UDC 004.9 **doi: 10.26906/SUNZ.2024.2.152**

B. Steblyanko, O. Ni, H. Kuchuk, D. Volk

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

FUZZY INTERACTIVE CLUSTERING METHOD

Abstract. The article examines an example of a system in which a large number of short texts are generated. In it, participants create strategic planning documents, within which key performance indicators are determined. The formulations of key performance indicators form a data set consisting of short texts. Within the framework of this system, there is an urgent task of forming and updating a classifier based on this set. A solution to this problem is presented using the fuzzy interactive clustering method. This method allows expert to perform clustering sets of short texts, issuing reverse communication based on the results of each step interactive clustering. Collection procedure reverse does not imply any connection availability of an expert special knowledge about work neural network and is assembled in human-readable form matrices reverse communications. Such an approach has advantages over clustering methods requiring adjustments metaparameters algorithm not related directly with the clustering results. Also important advantage the proposed method is opportunity realize clustering sets data related to various language domains that do not match the domain on which was produced education language models, due to proposed extension method dictionary language models This property allows use the proposed algorithm in a narrow way specialized domains, as well as in domains that do not allow you to obtain a full-fledged corpus of texts for yourself training language models.

Keywords: clustering, data, decision making, efficiency, neural net.

Introduction

Cluster analysis is one of the most important sections of system data analysis and is used in various problem areas - technical, natural science, social.

Clustering is an example of an unsupervised learning problem and comes down to dividing the original set of objects into subsets of classes in such a way that elements of one class are as similar as possible to each other, and elements of different classes are different.

Traditional cluster analysis methods work with objects specified as vectors signs [1–4]. When working with texts, the first step of the algorithm is clustering is definition space signs and construction in it vectors available texts [5, 6]. Typically received vectors have big dimensions and when working with them traditional cluster analysis methods do not provide sufficient efficiency[7–10].

When working with short texts dimension vectors does not decrease, but only is added property sparsity to feature vectors that creates additional difficulties with them processing by cluster analysis methods [11, 12]. Below the short texts in this research implied texts consisting from one or several sentences with a total number of words ranging from 5 to 100.

In addition, additional complicating factors solution tasks clustering for short texts are: synonymy, homonymy, more frequent, compared to ordinary texts, use abbreviations, slang expressions and neologisms and most the main thing is partial or complete absence context for short texts.

Swift height arrays information consisting from short text sets fragments, contributes intensification research in the field development methods processing texts using machine learning.

Problem annually dedicated to a significant number of studies. Big Part carried out research refers to texts in English language.

In the article an example of a system is considered in which is happening generation big number of short texts. In it the participants form documentation strategic planning, within the framework of which are determined key indicators efficiency.

Formulations key indicators efficiency form a data set consisting from short texts.

Within these systems acute the task is to form and update classifier based on this set.

This task can be solved with clustering.

Main Part

Modern methods clustering using neural networks are usually used neural network for preparation vectors signs and then used analytical method (based on formulas with hyperparameters) for clustering these signs.

As a result, the result of clustering due to quality received vectors signs, i.e. quality training neural network.

At the same time, in the last time appear methods allowing solve the clustering problem directly using neural network that allows combine process receiving vectors signs and actually clustering.

Sticking to ideas for the researcher most simple and accurate reverse communication there will be criticism received results clustering, in this work supposed reverse connection two types:

1 – "element X_i must belong to the cluster C_j";

2 – "element X_i should not be in cluster C_j".

Simultaneously may be received arbitrarily the number of such restrictions, in particular, is easily specified restriction "change" elements X_i and X_j in places" combination two restrictions first kind. In Table 1 shows an example of a formalized reverse connections from an expert in the form of a matrix reverse communications.

At the intersection of the line corresponding element (object) from the data set and cluster to which the an object was ranked with the highest degree confidence neural network (maximum value in the corresponding output component vector) is placed reverse connection two the above types in the form "Include" or "Exclude" respectively. in Table 1.

Table 1 – **Example of a matrix reverse connections from an expert**

	Cluster 1	**Cluster 2**	**…**	**Cluster K**
Element 1	+	+	…	turn on
Element 2	+	turn on	…	+
…	…	…	…	…
Element N	+	exclude	…	+

From generalized scheme methods clustering it is clear that treatment reverse communications at the level single target functions (functions loss, penalty functions, loss function) will allow simultaneously build performance objects in accordance with intention researcher influencing weights neural network and adjust errors clustering influencing the outcome of the next iterations clustering (for example, shifting clusters).

For the presented clustering method it is not difficult to notice that target function is aimed at q_{ij} was more p_{ij}.

If look at private derivatives for updating scales neural network and vectors centre's clusters

$$\frac{\partial L}{\partial z_i} = 2\sum_{j=1}^{J}\left(\begin{array}{l}\left(1+\left\|z_i-\mu_j\right\|^2\right)^{-1}\times \\ \times\left(p_{ij}-q_{ij}\right)\cdot\left(z_i-\mu_j\right)\end{array}\right); \quad (1)$$

$$\frac{\partial L}{\partial \mu_j} = -2\sum_{i=1}^{I}\left(\begin{array}{l}\left(1+\left\|z_i-\mu_j\right\|^2\right)^{-1}\times \\ \times\left(p_{ij}-q_{ij}\right)\cdot\left(z_i-\mu_j\right)\end{array}\right), \quad (2)$$

then it can be understood that in the case negative the difference (p_{ij} - q_{ij}) will be the element is "pushed out" from the cluster, despite the absence fine with sides target functions.

In this research for accounting reverse communications user when adjusting scales neural network by inverse method distribution errors offered use following formulas for calculation gradients target Features:

$$L = KL\left(P \| Q\right) = \\ = \sum_{i=1}^{I}\sum_{j=1}^{J} t_{ij}\cdot p_{ij}\cdot\log\frac{p_{ij}}{q_{ij}}; \quad (3)$$

$$\frac{\partial L}{\partial z_i} = 2\sum_{j=1}^{J}\left(\begin{array}{l}\left(1+\left\|z_i-\mu_j\right\|^2\right)^{-1}\times \\ \left(p_{ij}-q_{ij}\right)\cdot t_{ij}\cdot\left(z_i-\mu_j\right)\end{array}\right); \quad (4)$$

$$\frac{\partial L}{\partial \mu_j} = -2\sum_{i=1}^{I}\left(\begin{array}{l}\left(1+\left\|z_i-\mu_j\right\|^2\right)^{-1}\times \\ \left(p_{ij}-q_{ij}\right)\cdot t_{ij}\cdot\left(z_i-\mu_j\right)\end{array}\right), \quad (5)$$

where $T = \left(t_{ij}\right)$ is the matrix of feedback multipliers, in which the elements satisfy the conditions:

$$\left[\begin{array}{l} t_{ij} > 0, \text{ if element } i \text{ is included in the cluster } j; \\ t_{ij} < 0, \text{ if element } i \text{ is excluded from the cluster } j; \\ t_{ij} = 0, \text{ if element } i \text{ is remains in the cluster } j. \end{array}\right.$$

The matrix T is obtained trivially from matrices reverse communication from the expert presented in Table 1.

In the Table 2 shows an example with the result of the transformation matrices reverse connections from an expert to the matrix multipliers for target functions neural network.

Table 2 – **Feedback multiplier matrix for the objective function**

	Cluster 1	**Cluster 2**	**…**	**Cluster K**
Element 1	0	0	…	100
Element 2	0	100	…	0
…	…	…	…	…
Element N	0	–1000	…	0

Absolute meaning t_{ij} defines the speed at which cluster elements and centers will be strive for each other or push away from each other.

Also, on this speed influences exposed level learning rate) for a neural network. In this work in experiments used value 100 to enable element to the cluster and - 1000 to exclude element from the cluster.

These values were determined empirically, because experiments have shown that for the case pushing out element from the cluster has meaning use big absolute magnitudes than with attraction.

Described higher an approach allows enough effectively solve the clustering problem, and in the case of such multidimensional objects as text, to user necessary have opportunity render influence on the course process clustering for the purpose identifying hidden or explicit intentions.

One of main advantages proposed clustering method is that added restrictions are not rigid, do not lead to the need decide systems equations that can potentially turn out to be incompatible.

Any restrictions user, regardless of their internal there will be contradictions taken into account in fines sides target functions

Conclusions

In the article a method of interactive reverse clustering communications at the base modern methods clustering.

Used generalized construction scheme algorithm clustering and proposed architecture neural network allow combine advantages modern language models with high indicators accuracy and perplexity and most currently effective universal algorithms clustering.

Language blocks model and clustering in the proposed architecture artificial neural network with time can be easily replaced with blocks of more modern and efficient language models and methods clustering without the need changes proposed algorithm.

This method allows expert to perform clustering sets of short texts, issuing reverse communication based on the results of each step interactive clustering. Collection procedure reverse does not imply any connection availability of an expert special knowledge about work neural network and is assembled in human-readable form matrices reverse communications. Such an approach has advantages over clustering methods requiring adjustments metaparameters algorithm not related directly with the clustering results. Expert in such methods interacts with the algorithm as a "black box" model, which reduces efficiency man-machine interactions.

Also important advantage the proposed method is opportunity realize clustering sets data related to various language domains that do not match the domain on which was produced education language models, due to proposed extension method dictionary language models This property allows use the proposed algorithm in a narrow way specialized domains, as well as in domains that do not allow you to obtain a full-fledged corpus of texts for yourself training language models.

REFERENCES

1. McCann, B., Bradbury, J., Xiong, C., Socher R. (2017), Learned in Translation: Contextualized Word Vectors. *NIPS*. I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R.Fergus, S.V.N. Vishwanathan, R. Garnett (eds). P. 6297-6308.
2. Meier B.B., Elezi, I., Amirian, M., Dürr, O.. Stadelmann, T. (2018), "Learning Neural Models for End-to-End Clustering", *Artificial Neural Networks in Pattern Recognition*, Lecture Notes in Computer Science, Springer, Cham. / L. Pancioni, F. Schwenker, E. Trentin (eds.). Vol 11081.
3. Kammoun, N., Abassi, R., Guemara, S. (2019). Towards a new clustering algorithm based on trust management and edge computing for IoT. 2019 15th International Wireless Communications and Mobile Computing Conference, IWCMC 2019, 1570–1575, 8766492. doi: https://doi.org/10.1109/IWCMC.2019.8766492
4. Kovalenko, A., Kuchuk, H. (2022), Methods to Manage Data in Self-healing Systems. *Studies in Systems, Decision and Control*, 425, 113–171, doi: https://doi.org/10.1007/978-3-030-96546-4_3
5. Yang, J., Bao, L., Liu, W., Yang, R., Wu, C.Q. (2023). On a Meta Learning-Based Scheduler for Deep Learning Clusters. *IEEE Transactions on Cloud Computing*, 11(4), 3631–3642. doi: https://doi.org/10.1109/TCC.2023.3308161
6. Gomathi, B., Saravana Balaji, B., Krishna Kumar, V., Abouhawwash, M., Aljahdali, S., Masud, M. and Kuchuk, N. (2022), "Multi-Objective Optimization of Energy Aware Virtual Machine Placement in Cloud Data Center", *Intelligent Automation and Soft Computing*, Vol. 33(3), pp. 1771–1785, doi: http://dx.doi.org/10.32604/iasc.2022.024052
7. Zuev, A., Karaman, D., Olshevskiy, A. (2023). Wireless sensor synchronization method for monitoring short-term events. *Advanced Information Systems*, 7(4), 33–40. doi: https://doi.org/10.20998/2522-9052.2023.4.04
8. Petrovska, I., Kuchuk, H. (2023). Adaptive resource allocation method for data processing and security in cloud environment. *Advanced Information Systems*, 7(3), 67–73. doi: https://doi.org/10.20998/2522-9052.2023.3.10
9. Kuchuk, N., Mozhaiev, O., Semenov, S., Haichenko, A., Kuchuk, H., Tiulieniev, S., Mozhaiev, M., Davydov, V., Brusakova, O., Gnusov, Y. (2023). Devising a method for balancing the load on a territorially distributed foggy environment. *Eastern-European Journal of Enterprise Technologies*, 1(4 (121), 48–55. doi: https://doi.org/10.15587/1729-4061.2023.274177
10. Kuchuk, N., Kovalenko, A., Ruban, I., Shyshatskyi, A., Zakovorotnyi, O., Sheviakov, I. (2023). Traffic Modeling for the Industrial Internet of NanoThings. *2023 IEEE 4th KhPI Week on Advanced Technology*, KhPI Week 2023 - Conference Proceedings, 2023, doi: 194480. http://dx.doi.org/10.1109/KhPIWeek61412.2023.10312856
11. G. Khoroshun, O. Ryazantsev, and M. Cherpitskyi, "Clustering and anomalies of USA stock market volatility index data", *Advanced Information Systems*, vol. 7, no. 2, pp. 9–15, Jun. 2023. doi: 10.20998/2522-9052.2023.2.02.
12. Li, G., Liu, Y., Wu, J., Lin, D., Zhao, Sh. (2019). Methods of Resource Scheduling Based on Optimized Fuzzy Clustering in Fog Computing. Sensors, MDPI, 19(9). doi: https://doi.org/10.3390/s19092122

Received (Надійшла) 27.02.2024
Accepted for publication (Прийнята до друку) 17.04.2024

Метод нечіткої інтерактивної кластеризації

Б. О. Стеблянко, О. В. Ні, Г. А. Кучук, Д. М. Волк

Анотація. У статті досліджується приклад системи, де відбувається генерація великої кількості коротких текстів. У ній учасники формують документи стратегічного планування, у яких визначаються ключові показники ефективності. Формулювання ключових показників ефективності утворюють відповідний набір даних. У рамках цієї системи гостро стоїть завдання формування та актуалізації класифікатора, заснованого на даному наборі даних. Наведено розв'язання цієї задачі за допомогою методу нечіткої інтерактивної кластеризації. Даний метод дозволяє експерту проводити кластеризацію наборів коротких текстів, надаючи зворотний зв'язок за результатами кожного етапу інтерактивної кластеризації. Процедура зворотного зв'язку не передбачає наявності у експерта спеціальних знань про роботу нейронної мережі та збирається у вигляді матриці зворотного зв'язку, яку може прочитати людина. Такий підхід має переваги в порівнянні з методами кластеризації, що вимагають коригування метапараметрів алгоритму, не пов'язаних безпосередньо з результатами кластеризації. Також важливою перевагою запропонованого методу є можливість здійснювати кластеризацію наборів даних, що відносяться до різних мовних доменів, що не збігаються з доменом, на якому проводилося навчання мовної моделі, за рахунок запропонованого методу розширення словника мовної моделі. Ця властивість дозволяє використовувати запропонований метод у вузько спеціалізованих доменах, а також у доменах, що не дозволяють отримати повноцінний кортеж текстів для навчання мовної моделі.

Ключові слова: кластеризація, дані, прийняття рішень, ефективність, нейронна мережа.