УДК 681.3

К. П. СТОРЧАК, доктор техн. наук, доцент;
А. М. ТУШИЧ,
А. П. БОНДАРЧУК, канд. техн. наук, доцент,
Державний університет телекомунікацій, Київ

КЛАСТЕРНИЙ АНАЛІЗ ДАНИХ ІЗ ВИКОРИСТАННЯМ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

Запропоновано процес кластерного аналізу даних, що ґрунтується на об'єднанні об'єктів зі спільними властивостями у групи. Розглянуто та проаналізовано внесок науковців у вивчення кластерного аналізу даних із використанням штучних нейронних мереж та їх досягнення за останні роки, доцільність використання штучних нейронних мереж для процесу кластерного аналізу даних. Описано з використанням математичного апарату модель системи кластерного аналізу даних із використанням штучних нейронних мереж.

Ключові слова: кластер; кластеризація; кластерний аналіз даних; алгоритм; математична модель; штучна нейронна мережа; big data.

ВСТУП

Сьогодні набув широкого розповсюдження термін «big data», окресливши нову досить затребувану прикладну область — пошук способів автоматичного швидкого аналізу великих обсягів різноманітної інформації. Області застосування охоплюють доволі великий спектр послуг (від медицини до приватного бізнесу), оскільки «big data» — це результати наукових експериментів, метеорологічні спостереження, лог-файли банківських трансакцій, профілі користувачів у соціальних мережах тощо.

Одна з найвідоміших парадигм машинного навчання — штучна нейронна мережа. Зокрема щодо аналізу «big data» — це застосування так званого штучного інтелекту або машинного навчання, тобто набору методів, завдяки яким комп'ютер може знаходити в масивах початково невідомі взаємозв'язки та закономірності.

Одним із напрямків застосування штучних нейронних мереж є кластерний аналіз даних. Проблеми використання кластеризації є досить важливими, оскільки результати такого аналізу мають дуже великий вплив на формування подальших стратегій щодо дій підприємств.

Аналіз існуючих джерел. Дослідження кластерного аналізу у межах навчання без учителя останнім часом досягли значного прогресу. Багато науковців у підходах кластеризації використовують автокодер через його здатність навчитися представленню без нагляду. Наприклад, J. Xie, R. Girshick та A. Farhadi [5], використовували кодер шаруватого автокодера для отримання як прихованого подання простору, так і кластерного призначення одночасно. Окрім архітектур кодера-декодера, до задачі кластеризації також застосовані архітектури на основі нейромереж. Наприклад, у [6], запропонований J. Yang, D. Parikh та D. Batra, використовується агломераційний кластерний підхід, який ґрунтується на архітектурі на основі штучних нейронних мереж. Хоча цей підхід добре працює, він потребує чималих налаштувань гіперпараметрів, обмежуючи його застосовність у задачі кластеризації. Окрім того, ці підходи, перебуваючи без нагляду, можуть не вивчати семантично значуще подання, яке згодом впливатиме на продуктивність кластерів. Для оптимізації кластерних завдань використовували втрати розбіжності *KL* на додаток до терміну регуляризації для виявлення збалансованих кластерів. Така втрата має тенденцію обмежувати здатність працювати з нерівномірними розподілами класів. Незважаючи на те, що цей підхід показує узагальнення зразків, він потребує делікатного налаштування мережі з поперечною втратою ентропії.

ОСНОВНА ЧАСТИНА

Кластерний аналіз даних

Кластеризацією називають процес об'єднання об'єктів схожими властивостями у групи, причому заздалегідь їх кількість не визначається, а формується у процесі роботи системи. Оскільки завдання кластеризації є актуальним, тобто зростаюче нагромадження великого обсягу даних призводить до необхідності їх класифікувати з урахуванням все більшої кількості параметрів, тому постає завдання щодо розроблення і застосовування методів, які спеціалізуються на класифікації багатовимірних даних.

Реальні дані дуже відрізняються за характеристиками від досліджуваної вибірки. Отже, для оптимального аналізу даних доречніше обробляти різні вибірки різними методами.

Кластеризація — окремий клас задач машинного навчання, що, на відміну від класифікації, охоплює об'єкти навчальної вибірки, які не ма-

ють заздалегідь відомих відповідей учителя. Тому такий спосіб машинного навчання називають *навчанням без учителя*.

Постановка задачі. Розглянемо формальну постановку задачі: маємо простір об'єктів X, серед яких маємо скінченну множину об'єктів $X^l = \{x_i\}_{i=1}^l$, яку беремо як навчальну вибірку. Також для цих об'єктів задано функцію відстані $\rho: X \times X \to [0, \infty)$ (тобто відстань між цими об'єктами є заздалегідь відомою); необхідно так поділити нашу скінченну множину об'єктів на групи або кластери, щоб близькі об'єкти виявились всередині кожної групи, а між об'єктами різних груп відстані були достатньо великі.

Опис моделі системи кластерного аналізу даних

Розглядувану задачу можна розв'язувати чималою кількістю способів. Це зумовлено неоднозначністю подання початкових умов через неоднаковість (тобто відсутність загальноприйнятих) критеріїв якості кластеризації, різноманітними евристичними методами та підходами кластеризації, різними варіантами формування функції відстані між об'єктами, а також кількістю кластерів, яку заздалегідь неможливо передбачити.

Незважаючи на всі недоліки, існує безліч цілей, для яких розв'язуються задачі кластеризації, наприклад, спрощення подання про велику кількість об'єктів, розуміння їх внутрішньої структури (вони складаються з якихось груп однорідних об'єктів), з подальшим аналізом виокремлених груп.

Для алгоритму на основі нейронних мереж розглянемо відомі структури кластеризації k-середніх із розбіжністю KL на основі м'яких попарних обмеженнях.

Алгоритм обмеженого k-середнього значення використовує деякі марковані дані для керування неконтрольованою кластеризацією k-середніх. На відміну від випадкових ініціалізацій кластерних центрів у традиційних k-середніх, для ініціалізації центрів кластерів у обмежених k-середніх використовуються марковані зразки. Також при кожній ітерації k-середніх перепризначення кластера обмежується немаркованими зразками, а членство маркованих зразків фіксується. Ця процедура обмежених k-середніх показала поліпшення продуктивності за алгоритмом k-середніх.

Метод k-середніх. На вході буде подано $X^u = \{x_1^u, x_2^u, \ldots, x_m^u\}$ — набір немаркованих даних; $X_k^l = \{x_1^l, x_2^l, \ldots, x_p^l\}$ — набір маркованих даних у класі k, $X^l = \cup_{k=1}^K X_k^l$. На виході ми хочемо отримати розділені K набори $\{C_k\}_{k=1}^K$ з X^u, що мінімізує цільову функцію в k-середніх.

Задамо параметри:

1. $t = 0$.

2. Ініціалізація центрів кластерів:

$$\mu_k = \frac{1}{\left|x_k^l\right|} \sum_{x \in X_k^l} x .$$

3. Повторити до зближення:

• надати дані кластеру:

Для маркованих даних: $x \in X_k^l$ надати x до кластера C_k^{t+1}.

Для немаркованих даних: для $x_i^u \in X^u$ надати до C_k^{t+1} кластера, отриманого від $k = \arg\min_k \|x_i^u - \mu_k^t\|^2$.

Оновлення центрів:

$$\mu_k^{t+1} = \frac{1}{\left|C_k^l\right|} \sum_{x \in C_k^t} x.$$

• $t \leftarrow t+1$.

Інша складова алгоритму ґрунтується на розбіжності KL, що є мірою невідповідності між двома розподілами ймовірностей. З урахуванням K-вимірного вектора ймовірностей присвоєння кластерів p і q, що відповідають точкам відповідно x_p і x_q, розбіжність KL між p і q задається формулою

$$KL(p\|q) = \sum_{i=1}^{K} p_i \log \frac{p_i}{q_i}, \qquad (1)$$

де K — кількість кластерів. У даному підході використовуємо симетричний варіант розбіжності KL, оскільки ми маємо справу тільки з оптимізацією функції втрат для p і q одночасно:

$$L_{p,q} = KL(p\|q) + KL(q\|p). \qquad (2)$$

Втрати (2) отримують шляхом першої фіксації p і обчислення розбіжності q з p і навпаки.

ВИСНОВКИ

Запропонований метод дає можливість автоматизувати процес кластерного аналізу даних, особливо у разі, якщо кількість кластерів із самого початку невідома. Для цього на основі методів k-середніх та розбіжності KL було описано модель системи кластерного аналізу даних на основі нейронної мережі.

Список використаної літератури

1. ***Deep*** *clustering via joint convolutional autoencoder embedding and relative entropy minimization / K. G. Dizaji, A. Herandi, C. Deng [et al.] // In 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017. P. 5747–5756.*

2. ***Rich feature*** *hierarchies for accurate object detection and semantic segmentation / R. Girshick, J. Donahue, T. Darrell, J. Malik // In CVPR, 2014. P. 580–587.*

3. ***Improved*** *deep embedded clustering with local structure preservation / X. Guo, L. Gao, X. Liu, J. Yin // In International Joint Conference on Artificial Intelligence (IJCAI-17), 2017. P. 1753–1759,*

*4. **Deep** clustering with convolutional autoencoders / X. Guo, X. Liu, E. Zhu, J. Yin // In International Conference on Neural Information Processing, Springer, 2017. P. 373–382.*

*5. **Xie J., Girshick R., Farhadi A.** Unsupervised deep embedding for clustering analysis // In International conference on machine learning, 2016. P. 478–487.*

*6. **Yang J., Parikh D., Batra D.** Joint unsupervised learning of deep representations and image clusters // In CVPR, 2016. P. 5147–5156.*

Рецензент: доктор техн. наук, ст. наук. співробітник **Ю. В. Мельник,** Державний університет телекомунікацій, Київ.

К. П. Сторчак, А. Н. Тушич, А. П. Бондарчук

КЛАСТЕРНЫЙ АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

Предложен процесс кластерного анализа данных, основанного на объединении объектов, обладающих общими свойствами, в группы. Рассмотрены и проанализированы вклад ученых в изучение кластерного анализа данных с использованием искусственных нейронных сетей и их достижения ближайших лет, целесообразность использования искусственных нейронных сетей для процесса кластерного анализа данных. Описаны с использованием математического аппарата модель системы кластерного анализа данных с использованием искусственных нейронных сетей.

Ключевые слова: кластер; кластеризация; кластерный анализ данных; алгоритм; математическая модель; искусственная нейронная сеть; big data.

K. Storchak, A. Tushych, A. Bondarchuk

CLUSTER ANALYSIS OF DATA WITH THE USE OF ARTIFICIAL NEURAL NETWORKS

The article describes the process of cluster analysis of data - the analysis of data, which is based on the association of objects that have common properties in the group. The task of clustering is relevant, since the growing accumulation of a large number of data leads to the need to classify them in the light of an increasing number of parameters, so the task of developing and applying methods that specialize in the classification of multidimensional data sets. The article considers and analyzes the contribution of scientists to the study of cluster analysis of data using artificial neural networks and their achievement in the coming years, the feasibility of using artificial neural networks for the process of cluster analysis of data. The model of the system of cluster analysis of data using artificial neural networks is described with the use of mathematical apparatus. For a model of cluster analysis of data based on neural networks, well-known structures of k-medium clustering with differences in KL were considered on the basis of soft pairwise constraints. The proposed k-mean bounded algorithm uses some marked data to control the uncontrolled k-medium clustering. Unlike the random initialization of cluster centers in traditional k-averages, marking samples are used to initiate cluster centers in restricted k-averages. Also, with each iteration, the k-mean reassignment of the cluster is limited to unmarked samples, and the membership of the labeled samples is fixed. This procedure of limited k-averages showed performance improvements by the k-medium algorithm. The described method makes it possible to automate the process of cluster analysis of data, especially when the number of clusters from the beginning is unknown. To do this, based on known k-medium methods and KL differences, a model of cluster analysis system based on the neural network was described.

Keywords: cluster; clustering; cluster analysis; algorithm; mathematical model; artificial neural network; big data.