Глибинна Кластерувальна Нейро-Фаззі Система та її Послідовне Самонавчання

Анна Норцова
кафедра штучного інтелекту
Харківський національний університет
радіоелектроніки
Харків, Україна
anna.nortsova@nure.ua

Анастасія Дейнеко
кафедра штучного інтелекту
Харківський національний університет
радіоелектроніки
Харків, Україна
anastasiia.deineko@nure.ua

Deep Clustering Neuro-Fuzzy System and its Sequentional Self-Learning Algorithm

Anna Nortsova
Artificial Intelligence Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
anna.nortsova@nure.ua

Anastasiia Deineko
Artificial Intelligence Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
anastasiia.deineko@nure.ua

***Анотація*—В статті запропоновано новий метод послідовного кластерування потоків даних в умовах апріорно невідомої кількості класів. Пропонована модель містить в собі можливості нечітких самоорганізовних мап Кохонена та використовує ієрархічний підхід.**

***Abstract*—This paper proposed a new clustering method for serial data streams in the conditions of the apriori unknown number of classes. The proposed model contains the possibilities of fuzzy self-organising maps by Kohonen and uses a hierarchical approach.**

Ключові слова—глибинне навчання; індекс Ксі-Бені; інтелектуальний аналіз даних; кластерування; нейронна мережа; нейро-фаззі система; самоорганізовна мапа Кохонена; штучний інтелект

Keywords—deep learning; parametr clustering accuracy Xi-Beni; data mining;clustering; neural network;neuro-fuzzy system; Kohonen self-organised map; artificial intelligence

I. Вступ

Дуже актуальним напрямом є задача кластерування даних, що є одним із методів аналізу та обробки множин об'єктів. Його мета полягає у групуванні елементів на окремі підгрупи – кластери, що схожі за певними ознаками. Так як в умовах новітнього часу даних стає дедалі більше, потреба у адекватному групуванні зростає пропорційно.

На сьогоднішній день запропоновано багато підходів для отримання якісних результатів кластерування, але не існує однозначного рішення, яке б задовольняло всі вимоги. Серед підходів, що використовуються, особливо ефективними показали себе нейронні мережі та нейро-фаззі системи [1,2], завдяки, насамперед своїм універсальним апроксимуючим та екстраполюючим можливостям та здатності навчатися в умовах суттєвої структурної і параметричної невизначенності.

В задачах інтелектуального аналізу даних, що пов'язані з кластеруванням широкого розповсюдження набули самоорганізовні мапи Кохонена [1]. Властивості самоорганізації SOM пов'язані з тим, що налаштування синаптичних ваг відбувається за відсутністю зовнішнього навчального сигналу, тобто в режимі самонавчання, при цьому кожен вхідний образ викликає налаштування тих чи інших параметрів.

У випадку неопуклих та кластерів, що перетинаються, рішення про належність вхідного образу до одного з кластерів, що приймається згідно до правила «переможець отримує все», може давати неточне кластерування. Це зумовлено тим, що деякі образи можуть належати в той самий час до декількох кластерів з певним ступенем належності. Саме у зв'язку з цим доцільно забезпечити SOM можливості і властивості нечіткого кластерування.

II. КЛАССТЕРУВАЛЬНА НЕЙРО-ФАЗЗІ СИСТЕМА

A. *Конкурентні нейронні мережі*

Особливим видом самонавчання є, так зване, конкурентне навчання, коли всі нейрони мережі «змагаються» між собою за право бути активним, реалізуючи принцип «Переможець отримує все» (Winner takes all), що веде до того, що в мережі може активуватися тільки один нейрон. Саме ця особливість конкурентного навчання забезпечила йому широке використання в задачах класифікації та кластерування.

Процес конкуренції може бути представлений у формі наведеній у формулах (1) та (2):

$$y_q(k) = \begin{cases} 1, \text{якщо } w_q^T(k)\widetilde{x_k} > w_p^T(k)\widetilde{x_k} \text{ для всіх } p \neq q, \\ 0 \text{ в протилежному випадку,} \end{cases} \quad (1)$$

$$\widetilde{x_k} = \frac{\overrightarrow{x_k}}{\left\|\overrightarrow{x_k}\right\|}, \left\|w_q(k)\right\| = 1, \quad (2)$$

При цьому в кожен момент часу налаштовується тільки нейрон-переможець за допомогою алгоритму відомого як правило навчання Т. Кохонена [1], наведеного у формулі (3).

$$w_q(k+1) = \begin{cases} w_q(k) + \eta(k)\left(\widetilde{x_k} - w_q(k)\right), \text{якщо } q - \text{нейрон переміг} \\ w_q(k) \text{ в протидежному випадку} \end{cases} \quad (3)$$

Поточний нормований вектор $\widetilde{x_k}$ найближчий до вектора синаптичних ваг $w_2(k)$, в результаті чого «перемагає» третій нейрон який і налаштовує свої параметри за допомогою правила Кохонена, «підтягуючи» $w_2(k)$, до $\widetilde{x_k}$ на відстань пропорційну параметру кроку $\eta(k)$ як наведено у формулі (4):

$$\begin{cases} w_2^T(k)\widetilde{x_k} > w_1^T(k)\widetilde{x_k}, \\ w_2^T(k)\widetilde{x_k} > w_2^T(k)\widetilde{x_k}, \end{cases} \quad (4)$$

Конкурентне навчання лежить в основі ряду нейромереж, що набули широкого поширення в задачах обробки інформації та інтелектуального аналізу даних.

B. *Самоорганізовні мапи Т. Кохонена*

Мережа Кохонена [3] відноситься до самоорганізовних мереж. це означає, що вони не отримують бажаний вихідний сигнал при надходженн вхідного навчального вектора, а в результаті навчання мережу розділяє вхідні сигнали на класи, таким чином формуючи топологічні карти. Однією з найважливіших властивостей навченої мережі Кохонена є її здатність до узагальнення. Суть роботи мережі Кохонена полягає в тому, що відбувається кластерування вхідних векторів в групи схожих між собою векторів. При цьому ваги мережі налаштовуються так, щоб вхідні образи, які належать одному кластеру, активізували один і той ж вихідний нейрон.

Вектор кожного з нейронів мережі Кохонена може замінити групу відповідних йому векторів, що класифікуються.

Самоорганізовна мапа має дуже просту архітектуру з прямою передачею інформації. Крім нульового (рецепторного) шару, вона містить єдиний шар нейронів, який дуже часто називають шаром Кохонена.

Саме завдяки такій організації, кожен нейрон мережі отримує всю інформацію по аналізованому образу і генерує на свій вихід відповідний відгук.

Після цього в шарі Кохонена виникає режим конкуренції [4], в результаті якої визначається єдиний нейрон-переможець з максимальним вихідним сигналом. Цей сигнал по латеральним зв'язкам забезпечує збудження найближчих «сусідів» переможця і придушення реакції далеко віддалених вузлів.

Розглянемо докладніше архітектуру самоорганізовної мапи. На вхід мережі надходить n -мірний вхідний сигнал. мережа містить єдиний шар з m нейронів, які утворюють на площині прямокутні решітки. Нейрони характеризуються своїм місцем розташування в мережі. кожен нейрон шару Кохонена пов'язаний з кожним входом нульового (вхідного) шару прямими зв'язками, а також з усіма іншими нейронами поперечними (латеральними) зв'язками.

В процесі навчання сусідні нейрони впливають один на одного сильніше, ніж ті, які розташовані далі. Саме латеральні зв'язки в мережі забезпечують збудження одних нейронів і гальмування інших. Кожен нейрон з шару Кохонена формує зважену суму сигналів $f(x,w) = \sum_{i=1} w_i x_i$. При цьому, якщо синапси прискорюють, то $w_{ij} > 0$. Якщо ж синапси гальмуючі, то $w_{ij} < 0$.

Перед початком роботи алгоритму навчання мережі Кохонена вхідні вектори попередньо нормалізуються як показано у формулі (5):

$$\widetilde{x_i} = \frac{x_i}{\sqrt{\sum_i x_i^2}} = \frac{x_i}{\|x\|}, i = 1, 2, \ldots, N \quad (5)$$

Самоорганізовні мапи можуть мати різну топологію. Однак найбільш часто рецептори і нейрони розташовуються у вузлах одно- або двовимірної решітки.

C. *Нейронна мережа Т. Кохонена з нечітким висновком*

Властивості самоорганізації SOM пов'язані з тим, що налаштування синаптичних ваг відбувається без зовнішнього навчального сигналу, тобто в режимі самонавчання, при цьому кожен вступний образ викликає налаштування тих чи інших параметрів.

У разі пересічних і неопуклих кластерів рішення про приналежність вхідного образу до одного з кластерів, прийняте відповідно до правила «переможець отримує все», може давати неточне кластерування, так як деякі образи можуть належати одночасно до кількох кластерів з певним ступенем приналежності. У зв'язку з цим

розглядається нечітка самоорганізовна мапа, в якій нейрони SOM замінені нечіткими правилами і множинами. Ця мережа є, по суті, схемою реалізації Fuzzy C-means методу кластерування.

Процедура самонавчання, в такому випадку, теж базується на принципах конкурентного навчання. Загальноприйнята процедура самоорганізації складається з трьох основних етапів: конкуренції, кооперації і синаптичної адаптації [5]. Самоорганізація починається з аналізу образу, що надходить з рецепторного шару на всі нейрони шару Кохонена. Для кожного з нейронів обчислюється відстань як показано у прикладі (6)

$$D\left(\tilde{x}_k, w_q(k)\right) = \left\|\tilde{x}_k - w_q(k)\right\|, \quad (6)$$

до того ж, входи слід попередньо пронормувати, щоб $\|\tilde{x}_k\| = 1$. В якості відстані (6) використовується евклідова метрика, а мірою близькості між векторами $\tilde{x}_k$ і $w_q(k)$ може слугувати скалярний добуток, що обчислюється за формулою (7):

$$D\left(\tilde{x}_k, w_q(k)\right) = \tilde{x}_k^T w_q(k) = cos(\tilde{x}_k, w_q(k)) = cos\,\theta_q \quad (7)$$

Далі обирається нейрон-переможець (8), «найближчий» до вхідного образу такий, що:

$$(\tilde{x}_k, w^*(k)) = \min_q D\left(\tilde{x}_k, w_q(k)\right) \quad (8)$$

Після чого у найпростішому випадку, оминаючи етап кооперації, можна налаштувати синаптичні ваги мережі.

Однією з особливостей мапи Кохонена є наявність етапу кооперації в процесі самоорганізації, коли нейрон-переможець визначає так звану локальну область топологічного сусідства, в якій збуджується не тільки він сам, але і його найближче оточення, при цьому більш близькі до переможця нейрони збуджуються сильніше, ніж віддалені. Ця топологічна область визначається функцією сусідства $\varphi(q, l)$, що залежить від відстані $D\left(w_q(k), w_l(k)\right)$ між переможцем w_q^* та будь-яким нейроном з шару Кохонена $l = 1, 2, \ldots, n$ і деякого параметра, що задає ширину.

Використання функцій сусідства призводить до модифікованого правила навчання Кохонена (формула 9):

$$w_l(k+1) = w_l(k) + \eta(k)\phi(q, l, k)\left(\tilde{x}_k - w_l(k)\right), l = 1, 2, \ldots, m, \quad (9)$$

що реалізує принцип «Переможець отримує більше» (Winner takes more – WTM) замість традиційного «Переможець отримує все».

У разі, якщо в навчальній вибірці є класифікація спостережень, то для подальшого більш точного налаштування радіусів кластерів може бути використана проста градієнтна процедура оптимізації. Налаштування радіусів кластерів буде виражатися в отриманні більш точної нечіткої класифікації спостережень.

III. АНСАМБЛІ НЕЧІТКИХ МАП КОХОНЕНА

Якість рішення конкретної поставленої задачі (емуляції, прогнозування, розпізнавання образів, зворотного моделювання, управління тощо) може бути істотно підвищена за допомогою ансамблів (колективів, комітетів, сумішей, банків, груп) штучних нейромереж [6], в яких в даному випадку вхідні дані послідовно обробляються декількома нечіткими самоорганізовними мапами Т. Кохонена (FSOM), вихідні сигнали яких передаються на наступну ітерацію доти, доки не будуть задовольняти обраному критерію якості кластерування. Концептуальна схема нейромережевого ансамблю представлена на рис. 2.

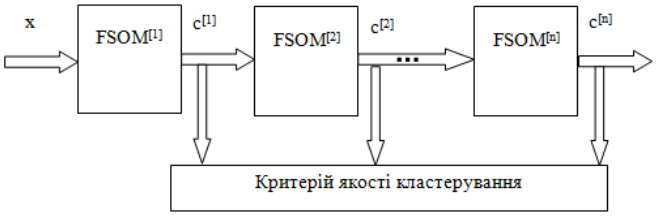

Рис. 3. Ансамбль нечітких нейронних мереж Кохонена

Системи машинного навчання, засновані на ансамблях, мають на увазі паралельне навчання багатьох моделей для вирішення одного завдання[7]. На відміну від традиційного підходу до машинного навчання, замість того, щоб намагатися сформулювати за даними єдину найкращу гіпотезу, ансамбль виробляє цілий набір гіпотез на основі різних моделей і далі комбінує їх в узагальнене рішення.

Так як в даній роботі розглядаються ієрархічні методи кластерування, то в цьому випадку метою використання ансамблів є реалізація саме ієрархічного підходу. Суть розробленого підходу полягає у тому, щоб поєднати властивості двох алгоритмів: самоорганізовних мап Кохонена і c-means з використанням нечіткості (FCM – Fuzzy C-means).

Як зазначалося раніше властивості SOM добре застосовуються для реалізації потокового кластеруваняя, тобто дають можливість роботи в online-режимі, а механізм c-means з нечіткою логікою використовується для вирішення проблеми кластерів, що перетинаються. Для оцінки результатів кластерування і вибору оптимальної кількості кластерів, прийнято рішення використовувати індекс Ксі-Бені(XB) наведеній у формулі (10).

$$XB = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ij}^m \|x_k - v_i\|^2}{n \min_{ii} \|v_j - v_i\|^2} \quad (10)$$

Чисельник у даній формулі відображає компактність нечіткого розбиття, а знаменник – якість розділення між кластерами. Згідно до цього, мінімальному значенню XB відповідає висока якість кластерування.

Спираючись на це система буде нарощувати свою архітектуру до того часу, поки не буде досягнутий деякий мінімум індекса Ксі-Бені.

Спираючись на вищесказане, запропонований ансабль при нарощенні своєї архітектури більше трьох шарів може розглядатися як глибинна система послідовно з'єднаних моделей нечітких самоорганізовних мап Т. Кохонена з метою досягнення оптимального результату кластерування.

IV. Імітаційне моделювання

Однією з основних переваг, притаманних пропонованій самонавчанній нейро-фаззі системі є те, що вона послідовно будує свою архітектуру, і автоматично визначає оптимальну кількість кластерів, спираючись на значення індексу Ксі-Бені на кожному етапі обробляння даних.

Серію експериментів було проведено на штучно зсинтезованих наборах даних з різним ступенем розмитості та перекриття класів, аби дослідити якість кластерування в послідовному режимі опрацювання інформації для різних варіантів розбиття даних.

Кожен з наборів даних містить вісімдесят спостережень з двома ознаками (для очності) у кожному спостереженні. Тестові дані були згенеровані таким чином, аби у першому наборі класи були чітко розподілені (crisp dataset), у другому наборі кластерні границі були дещо розмиті (fuzzy dataset), у третьому випадку класи сильно перетиналися (extra fuzzy dataset). Розроблена система видала кращі результати кластерування на наборі даних з чітко розподіленими класами і відповідно якість кластерування зменшувалась зі збільшенням розсіювання даних, тобто на вибірці, де границі класів спостережень є більш розмитими. Але не дивлячись на погіршення якості кластерування, запропонована система впоралася навіть з дуже розсіяними вибірками.

Спостереження надходили до нейро-фаззі мережі у послідовному режимі, вагові коефіцієнти нейронів були проініціалізовані, використовуючи online модифікацію обраного алгоритму кластерування на датасеті з довільних двадцятьох спостережень відповідного набору даних.

В Таблицях 1-3 наведено чисельні характеристики якості кластерування, що оцінювалися за допомогою індекса Ксі-Бені.

ТАБЛИЦЯ I. Результати кластерування Crisp dataset

I шар	**C = 2**	**III шар**	**C = 4**
Індекс Ксі-Бені	0.15687	***Індекс Ксі-Бені***	0.127232
II шар	**C = 3**	**IV шар**	**C = 5**
Індекс Ксі-Бені	0.052129	***Індекс Ксі-Бені***	0.15985

ТАБЛИЦЯ II. Результати кластерування Fuzzy dataset

I шар	**C = 2**	**III шар**	**C = 4**
Індекс Ксі-Бені	0.16668	***Індекс Ксі-Бені***	0.131235
II шар	**C = 3**	**IV шар**	**C = 5**
Індекс Ксі-Бені	0.07587	***Індекс Ксі-Бені***	0.27516

ТАБЛИЦЯ III. Результати кластерування Crisp dataset

I шар	**C = 2**
Індекс Ксі-Бені	0.12528
II шар	**C = 3**
Індекс Ксі-Бені	0.10584
III шар	**C = 4**
Індекс Ксі-Бені	0.020364
IV шар	**C = 5**
Індекс Ксі-Бені	0.510392
V шар	**C = 6**
Індекс Ксі-бені	0.93626

V. Висновки

Запропоновано новий метод послідовного кластерування потоків даних при апріорно невідомій кількості кластерів на основі нечітких самоорганізовних мап Т. Кохонена з використанням ієрархічного підходу.

Застосування запропонованого методу дає можливість визначення кількості кластерів за відсутності заздалегідь відомої інформації у випадках роздільних кластерів і кластерів, що перетинаються.

Реалізація розробленого методу досить проста і полягає в об'єднанні можливотей SOM для роботи з потоками даних в режимі online та ідеї алгоритму FCM для вирішення проблеми кластерів, що перетинаються у вибірках з розсіяними даними. Оцінка якості кластерування проводилася за допомогою індексу Ксі-Бені.

Література References

[1] Kohonen T. Self-Organizing Maps. Berlin: Springer-Verlag, 1995. 362 p.

[2] Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. N.Y.: Plenum Press, 1981. 272 p.

[3] Гібридна каскадна оптимізована нейронна мережа / І.П. Плісс, О.К. Тищенко, Д.С. Копаліані // Радіоелектроніка. Інформатика. Управління. Запоріжжя: ЗНТУ. 2014. 1(30). С. 129–134.

[4] Каскадные нейронные сети / Аведьян Э. Д., Баркан Г. В., Левин И. К. // Автоматика и телемеханика. 1999. No3. С. 38-55.

[5] Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems / Chin-Teng Lin, C. S. George Lee . Prentice Hall. 1996. 797 p.

[6] Schalkoff R. J. Artificial Neural Networks . N.Y. : The McGraw-Hill Comp. 1997. 528 p.

[7] The cascade neo-fuzzy neural network and its learning algorithm / Bodyanskiy Ye., Viktorov Ye., Pliss I. // Вісник Ужгородського національного університету. Серія «Математика і інформатика». 2008. Вип. 17. С. 48-58.