УДК 004.032.26 **https://doi.org/10.33082/td.2017.1-1.10**

ОНЛАЙН МОДИФІКАЦІЯ МЕТОДУ Х-СЕРЕДНІХ НА ОСНОВІ АНСАМБЛЮ САМООРГАНІЗОВНИХ МАП Т. КОХОНЕНА

ОНЛАЙН МОДИФИКАЦИИ МЕТОДА Х-СРЕДНИХ НА ОСНОВЕ АНСАМБЛЯ САМООРГАНИЗУЮЩИХСЯ КАРТ Т. КОХОНЕНА

ONLINE MODIFICATION OF THE METHOD OF X-MEDIUM ON THE BASIS OF ANSAMBLY OF SELORGANIZED MAP T. KOHONEN

Є.В. БОДЯНСЬКИЙ, докт.техн.наук,
А.О. ДЕЙНЕКО, канд.техн.наук,
П.Є. ЖЕРНОВА, В.О. РЄПІН
Харківський національний університет радіоелектроніки, Україна

Розглянуто послідовну модифікацію методу кластерування Х-середніх. Цей підхід базується на ідеї ансамблю кластерувальних нейронних map Т. Кохонена. При цьому кожна кластерувальна мережа містить різну кількість нейронів, яка визначається можливою кількістю кластерів. Всі члени ансамблю працюють паралельно, а якість кластерування визначається за допомогою індексу Цаліньського-Харабаша. Проведені експериментальні дослідження на даних з репозіторія UCI підтвердили ефективність запропонованого підходу.

Ключові слова: *кластерування, метод Х-середніх, ансамбль нейронних мереж, самоорганізовна мапа, самонавчання, нейронна мережа Т. Кохонена, міра схожості.*

Рассмотрена последовательная модификация метода кластеризации Х-средних. Этот подход основывается на идее ансамбля кластеризующих нейронных сетей, в качестве которых используются самоорганизующиеся карты Т. Кохонена. При этом каждая кластеризующая сеть содержит разное количество нейронов, определяемое возможным числом кластеров. Все члены ансамбля работают параллельно, а качество кластеризации определяется с помощью индекса Цалиньского-Харабаша. Проведенные эксперименты на репозиторных данных подтвердили эффективность развиваемого подхода.

Ключевые слова: *кластеризация, метод Х-средних, ансамбль нейронных сетей, самоорганизующаяся карта, самообучение, нейронная сеть Т. Кохонена, мера схожести.*

The modified X-means method for clustering in the case when observations are sequentially fed to processing the proposed. This approach's based on the ensemble of the clustering neural networks, proposed ensemble contains the T. Kohonen's self-organizing maps. Each of the clustering neural networks consist of different number of neurons, where number of clusters is connected with the quality of there neurons. All ensemble members process information that siquentionally is fed to the system in the parallel mode. The effectiveness of clustering process is determined using Caliński-Harabasz index. The self-learning algorithm uses similarity measure of special type that. The feature of proposed method is absent of the competition step, i.e. neuron-winner is not determined. A number of experiments has been held in order to investigate the proposed system's properties. Experimental results have proven the fact that the system under consideration could be used to solve a wide range of Data Mining tasks when data sets are processed in an online mode. The proposed ensemble system provides computational simplicity, and data sets are processed faster due to the possibility of parallel tuning.

Keywords: *clustering, X-means method, ensemble of neural network, self-organization map, self-learning, T. Kohonen's neural network, similarity measure.*

Вступ. Завдання кластерування масивів даних є важливою частиною загальної проблеми Data Mining, а для її вирішення на сьогодні розроблено безліч різних методів [1,2].

При обробці великих обсягів інформації на перший план виходять вимоги по швидкодії і простоті чисельної реалізації використовуваних алгоритмів кластерування.

Одним з найбільш популярних алгоритмів є метод К-середніх, завдяки своїй простоті, наочності результатів і можливості їх ясної інтерпретації.

Цей метод відноситься до алгоритмів, заснованих на обчисленні прототипів-центроїдів, в результаті чого масив вихідних даних

$$X = \{x(1), x(2),..., x(k),..., x(N)\} \subset R^n,$$
$$x(k) = (x_1(k),..., x_i(k),..., x_u(k))^T, k = 1,2,..., N$$

розбивається на m кластерів, де їх кількість m задається апріорно або вибирається, як правило, виходячи з суто емпіричних міркувань.

Для формального знаходження числа кластерів m був розроблений метод X -середніх [3], заснований на статистичному аналізі розподілу даних у вихідному масиві X .

Якщо при роботі з K -середніми число кластерів m було вибрано правильно, то отримані результати повністю збігаються з результатами X -середніх.

Останні роки в зв'язку з інтенсивним розвитком Data Stream Mining [4] природно виникла необхідність вирішення завдань кластерування в online режимі, коли дані на обробку послідовно надходять спостереження за спостереженням, обсяг масиву N не обмежений і зростає з часом, а k набуває сенсу поточного дискретного часу.

У подібній ситуації стандартні K-середні неефективні, проте з успіхом можуть бути використані кластерувальні нейронні мережі Т. Кохонена (SOM) [5], що вирішують завдання в online режимі, а одержаний результат повністю збігається з K-середніми в силу використання загального критерію кластеризації-самонавчання, заснованого на евклідовій метриці.

При цьому проблема вибору m тут залишається відкритою, включення додаткових «мертвих» нейронів в мережу, як правило, її не вирішує, а використання X-середніх в online режимі в їх традиційній формі принципово неможливо.

Альтернативою стандартним X-середнім може бути використання ідеї кластерувальних ансамблів, при цьому нами пропонується формувати ансамбль на основі паралельно з'єднаних входами SOM^m, кожна з яких апріорно орієнтована на різну кількість можливих кластерів $m = 1,2,3...,M$. Таким чином, перша кластерувальна мережу ансамблю працює в припущенні $m = 2$, тобто в шарі Кохонена містить всього два нейрона з синаптичними вагами-центроїдами w_1^2 і w_2^2.

Другий елемент ансамблю містить три нейрона з векторами синаптичних ваг w_1^3, w_2^3, w_3^3, і, нарешті, остання SOM^M ансамблю працює в припущенні, що число можливих кластерів дорівнює M, тобто містить M нейронів − адаптивних лінійних асоціаторів.

Мета роботи. Мета роботи полягає у дослідженні результатів модифікації методу x-середніх на основі ансамблю самоорганізовних мап Т. Кохонена.

Матеріали дослідження.

1. Алгоритм настроювання нейронних мереж ансамблю

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

Розглянемо процес самонавчання m-й мережі Кохонена SOM^m, що містить m нейронів з синаптичними вагами

$$\left\{w_1^m, w_2^m, ..., w_m^m\right\} \subset R^n.$$

В основі алгоритму налаштування синаптичних вагів полягає принцип конкурентного самонавчання, який реалізується в три основні етапи (конкуренція, кооперація, синаптична адаптація) і починається з

аналізу вхідного вектора-образу $x(k)$, що надходить з рецепторного (нульового) шару на всі нейрони шару Кохонена.

Для кожного з нейронів обчислюється відстань

$$D\left(x(k), w_j^m(k-1)\right) = \left\| x(k) - w_j^m(k-1) \right\|, \; j = 1,2,...,m,$$

при цьому, якщо вхідні сигнали попередньо пронормовані за допомогою перетворення

$$\tilde{x}(k) = \frac{x(k)}{\|x(k)\|} \tag{1}$$

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

$$sim\left(\tilde{x}(k), w_j^m(k-1)\right) = x^{\sim T}(k) w_j^m(k-1) = \cos\left(\tilde{x}(k), w_j^m(k-1)\right). \tag{2}$$

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

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

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

$$w_j^m(k) = \begin{cases} w_j^m(k-1) + \eta(k) \times \\ \times \left(\tilde{x}(k) - w_j^m(k-1)\right), \quad якщо \quad w_j^m(k-1) = w^{m^*}(k-1), \\ w_j^m(k-1) \quad у\ протилежному\ \ випадку. \end{cases} \tag{3}$$

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

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

Регулювання кроку $\eta(k)$ зазвичай проводиться, виходячи з емпіричних міркувань, а загальна рекомендація полягає в тому, що він повинен монотонно зменшуватися в процесі самонавчання. У найпростішому випадку для регулювання кроку можуть бути використані співвідношення

$$\eta(k)=r^{-1}(k), r(k)=\alpha r(k-1)+\|x(k)\|^2, 0\le\alpha\le 1,$$

або

$$r(k)=\alpha r(k-1)+1, 0\le\alpha\le 1 \quad (4)$$

для входів, нормованих відповідно до (1).

Зрозуміло, що при $\alpha=1, \eta(k)=k^{-1}$, тобто задовільняє умовам стохастичної апроксимації.

Важливою особливістю нейронної мережі Кохонена є наявність етапу кооперації, коли нейрон-переможець $w^{m^*}(k-1)$ визначає локальну область топологічного сусідства, в якому збуджується не тільки він сам, але і його оточення, при цьому більш «схожі» на переможця нейрони збуджуються сильніше ніж більш віддалені «сусіди». Ця область описується функцією сусідства $\varphi(j,l), l=1,2,...,m$, що залежить від відстані

$$D\left(w^{m^*}(k-1), w_l^m(k-1)\right)=D\left(w_j^m(k-1), w_l^m(k-1)\right),$$

між переможцем і будь-яким з нейронів $w_l^m(k-1)$ шару Кохонена. Як правило $\varphi(j,l)$ – це ядерна функція симетрична щодо максимуму в точці з $D\left(w_j^m(k-1), w_l^m(k-1)\right)$ і приймаюча в ній одиничне значення $\varphi(j,l)=1$. Зі збільшенням відстані $D\left(w_j^m(k-1), w_l^m(k-1)\right)$ ця функція монотонно зменшується.

У переважній більшості випадків в якості функції сусідства використовується ґавсіан

$$\varphi(j,l)=\exp\left(-\frac{\left\|w_l^m(k-1)-w^{m^*}(k-1)\right\|^2}{2\sigma^2}\right).$$

Використання функції сусідства призводить до алгоритму самонавчання

$$w_l^m(k)=w_l^m(k-1)+\eta(k)\varphi(j,l)\left(\tilde{x}(k)-w_l^m(k-1)\right)\forall l=1,2,...,m, \quad (5)$$

що реалізує правило «переможець отримує більше» (WTM), при цьому при $l=j$ цей алгоритм збігається зі співвідношенням (3).

В принципі, можна взагалі відмовитися від етапу конкуренції та визначення переможця як такого.

При цьому в ролі переможця в даному випадку виступає сам вхідний вектор-образ, а в якості функції сусідства використовується міра схожості (2).

При цьому алгоритм самонавчання m-го елемента ансамблю набуває вигляду

$$\begin{aligned} w_l^m(k) &= w_l^m(k-1)+\eta(k)\left[\cos\left(\tilde{x}(k), w_l^m(k-1)\right)\right]_+\left(\tilde{x}(k)-w_l^m(k-1)\right)= \\ &= w_l^m(k-1)+\eta(k)\left[\tilde{x}^T(k)w_l^m(k-1)\right]_+\left(\tilde{x}(k)-w_l^m(k-1)\right)= \\ &= w_l^m(k-1)+\eta(k)\left[y_l^m(k)\right]_+\left(\tilde{x}(k)-w_l^m(k-1)\right), \end{aligned} \tag{6}$$

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

Зрозуміло, що процедура (6) з обчислювальної точки зору набагато простіше стандартних алгоритмів (3), (5), завдяки виключенню етапу конкуренції і має ясний фізичний зміст.

2. Визначення кількості кластерів

В процесі роботи ансамблю постійно проводиться оцінка якості кластерування за допомогою критерію Цаліньського-Харабаша [2] або в його стандартній формі, або за допомогою його online модифікації. При цьому критерій в загальному вигляді має форму

$$CH(m)=\frac{1}{m-1}TrS_B^m\left(\frac{1}{N-m}TrS_w^m\right)^{-1} \tag{7}$$

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

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

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

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

$$u_j=\begin{cases}1, & \text{якщо } x(k) \text{ належить } j-\text{му кластеру} \\ 0-y & \text{протилежному випадку}\end{cases}$$

– чітка функція належності *k*-го спостереження *j*-му кластеру.

Переписавши вираз для TrS_B^m у формі

$$TrS_B^m = \frac{1}{N}\sum_{j=1}^{m} N_j^m \left\| w_j^m - w^{-m} \right\|^2,$$

а TrS_w^m – $TrS_w^m = \frac{1}{N}\sum_{j=1}^{m} u_j(k)\left\| x(k) - w_j^m \right\|^2$, критерій (7) можна представити у вигляді

$$CH(m) = \frac{\frac{1}{m-1}\sum_{j=1}^{m}\sum_{\tau=k-s+1}^{k} N_j^m \left\| w_j^m - w^{-m} \right\|^2}{\frac{1}{N-m}\sum_{j=1}^{m}\sum_{\tau=k-s+1}^{k} u_j(k)\left\| x(k) - w_j^m \right\|^2} \tag{8}$$

більш зручному з точки зору обчислювальної реалізації.

При аналізі даних, що надходять на обробку в online режимі, розрахунок критерію (8) доцільно організувати на ковзному вікні розмірності s $(s = 1,2,..,N)$, при цьому в поточний момент часу k $CH(m)$ можна записати як

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

де

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

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

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

Пропонована процедура ансамблевого online кластерування на основі системи нейронних мереж Т. Кохонена є за суттю адаптивною модифікацією методу X-середніх, орієнтованою на обробку потоків даних, досить проста в чисельній реалізації і дозволяє вирішити задачу чіткого кластерування в умовах апріорно невідомого або змінного числа кластерів.

3. Імітаційне моделювання

Для підтвердження працездатності розробленого ансамблю самоорганізовних мап Т. Кохонена була вирішена задача кластерування на основі штучно згенерованої вибірки і тестових вибірок з UCI-репозіторія [14]. Було взято набори даних:

1. Штучна вибірка Random Matrix, яка наочно відображає три лінійно розділимих кластери.

Вибірка Random Matrix, як представлено на рис. 1, містить три лінійно розділимих класів, де кожен елемент вибірки має три випадкові параметри.

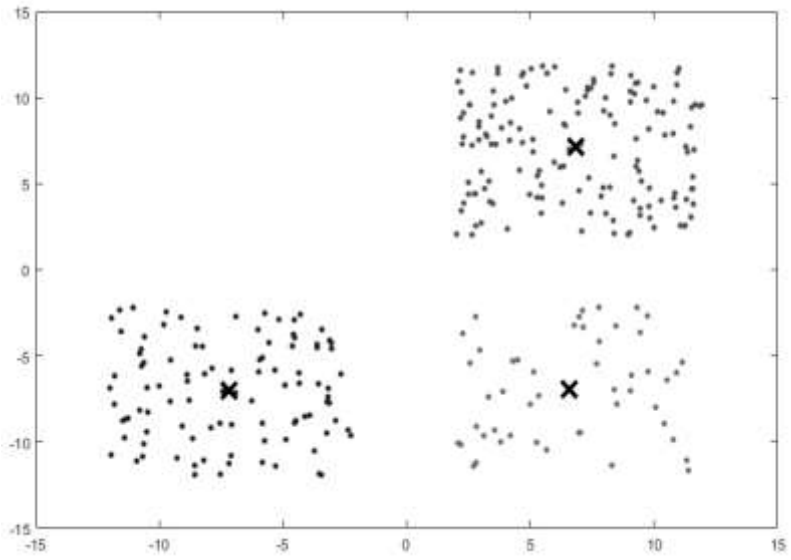


Рис. 1. Штучно згенерована лінійно розділима вибірка Random Matrix

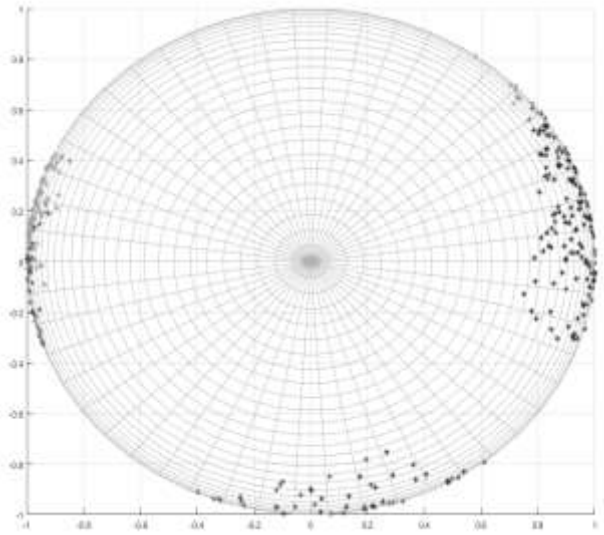
2. Тестова вибірка «Iris». Вибірка складається з даних про квітки ірису, по 50 примірників з трьох видів − Ірис щетиністий (Irissetosa), Ірис віргінський (Irisvirginica) і Ірис різнокольоровий (Irisversicolor).

Для кожного екземпляра вимірювалися чотири характеристики (в сантиметрах): довжина чашолистки (sepallength); ширина чашолистки (sepalwidth); довжина пелюстки (petallength); ширина пелюстки (petal-width).

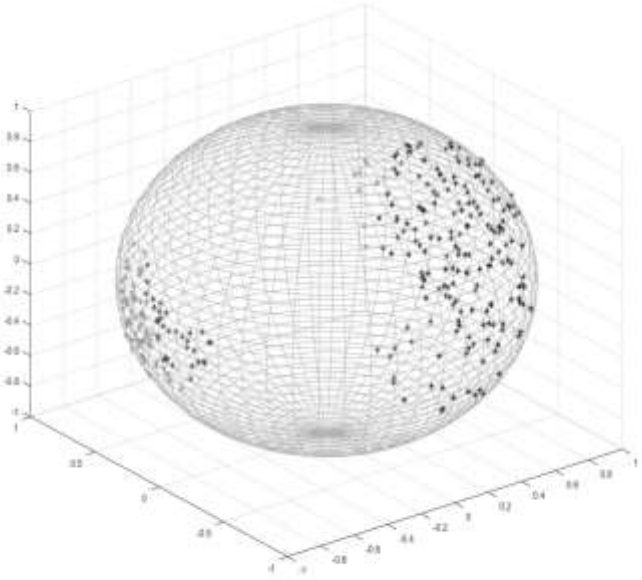
Всі дані були спочатку пронормовані на гіперкулю в інтервалі [-1,1] і відцентровані щодо середнього значення.

З метою оцінки ефективності ансамблю самоорганізовних мап Т. Кохонена (SOM^m) результати кластерування порівнювались зі стандартним методом кластерування К-середніх.

Для підтвердження якості кластерування було взято індекс Цаліньського-Харабаша, який наведено в табл. 1 для вибірок Random Matrix та Iris. Візуалізація результатів кластерування наведена на рис. 2 та рис. 3.

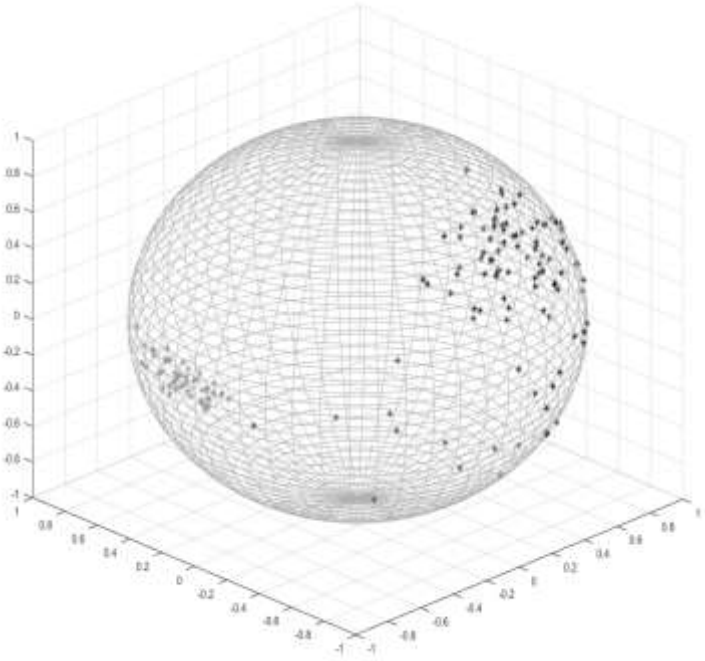


а)

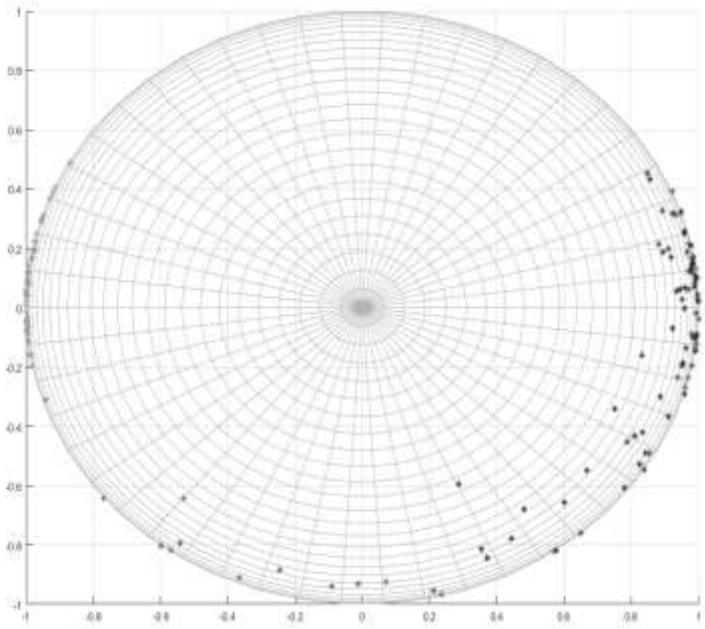


б)

Рис. 2. Візуалізація вибірки Random Matrix:
а) вид збоку; б) вид зверху



а)



б)

Рис. 3. Візуалізація вибірки Iris:
а) вид збоку; б) вид зверху

Таблиця 1

Індекс Цаліньського-Харабаша для выбірок Random Matrix та Iris

RANDOM MATRIX		
Метод	SOM^m	k-means
Індекс CH для 2 кластерів	650,119137400817	49,3904185658740
Індекс CH для 3 кластерів	782,603215072022	611,890289394461
Індекс CH для 4 кластерів	585,205208331037	411,869958970689
IRIS		
Метод	SOM^m	k-means
Індекс CH для 2 кластерів	506,384020879337	24,1478668157212
Індекс CH для 3 кластерів	521,993404839107	95,9506726585689
Індекс CH для 4 кластерів	463,871189144183	74,4873397342681

Висновки. У статті запропоновано online аналог методу X-середніх призначений для вирішення завдання кластерування потоку даних в умовах, коли число кластерів апріорно невідомо.

В основі запропонованого підходу лежить ансамбль паралельно включених кластеру вальних нейронних мереж Т. Кохонена, що містять різну кількість нейронів. Оптимальна кількість кластерів визначається числом нейронів у найкращій в сенсі якості з кластерувальних нейронних мереж.

Введена модифікація методу X-середніх призначена для вирішення завдань в рамках інтелектуального аналізу потоків даних (Data Stream Mining).

ЛІТЕРАТУРА

1. *Gan G. Data Clustering: Theory, Algorithms and Application / G. Gan, Ch. Ma, J.Wu. – Philadelphia: SIAM, 2007. – 455 p.*
2. *Xu R. Clustering, IEEE Press Series on Computational Intelligence / R. Xu, D.C. Wunsch. – Hoboken, NJ: John Wiley & Sons, Inc., 2009. – 370 p.*
3. *Pelleg D. X-means: extending K-means with efficient estimation of the number of clusters / D. Pelleg, A. Moor // Proc. 17th Int. Conf. on Machine Learning. – San Francisco: Morgan Kaufmann. – 2000. – P.727-730.*
4. *Ishioka T. An expansion of X-means for automatically determining the optimal number of clusters / T. Ishioka // Proc. 4th IASTED Int. Conf. Computational Intelligence. – Calgary, Alberta. – 2005. – P.91-96.*

5. *Bifet A. Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams / A. Bifet. – Amsterdam: IOS Press, 2010. – 224 p.*
6. *Kohonen, T. Self-OrganizingMaps / T.Kohonen. –Berlin: Springer-Verlag, 1995. –362 p.*
7. *Strehl A. Cluster Ensembles – A knowledge reuse framework for combining multiple partitions / A. Strehl, J. Ghosh // Journal of Machine Learning Research. – 2002. – № 2. – P.583-617.*
8. *Topchy A. Clustering ensembles: models of consensus and weak partitions / A. Topchy, A.K. Jain, W. Punch // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2005. – № 27. – P.1866-1881.*
9. *Alizadeh H. To improve the quality of cluster ensembles by selecting a subset of base clusters / H. Alizadeh, B.Minaei-Bidgoli, H.Parvin // Journal of Experimental & Theoretical Artificial Intelligence. – 2013. – № 26. – P.127-150.*
10. *Charkhabi M. Cluster ensembles, majority vote, voter eligibility and privileged voters / M. Charkhabi, T. Dhot, S.A. Mojarad // Int. Journal of Machine Learning and Computing. – 2014. – № 4. – P.275-278.*
11. *Bodyanskiy Ye. Computational intelligence techniques for data analysis / Ye. Bodyanskiy // Lecture Notes in Informatics. – Bonn: GI. – 2005. – P.15-36.*
12. *Бодянский Е.В. Искусственные нейронные сети: архитектуры, обучение, применения / Е.В. Бодянский, О.Г. Руденко. – Харьков: ТЕЛЕТЕХ, 2004. – 372 с.*
13. *Аналіз та обробка потоків даних засобами обчислювального інтелекту: Монографія / Є.В. Бодянський, Д.Д. Пелешко, О.А. Винокурова, С.В. Машталір, Ю.С. Іванов. – Львів: Вид-во Львів. політехніки, 2016. – 235 с.*
14. *Murphy P.M. UCI Repository of machine learning databases / P.M. Murphy, D. Aha. – URL: http://www.ics.uci.edu/mlearn/ MLRepository.html.CA:University of California, Department of Information and Computer Science, 1994.*

Стаття надійшла до редакції 29.09.2017