

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

КОЛЧИГІН БОГДАН ВЛАДЛЕНОВИЧ

УДК 004.032.26

**АДАПТИВНІ НЕЙРО-ФАЗЗИ СИСТЕМИ ДЛЯ НЕЧІТКОГО КЛАСТЕРНОГО
АНАЛІЗУ В УМОВАХ НЕВИЗНАЧЕНОСТІ**

05.13.23 – системи та засоби штучного інтелекту

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2014

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
Гороховатський Володимир Олексійович,
Харківський інститут банківської справи
Університету банківської справи Національного
банку України, завідувач кафедри інформаційних
технологій;

доктор технічних наук, доцент
Литвиненко Володимир Іванович,
Херсонський національний технічний університет,
завідувач кафедри інформатики та комп'ютерних
наук.

Захист відбудеться «___» _____ 2014 р. о ___ годині на засіданні спеціалізованої вченої ради Д 64.052.01 у Харківському національному університеті радіоелектроніки за адресою: Україна, 61166 м. Харків, пр. Леніна, 14.

Із дисертацією можна ознайомитися в бібліотеці Харківського національного університету радіоелектроніки за адресою: Україна, 61166 м. Харків, пр. Леніна, 14.

Автореферат розісланий «___» _____ 2014 р.

Учений секретар
спеціалізованої вченої ради

О.А. Винокурова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Задачі інтелектуального аналізу даних в умовах структурної та параметричної невизначеності є однією з найважливіших проблем штучного інтелекту. Найбільш ефективні методи вирішення цієї проблеми пов'язані з досягненнями в галузі обчислювального інтелекту, що будуються на основі штучних нейронних мереж і систем нечіткого виведення. При цьому в рамках аналізу має вирішуватися досить широкий клас задач, таких як емуляція, ідентифікація, прогнозування, класифікація, діагностика та кластеризація.

Остання задача особливо важлива через те, що її вирішення здійснюється в режимі навчання без вчителя. Водночас задача кластеризації ускладнюється тим, що в реальних даних кластери, як правило, мають досить складну форму, істотно перекриваються, дані в цих кластерах розподілені з різною щільністю, а самі кластери можуть мати різний об'єм. Традиційні методи вирішення завдань кластеризації, що засновані на нечіткому підході, базуються на деяких апріорних припущеннях щодо характеру розподілення даних та інших властивостей вибірки, і тому для ситуацій, у яких потрібна поточна обробка зростаючого масиву даних, властивості якого змінюються з часом, стає необхідним застосування складних гібридних і нейромережових методів, методів колективного виведення, методів нечітких систем типу-2 (type-2 fuzzy systems), які здатні працювати з більш високим ступенем невизначеності вихідних даних.

Різноманітність таких методів і велика кількість настроюваних параметрів потребують від аналітика глибокого знання математичних засад нечіткої логіки та методів кластеризації, що уповільнює їх впровадження. Крім того, багато ефективних методів кластеризації розроблені для розв'язання окремих специфічних задач і не узагальнені на інші випадки (як правило, це стосується роботи в послідовному режимі та використання розширених підходів до нечіткої логіки). У зв'язку з цим існує потреба досліджень у напрямку універсалізації та уніфікації методів адаптивної кластеризації, призначених для роботи в послідовному режимі та в умовах істотно більш високого рівня апріорної і поточної невизначеності щодо характеру розподілу даних, ніж це відомо на теперішній час, а також потреба в побудові на засаді таких методів нейро-фаззі систем, здатних до гнучкої автоматичної адаптації к умовам задачі.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційну роботу виконано в рамках держбюджетних НДР: «Еволюційні гібридні системи обчислювального інтелекту зі змінною структурою для інтелектуального аналізу даних» (№ДР 0110U000458); «Нейро-фаззі системи для поточної кластеризації та класифікації послідовностей даних за умов їх викривленості відсутніми та аномальними спостереженнями» (№ДР 0113U000361), які виконувалися згідно з тематичними планами науково-дослідних робіт Харківського національного університету радіоелектроніки, затвердженими Міністерством освіти і науки України за результатами конкурсного відбору проектів наукових досліджень. У межах наведених тем здобувач як виконавець запропонував методи адаптивного кластерного аналізу даних на основі рекурентних мап Кохонена, адаптивних нейро-фаззі систем і методів їх навчання для обробки даних за умов апріорної та поточної невизначеності.

Мета і задачі дослідження. Метою дисертаційної роботи є розробка методів адаптивної кластеризації даних на основі нейро-фаззі моделей і методів навчання таких моделей, які дозволять підвищити якість кластеризації в умовах апріорної та поточної невизначеності щодо розподілу даних.

Для досягнення поставленої мети необхідно розв'язати такі наукові задачі:

- проаналізувати існуючі методи та підходи до розробки систем нечіткої адаптивної кластеризації;
- розробити метод адаптивного самонавчання складеної кластеризуючої нейро-фаззі мережі Кохонена, який об'єднує в собі можливісний і ймовірнісний підходи до кластеризації;
- синтезувати адаптивні методи кластеризації даних на основі імовірнісного і можливісного підходів до нечіткої логіки;
- синтезувати адаптивні методи кластеризації, стійкі до різних типів апріорної невизначеності розподілу даних;
- розробити архітектури нейромережових систем для сумісної роботи розроблених методів;
- вирішити за допомогою розроблених методів і моделей низку практичних завдань нечіткої кластеризації.

Об'єкт дослідження. Процес кластеризації даних в умовах апріорної та поточної невизначеності характеру розподілу даних.

Предмет дослідження. Адаптивні нейро-фаззі системи та методи нечіткого кластерного аналізу.

Методи дослідження. Основні результати роботи отримані на основі теорії обчислювального інтелекту, а саме на методах самоорганізації і самонавчання, теорії нечіткої логіки і нечіткої логіки типу-2, теорії штучних нейронних мереж і теорії прийняття колективних рішень для побудови архітектур ансамблів нейро-фаззі мереж для нечіткої кластеризації; теорії оптимізації та статистичного аналізу для синтезу ефективних методів нечіткої кластеризації, методів навчання та самоорганізації ансамблів нейро-фаззі мереж. Імітаційне моделювання застосовується для перевірки якості кластеризації з використанням синтезованих методів і архітектур.

Наукова новизна результатів дисертаційної роботи.

1. Вперше запропоновано метод адаптивного самонавчання складеної кластеризуючої нейро-фаззі мережі Кохонена, який об'єднує в собі можливісний і ймовірнісний підходи до кластеризації, що дозволило виконувати адекватну кластеризацію за наявності аномальних спостережень у даних.
2. Вперше запропоновано адаптивний варіант методу нечіткої еліпсоїдальної кластеризації на основі алгоритму Густафсона-Кесселя, що дозволило розширити набір форм скупчень, які можна виділити в даних, що надходять послідовно.
3. Отримали подальшого розвитку методи адаптивної можливісної та ймовірнісної кластеризації зі змінним фаззіфікатором шляхом виділення параметрів розмиття та придушення результатів нечіткої кластеризації, що дозволило виділити компоненти розділимості та компактності кластерів і завдяки цьому проводити виділення кластерів, які істотно відрізняються щільністю, в даних, що надходять послідовно.
4. Удосконалено метод адаптивної робастної кластеризації шляхом введення

критерія спеціального виду, який відрізняється тим, що є узагальненням правил Кохонена «переможець отримує все» і «переможець отримує більше» та може бути застосований для навчання в on-line режимі, що дозволило поліпшити якість кластеризації даних, які містять викиди.

Практичне значення отриманих результатів. Запропоновані в роботі методи навчання самоорганізованих мап для адаптивної обробки даних в умовах апіорної невизначеності, засновані на нейромережових і нейро-фаззі методах та архітектурах, доведені до програмної реалізації, що дозволило автоматизувати процес обробки нестаціонарних даних для вирішення завдання кластеризації. Запропоновані методи та архітектури, покладені в основу розроблених програмних модулів, надали можливість підвищити якість обробки даних в задачах кластеризації у режимі їх послідовної обробки, що підтвердило доцільність дослідження для вирішення практичних завдань.

Запропоновані в дисертаційній роботі методи адаптивної робастної кластеризації впроваджено в ТОВ «Видавництво «Ранок» (м. Харків) для вирішення завдання автоматичної класифікації та пошуку ілюстративних матеріалів, що значно прискорило підготовку до друку навчальних матеріалів у порівнянні з ручним пошуком, який застосовувався раніше (акт від 15.08.2013 р.). Запропоновані можливісні та робастні системи кластеризації впроваджені в Міжнародному фонді пам'яті Б.А. Чичибабіна (м. Харків), та використовуються для автоматичної обробки художнього тексту, а саме для розв'язання задач інтелектуального текстового пошуку, аналізу лексикографії та прагматики тексту (акт від 25.06.2013 р.)

Також результати, що пов'язані з розробкою методів адаптивного кластерного аналізу на основі рекурентних мап Кохонена, ансамблів адаптивних нейро-фаззі мереж та методів їх навчання для обробки даних за умов апіорної та поточної невизначеності, впроваджені в навчальний процес на кафедрі штучного інтелекту Харківського національного університету радіоелектроніки в курсах «Нейромережові методи обчислювального інтелекту» та «Штучні нейронні мережі (акт від 21.11.2013 р.), а також у науково-дослідні роботи Харківського національного університету радіоелектроніки (акт від 30.08.2013 р.).

Особистий внесок здобувача. Усі положення, що виносяться на захист, основні результати теоретичних та експериментальних досліджень отримані здобувачем особисто. Їх основний зміст викладено в роботах [1–15]. У публікаціях, написаних у співавторстві, здобувач зробив такий внесок: [1] – розробив рекурентні можливісний і ймовірнісний методи самонавчання нейро-фаззі мережі Кохонена; [2] – запропонував метод адаптивного самонавчання складеної нейро-фаззі мережі Кохонена для кластеризації, який об'єднує в собі можливісний і ймовірнісний підходи до кластеризації; [3] – розробив рекурентний метод самонавчання нейро-фаззі мережі Кохонена зі змінним фаззіфікатором; [4] – розробив рекурентний варіант алгоритму нечіткої кластеризації Густафсона-Кесселя; [5] – запропонував рекурентні методи самонавчання нечітких самоорганізованих мап Кохонена зі змінними параметрами розмиття та придушення; [6] – запропонував спосіб використання адаптивної можливісної кластеризації для обробки інформації в текстовому вигляді; [7] – розробив архітектуру модифікованої мапи Кохонена; [8] – запропонував архітектуру складеної мапи Кохонена для використання поєданого можливісного та ймовірнісного підходів

при кластеризації даних у вигляді тексту; [9] – розробив модель нечіткої кластеризації типу-2 на основі архітектури складеної нейро-фаззі системи; [10] – розробив модель нечіткої кластеризації типу-2 на засаді множини мап Кохонена; [11] – запропонував модель нечіткої мапи Кохонена з лінійним параметром фаззіфікації; [12] – запропонував модифіковану нечітку мапу Кохонена з нечіткою функцією фаззіфікації; [13] – розробив складену нейро-фаззі систему для адаптивної нечіткої кластеризації типу-2; [14] – запропонував модель нечіткої самоорганізовної мапи Кохонена для кластеризації з лінійним фаззіфікатором, що змінюється; [15] – розробив спосіб об'єднання адаптивних мап Кохонена спеціального виду в ансамблі та адаптивного формування нечіткого розбиття типу-2 на засаді колективного виведення.

Апробація результатів дисертації. Основні результати дисертаційної роботи були представлені та обговорені на 2-й Міжнародній науково-практичній конференції «Математичне та програмне забезпечення інтелектуальних систем "MPZIS-2008"» (Дніпропетровськ, 2008), Міжнародній науково-практичній молодіжній школі-семінарі студентів, аспірантів та молодих вчених «Інформаційні інтелектуальні системи» (Харків, 2008, 2009), 12-й Всеукраїнській (7-й міжнародній) студентській науковій конференції з прикладної математики та інформатики (Львів, 2009), Міжнародній науковій конференції «Автоматизація: проблеми, ідеї, рішення» (Севастополь, 2010), Міжнародній науковій конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (Євпаторія, 2011), Міжнародній науково-практичній конференції «Інформаційні технології та інформаційна безпека в науці, техніці та освіті "Інфотех- 2011"» (Севастополь, 2011), Міжнародній конференції з комп'ютерних наук і інформаційних технологій «CSIT`2011» (Львів, 2011), 6-й міжнародній школі-семінарі «Теорія прийняття рішень» (Ужгород, 2012), 2-й міжнародній конференції «Теоретичні та прикладні аспекти кібернетики » (Київ, 2012).

Публікації. За результатами дисертаційних досліджень опубліковано 15 наукових праць, із них 5 статей, серед яких 4 у фахових виданнях України з технічних наук та 1 за кордоном (4 статті входять до міжнародних наукометричних баз), 10 публікацій матеріалів і тез доповідей на міжнародних та всеукраїнських науково-технічних конференціях.

Структура дисертації. Дисертація складається зі вступу, чотирьох розділів, висновків, списку використаних джерел і додатку. Загальний обсяг дисертації складає 144 сторінки (з них 121 стор. основного тексту), 27 рисунків, 1 таблиця, список використаних джерел зі 120 найменувань на 15 сторінках та 1 додаток на 5 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету і завдання досліджень, наведено відомості щодо публікацій, апробації роботи та особистого внеску здобувача.

У **першому розділі** виконано аналіз стану проблеми створення систем, що самонавчаються, а саме розробки методів кластеризації даних. Описано переваги підходів і методів, що базуються на нечіткій (фаззі) логіці, виявлені перспективи застосування нечіткої логіки вищих порядків для вдосконалення методів кластеризації.

Розглянуто різновиди нейронних мереж для кластеризації, основна увага приділена нейро-фаззі мережам Кохонена.

Наголошено, що, оскільки задача кластеризації є недовизначеною, неможливо сформулювати загальний критерій успішності кластеризації. Використання цільових функцій при розробці процедур кластеризації передбачає задання таких критеріїв у неявному вигляді, виходячи з деяких апіорних уявлень про характер даних. Зважаючи на це, припускається, що ефективним способом обробки даних, природа яких невідома заздалегідь або змінюється з часом, є обробка даних ансамблем кластеризаторів, кожен з яких працює з власних передумов і впливає на кінцеву матрицю розбиття відповідно до своєї доцільності. У результаті проведеного аналізу сформульовані основні задачі дисертаційного дослідження, пов'язані з побудовою та вдосконаленням методів кластеризації, що придатні до роботи в ансамблях кластеризаторів і охоплюють найбільш поширені типи невизначеності в даних.

Другий розділ присвячено дослідженню та розвиненню методів адаптивної нечіткої кластеризації. Описано базові різновиди цільових функцій, способи їх удосконалення для врахування властивостей даних, описано основні аналітичні та числові методи оптимізації таких цільових функцій, оглянуто основані на них пакетні та рекурентні методи кластеризації. Розглянуто низку підходів і заснованих на них методів, спрямованих на обробку основних видів невизначеностей у даних. Узагальнення та універсалізація параметрів, які використовуються для отримання цих методів, дозволяють створювати моделі, що здатні гнучко пристосовуватися до змін характеристик досліджуваних даних в широких межах і мають мінімальне число параметрів, які налаштовуються вручну. Окрему увагу приділено подібності схем роботи основних методів, що дозволяє згодом розглядати ансамблі подібних моделей, які працюють паралельно для досягнення консенсусу про властивості даних.

Розглянуто засновані на прототипах методи кластеризації. Такі методи передбачають виділення з вибірки невеликої кількості найбільш типових, усереднених спостережень (званих прототипами або центроїдами) і розбиття решти вибірки на кластери на підставі близькості до них. Використовуючи підходи нечіткої логіки, можна розглядати кластери як нечіткі множини, а задачу кластеризації як розрахунок оптимальної функції належності до цих множин. Найпростіша цільова функція нечіткої кластеризації у загальному випадку має вигляд:

$$E(u_j, c_j) = \sum_{k=1}^N \sum_{j=1}^m u_j^\beta(k) d(x(k), c_j) \quad (1)$$

при обмеженні

$$\sum_{j=1}^m u_j(k) = 1, \quad (2)$$

де $k=1, 2, \dots, N$ – номер спостереження у вибірці, $u_j(k) \in [0, 1]$ – рівень належності вектора $x(k)$ до j -го кластеру, c_j – центроїд (прототип) j -го кластеру, β – невід'ємний параметр фаззіфікації (фаззіфікатор), що визначає розмитість границь між кластерами, N – обсяг вибірки даних, m – кількість кластерів.

Отримана аналітично процедура оптимізації такої цільової функції при використанні

евклідової метрики та при $\beta=2$ приводить до алгоритму Fuzzy C-Means (FCM), найрозповсюдженішому методу нечіткої кластеризації.

Обговорюється вплив значення фаззифікатора на результуюче розбиття, а також альтернативні форми введення компоненти фаззифікації, окрім степеневого параметру. Пропонується використовувати альтернативну цільову функцію з лінійною формою фаззифікатора:

$$E(u_j, c_j) = \sum_{k=1}^N \sum_{j=1}^m (\alpha u_j^2(k) + (1-\alpha)u_j(k)) \|x(k) - c_j\|^2, \quad (3)$$

де $\alpha \in [0,1]$ – параметр фаззифікації.

На її засаді виводиться процедура кластеризації з більш прозорим та толерантним до помилок настроювання параметром:

$$\begin{cases} u_j(k) = -\frac{1-\alpha}{2\alpha} + \left(1 + m \frac{1-\alpha}{2\alpha}\right) / \sum_{l=1}^m \frac{\|x(k) - c_j\|^2}{\|x(k) - c_l\|^2}, \\ c_j = \left(\sum_{k=1}^N (\alpha u_j^2(k) + (1-\alpha)u_j(k)) x(k) \right) / \left(\sum_{k=1}^N (\alpha u_j^2(k) + (1-\alpha)u_j(k)) \right), \\ \lambda(k) = -\left(1 + m \frac{1-\alpha}{2\alpha}\right) / \sum_{l=1}^m (2\alpha \|x(k) - c_l\|^2)^{-1}, \end{cases} \quad (4)$$

яка при $\alpha=1$ зводиться до алгоритму FCM, а також допускає ще деякі спрощення.

Використання розглянутих вище процедур має на увазі, що вибірка, яка підлягає кластеризації, задана заздалегідь у розмірі N спостережень і не може змінюватися в процесі обробки. Це часто є неприйнятним для роботи з великим обсягом даних та з даними, що поповнюються в реальному часі. У таких випадках стає на пригоді парадигма інкрементного навчання та адаптивні процедури, що дозволяють вирішувати задачу кластеризації в on-line режимі.

Для отримання інкрементної процедури за критерієм (3) оптимізуємо лагранжیان

$$L(u_j(k), c_j, \lambda(k)) = \sum_{k=1}^N \sum_{j=1}^m (\alpha u_j^2(k) + (1-\alpha)u_j(k)) \|x(k) - c_j\|^2 + \sum_{k=1}^N \lambda(k) \left(\sum_{j=1}^m u_j(k) - 1 \right) \quad (5)$$

за допомогою процедури нелінійного програмування Ерроу-Гурвіца-Удзави. В результаті отримуємо систему рівнянь, що здійснюють рекурентну мінімізацію цільової функції:

$$\begin{cases} u_j(k+1) = -\frac{1-\alpha}{2\alpha} + \left(1 + m \frac{1-\alpha}{2\alpha}\right) / \sum_{l=1}^m \frac{\|x(k) - c_j\|^2}{\|x(k) - c_l\|^2}, \\ c_j(k+1) = c_j(k) + \eta(k) (\alpha u_j^2(k+1) + (1-\alpha)u_j(k+1)) (x(k+1) - c_j(k)), \end{cases} \quad (6)$$

де $\eta(k)$ – параметр кроку налаштування, який визначає швидкість збіжності та

обирається зазвичай з емпіричних міркувань.

Далі у розділі обговорено обмеження (2) та його наслідків. Хоча воно полегшує розрахунки та інтерпретацію, але також призводить до чутливості методів до викидів, адже алгоритм буде «дотягувати» рівні належності до одиниці, навіть якщо спостереження лежить дуже далеко від центроїдів кластерів і об'єктивно повинно отримати мінімальну належність до кожного із них. Крім того, це обмеження призводить до конкуренції близьких кластерів, коли одне спостереження впевнено належить до кількох з них, але це не може бути описано так, щоб сума належностей не перевищила одиницю. Це унеможливорює розпізнавання кластерів, які перетинаються, а також вкладених кластерів. Пропонується подолати ці незручності, використовуючи можливісний підхід до нечіткої логіки, який не потребує використання обмеження (2). Цільова функція можливісної кластеризації має вигляд

$$E(u_j, c_j) = \sum_{k=1}^N \sum_{j=1}^m u_j^\beta(k) d(x(k), c_j) + \sum_{j=1}^m \mu_j \sum_{k=1}^N (1 - u_j(k))^\beta, \quad (7)$$

де скалярний параметр $\mu_j > 0$ визначає нечітку границю кластера.

Використовуючи підходи, описані раніше, запропонована адаптивна процедура можливісної кластеризації з лінійним фаззифікатором:

$$\begin{cases} u_j(k+1) = \frac{1}{2} \left(\mu_j(k) + \left(\alpha \mu_j(k) + (1-\alpha) \|x(k+1) - c_j(k)\|^2 \right) \right) / \left(\alpha \|x(k+1) - c_j(k)\|^2 + \mu_j(k) \right), \\ c_j(k+1) = c_j(k) + \eta(k) \left(\alpha u_j^2(k+1) + (1-\alpha) u_j(k+1) \right) (x(k+1) - c_j(k)), \\ \mu_j(k+1) = \left(\sum_{p=1}^{k+1} \left(\alpha u_j^2(p) + (1-\alpha) u_j(p) \right) \|x(p) - c_j(k+1)\|^2 \right) / \sum_{p=1}^{k+1} \left(\alpha u_j^2(p) + (1-\alpha) u_j(p) \right). \end{cases} \quad (8)$$

Зазначено, що формули розрахунку центроїдів у виразах (6) та (8) збігаються, що полегшує паралельне опрацювання даних обома методами.

Крім того, процедура легко доповнюється деякими методами постобробки результату, що в on-line режимі значно підвищує адаптивність. Зокрема механізмом виявлення нових кластерів на основі появи великої кількості спостережень з малими належностями до всіх існуючих кластерів (що можливо тільки при використанні саме адаптивної можливісної процедури), а також введення процедури придушення (suppression), такої, що

$$u_p^s(k) = \sum_{j=1}^m u_j(k) - \sigma \sum_{\substack{j=1 \\ j \neq p}}^m u_j(k) = \sum_{j=1}^m u_j(k) - \sigma \left(\sum_{j=1}^m u_j(k) - u_p(k) \right) = \sigma u_p(k) + (1-\sigma) \sum_{j=1}^m u_j, \quad (9)$$

де $u_p(k)$ – найбільше значення належності поточного спостереження, а σ – настроюваний параметр. Нескладно помітити, що зі зменшенням σ результати стають «більш чіткими», а зі збільшенням σ прямують до рішення, одержаного при фаззифікаторі $\beta = 2$. Видно, що параметр σ має той самий зміст регулювання рівня нечіткості, що й α у цільовій функції (3).

Далі в розділі розглянуті обмеження FCM-подібних алгоритмів, пов'язані з ізотропністю функції відстані. Це призводить до того, що, якщо скупчення даних розподілено не рівномірно, а наприклад у формі еліпсу, отриманий кластер все одно буде мати сферичну форму з радіусом довгої осі еліпсу. Це обмежує використання методів у реальних задачах. Для подолання цієї проблеми запропоновано можливісний варіант алгоритму Густафсона-Кесселя. Для того щоб отримати його адаптивну форму, використано формулу Шермана-Моррісона для інкрементного обернення матриці, в результаті чого отримано адаптивну можливісну процедуру з еліпсоїдальними кластерами:

$$\left\{ \begin{array}{l} u_j(k+1) = \left(1 + \left(((x(k+1) - c_j(k))^T A_j(k)(x(k+1) - c_j(k))) \mu_j(k)^{-1} \right)^{\frac{1}{1-\beta}} \right)^{-1}, \\ S_j(k+1) = S_j(k) + u_j^\beta(k+1) (x(k+1) - c_j(k)) (x(k+1) - c_j(k))^T, \\ S_j^{-1}(k+1) = S_j^{-1}(k) - \frac{u_j^\beta(k+1) S_j^{-1}(k) (x(k+1) - c_j(k)) (x(k+1) - c_j(k))^T S_j^{-1}(k)}{1 + u_j^\beta(k+1) (x(k+1) - c_j(k))^T S_j^{-1}(k) (x(k+1) - c_j(k))}, \\ |S_j(k+1)| = |S_j(k)| (1 + u_j^\beta(k+1) \|x(k+1) - c_j(k)\|^2), \\ A_j(k+1) = |S_j(k+1)|^{-\frac{1}{n}} S_j^{-1}(k+1), \\ c_j(k+1) = c_j(k) + \eta(k) u_j^\beta(k+1) A_j(k+1) (x(k+1) - c_j(k)), \\ \mu_j(k+1) = \left(\sum_{p=1}^{k+1} u_j^\beta(p) (x(p) - c_j(k+1))^T A_j(k+1) (x(p) - c_j(k+1)) \right) / \sum_{p=1}^{k+1} u_j^\beta(p). \end{array} \right. \quad (10)$$

Також запропоновано спосіб сумісної роботи можливісного та ймовірнісного адаптивних процедур Густафсона-Кесселя для взаємної компенсації недоліків обох підходів.

Остання глава розділу присвячена робастним процедурам кластеризації. Наявність викидів відбивається на розподілі даних повільним (порівняно з нормальним) убуттям його «хвостів». Основний використовуваний підхід для обробки даних із великою дисперсією – заміна метрики в цільовій функції на критерій близькості, що спадає повільніше, ніж квадрат евклідової відстані. Прикладами таких критеріїв можуть слугувати функція Гріна, функція Гауса, функція Коші тощо. Окремо зупинимось на використанні в якості критерію близькості функції Коші, яка є простою в розрахунках і легко виражається в інкрементній формі.

Дистанція Коші має вигляд

$$d_R^2(x, c) = \frac{1}{2} \sigma_c^2 \ln \left(1 + \left(\|x(k) - c\|^2 / \sigma_c^2 \right) \right), \quad (11)$$

де σ – настроюваний параметр. На основі цієї метрики можуть бути побудовані ймовірнісна та можливісна цільові функції, виведені пакетні та адаптивні процедури кластеризації, що й було зроблено. Найуніверсальнішою є адаптивна процедура нечіткої можливісної кластеризації за критерієм Коші:

$$\left\{ \begin{array}{l} u_j(k+1) = \left(1 + \left(d_R^2(x(k+1), c_j(k)) \mu_j(k)^{-1} \right)^{\frac{1}{\beta-1}} \right)^{-1}, \\ c_j(k+1) = c_j(k) + \eta(k) u_j^\beta(k+1) \left(\sigma_C^2(x(k+1) - c_j(k)) \right) / \left(\sigma_C^2 + \|x(k+1) - c_j(k)\|^2 \right), \\ \mu_j(k+1) = \sum_{p=1}^{k+1} u_j^2(p) \|x(p) - c_j(k+1)\|^2 / \sum_{p=1}^{k+1} u_j^2(p). \end{array} \right. \quad (12)$$

У розділі запропоновано низку процедур, кожна з яких є стійкою або адаптивною до окремих типів невизначеності, що зустрічаються в даних. Однак зрозуміло, що жодна з подібних процедур неспроможна охопити усі види невизначеності та нестационарності даних. Стає очевидною необхідність застосування систем колективного висновку, які включають до себе паралельні процедури з різними апіорними припущеннями щодо характеру даних. Кожна з таких процедур буде успішною на своїй підмножині даних чи окремий проміжок часу, а система колективного висновку повинна розпізнати цей момент і надати їй пріоритет.

Третій розділ присвячений архітектурам нейромережових систем та їх ансамблів, адаптованих для роботи з процедурами, описаними в попередньому розділі. Базовою структурою серед них є нейронна мережа Кохонена.

Зазначено, що базова процедура роботи нейромережі Кохонена за схемою «переможець отримує більше» (Winner Takes More, WTM) має форму

$$w_l(k+1) = w_l(k) + \eta(k) \phi(l, w^*) (x(k) - w^*(k)) \quad (13)$$

(де w^* – вектор ваг нейрона-переможця, а ϕ – функція сусідства між нейронами) дуже схожу на розрахунок центроїдів усіх розглянутих раніше процедур, якщо сприйняти функцію належності як функцію сусідства. Це відкриває широкі можливості як для удосконалення нейромережових архітектур, так і для організації продуктивної взаємодії процедур кластеризації на засаді нейромережевого підходу.

Архітектура адаптивної кластеризуючої нейро-фаззі мережі Кохонена наведена на рис. 1. Мережа здатна здійснювати як імовірнісну, так і можливісну адаптивну кластеризацію. Нейрони шару Кохонена розташовані в тому ж просторі, що входні дані, і під час кластеризації інтерпретуються як центроїди кластерів. Це означає, що

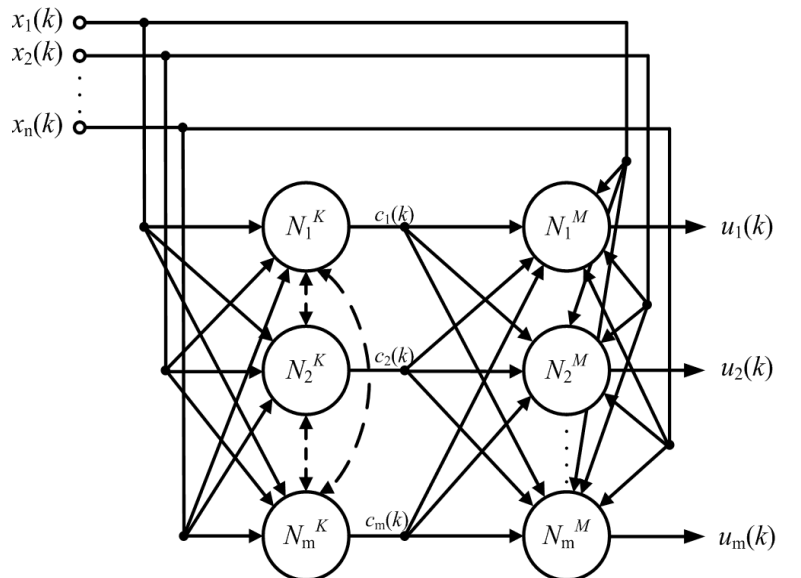


Рисунок 1 – Узагальнена архітектура адаптивної кластеризуючої нейро-фаззі мережі Кохонена

кілька мереж, що обробляють ті самі дані, здатні обмінюватися обчисленими центроїдами, тобто, по суті, мати об'єднаний перший шар нейронів.

Метод кластеризації комбінованою адаптивною нейро-фаззі мережею Кохонена можна представити таким чином:

$$\left\{ \begin{array}{l} c_j^{PR}(k+1) = c_j^{PS}(k) + \eta(k) (u_j^{PR}(k))^\beta (x(k+1) - c_j^{PS}(k)), \\ u_j^{PR}(k+1) = \|x(k+1) - c_j^{PR}(k+1)\|^{\frac{2}{1-\beta}} / \sum_{l=1}^m \|x(k+1) - c_l^{PR}(k+1)\|^{\frac{2}{1-\beta}}, \\ c_j^{PS}(k+1) = c_j^{PR}(k+1) + \eta(k) (u_j^{PS}(k))^\beta (x(k+1) - c_j^{PR}(k+1)), \\ u_j^{PS}(k+1) = \left(1 + \left(\|x(k) - c_j^{PS}(k+1)\|^2 \mu_j(k)^{-1} \right)^{\frac{1}{1-\beta}} \right)^{-1}, \\ \mu_j(k+1) = \sum_{p=1}^{k+1} (u_j^{PS}(p))^\beta \|x(p) - c_j^{PS}(k+1)\|^2 / \sum_{p=1}^{k+1} (u_j^{PS}(p))^\beta, \end{array} \right. \quad (14)$$

де індекс PR позначає розрахунок у рамках імовірнісного (probabilistic) підходу, а PS – можливісного (possibilistic).

За структурою методу (14) видно, що результатом кластеризації при такому підході є дві матриці розбиття – за версіями імовірнісного та можливісного методів. Зазвичай пропонується приймати як остаточну відповідь можливісне розбиття як більш універсальне, однак повне ігнорування результатів ймовірнісної кластеризації – явно нераціональне використання обчислювальних ресурсів. Використовуючи техніки колективного виведення, стало можливим поліпшити загальний результат за рахунок використання альтернативного розбиття U^{PR} .

Розглянуто окремі методи формування колективного виведення, зокрема ансамблів нейромереж, виділено деякі загальні принципи подібних складених систем. На основі теорії колективного вибору сформульовані принципи формування таких ансамблів і критерії їх ефективності. Показано, що з використанням ансамблів навіть однотипних систем за дотримання низки умов можна одержувати результат, який кращий, ніж будь-яка окрема система, що входить до такого ансамблю.

Зважена сума матриць, ваги якої відображають доречність кожного розбиття, дозволяє отримати результат кращий кожного окремого. Однак об'єктивна оцінка якості розбиття є принципово невирішуваною проблемою за природою задачі кластеризації. Проте існує низка індексів якості кластеризації, що можуть використовуватися як деякі аналоги навчального сигналу. У роботі пропонується використовувати так званий «індекс розбиття» як такий, що може бути розрахований адаптивно. Він має простий вигляд:

$$Q_U = \frac{1}{N} \sum_{j=1}^m \sum_{k=1}^N u_j^2(k)$$

та ясний фізичний зміст: чим краще виражені кластери, тим більше буде коефіцієнт розбиття, а його мінімум $Q \rightarrow 1/m$ досягається, якщо дані належать усім кластерам

рівномірно (що є очевидно тривіальним рішенням). Цей індекс дуже просто отримати в ітеративній формі, використовуючи формулу:

$$Q_{u_j}(k) = Q_{u_j}(k-1) + \frac{1}{k} (u_j^2(k) - Q_{u_j}(k-1)),$$

що в підсумку дозволяє отримати колективний висновок як зважене середнє:

$$u_j(k) = \left(Q_{u_j^{PS}}(k) u_j^{PS}(k) + Q_{u_j^{PR}}(k) u_j^{PR}(k) \right) / \left(u_j^{PS}(k) + u_j^{PR}(k) \right).$$

Аналогічно можна об'єднувати для колективної роботи будь-які описані в роботі адаптивні методи. При цьому поява в системі множини нечітких функцій належності, а також нечіткої функції їх якості дозволяє говорити про систему в цілому з точки зору нечіткої логіки типу-2. У рамках цього підходу, процедура є простим варіантом редукції моделі, до того ж таким, який виконується ітеративно.

Крім того, характер поведінки систем з лінійним фаззифікатором дозволяє стверджувати, що результати роботи таких систем з різним значенням α при однакових параметрах ініціалізації є зрізами по порозу α деякої неявно заданої вторинної функції належності.

Робота в ансамблі вимагає диверсифікації мереж, що його складають, за параметрами і, відповідно, за результатами. Ця обставина, а також те, що при паралельній роботі великої кількості мереж обмін інформацією про центроїди по типу процедури для кожної їхньої пари стає обчислювально складним, призводить до необхідності працювати з великим числом центроїдів, незалежно породженими різними мережами. Це призводить до необхідності ввести для центроїдів процедуру редукції, яка є аналогічною редукції функції належності і в цілому є частиною процедури редукції моделі, необхідної у будь-якій системі нечіткої логіки типу-2.

Зважаючи на те що в разі наявності в даних скупчень будь-яка процедура кластеризації, заснована на прототипах, прагнучиме розташувати центроїди в ці скупчення, було припущено, що множина центроїдів, розрахована множиною паралельно працюючих з тими самими даними нейромережами Кохонена, буде наближатися до скупчень даних незалежно від параметрів настройки та умов ініціалізації. У цьому випадку редукція центроїдів може бути зведена до виділення кластерів серед усієї множини центроїдів, розрахованих ансамблем, і усереднення цих скупчень за тим чи іншим правилом колективного виведення. Найпростіше рішення, яке задовольняє цьому формулюванню – провести процедуру кластеризації над вибіркою, що складається виключно із центроїдів, з тією лише різницею, що головним результатом цієї кластеризації буде не матриця розбиття, а отримані прототипи. Логічно використовувати для цього адаптивну нейромережу Кохонена, аналогічну введеним раніше. Загальна архітектура ансамблю, що реалізує повну процедуру адаптивної нечіткої кластеризації типу-2, наведена на рис. 2. Така архітектура може включати довільне число паралельних нейромереж Кохонена, що здійснюють кластеризацію будь-якими методами, розглянутими у другому розділі. Отримані цими методами центроїди в свою чергу кластеризуються вихідною мапою Кохонена, щоб знайти стійкі скупчення, а функції належності зважено сумуються в рамках підходу колективного виведення.

Здійснювана таким чином рекурентна нечітка кластеризація типу-2 дозволяє отримувати адекватні результати кластеризації після надходження кожного нового спостереження навіть в умовах невизначеності та мінливості вхідних даних. Система в

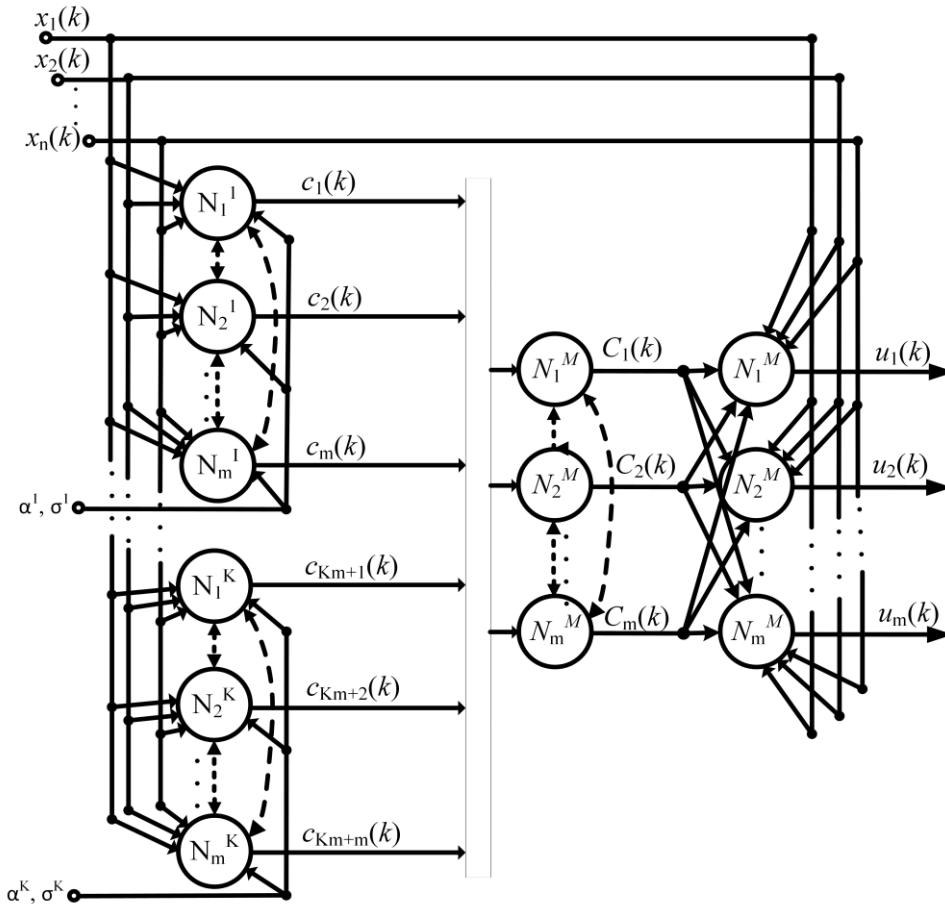


Рисунок 2 – Архітектура ансамблю нейро-фаззі мереж Кохонена для нечіткої кластеризації типу-2

експерименти для порівняння запропонованих архітектур з відомими раніше, для виявлення впливу на результати кластеризації кожного з настроюваних параметрів, для оцінки роздільної здатності окремих методів кластеризації у складних умовах (за наявності викидів, кластерів різної щільності, складної форми, тощо). Представлені також результати впровадження результатів дисертаційної роботи з метою їх практичного використання у прикладних системах.

У першій серії експериментів було досліджено вплив параметру фаззифікації на успіх кластеризації штучної вибірки з 5 частково перетнутих кластерів. На рис. 3 подані середні результати та дисперсія в серії із 50 експериментів для кожного значення параметра фаззифікації ($\alpha \in [0,1]$ для адаптивної нечіткої кластеризуючої мережі Кохонена (AFKCN), $1 < \beta \leq 5$ для FCM).

Можна бачити, що навіть на добре розділимій вибірці FCM є дуже чутливим до вибору степеневого фаззифікатора β , у той час як лінійний фаззифікатор майже не потребує налаштування. Зі збільшенням α в режимі без придушення спостерігається збільшення дисперсії результатів, однак на середній результат серії це не впливає.

on-line режимі пристосовується до змін характеру даних, а в залежності від вибору початкового поєднання кластеризаторів для побудови ансамблю може бути спочатку підготовлена до тих чи інших основних видів невизначеності (різного ступеня забрудненості викидами, різниці в характерних масштабах кластерів, їх числа, форми, щільності тощо) без шкоди для якості кластеризації.

Четвертий розділ присвячено експериментальним дослідженням та отриманню якісних характеристик запропонованих методів та архітектур. Були проведені

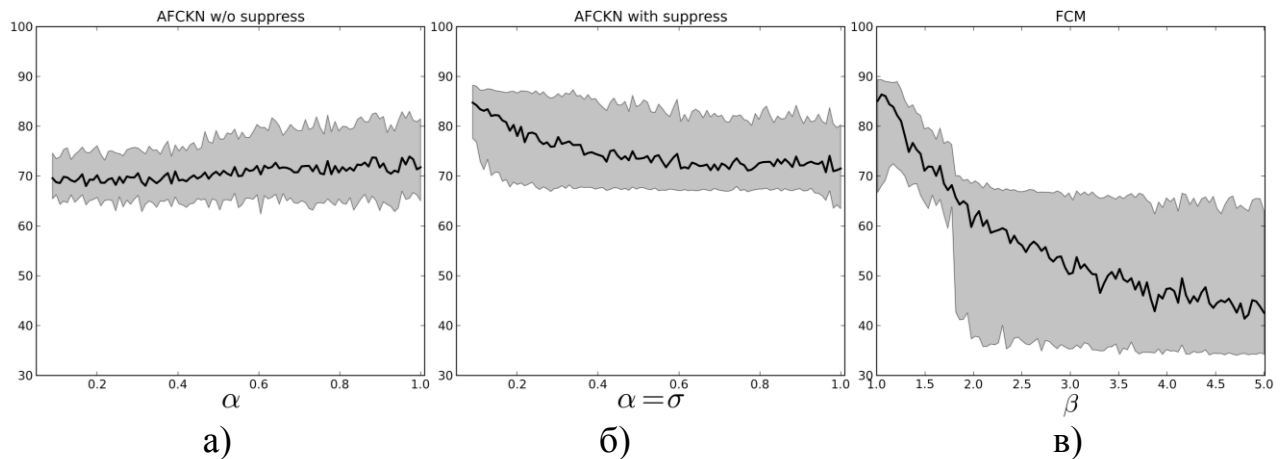


Рисунок 3 – Вплив параметра фаззифікації на точність кластеризації: а – AFKCN без придушення; б – AFKCN з придушенням; в – FCM

При роботі з придушенням характер впливу параметрів AFKCN на результат схожий на вплив фаззифікатора FCM, однак для AFKCN точність падає значно повільніше та залишається в середньому вище, ніж у режимі з придушенням на всьому діапазоні значень параметрів.

Результати наступної серії експериментів наведено в таблиці 1. Отримано порівняння точності кластеризації на стандартних розмічених вибірках для методів чітких К-середніх (НКМ), нечітких С-середніх (FCM) при $\beta = 2$ та адаптивної нечіткої кластеризуючої мережі Кохонена (AFKCN) з різними значеннями фаззифікатора та параметра придушення. Для кожної вибірки наведені середній, максимальний та мінімальний результати кожного метода, запущеного 50 разів із випадковою ініціалізацією.

Таблиця 1 – Порівняння точності кластеризації на стандартних вибірках

Досліджувані методи	Iris			Wine			WDBC		
	avg	max	min	avg	max	min	avg	max	min
НКМ	87	96	60	68	74	54	83	91	66
FCM ($\beta=2$)	70	72	33	69	74	33	86	87	86
AFKCN без придушення	72	96	68	63	75	27	87	93	63
AFKCN з придушенням ($\alpha=\sigma=0.1$)	85	93	66	70	73	66	89	91	83
AFKCN з придушенням ($\alpha=\sigma=0.3$)	76	63	66	68	70	56	89	91	86
AFKCN з придушенням ($\alpha=\sigma=0.5$)	75	90	66	68	70	66	90	91	88

Результати показують, що ефективність AFKCN з придушенням у цілому дещо вище та стабільніше, ніж у НКМ і FCM. Параметр α чинить тим більший вплив на ефективність, чим вище ступінь перекриття класів, і є майже незначущим у разі лінійно розділимих класів, як у вибірці WDBC.

У попередніх експериментах використовувалися рівні значення параметрів придушення та фаззифікації. Щоб оцінити механізм придушення окремо, була проведена серія експериментів, типовий результат яких наведений на рисунку 4.

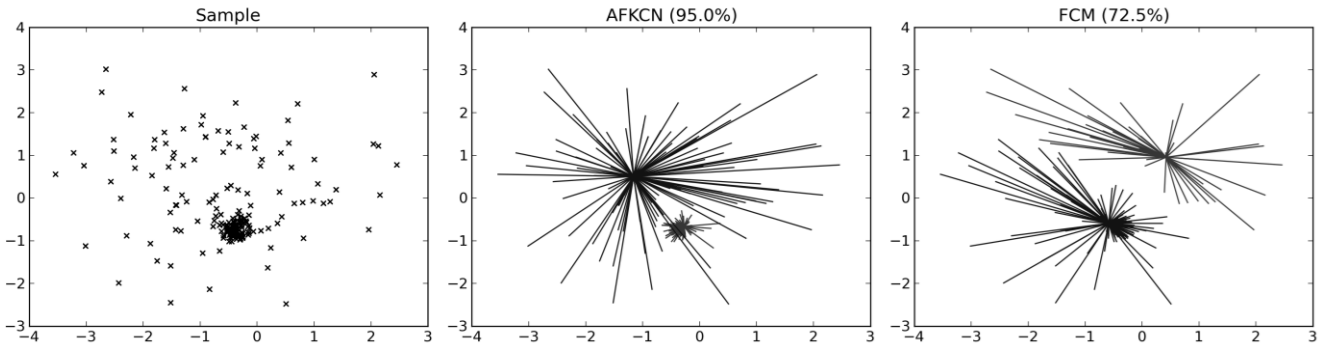


Рисунок 4 – Порівняння результатів на вибірці зі змінною щільністю спостережень

Для ілюстрації була створена синтетична вибірка, що складається зі спостережень, нормально розподілених відносно двох центрів з різними параметрами розподілу. AFKCN правильно проводить кластеризацію, якщо $\sigma < \alpha$. На рисунку 4 показаний результат при $\alpha = 0.5$ та $\sigma = 0.1$ (точність 95%), для порівняння наведено результат FCM при $\beta = 2$ (точність 72,5%).

Продемонстровано, що параметри α і σ доцільно розділити, якщо в даних очікуються кластери істотно різної щільності. Цей результат може бути корисним при роботі із сильно зашумленими даними, коли доцільно виділити викиди або фон в окремий кластер.

Окрема серія експериментів була проведена для перевірки робастності методу, заснованого на критерії Коші. Для цього була створена штучна вибірка з додаванням 20% викидів. На рис. 5 наведено центральну частину вибірки та її типове розбиття: методом FCM та адаптивною кластеризуючою за критерієм Коші системою з лінійним фаззифікатором.

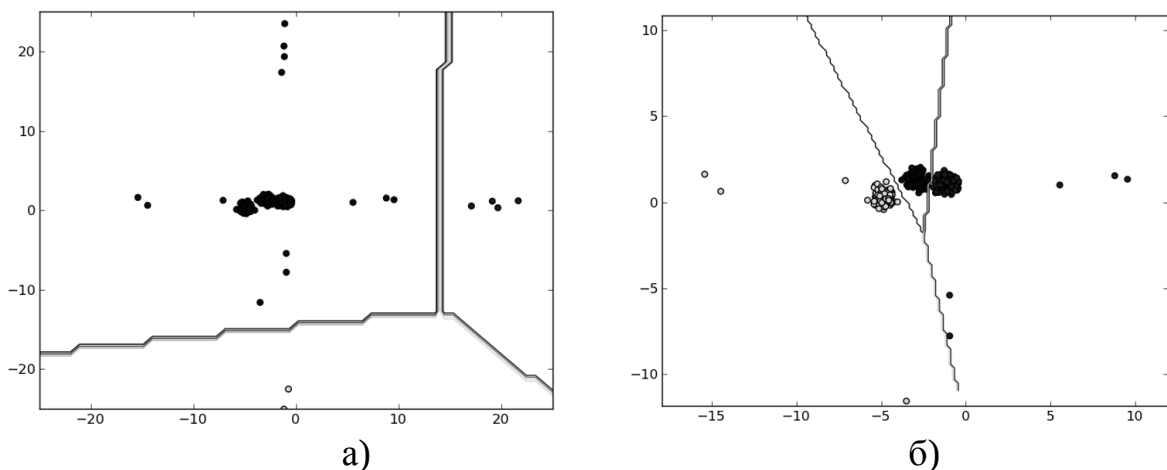
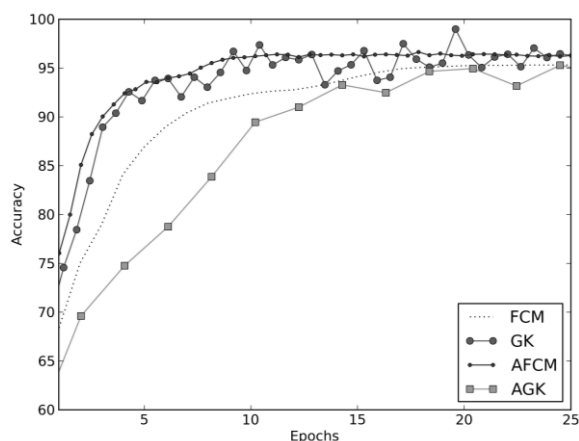
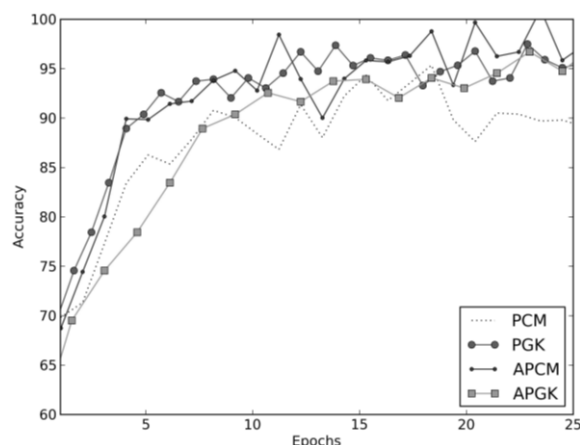


Рисунок 5 – Порівняння робастності методів на штучній вибірці: а – розбиття методом FCM; б – кластеризація за критерієм Коші

Для оцінки якості кластеризації методу Густафсона-Кесселя окрім перевірки, що його результати ідентичні в пакетній та адаптивній формах, була вивчена динаміка якості кластеризації залежно від кількості оброблених даних. Це було реалізовано через багатоепошне навчання (повторні проходи по вибірці), що дозволило порівняти як пакетні, так і адаптивні алгоритми. Результати наведено на рис. 6.



а)



б)

Рисунок 6 – Точність кластеризації залежно від числа ітерацій по вибірці: а - ймовірнісні методи; б – можливісні методи. Аббревіатурами позначені методи: FCM – нечітких С-середніх, GK – Густафсона-Кесселя, PCM – можливісний нечітких С-середніх, PGK – можливісний Густафсона-Кесселя. Літера «А» позначає адаптивні варіанти відповідних алгоритмів

Цікаво відзначити, що через велику кількість обчислюваних параметрів, адаптивні форми як можливісного, так і ймовірнісного метода Густафсона-Кесселя вимагають для якісного налаштування навіть більшої кількості спостережень, ніж їх пакетні форми, що зазвичай не є характерним для адаптивних алгоритмів, проте є закономірним з точки зору складності систем. Разом з тим значно важливіше, що, маючи набагато гнучкіші можливості адаптації до вхідних спостережень, адаптивні форми методу Густафсона-Кесселя зберігають монотонність зростання якості кластеризації з числом прийнятих спостережень навіть краще, ніж більшість пакетних алгоритмів, що особливо вигідно виділяє адаптивний можливісний метод Густафсона-Кесселя на тлі інших можливісних алгоритмів, які традиційно відрізняються нестабільністю результатів.

Наведено результати впровадження дисертаційних досліджень. Вирішені важливі практичні завдання: 1) використання розроблених методів для автоматичної сегментації та класифікації візуальних, зокрема ілюстративних, матеріалів за різними критеріями пошуку дозволило врахувати особливості задачі підготовки друкованої продукції, класифікації та пошуку ілюстрацій, що в цілому привело до підвищення ефективності процесу підготовки друкованої продукції порівняно з методами, що застосовувалися раніше, та до зниження витрат, що виникають у разі ручного пошуку ілюстрацій; 2) використання розроблених методів для вирішення задач нечіткого пошуку, лексикографії та прагматики тексту дозволило врахувати особливості задачі лінгвістичного та літературознавчого аналізу художніх творів, автоматичної лексикографічної обробки природномовного тексту; обробляти інформацію, задану у формі художнього тексту, що сприяло підвищенню ефективності та розширенню можливостей літературознавчих досліджень порівняно з методами, що застосовувалися раніше. У додатку наведено документи щодо впровадження результатів досліджень та їх практичного використання в прикладних системах.

ВИСНОВКИ

У дисертаційній роботі представлені результати, які згідно з поставленою метою є рішеннями актуальної науково-технічної задачі підвищення ефективності нечіткої адаптивної кластеризації даних в умовах поточної та апіорної невизначеності. Автором одержано такі наукові та практичні результати.

1. Новий метод адаптивного самонавчання складеної кластеризуючої нейро-фаззи мережі Кохонена, який об'єднує в собі можливісний і ймовірнісний підходи до кластеризації, що дозволило виконувати адекватну кластеризацію за наявності аномальних спостережень у даних.

2. Розвинено методи адаптивної можливісної та ймовірнісної кластеризації зі змінним фаззифікатором шляхом виділення параметрів розмиття і придушення результатів нечіткої кластеризації, що дозволило виділити компоненти розділимості та компактності кластерів і завдяки цьому проводити виділення кластерів, які істотно відрізняються за щільністю, в даних, що надходять послідовно.

3. Вперше розроблено адаптивний варіант методу нечіткої еліпсоїдальної кластеризації, що дозволило розширити набір форм кластерів в даних, що надходять послідовно.

4. Удосконалено метод адаптивної робастної кластеризації шляхом введення критерію спеціального виду, який придушує аномальні спостереження, що дозволило поліпшити якість кластеризації даних із викидами.

5. Удосконалено архітектури нейромережових систем колективного виведення, що дозволило застосовувати їх для здійснення адаптивної колективної кластеризації у послідовному режимі, стійкої до різних видів невизначеності та нестаціонарності у даних.

6. Проведене імітаційне моделювання запропонованих методів і систем показало їх переваги в порівнянні з існуючими методами в задачах обробки нестаціонарних даних. Із використанням запропонованих моделей і методів були вирішені актуальні практичні завдання сегментації та автоматизованої класифікації зображень, а також задачі лексикографічного, прагматичного аналізу тексту та інтелектуального пошуку. Результати впроваджені в навчальний процес та в науково-дослідні роботи Харківського національного університету радіоелектроніки.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Бодянский Е. В. Самообучающаяся нейро-фаззи система для адаптивной кластеризации текстовых документов [Текст] / Е. В. Бодянский, В. В. Волкова, Б. В. Колчигин // Бионика интеллекта: науч.-техн. журнал. – Харьков: ХНУРЭ. – 2009. – №1(70). – С. 34-38.

2. Колчигин Б. В. Адаптивная нейро-фаззи сеть Кохонена [Текст] / Б. В. Колчигин, В. В. Волкова, Е. В. Бодянский // Радиоэлектроника, информатика, управление: науч. журнал. – Запорожье: ЗНТУ. – 2011. – №1 (24). – С. 99-104. (Входить до міжнародних наукометричних баз INSPEC, Index Copernicus, Ulrich's Periodical Directory, WorldCat, EBSCO, BASE та ін.)

3. Bodyanskiy Ye. Adaptive neuro-fuzzy Kohonen network with variable fuzzifier [Текст]

/ Ye. Bodyanskiy, B. V. Kolchygin, I. P. Pliss // International Journal «Information Theories and Applications». – Sofia: ITHEA, 2011. – Vol.18., №3. – P. 215-223. (Входит до міжнародної наукометричної бази BASE.)

4. Бодянский Е. В. Адаптивная нечёткая кластеризация на основе алгоритма Густафсона-Кесселя [Текст] / Е. В. Бодянский, Б. В. Колчигин, В. В. Волкова, И. П. Плисс // Управляющие системы и машины. – 2013. – №2 (244). – С.40-46. (Входит до міжнародної наукометричної бази SCOPUS та ін.)

5. Колчигин Б. В. Адаптивная нечёткая кластеризация с переменным фаззификатором [Текст] / Б. В. Колчигин, Е. В. Бодянский // Кибернетика и системный анализ. – 2013. – №3 (49). – С. 47-55. (Входит до міжнародних наукометричних баз SCOPUS, INSPEC, EBSCO, ProQuest, Computer Science Index та ін.)

6. Бодянский Е. В. Нечёткая возможностная кластеризация текстовых документов на основе самоорганизующейся карты Кохонена / Е. В. Бодянский, Б. В. Колчигин, В. В. Волкова // Математическое и программное обеспечение интеллектуальных систем (MPZIS-2008): II Международная научно-практическая конференция, Днепрпетровск: 12-14 ноября 2008 г.: тезисы докладов – Днепрпетровск: ДНУ, 2008. – С. 47-48.

7. Волкова В. В. Возможностная фаззи-кластеризация текстовых массивов в реальном времени на основе самообучающейся нейронной сети / В. В. Волкова, Б. В. Колчигин // Информационные интеллектуальные системы: Научно-практическая молодёжная школа-семинар студентов, аспирантов и молодых учёных, Харьков, 2-4 декабря 2008 г.: тезисы докладов – Харьков: ХНУРЭ, 2008. – С. 22-25.

8. Волкова В. В. Нейро-фаззі системи, що самонавчаються, для адаптивної кластеризації текстових документів / В. В. Волкова, Б. В. Колчигін // XII Всеукраїнська (VII міжнародна) студентська наукова конференція з прикладної математики та інформатики, Львів, 28-29 квітня 2009 р.: тези доповідей, – Львів: ЛНУ ім. І. Франка, 2009. – С. 135-136.

9. Колчигин Б. В. Нейро-фаззи система второго типа для пакетной кластеризации / Б. В. Колчигин // Информационные интеллектуальные системы: II научно-практическая молодёжная школа-семинар студентов, аспирантов и молодых учёных, Харьков, 8-9 декабря 2009 г.: тезисы докладов – Харьков: ХНУРЭ, 2009. – С. 230-233.

10. Бодянский Е. В. Адаптивная нейро-фаззи сеть Кохонена типа-2 / Е. В. Бодянский, Б. В. Колчигин // Автоматизация: проблемы, идеи, решения: международная научная конференция, Севастополь, 6-10 сентября 2010 г.: материалы конференции. – Севастополь: СевНТУ, 2010. – Том 1. – С. 135-137.

11. Бодянский Е. В. Адаптивная самоорганизующаяся нейро-фаззи сеть с настраиваемым фаззификатором / Е. В. Бодянский, Б. В. Колчигин // Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта: международная научная конференция, Евпатория, 2011 г.: материалы конференции. – Херсон: ХНТУ, 2011. – Том 1. – С. 264-268.

12. Бодянский Е. В. Нечёткая кластеризация с нечётким фаззификатором / Е. В. Бодянский, Б. В. Колчигин, И. П. Плисс // Информационные технологии и информационная безопасность в науке, технике и образовании «Инфотех-2011»: международная научно-практическая конференция, Севастополь, 5-10 сентября 2011 г.: материалы конференции. – Севастополь: СевНТУ, 2011. – С. 5-6.

13. Bodyanskiy Ye. Type-2 neuro-fuzzy system for clustering with variable fuzzifier / Ye. Bodyanskiy, B.V. Kolchygin // Computer Science and Information Technologies «CSIT`2011»: International Conference, Lviv, November 16-19, 2011: proceedings. – Lviv: «Lvivska Politechnika», 2011. – P. 15-16.

14. Бодянський Є. В. Адаптивна нечітка кластеризація зі змінним фаззифікатором / Є.В. Бодянський, Б. В. Колчигін, О. В. Колесникова // Теорія прийняття рішень: VI Міжнародна школа-семінар, Ужгород, 1-6 жовтня 2012 р.: праці конференції – С. 38-39.

15. Kolchygin B. V. Ensemble of neuro-fuzzy Kohonen networks for adaptive clustering / B. V. Kolchygin // Theoretical and Applied Aspects of Cybernetics: 2nd International Scientific Conference of Students and Young Scientists, Kyiv, November 12-16, 2012: proceedings. – Kyiv: Bukrek. – 2012. – P. 176-181.

АНОТАЦІЯ

Колчигін Б. В. Адаптивні нейро-фаззі системи для нечіткого кластерного аналізу в умовах невизначеності. – 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, 2014.

Дисертація присвячена дослідженню та розробці методів адаптивної нечіткої кластеризації з використанням нейромереж Кохонена та їх ансамблів. Розглядаються засновані на прототипах методи кластеризації та їх модифікації. Зазначено, що, як усі методи навчання без вчителя, будь-які методи кластеризації мають деякі апіорні припущення про характер розподілу даних в оброблюваній вибірці. Побудова системи кластеризації, яка вдало працює в умовах заздалегідь невідомих або мінливих характеристик даних, можлива тільки при використанні методів колективного виведення. Побудова такої системи на основі кластеризаторів пов'язана зі складнощами, що виходять з самої природи задачі кластеризації, а саме з відсутності об'єктивних критеріїв якості розбиття, наявності великого числа параметрів, що настраюються, та використанням переважно пакетних методів роботи з даними.

У роботі пропонується ряд методів кластеризації, які охоплюють найпоширеніші методи обробки джерел нестаціонарності в даних і модифіковані для можливості роботи в on-line режимі. Всі методи побудовані на загальному підході, який полягає в мінімізації числа параметрів, що настраюються, та наданні їм ясного фізичного сенсу, а також у забезпеченні можливості роботи цих методів з вибіркою, що поповнюється. Отримані таким чином методи кластеризації працюють у деякому сенсі одноманітно, що дозволяє об'єднувати їх в нейромережеві ансамблі для колективного отримання матриці розбиття, кращої кожної із тих, що отримані окремими методами.

Використовуючи особливості роботи нечітких методів кластеризації, побудованих на прототипах, уперше була реалізована система кластеризації в парадигмі нечіткої логіки типу-2, що працює повністю в on-line режимі. Використання такого потужного інструменту продемонструвало свою ефективність на зростаючих вибірках даних, у яких спостерігається суттєвий дрейф характеристик із часом: поява й зникнення кластерів, зміна їх характерних масштабів і щільності, міри перетину тощо.

Проведено експериментальне дослідження властивостей і характеристик розроблених методів, вирішена низка практичних задач.

Ключові слова: нейронні мережі Кохонена, нейро-фаззі мережі, методи самонавчання, функція належності, нечітка логіка типу-2, адаптивна нечітка кластеризація.

АННОТАЦИЯ

Колчигин Б. В. Адаптивные нейро-фаззи системы для нечёткого кластерного анализа в условиях неопределённости. – На правах рукописи.

Диссертация на соискание учёной степени кандидата технических наук по специальности 05.13.23 – системы и средства искусственного интеллекта. – Харьковский национальный университет радиоэлектроники, Министерство образования и науки Украины, Харьков, 2014.

Диссертация посвящена исследованию и разработке методов адаптивной нечёткой кластеризации с помощью нейросетей Кохонена и их ансамблей. Рассматриваются основанные на прототипах методы кластеризации и их модификации. Отмечено, что как методы обучения без учителя, любые методы кластеризации имеют некоторые априорные представления о характере распределения данных в обрабатываемой выборке – таких, как форма, размер, плотность и число кластеров, количество выбросов, степень наложения кластеров и т. д., – и работают успешно только в случае оправдания этих ожиданий. Построение кластеризирующей системы, успешно работающей в условиях неизвестных заранее или меняющихся в процессе характеристик данных, возможно только при использовании методов коллективного вывода, которые подразумевают синтез актуальной модели в on-line режиме на основе параллельно работающих более простых систем обработки, каждая из которых успешна только на некоторой части выборки. Построение такой системы на основе кластеризаторов сопряжено со сложностями, связанными с самой постановкой задачи кластеризации, в первую очередь отсутствием объективного критерия оценки качества разбиения, а также с большим числом настраиваемых параметров и использованием преимущественно пакетных методов работы с данными.

В работе предлагается ряд методов кластеризации, охватывающих самые распространённые методы обработки источников нестационарности в данных и модифицированные для возможности работы в on-line режиме. Все методы построены на общем подходе, заключающемся в минимизации числа настраиваемых параметров и придании им ясного физического смысла, а также в обеспечении возможности работы методов с пополняющейся выборкой. Полученные таким образом методы кластеризации работают в некотором смысле единообразно, что позволяет объединять их в нейросетевые ансамбли для коллективного вывода матрицы разбиения, которая лучше каждой из полученных отдельными методами.

Используя особенности работы нечётких методов кластеризации, построенных на прототипах, впервые была реализована система кластеризации в парадигме нечёткой логики типа-2, работающая полностью в on-line режиме. Использование столь мощного инструмента продемонстрировало свою эффективность на растущих выборках данных, в которых наблюдается существенный дрейф характеристик со временем: появление и исчезновение кластеров, смена их характерных масштабов и

плотности, доли выбросов, степени наложения и т. д.

Проведено экспериментальное исследование свойств и характеристик разработанных методов, а также сравнительный анализ с некоторыми ранее известными методами. С использованием предложенных подходов решены практические задачи сегментации и автоматизированной классификации иллюстративного материала для учебно-методической литературы, а также задачи лексикографического и прагматического анализа текстов и интеллектуального поиска в корпусе текстов.

Ключевые слова: нейронные сети Кохонена, нейро-фаззи сеть, методы самообучения, функция принадлежности, нечёткая логика типа-2, адаптивная нечёткая кластеризация.

ABSTRACT

Kolchygin B.V. Adaptive neuro-fuzzy systems for fuzzy cluster analysis under conditions of uncertainty. – Manuscript.

The thesis for the degree of candidate of technical sciences, specialty 05.13.23 – systems and methods of artificial intelligence. – Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2014.

The thesis is devoted to research and development of adaptive fuzzy clustering systems using Kohonen neural networks and their ensembles. Prototype-based clustering methods and their modifications are considered. It is noticed that as the methods of unsupervised learning, any clustering methods have some a priori view about the nature of the data distribution in the treated sample, and work successfully only if data meets these expectations. Building a clustering system, successfully operating in unknown in advance conditions or in cases of changing characteristics of the data during work is possible only with using of methods of collective reasoning. It involves the synthesis of the current model in on-line mode based on parallel processing systems, each of which is successful only on some part of the sample. The construction of such a system on the basis of clustering systems fraught with difficulties due to the formulation of the problem of clustering, primarily the lack of objective criteria for estimating the quality of the partition, as well as a large number of adjustable parameters and using mainly batch methods to process the data.

The thesis proposes a number of clustering methods, covering the most common methods of processing the sources of nonstationarity in data and modified to be able to work in on-line mode. All methods are based on the common approach to minimize the number of adjustable parameters, and give them a clear physical meaning, as well as providing opportunity to work with a growing sample. Thus obtained clustering methods work in some sense uniformly, that allows to combine them into neural network ensembles for the collective result of the partitioning matrix, the best of each of those obtained by individual methods.

Using the features of the fuzzy prototype-based clustering methods, the clustering system in the paradigm of fuzzy logic type-2 working completely in on-line mode was implemented first. Using such a powerful tool has demonstrated its effectiveness in the growing data samples in which there is a significant properties' drift over time: the appearance and disappearance of clusters, changing their characteristic scale and density of outliers, the degree of overlapping, etc.

An experimental study of the properties and characteristics of the developed methods is carried out and recommendations on their use in solving practical problems are proposed.

Keywords: Kohonen neural network, neuro-fuzzy network, self-learning algorithms, membership function, type-2 fuzzy logic, adaptive fuzzy clustering.