойная гибридная нейро-фаззи система вычислительного интеллекта на основе системы Ванга-Менделя и нечеткой кластеризующей самоорганизующей сети, которая позволяет в процессе самообучения настраивать не только свои параметры, но и архитектуру в on-line режиме и решать задачи кластеризации потока данных при условии априорно неизвестной формы кластеров и уровней их перекрытия.

Модифицирован метод кластеризации EM (ожидание-максимизация) путем использования ядерных функций специального вида, что позволяет в отличие от стандартного подхода, решать задачу кластеризации в условиях, перекрывающихся кластеров с расчетом оценки принадлежности каждого наблюдения к каждому кластеру.

Модифицировано искусственную нейронную сеть для анализа главных компонент путем введения дополнительных слоев ядерных функций для повышения размерности исходного пространства, что позволило обрабатывать информацию, которая содержится в классах произвольной формы.

Предложенные в работе методы нечеткой ядерной кластеризации позволяют решать задачи обработки данных в on-line режиме, когда данные поступают на обработку последовательно, одно за другим, а кластеры могут перекрываться. Синтезированные в работе методы подтвердили свою эффективность в задачах обработки информации в системах интеллектуального анализа данных. Разработанные методы кластеризации позволили улучшить эффективность решения задач в системах интеллектуального анализа данных в ООО «Академия СМАРТ».

Ключевые слова: ядерная нейронная сеть, нейро-фаззи системы, самоорганизующаяся карта Кохонена, радиально-базисная нейронная сеть, обобщенная регрессионная нейронная сеть, EM-вероятностный алгоритм кластеризации, ядерная функция.

ABSTRACT

Khaustova Y.V. (Kutsenko Y.V.) Fuzzy clustering methods based on kernel

functions in data mining tasks. – Manuscript.

A thesis for the candidate degree in technical 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, 2017.

The clustering system based on the evolving general regression neural network and self-organizing map of T.Kohonen, is proposed in the thesis. An on-line neuro-fuzzy system for solving data stream fuzzy clustering task and its self-learning procedures based on T. Kohonen's rule are proposed in the thesis. During a learning procedure in on-line mode, the proposed system tunes both its parameters and its architecture. For tuning of membership functions parameters of neuro-fuzzy system the method based on competitive learning is proposed. In the thesis soft probabilistic clustering algorithm of multidimensional data sets that are sequentially fed to processing in on-line mode is investigated. The proposed system solves the tasks of Data Stream Mining when classes are overlapped.

Key words: kernel neural network, neuro-fuzzy systems, self-organizing Kohonen map, radial-basis functions neural network, general regression network, EM – probabilistic clustering algorithm, kernel activation function.