ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2018. No. 4 (6)

УДК 004.032.26 DOI: https://doi.org/10.30837/2522-9818.2018.6.042

П. Є. Жернова, Є. В. Бодянський

ЯДЕРНА НЕЧІТКА КЛАСТЕРИЗАЦІЯ ПОТОКІВ ДАНИХ НА ОСНОВІ АНСАМБЛЮ НЕЙРОННИХ МЕРЕЖ

Предметом дослідження в статті є кластерування даних на основі ансамблю нейронних мереж. **Мета** роботи – створення нового підходу в задачах кластерування потоків даних, коли інформація надходить в online режимі спостереження за спостереженням. У статті вирішуються наступні **завдання**: формування моделі ансамблів нейронних мереж для кластерування даних, розробка методів кластерування даних при роботі з даними великих розмірностей, розробка методів онлайн кластерування даних з використанням ансамблів нейронних мереж, які працюють у паралельному режимі. Отримані наступні **результати**: сформульовано принципи роботи ансамблів нейронних мереж Т. Кохонена, і визначені практичні вимоги для роботи з даними великої розмірності. Показані можливі підходи для вирішення перерахованих завдань. Досліджено принцип роботи ансамблю паралельно налаштованих кластерувальних мереж Т. Кохонена. Для навчання шарів ансамблю нейронних мереж використовувалися процедури, які працюють за принципом WTA і WTM. Були використані радіально-базисні функції для підвищення розмірностей вхідного простору. Розроблено математичну модель для вирішення задачі кластерування даних в online режимі. Розроблено математичну модель для визначення якості кластерування з використанням індексу Девіса-Булдена, який був переформульований для online режиму. **Висновки**: В роботі запропоновано новий підхід до завдання кластерування потоків даних, коли інформація надходить в online режимі спостереження за спостереженням за умов, що кількість і форма кластерів заздалегідь невідома. Основна ідея підходу базується на ансамблі нейронних мереж, який складається з самоорганізованих мап Кохонена. Всі члени ансамблю обробляють інформацію, яка послідовно подається в систему в паралельному режимі. Експериментальні результати підтвердили той факт, що система може бути використана для вирішення широкого кола завдань Data Stream Mining.

Ключові слова: кластерування; метод *X*-середніх; самоорганізована мапа Кохонена; ансамбль нейронних мереж; самонавчання.

Вступ

Проблема кластеризації масивів даних є складовою частиною комплексу завдань, що вирішуються в рамках Data Mining, а для її вирішення на сьогодні розроблено безліч підходів, методів та алгоритмів [1–3], що відрізняються один від одного як вихідними припущеннями, так і використовуваним математичним апаратом. Одним з найбільш популярних методів кластерування є метод К-середніх, завдяки наочності отриманих результатів, простоті використовуваного математичного апарату і чисельної реалізації. В рамках цього підходу передбачається, що вихідний масив даних $X = \{x(1),...,x(k),...,x(N)\}$, $x(k) = (x_1(k),...,x_i(k),...,x_n(k))^T \in R^n$, $k = 1,2,...N$, повинен бути розбитий на m неперетинних опуклих лінійно роздільних класів, при цьому число цих класів m задається апріорно, виходячи з тих чи інших, як правило, емпіричних припущень.

Альтернативою емпіричному підходу є досить формалізований метод *X*-середніх [4, 5], вельми громіздкий з обчислювальної точки зору і пов'язаний з досить суворими апріорними статистичними припущеннями про характер розподілу вихідних даних. Крім того, обидва ці методи вимагають багаторазового перебору даних у вихідному масиві *X*, що обмежує їх можливості в задачах обробки великих масивів інформації (Big Data) і потоків даних, коли інформація подається на вхід кластерувальної системи послідовно спостереження за спостереженням в online режимі (Data Stream Mining). У цій ситуації номер спостереження *k* набуває сенсу поточного дискретного часу, а обсяг даних *N* практично не обмежений.

У подібних ситуаціях дуже ефективними показали себе кластерувальні самонавчанні штучні нейронні мережі [6–9] і, перш за все, самоорганізовні мапи Т. Кохонена (*SOM*) [10], що обробляють дані в послідовному режимі. Результат роботи *SOM* збігається з результатами *K*-середніх, при цьому число кластерів *m* також задається апріорно.

Зберегти можливість online обробки за допомогою *SOM* і встановлення числа кластерів *m* за допомогою *X*-середніх можна, скориставшись ідеєю кластерувальних ансамблів [11–14], при цьому в якості елементів ансамблю використовувати кластерувальні нейронні мапи Кохонена SOM^m [15], кожна з яких розрахована на різне число можливих класів $m = 2,3,...,M$. В рамках такого підходу перший член ансамблю SOM^2 в шарі Кохонена містить всього два нейрони з векторами синаптичних ваг w_1^2, w_2^2, а останній SOM^m – M нейронів з вагами-центроїдами $w_1^M, w_2^M, ..., w_M^M$.

В процесі роботи ансамблю всі SOM^m функціонують паралельно, а в якості фінального результату обирається кластерувальна мережа-переможець, яка показала найкращий результат у сенсі застосовуваного критерію якості кластеризації [2, 16].

Відзначимо, що подібно до того, як в кожній з SOM^m на кожному кроці обробки інформації *k* обирається свій нейрон-переможець, так і в ансамблі на кожному кроці обирається нейронна мережа-переможець, яка забезпечує найкращий результат кластеризації.

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2018. № 4 (6) ISSN 2524-2296 (online)

Суттєвим обмеженням, що знижує можливості подібного підходу, є вимога лінійної роздільності і опуклості формованих кластерів, в той час як реальні дані можуть утворювати класи абсолютно довільної форми. У подібних ситуаціях дуже корисним може виявитися використання теореми Ковера (T. Cover) про лінійну роздільність в просторах ознак підвищеної розмірності [17] і ядр Мерсера (J. Mercer) [18], що забезпечують це підвищення. На основі такого підходу були розроблені, так звані, ядерні самоорганізовані мапи (*KSOM*) [19,20], які показали гарні результати в умовах класів досить довільної форми при відомій їх кількості m і в умовах фіксованого обсягу оброблюваної вибірки N.

У зв'язку з цим є доцільною розробка ансамблю ядерних кластерувальних нейронних мереж, призначеного для online обробки потоків даних в умовах невідомої або змінної кількості класів.

Архітектура ансамблю ядерних кластерувальних нейронних мереж

На рис. 1 наведена архітектура ансамблю ядерних кластерувальних нейронних мереж, яка містить п'ять шарів обробки інформації.

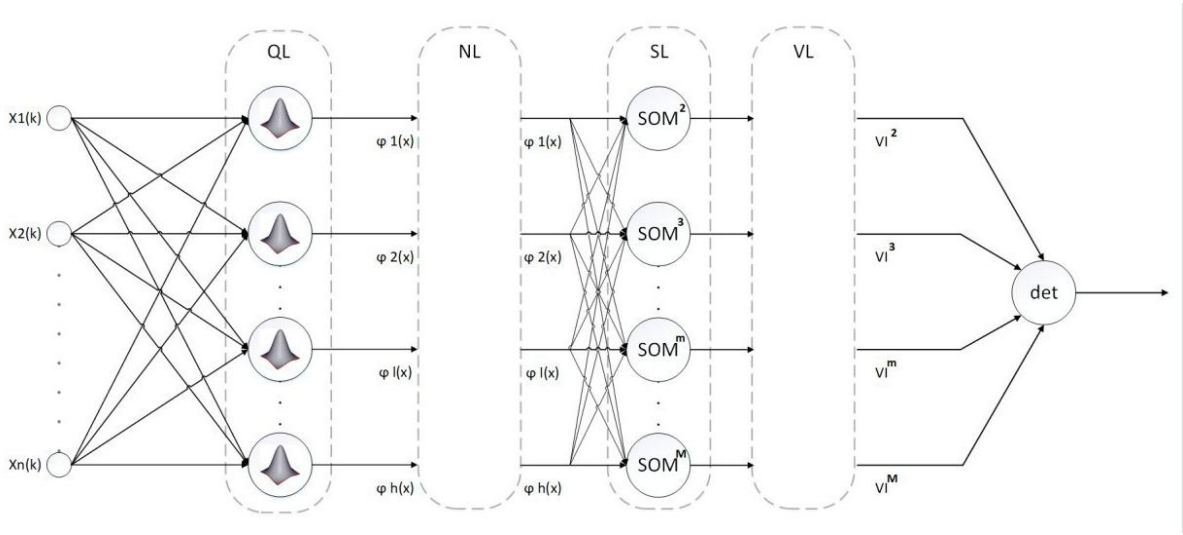


Рис. 1. Архітектура ансамблю ядерних кластерувальних нейронних мереж

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

$$\varphi_l(x) = \left(1 + \frac{\|x - c_l\|^2}{\gamma_\varphi}\right)^{-1} = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - c_l\|^2},$$

де $c_l - (n \times 1)$ – вектор, що задає "центр" радіально-базисної функції $\varphi_l(x)$, γ_φ – скалярний параметр, що визначає область рецепторного поля – "ширину" цієї функції.

Таким чином, при надходженні на вхід системи векторного сигналу $x(k) = (x_1(k), ..., x_n(k))^T \in R^n$, на виході першого прихованого шару *RL* формується векторний сигнал $\varphi(x(k)) - (\varphi_1(x(k)), ..., \varphi_l(x(k)), ..., \varphi_h(x(k)))^T \in R^h$, $h > n$.

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

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

І, нарешті, у вихідному шарі, що містить єдиний вузол - детектор оптимуму, визначається конкретна SOM^{m*}, що забезпечує найкращу якість кластеризації, при цьому вважається, що в аналізованому масиві даних міститься m^* кластерів.

Самонавчання ядерної кластерувальної системи на основі ансамблю нейронних мереж

Процес самонавчання даної системи реалізується на рівні першого шару RL, де налаштовуються центри c_l, $l = 1, 2, ..., h$ ядерних функцій $\varphi_l(x)$, і третього прихованого шару SL, де уточнюються синаптичні

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2018. No. 4 (6)

ваги w_j^m, $m = 2,3,...,M$, $j = 1,2...,m$ кожної нейронної мережі SOM^m ансамблю.

Розглянемо спочатку процес налаштування центрів ядерних функцій, що складається з послідовності наступних кроків [21]:

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

Крок 1: при подаванні на вхід системи першого вектора – спостереження $x(1)$ формується перший центр c_1 і перша радіально-базисна функція

$$\varphi_1(x) = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - c_1\|^2},$$

де $c_1 = x(1)$.

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

$$\|x(2) - c_1\| \le \Delta$$

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

$$\Delta < \|x(2) - c_1\| \le 2\Delta, \qquad (1)$$

то c_1 коригується відповідно до правила самонавчання Т. Кохонена "Переможець отримує все" [10]:

$$c_1(2) = c_1(1) + \eta(2)(x(2) - c_1(1)), \qquad (2)$$

де $c_1(1) = x(1)$, $0 < \eta(2) < 1$ – параметр кроку навчання.

Якщо ж виконується умова

$$2\Delta < \|x(2) - c_1\|,$$

то формується нова ядерна функція

$$\varphi_2(x) = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - c_2\|^2} = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - x(2)\|^2}.$$

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

Процес налаштування третього прихованого шару також складається з трьох етапів [10]: конкуренції, кооперації і синаптичної адаптації і реалізується для кожної SOM^m ансамблю, при цьому вектори синаптичних ваг w_j^m описують h-вимірні центроїди формованих кластерів.

На етапі конкуренції сигнал c виходу другого прихованого шару NL $\varphi(x(k)) \in R^h$ надходить на входи всіх SOM^m, де порівнюється з кожним з векторів синаптичних ваг $w_j^m(k-1)$ в сенсі відстані

$$D(\varphi(x(k)), w_j^m(k-1)) = \left\|\varphi(x(k) - w_j^m(k-1)\right\|, \qquad (3)$$

$j = 1,2,...m$; $m = 2,3,...M$. Оскільки $\left\|\varphi(x(k))\right\| = 1$, то замість евклідової метрики (3) набагато простіше використовувати косинусну міру подібності

$$sim(\tilde{\varphi}(x(k)), w_j^m(k-1)) = \tilde{\varphi}^T(x(k)) w_j^m(k-1),$$

за допомогою якої для кожної SOM^m визначається свій нейрон-переможець, для якого

$$\tilde{\varphi}^T(x(k)) w_j^{m*}(k-1) = \max_j \tilde{\varphi}^T(x(k)) w_j^m(k-1).$$

На етапі кооперації все M–1 нейронів-переможців ансамблю формують області топологічного сусідства, в яких налаштовуються не тільки ці переможці, а й їхні найближчі сусіди. Ця область описується функціями сусідства $\varphi(j,l)$, в якості яких можуть бути використані ядерні функції аналогічні радіально-базисним функціям першого прихованого шару:

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

На етапі синаптичної адаптації відбувається уточнення синаптичних ваг-центроїдів кожної з SOM^m за допомогою правила самонавчання Т. Кохонена "Переможець отримує більше":

$$w_l^m(k) = w_l^m(k-1) + \eta(k)\varphi(j,l)(\varphi(x(k)) - w_l^m(k-1). \qquad (4)$$

Нескладно бачити, що для переможця w_j^{m*} (4) збігається з правилом навчання (2). Необхідно зауважити, що в правилі самонавчання (4) параметри кроку $\eta(k)$ і γ обираються, як правило, виходячи з емпіричних міркувань і повинні монотонно зменшуватися в процесі налаштування.

Цей процес зручно організувати за допомогою системи співвідношень:

$$\begin{cases} \eta(k) = r^{-1}(k);\ r(k) = \alpha r(k-1) + \left\|\tilde{\varphi}(x(k))\right\|^2 = \alpha r(k-1) + 1, \\ \gamma(k) = \eta(k)\gamma(k-1),\ 0 < \alpha \le 1, \end{cases}$$

які при $\alpha = 1$ автоматично перетворюються в процедуру стохастичної апроксимації.

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2018. № 4 (6) ISSN 2524-2296 (online)

Нескладно помітити, що перший і третій шари системи фактично навчаються згідно однотипним процедурам типу WTA і WTM [10].

Налаштування четвертого прихованого шару

У четвертому прихованому шарі системи проводиться оцінка якості кластеризації за допомогою того чи іншого індексу валідації VI^m [1], при цьому цей індекс розраховується для кожної з мап Кохонена SOM^m, $m = 2,3,...M$. В якості такого індексу зручно використовувати критерій Девіса-Булдена (Davies DL, Bouldin DW) [22], за допомогою якого можна оцінювати якість кластеризації навіть у разі несферичних класів. Для випадку m кластерів цей індекс може бути записаний у вигляді

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

де $D(w_j^m(k), w_q^m(k))$ – відстань між центроїдами: $D(w_j^m(k), w_q^m(k)) = \left\| w_j^m(k) - w_q^m(k) \right\|$, $s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k)))$ – характеристики внутрішньокластерного розсіювання для j-го кластеру:

$$s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k))) = \left(\frac{\sum_{k=1}^{N} u_j(k) \left\| \tilde{\varphi}(x(k)) - w_j^m(k) \right\|^2}{\sum_{k=1}^{N} u_j(k)} \right)^{\frac{1}{2}},$$

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

$$u_j(k) = \begin{cases} 1, \text{ якщо } \tilde{\varphi}(x(k)) \text{ віднесен до } j\text{-го кластеру}, \\ 0 \text{ в іншому випадку.} \end{cases}$$

В якості оптимальної кількості кластерів m^* обирається значення, що забезпечує мінімум DB (m), тобто $DB(m^*) = \min_m \{DB(2), DB(3), ..., DB(M)\}$,

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

Результати моделювання

Ми випробували запропонований метод із двома різними навчальними наборами даних. Перший набір даних штучно створено так, що він містить 3 кластери, 300 спостережень, кожне спостереження має 3 функції. Другий набір даних "Ірис" взято з UCI-репозиторію [24]. Цей набір даних складається з 150 спостережень, які поділяються на 3 класи, де кожне спостереження має 3 випадкові функції. Кластери чітко видно в штучному створеному наборі даних і показані на рисунку 2.

Обчислювальна точність запропонованого методу була порівняна з відомим алгоритмом K-means та наведена у таблиці 1.

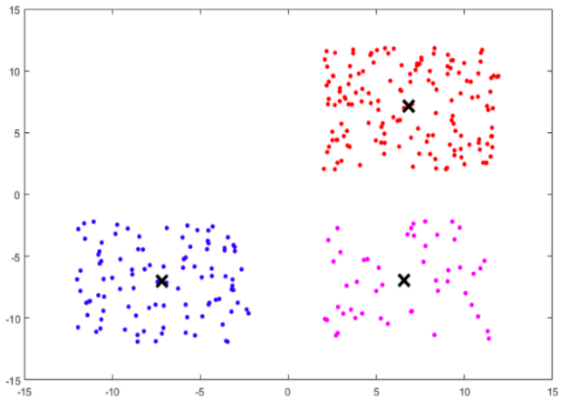


Рис. 2. Штучно згенерований набір даних

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

$$s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k)), s) = \left(\frac{\sum_{\tau=k-s+1}^{k} u_j(\tau) \left\| \tilde{\varphi}(x(\tau)) - w_j^m(k) \right\|^2}{\sum_{\tau=k-s+1}^{k} u_j(\tau)} \right)^{\frac{1}{2}};$$

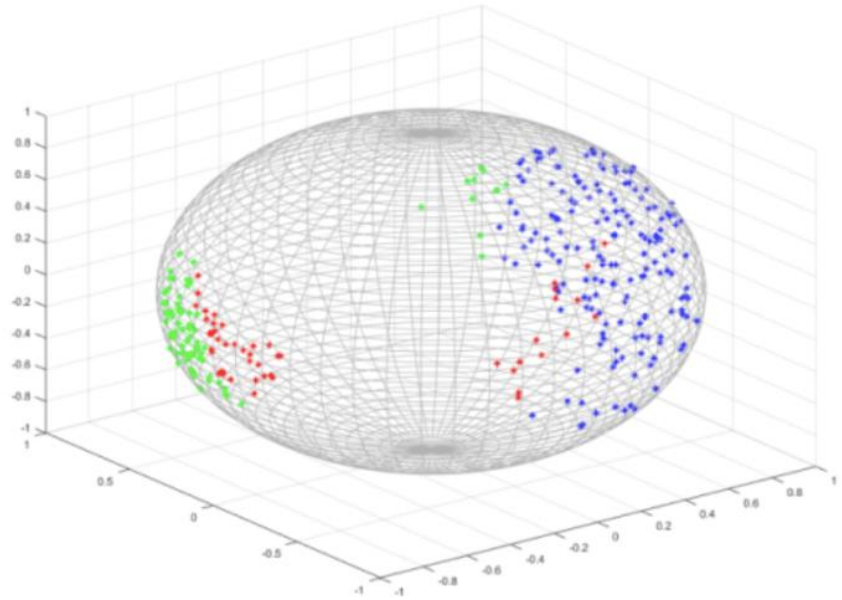
який розраховується у вихідному шарі.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2018. No. 4 (6)

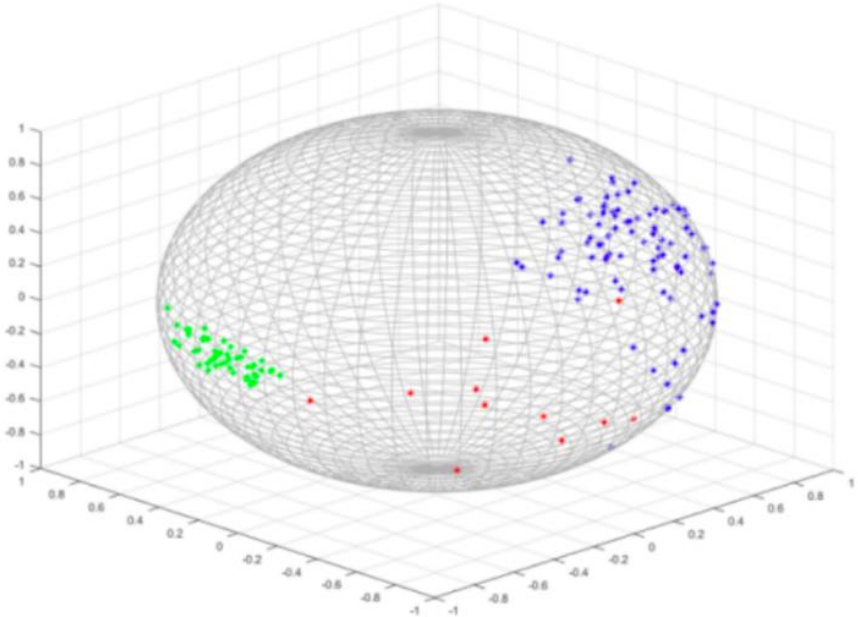
Таблиця 1. *Результати кластерування для різної кількості кластерів*

	SOM^m	***k-means***
Штучно згенерований набір даних		
точність кластерізації для 2-х кластерів	0,71	0,70
точність кластерізації для 3-х кластерів	**0,89**	0,76
точність кластерізації для 4-х кластерів	0,68	0,67
"Іриси Фішера"		
точність кластерізації для 2-х кластерів	0,84	0,83
точність кластерізації для 3-х кластерів	**0,91**	0,87
точність кластерізації для 4-х кластерів	0,72	0,73

Для візуалізації взяті набори даних проектували за допомогою методу PCA (аналіз основних компонент) на три основні компоненти. Результати роботи ансамблю нечіткої кластерування потоків даних наведено на рисунку 3.



а) Штучно згенерований набір даних



б) "Іриси Фішера"

Рис. 3. Результати візуалізації запропонованого ансамблю

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2018. № 4 (6) *ISSN 2524-2296 (online)*

Висновки

У статті пропонується нейромережевий підхід до задачі кластеризації потоків даних, які в online режимі надходять на обробку, в припущенні, що заздалегідь невідомі кількість кластерів та їх форма. В основі підходу лежить ідея ядерної кластеризації та ансамблю нейронних мереж, елементами якого є самоорганізовні мапи Т. Кохонена. Запропонована система характеризується простотою чисельної реалізації, високою швидкодією і може бути використана для вирішення широкого класу задач обробки потоків даних в умовах апріорної невизначеності про їх властивості.

Список літератури

1. Gan G., Ma Ch., Wu J. Data Clustering: Theory, Algorithms and Applications. Philadelphia : SIAM, 2007.
2. Xu R., Wunsch D. C. Clustering. Hoboken, NJ : John Wiley & Sons, Inc., 2009. IEEE Press Series on Computational Intelligence,
3. Aggarwal C. C., Reddy C. K. Data Clustering. Algorithms and Application. Boca Raton : CRC Press, 2014.
4. Pelleg D., Moor A. X-means: extending K-means with efficient estimation of the number of clusters. In: Proc. 17th Int. Conf. on Machine Learning. San Francisco : Morgan Kaufmann, 2000. P. 727–730.
5. Ishioka T. An expansion of X-means for automatically determining the optimal number of clusters. In: Proc. 4th IASTED Int. Conf. Computational Intelligence. Calgary, Canada : Proceedings of International Conference on Computational Intelligence, 2005. P. 91–96.
6. Rutkowski L. Computational Intelligence. Methods and Techniques. Berlin-Heidelberg : Springer-Verlag, 2008.
7. Mumford C., Jain L. Computational Intelligence. Collaboration, Fuzzy and Emergence. Berlin : Springer-Vergal, 2009.
8. Kruse R., Borgelt C., Klawonn F., Moewes C., Steinbrecher M., Held P. Computational Intelligence. A Methodological Introduction. Berlin : Springer, 2013.
9. Du K. L., Swamy M. N. S. Neural Networks and Statistical Learning. London : Springer-Verlag, 2014.
10. Kohonen T. Self-Organizing Map. Berlin : Springer-Verlag, 1995.
11. Strehl A., Ghosh J. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*. 2002. P. 583–617.
12. Topchy A., Jain A. K., Punch W. Clustering ensembles: models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005.No. 27. P. 1866–1881.
13. Alizadeh H., Minaei-Bidgoli B., Parvin H. To improve the quality of cluster ensembles by selecting a subset of base clusters. *Journal of Experimental & Theoretical Artificial Intelligence*. 2013. No. 26. P. 127–150.
14. Charkhabi M., Dhot T., Mojarad S. A. Cluster ensembles, majority vote, voter eligibility and privileged voters. *Int. Journal of Machine Learning and Computing*. 2014. No. 4. P. 275–278.
15. Bodyanskiy Ye. V., Deineko A. A., Zhernova P. Ye., Riepin V. O. Adaptive modification of X-means method based on the ensemble of the T. Kohonen's clustering neural networks. Odessa : Materials of the VI Int. Sci. Conf. "Information Managements Systems and Technologies", 2017. P. 202–204.
16. Bezdek J. C., Keller J., Krishnapuram R., Pal N. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. The Handbook of Fuzzy Sets. Dordrecht, Netherlands : Springer, 1999. Vol. 4.
17. Cover T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. on Electronic Computers. 1965. No. 14. P. 326–334.
18. Girolami M. Mercer kernel-based clustering in feature space. IEEE Trans. on Neural Networks. 2002. No. 3. Vol. 13. P. 780–784.
19. MacDonald D., Fyfe C. Clustering in data space and feature space. Belgium : ESANN'2002 Proc. European Symp. on Artificial Neural Networks. Bruges (24-26 April 2002), 2002. P. 137–142.
20. Camastra F., Verri A. A novel kernel method for clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence. 2005. No. 5. P. 801–805.
21. Bodyanskiy Ye. V., Deineko A. A., Kutsenko Y. V. On-line kernel clustering based on the general regression neural network and T. Kohonen's self-organizing map. Automatic Control and Computer Sciences. 2017. No. 51. P. 55–62.
22. Davies D. L., Bouldin D. W. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979. No. 4. P. 224–227.
23. Murphy P. M., Aha D. UCI Repository of machine learning databases. Department of Information and Computer Science, CA : University of California, 1994. URL : http://www.ics.uci.edu/mlearn/MLRepository.html.

References

1. Gan, G., Ma, Ch., Wu, J. (2007), *Data Clustering: Theory, Algorithms and Applications*, Philadelphia : SIAM.
2. Xu, R., Wunsch, D. C. (2009), *Clustering*, Hoboken, NJ : John Wiley & Sons, Inc., IEEE Press Series on Computational Intelligence.
3. Aggarwal, C. C., Reddy, C. K. (2014), *Data Clustering*, Algorithms and Application, Boca Raton : CRC Press.
4. Pelleg, D., Moor, A. (2000), "X-means: extending K-means with efficient estimation of the number of clusters", *In: Proc. 17th Int. Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, P.727–730.
5. Ishioka, T. (2005), "An expansion of X-means for automatically determining the optimal number of clusters", *In: Proc. 4th IASTED Int. Conf. Computational Intelligence*, Calgary, Alberta, P. 91–96.
6. Rutkowski, L. (2008), *Computational Intelligence. Methods and Techniques*, Berlin-Heidelberg: Springer-Verlag.
7. Mumford, C. and Jain, L. (2009), *Computational Intelligence. Collaboration, Fuzzy and Emergence*, Berlin : Springer-Vergal.
8. Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M. and Held, P. (2013), *Computational Intelligence. A Methodological Introduction*, Berlin : Springer.
9. Du, K. L. and Swamy, M. N. S. (2014), *Neural Networks and Statistical Learning*, London : Springer-Verlag.
10. Kohonen, T. (1995), *Self-Organizing Maps*, Berlin : Springer-Verlag.
11. Strehl, A., Ghosh, J. (2002), "Cluster ensembles – A knowledge reuse framework for combining multiple partitions", *Journal of Machine Learning Research*, P. 583–617.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2018. No. 4 (6)

12. Topchy, A., Jain, A. K., Punch, W. (2005), "Clustering ensembles: models of consensus and weak partitions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No. 27, P. 1866–1881.
13. Alizadeh, H., Minaei-Bidgoli, B., Parvin, H. (2013), "To improve the quality of cluster ensembles by selecting a subset of base clusters", *Journal of Experimental & Theoretical Artificial Intelligence*, No. 26, P. 127–150.
14. Charkhabi M., Dhot T., Mojarad S. A. (2014), "Cluster ensembles, majority vote, voter eligibility and privileged voters", *Int. Journal of Machine Learning and Computing*, No. 4, P. 275–278.
15. Bodyanskiy, Ye. V., Deineko, A. A., Zhernova, P. Ye., Riepin, V. O. (2017), "Adaptive modification of X-means method based on the ensemble of the T. Kohonen's clustering neural networks", *Materials of the VI Int. Sci. Conf. "Information Managements Systems and Technologies"*, Odessa, P. 202–204.
16. Bezdek, J. C., Keller, J., Krishnapuram, R., Pal N. (1999), *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, The Handbook of Fuzzy Sets, Kluwer, Dordrecht, Netherlands : Springer, Vol. 4.
17. Cover, T. M. (1956), "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition", *IEEE Trans. on Electronic Computers*, No. 14, P. 326–334.
18. Girolami, M. (2002), "Mercer kernel-based clustering in feature space", *IEEE Trans. on Neural Networks*, Vol. 13, No. 3, P. 780–784.
19. MacDonald, D., Fyfe, C. (2002), "Clustering in data space and feature space", *ESANN'2002 Proc. European Symp. on Artificial Neural Networks*, Bruges (Belgium), P. 137–142.
20. Camastra, F., Verri, A. (2005), "A novel kernel method for clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, No. 5, P. 801–805.
21. Bodyanskiy, Ye. V., Deineko, A. A., Kutsenko, Y. V., "On-line kernel clustering based on the general regression neural network and T. Kohonen's self-organizing map", *Automatic Control and Computer Sciences*, No. 51 (1), P. 55–62.
22. Davies, D. L., Bouldin, D. W. (1979), "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No. 4, P. 224–227.
23. Murphy, P. M., Aha, D. (1994), *UCI Repository of machine learning databases*, available at : http://www.ics.uci.edu/mlearn/MLRepository.html, Department of Information and Computer Science, CA : University of California.

Надійшла (Received) 26.11.2018

Відомості про авторів / Сведения об авторах / About the Authors

Жернова Поліна Євгеніївна – Харківський національний університет радіоелектроніки, асистент кафедри системотехніки, Харків, Україна, e-mail: polina.zhernova@gmail.com; ORCID ID: https://orcid.org/0000-0002-2154-4766.

Жернова Полина Евгеньевна – Харьковский национальный университет радиоэлектроники, ассистент кафедры системотехники, Харьков, Украина.

Zhernova Polina – Kharkiv National University of Radio Electronics, Assistant Lecturer at the Department of System Engineering, Kharkiv, Ukraine.

Бодянський Євгеній Володимирович – доктор технічних наук, професор, Харківський національний університет радіоелектроніки, професор кафедри штучного інтелекту, науковий керівник ПНДЛ АСУ, Харків, Україна, e-mail: yevgeniy.bodyanskiy@nure.ua; ORCID ID: https://orcid.org/0000-0001-5418-2143.

Бодянский Евгений Владимирович – доктор технических наук, профессор, Харьковский национальный университет радиоэлектроники, профессор кафедры искусственного интеллекта, научный руководитель ПНДЛ АСУ, Харьков, Украина.

Bodyanskiy Yevgeniy – Doctor of Sciences (Engineering), Professor, Kharkiv National University of Radio Electronics**,** Professor at the Department of Artificial Intelligence, Scientific Head at the CSRL**,** Kharkiv, Ukraine.

ЯДЕРНАЯ НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ ПОТОКОВ ДАННЫХ НА ОСНОВЕ АНСАМБЛЯ НЕЙРОННЫХ СЕТЕЙ

Предметом исследования в статье является кластеризация данных на основе ансамбля нейронных сетей. **Цель** работы – создание нового подхода в задачах кластеризации потоков данных, когда информация поступает в online режиме наблюдения за наблюдением. В статье решаются следующие **задачи**: формирование модели ансамблей нейронных сетей для кластеризации данных, разработка методов кластеризации данных при работе с данными больших размерностей, разработка методов онлайн кластеризации данных с использованием ансамблей нейронных сетей, работающих в параллельном режиме. Получены следующие **результаты:** сформулированы принципы работы ансамблей нейронных сетей Т. Кохонена, и определены практические требования для работы с данными большой размерности. Показаны возможные подходы для решения перечисленных задач. Исследован принцип работы ансамбля параллельно настроенных кластеризующих сетей Т. Кохонена. Для обучения слоев ансамбля нейронных сетей использовались процедуры, работающие по принципу WTA и WTM. Были использованы радиально-базисные функции для повышения размерностей входного пространства. Разработана математическая модель для решения задачи кластеризации данных в online режиме. Разработана математическая модель для определения качества кластеризации с использованием индекса Дэвиса-Булдена, который был переформулирован для online режима. **Выводы:** В работе предложен новый подход к задаче кластеризации потоков данных, когда информация поступает в online режиме наблюдение за наблюдением при условии, что количество и форма кластеров заранее неизвестны. Основная идея этого подхода базируется на ансамбле нейронных сетей, который состоит из самоорганизующихся карт Кохонена. Все члены ансамбля обрабатывают информацию, которая последовательно подается в систему в параллельном режиме. Экспериментальные результаты

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2018. № 4 (6) ISSN 2524-2296 (online)

подтвердили тот факт, что рассматриваемая система может быть использована для решения широкого круга задач Data Stream Mining.

Ключевые слова: кластеризация; метод X-средних; самоорганизующаяся карта Кохонена; ансамбль нейронных сетей; самообучение.

KERNEL FUZZY CLUSTERING OF DATA streams BASED ON THE ENSEMBLE OF NEURAL NETWORKS

The **subject matter** of the study is data clustering based on the ensemble of neural networks. The **goal** of the work is to create a new approach to solving the tasks of clustering in data streams when information is fed observation-by-observation in online mode. The following **tasks** were solved in the article: the model of neural network ensembles for data clustering was created, the methods of data clustering to process mass data were developed, the methods of online data clustering of data using neural network ensembles working in the parallel mode were developed. The following **results** were obtained: the operation principles of the ensembles of the Kohonen neural network were formulated and practical requirements for dealing with mass data were specified. The probable approaches to solving these problems were indicated. The operation principle of the ensemble of parallel tuned Kohonen clustering networks was studied. The procedures based on the WTA and WTM principles were used to train layers of the neural network ensemble. Radial basis functions were used to increase the dimension of the input space. The mathematical model was developed for solving the problem of data clustering in online mode. The mathematical model was developed to determine the quality of clustering using the Davies-Bouldin index, which was rewritten for online mode. **Conclusions**. The paper proposes a new approach to solving the problem of clustering data streams when information is fed observation-by-observation in online mode, provided that the number and shape of clusters are unknown in advance. The main idea of this approach is based on the ensemble of neural networks, which consists of Kohonen self-organizing maps. All members of the ensemble process information that is sequentially fed into the system in parallel mode. Experimental results confirmed the fact that the considered system can be used to solve a wide range of Data Stream Mining tasks.

Keywords: clustering; X-means method; Kohonen self-organizing map; neural network ensemble; self-learning.