УДК 004.7.052:004.414.2 **DOI:** 10.31673/2412-9070.2020.031316

О. М. ТКАЧЕНКО, доктор техн. наук, доцент;
Н. В. РУДЕНКО, здобувач;
С. Р. КУФТЕРІНА, ст. викладач;
А. В. ЛЕМЕШКО, ст. викладач;
А. Г. ЗАХАРЖЕВСЬКИЙ, здобувач,
Державний університет телекомунікацій, Київ

АЛГОРИТМ ВИЗНАЧЕННЯ ОПТИМАЛЬНОЇ КІЛЬКОСТІ КЛАСТЕРІВ НА БАЗІ НЕЙРОННОЇ МЕРЕЖІ КОХОНЕНА

Розглянуто можливості використання систем штучного інтелекту для розв'язання задач кластеризації. Визначено значення критерію оптимальності для різних сполучень кількості кластерів і кількості нейронів вихідного шару мережі. Сформульовано загальну послідовність дій для обчислення оптимальної кількості нейронів вихідного шару мережі Кохонена.

Ключові слова: дані; аналіз; кластер; нейрон; мережа; множина; критерій; оптимальний вектор; навчання; інтелектуальний.

Вступ

Інтелектуальний аналіз даних — це процес виявлення раніше невідомих, нетривіальних, практично корисних і доступних інтерпретацій знань, необхідних для ухвалення рішень у різних сферах людської діяльності. Методи інтелектуального аналізу даних (або, що те саме, *knowledge discovery in data* — виявлення знань у базах, або *data mining*) перебувають на стику баз даних, статистики та штучного інтелекту.

У рамках інтелектуального аналізу одним із вирішуваних завдань є кластеризація — поділ множини вхідних векторів на групи (кластери) за ступенем «схожості» один на одного.

Основна частина

Обов'язковою вимогою для використання нейронних мереж, зокрема мережі Кохонена, є точне задання кількості нейронів у вхідному і вихідному шарах. Оскільки кожний нейрон вихідного шару відповідає за належність зразка до певного кластера, то кількість кластерів завжди дорівнює кількості вихідних нейронів. Ця властивість вступає в протиріччя з тим, що мережа Кохонена навчається без учителя. Для того щоб повністю вивести користувача з процесу кластеризації, необхідна можливість автоматичного визначення кількості кластерів, оптимальної для заданої множини вхідних векторів [3].

Належність об'єкта до певного класу визначається за максимальним значенням виходу нейронів останнього шару мережі. Чим вище значення виходу, тим «впевненіша» мережа в належності об'єкта, поданого на вхід, до відповідного класу.

Стосовно мережі Кохонена, значення максимального виходу нейрона тим більше, чим ближче до ядра кластера перебуває точка, координати якої складають вхідний вектор. Результат навчання нейронної мережі даного виду на одній і тій самій навчальній множині для різних значень кількості нейронів вихідного шару зображено на рис. 1. Для кожної точки вказано значення максимального виходу нейронів вихідного шару.

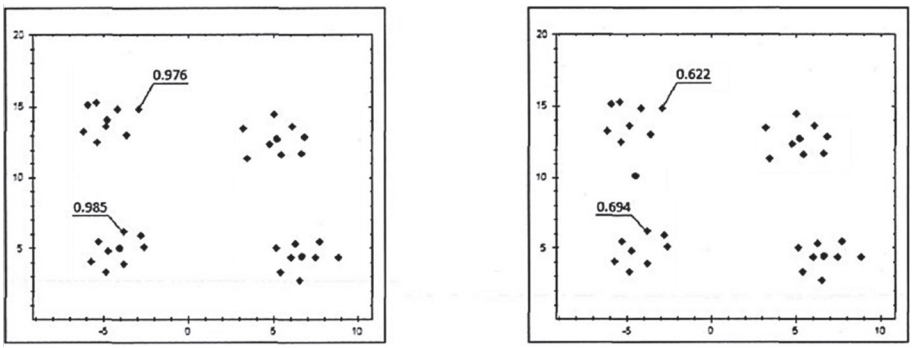

Рис. 1. Розміщення ядер кластерів після навчання нейронної мережі Кохонена для різних значень кількості нейронів вихідного шару

Навчальна множина складається з груп двокомпонентних векторів (що досить чітко вирізняються), поданих точками на площині. У результаті навчання за відповідності кількості виходів нейронної мережі реальній кількості кластерів ядра встановлюються точно в центр кластера, у разі нестачі виходів ядро встановлюється в центр між двома кластерами. При цьому збільшується відстань від нього до точок, що належать найближчим кластерам, а отже, для кожної із зазначених точок зменшується значення, здобуте на виході нейронної мережі [15].

Скориставшись даною властивістю нейронних мереж, можна визначити наскільки конфігурація мережі відповідає завданню, в якій вона застосовується. Оптимальна кількість нейронів при цьому буде такою, що для кожного кластера середнє значення максимального виходу нейрона вихідного шару, який відповідає за належність до даного кластера, буде максимальне. Критерієм оптимальності вважатимемо середнє значення максимального виходу нейрона вихідного шару.

Формулу розрахунку середніх значень можна подати у вигляді

$$\text{mid} = \frac{\sum_{i=0}^{k_1} \text{Out}_{1i} + \sum_{i=0}^{k_2} \text{Out}_{2i} + \ldots + \sum_{i=0}^{k_n} \text{Out}_{ni}}{k_1 + k_2 + \ldots + k_n}, \quad (1)$$

де k_1— кількість зразків, що належать до 1-го кластера; k_n — кількість зразків, що належать до n-го кластера; Out_{ij} — максимальне значення виходу нейронів останнього шару для j-го зразка, що належить до i-го кластера.

Можливий і другий варіант, де розраховується середня різниця між максимальним і мінімальним виходами нейронів останнього шару:

$$\text{mid} = \frac{\sum_{i=0}^{k_1} (\max_{1i} - \min_{1i}) + \sum_{i=0}^{k_2} (\max_{2i} - \min_{2i}) + \ldots + \sum_{i=0}^{k_n} (\max_{ni} - \min_{ni})}{k_1 + k_2 + \ldots + k_n}, \quad (2)$$

де k_1— кількість зразків, що належать до 1-го кластера; k_n — кількість зразків, що належать до n-го кластера; $\max_{ij}$ — максимальне значення виходу нейронів останнього шару для j-го зразка, що належить до i-го кластера; $\min_{ij}$ — мінімальне значення виходу нейронів останнього шару для j-го зразка, що належить до i-го кластера.

Оскільки однією з умов для «впевненої» належності до певного класу є наявність досить малих значень виходів інших нейронів останнього шару, то можна скористатися такою формулою:

$$\text{mid} = \frac{\sum_{i=0}^{k_1} \left(\max_{1i} - \frac{\sum_m \text{Out}_{m1}}{n_c - 1} \right) + \sum_{i=0}^{k_2} \left(\max_{2i} - \frac{\sum_m \text{Out}_{m2}}{n_c - 1} \right) + \ldots + \sum_{i=0}^{k_n} \left(\max_{ni} - \frac{\sum_m \text{Out}_{mj}}{n_c - 1} \right)}{k_1 + k_2 + \ldots + k_n}, \quad (3)$$

де k_1— кількість зразків, що належать до 1-го кластера; k_n — кількість зразків, що належать до n-го кластера; n_c — кількість нейронів вихідного шару (кількість кластерів); $\max_{ij}$ — максимальне значення виходу нейронів останнього шару для j-го зразка, що належить до i-го кластера; Out_{mj} — значення виходу m-го нейрона ($m \neq i$).

Результати кластеризації мережею Кохонена з різною кількістю вихідних нейронів

Кількість реальних кластерів	Кількість виходів мережі Когонена	Результат за першою формулою	Результат за другою формулою	Результат за третьою формулою
2	1	0,64	0,55	0,61
	2	0,98	0,76	0,82
	3	0,98	0,76	0,77
	4	0,98	0,76	0,68
3	1	0,26	0,37	0,54
	2	0,73	0,59	0,63
	3	0,98	0,81	0,85
	4	0,98	0,81	0,78
4	2	0,42	0,30	0,52
	3	0,71	0,48	0,75
	4	0,97	0,88	0,89
	5	0,97	0,88	0,80

Швидкість роботи сучасних комп'ютерів і швидкість роботи нейронної мережі Кохонена порівняно з іншими типами нейронних мереж дає можливість здійснювати велику кількість навчань мережі за короткий час, отже, можна скористатися одним із чисельних методів визначення максимального значення функції.

Результати експериментальних досліджень щодо визначення критерію оптимальності за наведеними раніше формулами для різних сполучень кількості кластерів і кількості нейронів вихідного шару мережі подано в таблиці на с. 9.

Оскільки після досягнення оптимальної кількості кластерів наступні нейрони, що додаються до мережі, не будуть активуватися, то вони не впливають на результат обчислення середнього значення. Таким чином, графіки залежності середнього максимального значення від кількості нейронів у вихідному шарі мають вигляд, як це зображено на рис. 2.

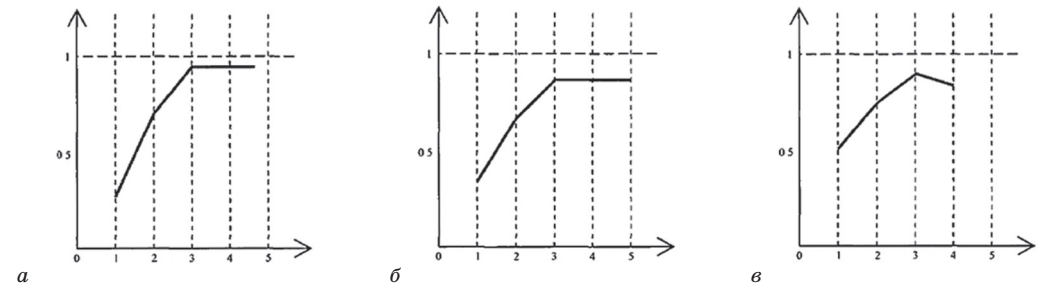

Рис. 2. Графіки залежності різних критеріїв оптимальності від кількості нейронів у вихідному шарі мережі

Для використання як критерій оптимальності в розглянутій задачі необхідно вибрати найбільш помітні значення вихідних сигналів. Із таблиці випливає, що такі результати дістаємо у разі використання формули (3). На графіку (див. рис. 2, *в*) також можна побачити наявність максимуму в здобутій функції.

Ґрунтуючись на викладеному, сформулюємо загальну послідовність дій для обчислення оптимальної кількості нейронів вихідного шару мережі Кохонена:

1. Підготовка множини вхідних векторів.

2. Навчання нейронної мережі Кохонена з використанням вхідної множини.

3. Кластеризація вхідної множини навченої мережі.

4. Підрахунок критерію оптимальності за формулою (3).

5. Визначення чисельним методом кількості нейронів вихідного шару, при якому значення критерію буде максимальне.

Для перевірки описаної методики доцільно скористатися програмною реалізацією нейронної мережі Кохонена в програмі Mathcad, що дає можливість задавати кількість вхідних і вихідних нейронів.

Відповідно до методики на початковому етапі сформовано множину вхідних векторів, кожний з яких має три значення. Графічне зображення даної множини ілюструє рис. 3.

За початкову кількість нейронів вихідного шару було взято три і за формулою (3) для нього розраховане значення середньої різниці між максимальним виходом нейрона і середнім значенням виходів інших нейронів. У результаті здобуто значення 0,674.

Далі кількість нейронів було збільшено і стало дорівнювати чотирьом. Для даної кількості так само було обчислено значення критерію оптимальності, яке становило 0,968.

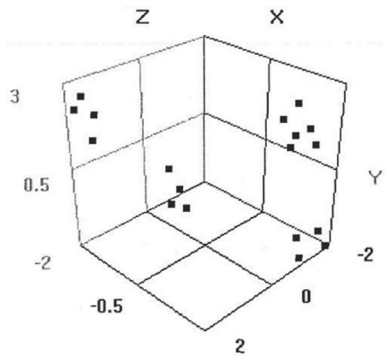

Рис. 3. Множина вхідних векторів із чотирма виокремленими кластерами

На наступному кроці кількість нейронів вихідного шару було збільшено до п'яти, і для даного значення так само було розраховано критерій оптимальності, який становив 0,76.

Отже, максимальне значення середньої різниці між максимальним виходом нейрона і середнім значенням виходів інших нейронів досягається у разі кількості нейронів вихідного шару, що дорівнює чотирьом. Саме така кількість є оптимальною для даної вхідної множини (див. рис. 3).

Висновки

Розглянутий метод дає можливість визначати кількість кластерів, оптимальну для заданої множини і для вихідних значень, заданих векторами більшої розмірності.

Запропонована методика є подальшим розвитком методів навчання без учителя. Вона дає змогу уникнути необхідності задання кількості виходів нейронної мережі Кохонена й може знайти широке застосування як у розв'язанні задач інтелектуального аналізу даних, так і в процесі розпізнавання нових невідомих класів і ситуацій у різних сферах діяльності.

Список використаної літератури

*1. **Hastie T., Tibshirani R., Friedman J.** Chapter 14.4 Self-Organizing Maps // The Elements of Statistical Learning. 2009. P. 528–534.*

*2. **Ткаченко О. М., Голубенко О. І.** Завдання систем штучного інтелекту в смарт-містах // Сучасні інфокомунікаційні технології: зб. тез наук.-техн. конф. Київ, 2019. С. 239.*

*3. **Ткаченко О. М., Підмогильний О. О.** Інтервальні нейронні мережі як детектори нестабільності для реконструкції астрономічних зображень екзопланет // XI Міжнар. наук.-техн. конф. «Інформаційно-комп'ютерні технології – 2020 (ІКТ-2020)», Житомир, 09-11 квітня 2020. Житомир: Житомирська політехніка, 2020. С. 80–81.*

О. М. Ткаченко, Н. В. Руденко, С. Р. Куфтерина, А. В. Лемешко, А. Г. Захаржевский

АЛГОРИТМ ОПРЕДЕЛЕНИЯ ОПТИМАЛЬНОГО КОЛИЧЕСТВА КЛАСТЕРОВ НА БАЗЕ НЕЙРОННОЙ СЕТИ КОХОНЕНА

Рассмотрены возможности использования систем искусственного интеллекта для решения задач кластеризации. Определено значение критерия оптимальности для различных сочетаний числа кластеров и количества нейронов выходного слоя сети. Сформулирована общая последовательность действий для вычисления оптимального количества нейронов выходного слоя сети Кохонена.

Ключевые слова: данные; анализ; кластер; нейрон; сеть; множество; критерий; оптимальный вектор; обучение; интеллектуальный.

O. Tkachenko, N. Rudenko, S. Kufterina A. Lemeshko, A. Zakharzhevskyi

ALGORITHM FOR DETERMINING THE OPTIMUM NUMBER OF CLUSTERS ON THE BASIS OF THE KOHONEN NEURAL NETWORK

The article discusses the possibilities of using artificial intelligence systems to solve clustering problems. The value of the optimality criterion for various combinations of the number of clusters and the number of neurons of the output network layer is determined. Self-organizing maps (SOM, Self Organizing Maps), developed by T. Kohonen and representing a powerful tool combining two important paradigms of data analysis - clustering and projecting, visualization of multidimensional data on a plane are considered. An example of the location of cluster nuclei after training the Kohonen neural network for different values of the number of neurons in the source layer is given. Comparing the speed of modern computers with the speed of the Kohonen neural network, with other types of neural networks, allows you to conduct a large number of network exercises in a short time, so you can use one of many methods to determine the maximum value of the function. The results of experimental studies to determine the criterion of optimality are presented in the article for different combinations of the number of clusters and the number of neurons in the original layer of the network. According to the method at the initial stage, a set of input vectors is formed, each of which includes three values. A general sequence of actions is formulated to calculate the optimal number of neurons in the output layer of the Kohonen network. The methodology presented in the article is a further development of teaching methods without a teacher. The technique proposed in the article avoids the need to specify the number of outputs of the Kohonen neural network and can be widely used both in solving data mining problems and in recognizing new unknown classes and situations in different fields.

Keywords: data; analysis; cluster; neuron; network; set; criterion; optimal vector; learning; intellectual; methodology; Kohonen, vectors.