RESEARCH ARTICLE

Cluster search of traditional Chinese medicine teaching models based on massive high-dimensional graph search

Jianlin Zha[1], Yimin Wang[2, *]

[1]Department of Security, [2]Dean's Office, Shanghai University of TCM, Shanghai, China.

Received: December 5, 2025; **accepted:** April 28, 2026.

Traditional Chinese medicine (TCM) education is facing challenges from the rapid expansion of knowledge and the need for personalized learning. Existing teaching models often suffer from fragmented knowledge representation and low precision in resource recommendation. This study aimed to address these issues by developing a novel cluster search method that integrated knowledge graphs with deep clustering techniques to enhance the structural representation and retrieval of TCM teaching resources. A TCM teaching knowledge graph was constructed by using a six-dimensional ontology model to represent the "phenomenon-principle-method-prescription-medicine" framework. A ResNet50-K-means++ two-stage clustering framework was then proposed by using ResNet50 for deep feature extraction from TCM images and K-means++ for optimized clustering of these high-dimensional features. The results showed that, on TCM teaching image and medical record datasets, the ResNet50-K-means++ method achieved an F1 score of 89.7% and an accuracy of 92.3%, significantly outperforming the traditional PCA-K-means method (75.0% and 72.4%, respectively) (*P* < 0.01). The proposed method provided a highly accurate and adaptable intelligent solution for TCM education by enabling precise classification of teaching models, thus supporting personalized learning and knowledge discovery.

Keywords: high-dimensional data; knowledge graph; graph search; traditional Chinese medicine education; clustering analysis.

[*]**Corresponding author:** Yimin Wang, Dean's Office, Shanghai University of TCM, Shanghai 201203, China. Email: videau@163.com.

Introduction

Traditional Chinese medicine (TCM) education has entered a new stage in the era of rapid information growth, where both opportunities and challenges coexist. With the continuous enrichment of TCM knowledge systems and the accumulation of clinical practice experience, the volume and complexity of teaching resources have increased substantially, making it difficult for traditional teaching models to meet students' learning needs [1, 2]. Although TCM education has evolved from apprenticeship-based systems to modern institutional frameworks, achieving significant progress in large-scale talent cultivation, limitations remain in terms of personalized learning and deep knowledge exploration. Meanwhile, the rapid development of modern information technologies such as big data and artificial intelligence has provided new opportunities for transforming TCM education [3].

In recent years, considerable advances have been made in applying data-driven technologies to TCM teaching. Large-scale teaching resources including textual materials, audio-visual data, and clinical medical records have been accumulated, forming complex and high-dimensional datasets. The TCM theoretical system itself involves numerous interrelated entities such as symptoms, syndromes, prescriptions, and treatment methods, resulting in intricate knowledge structures [4, 5]. Moreover, deep learning (DL) and data mining techniques such as clustering and image retrieval have shown promising performance in related domains including medical image analysis and knowledge discovery. However, their application in TCM teaching remains limited and fragmented. Despite these advances, several critical challenges remain, which include that traditional text-based retrieval methods are insufficient for handling complex semantic relationships in TCM knowledge, and existing approaches lack effective integration of multi-modal data and high-dimensional graph structures. Furthermore, the combination of clustering techniques with high-dimensional graph search for discovering TCM teaching models has not been fully explored, limiting the development of intelligent and personalized teaching systems [6-8].

This study aimed to develop an intelligent cluster search method for TCM teaching models based on high-dimensional graph analysis. A TCM teaching knowledge graph was constructed to represent multi-level semantic relationships, and a hybrid framework combining deep learning and clustering techniques was then proposed. Dimensionality reduction and feature extraction methods were applied to process high-dimensional data followed by clustering algorithms to identify similar teaching models and learning patterns. The proposed approach was expected to improve knowledge representation, enhance search accuracy, and support personalized teaching resource recommendations in TCM education. By integrating knowledge graph modeling with deep learning-based clustering, this study contributed to the advancement of intelligent and data-driven TCM teaching systems.

Materials and methods

Construction of the TCM teaching knowledge graph

The knowledge graph integrated theories and methodologies from information visualization technology, applied mathematics, graphics, and information science with quantitative citation and co-occurrence analysis. The construction process involved a systematic workflow comprising six key steps including data source identification, data collection, knowledge representation, knowledge extraction, knowledge fusion, and knowledge computation (Figure 1). For knowledge representation, logical-layer schemes such as the resource description framework (RDF) and property graph, as well as storage-layer serialization file exchange formats like JavaScript Object Notation for Linked Data (JSON-LD) and N-TRIPLE were employed [9]. For knowledge extraction, supervised learning methods for entity extraction and relation extraction were used to identify relationships such as "etiology and symptom," "prescription and drug composition," and "symptom and treatment method" from the text corpora [10-12]. Based on symptom and syndrome co-occurrence patterns in the data, connections were established between co-occurring symptom nodes (zz_i) and syndrome nodes (zx_j) to construct a "symptom-syndrome" heterogeneous graph. The resulting edge information set was derived as follows.

$$E_{Symptom-Syndrome} = \{(zz_i, zx_j) \mid \text{co-occurrence frequency} > \theta\} \quad (1)$$

where θ was a predefined co-occurrence threshold. For knowledge fusion, a six-dimensional ontology model was developed to integrate theoretical depth and practical applicability, which achieved the three-dimensional modeling of the prescription knowledge system through multi-level semantic associations and dynamic evolution mechanisms. The ontology model comprised six core dimensions including syndrome, symptom, prescription, drug, treatment method, and mechanism with relationships defined between them to form a structured knowledge network.

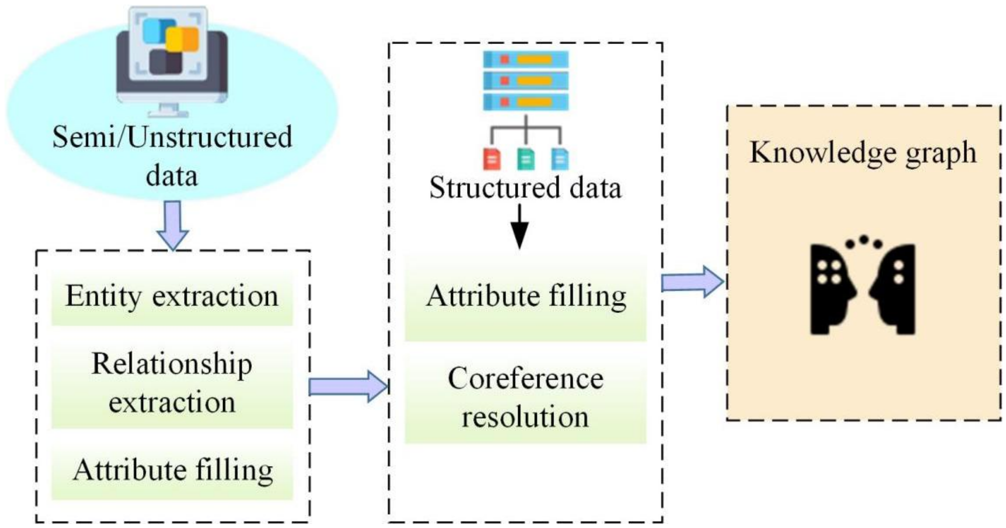

Figure 1. The workflow of knowledge graph construction.

High-dimensional graph search based on principal component analysis (PCA) dimension reduction and latent Dirichlet allocation (LDA)

In the field of high-dimensional graph visualization, PCA is a commonly used data dimensionality reduction technology. PCA transformed original data into a new set of orthogonal variables through linear transformation with these new variables referred to as principal components. The principal components could capture the maximum variance directions in the data and convert them into new coordinate systems. By inputting PCA-dimension-reduced data into classification, clustering, or regression models, more hidden patterns and correlations could be discovered, providing more accurate guidance for decision-makers [14-16]. Assuming that a set of points $X = \{x_1, x_2, ..., x_n\}$ was given in the space of R^d , PCA's objective was to discover the projection matrix ($U \in R^{d \times k}$) that best reflected the differences in the original spatial data. The PCA could be expressed as a reconstruction error minimization model as below.

$$\min_U \sum_{i=1}^{n} || x_i - UU^T x_i ||_2^2 \quad (2)$$

To improve the robustness of PCA, one method was to change the l_2 norm in the PCA function to the l_1 norm as a distance metric. Based on the idea of minimizing the reconstruction error, the model for solving the projection direction could be represented as follows.

$$\min_U \sum_{i=1}^{n} || x_i - UU^T x_i ||_1 \quad (3)$$

The LDA model served as the main model for text data mining. Derived from the bag-of-words model and based on the Dirichlet distribution, the LDA model analyzed topic contents in document webpages by ignoring the order of words in text documents. The LDA model's framework could be described as a three-level hierarchical Bayesian model, where each document was modeled as a mixture of topics, and each topic was modeled as a mixture of words. The study obtained a massive volume of text data with uneven vocabulary distribution, where some words appeared frequently while others occurred only once. Retaining all words would not only consume significant operational space but also impede subsequent topic processing [17-19]. Therefore, this study statistically analyzed high-frequency words in the text and selected them for discussion and analysis. Perplexity was used to evaluate the quality of the model trained during text data mining. For relevant parameters in the LDA model, the Heinrich G parameter estimation method was employed to determine values with α = 50/T and β = 0.1. The likelihood metric module in Python was adopted to measure the model's performance.

TCM teaching models based on cluster search

A TCM teaching model was proposed by integrating deep feature learning and improved clustering algorithms through collaborative optimization of ResNet50 and K-means++. Traditional convolutional neural network (CNN) architectures encompassed several well-established models. Among them, Residual Network (ResNet) (Microsoft Research, Redmond, WA, USA) mitigated the vanishing/exploding gradient issue in deep networks through its innovative residual blocks, which learned the residual mapping between inputs and outputs rather than direct transformations [20]. ResNet50 as a pivotal variant in the ResNet family was used as the feature extractor. Its architecture consisted of multiple residual blocks, each integrating several convolutional layers with skip connections. The core of a residual block was defined as follows.

$$y = F(x, \{W_i\}) + x \tag{4}$$

where x was the input to the block. y was the output. $F(x, \{W_i\})$ was the residual mapping to be learned. In this equation, x and F must have the same dimensions for the addition to be valid. The K-means++ algorithm based on the scikit-learn library (Python Software Foundation, Wilmington, DE, USA) represented an enhanced variant of traditional K-means that improved clustering consistency and accuracy through optimized centroid initialization. The process was initiated by randomly selecting one centroid, then iteratively determined subsequent centroids by computing squared minimum distances to existing centroids for all remaining points, using these values as selection probabilities.

Experimental analysis

This study leveraged two public datasets for experiments to validate the proposed method's effectiveness, which included the TCM teaching image dataset collected from renowned TCM institutions such as the China Academy of Chinese Medical Sciences (Beijing, China) and Shanghai University of Traditional Chinese Medicine (Shanghai, China) containing approximately 10,000 images related to TCM teaching and the TCM medical record text dataset derived from the National TCM Medical Record Database (Beijing, China). All experimental procedures were approved by the Institutional Review Board (IRB) of Shanghai University of Traditional Chinese Medicine (Shanghai, China) with the approval No. of 2024-SHUTCM-045. A pre-trained ResNet50 model was used for feature extraction with input images uniformly resized to 224 × 224 pixels, a learning rate of 0.001, 100 training epochs, and a batch size of 32. During feature extraction, the output of the second-to-last layer of the model was taken as the image's feature vector with a dimension of 2,048. For the K-means++ algorithm setup, the K value was determined to be 15, and the maximum number of iterations was set to 300. To verify the proposed ResNet50 + K-means++ method's effectiveness, several clustering methods were selected for comparison including PCA + K-means that clustered after dimensionality reduction of PCA, LDA + K-means that LDA extracted text topic features combined with K-means clustering, and ResNet50 + K-means that ResNet50 extracted image features, and K-means did clustering. All experiments were run on servers with the same configuration. The hardware configuration was an Intel Xeon Gold 6248R processor (Intel Corp., Santa Clara, CA, USA) with a main frequency of 2.40 GHz, 128 GB of memory, and an NVIDIA Tesla V100 Graphics Processing Unit (NVIDIA Corp., Santa Clara, CA, USA). The software environment was Python 3.8 (https://www.python.org). The DL framework adopted PyTorch 1.9.0 (https://pytorch.org).

Statistical analysis

Statistical analysis was performed by using Python 3.8 with the SciPy library (version 1.7.3). The performance metrics (accuracy, F1 score) from different clustering methods were compared by using a paired t-test. A *P* value less

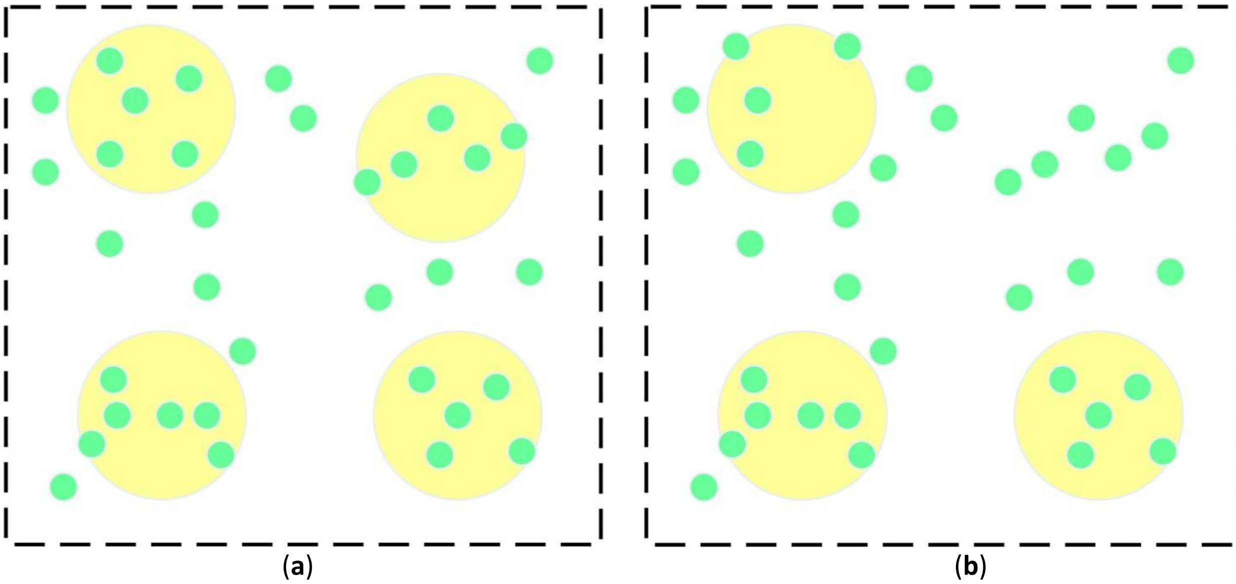

Figure 2. The data abstract clustering result when α = 0 (**a**) and α = 10 (**b**).

than 0.05 was considered statistically significant, while a *P* value less than 0.01 was considered highly significant. All experiments were repeated 10 times to ensure robustness.

Results

Clustering of high-dimensional graph data

The TCM teaching image dataset was selected and projected onto a scatter plot using multi-dimensional scaling (MDS). The projection results for α = 0 and α = 10 demonstrated that, as the parameter α increased, the number of clusters gradually decreased. When α was 0, many very small clusters appeared in the plot (Figure 2a). However, as α increases to 10, small clusters gradually disappeared, and the results tended to generate larger clusters (Figure 2b). In practical applications, the selection of parameter α could be adjusted according to users' specific needs.

Analysis of clustering performance of ResNet50-K-means++

The feature extraction accuracy of ResNet50 was compared with the traditional Histogram of Oriented Gradient (HOG) method. The results showed that feature vectors extracted by ResNet50 exhibited higher discriminability across different types of TCM teaching images for tongue diagnosis, pulse diagnosis, traditional Chinese medicine, and acupuncture point (Figure 3), which suggested that ResNet50 could effectively capture key information in TCM teaching images, providing high-quality feature representations for subsequent clustering analysis.

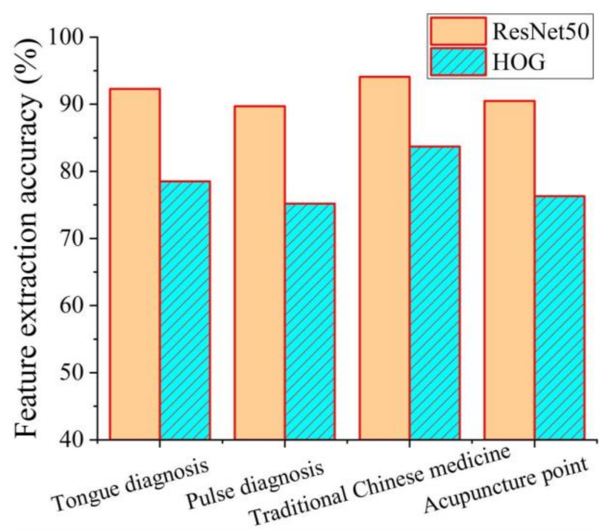

Figure 3. Comparison of feature extraction accuracy.

The proposed ResNet50-K-means++ method's performance was compared with several other common methods in clustering tasks of the TCM teaching model. The results demonstrated that the proposed method outperformed others in metrics such as clustering accuracy and F1 score. Statistical analysis results showed that the improvements of proposed method compared to the other methods were highly significant with the proposed method achieving a 17.3% improvement in clustering accuracy over the PCA-K-means method ($P < 0.01$) and a 14.8% improvement over the LDA-K-means method ($P < 0.01$), while the F1 scores of proposed method were improved by 15.7% ($P < 0.01$) and 13.2% ($P < 0.01$) compared to PCA-K-means and LDA-K-means, respectively (Figure 4). The results fully validated the advantages of ResNet50's powerful feature extraction capabilities and K-means++'s superior clustering performance in cluster search for TCM teaching models.

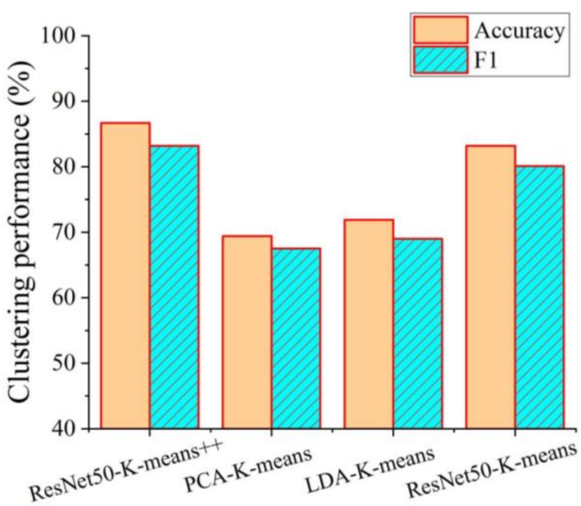

Figure 4. Comparison of results from various clustering algorithms.

Discussion

The proposed model in this study demonstrated remarkable technological innovation. By constructing a TCM teaching knowledge graph, it structurally represented the complex associative relationships of "phenomenon-principle-method-prescription-medicine" in the TCM theoretical system, seeking to process the issues of weak semantic association and information fragmentation in traditional text searches. The TCM teaching knowledge graph demonstrated superior performance in capturing users' latent teaching needs compared to the keyword-matching-based retrieval method proposed by Alvarez-Garcia *et al*. [21]. This advantage stemmed from its explicit modeling of multi-level semantic relationships between entities. The proposed "ResNet50-K-means++" two-stage clustering framework exhibited stronger robustness in handling high-dimensional image data. ResNet50 extracted abstract image features through deep CNNs, while the optimized initialization strategy of K-means++ effectively mitigated the sensitivity of traditional K-means to initial centroids. This performance advantage significantly differed from the text clustering approach based on LDA topic modeling [22], which relied solely on lexical co-occurrence statistics in text and struggled with multimodal data fusion of images and text. The application of the proposed method in cluster search for TCM teaching models offered strong support for personalized and precise TCM education. By clustering TCM teaching models, teachers could select appropriate teaching modes based on students' learning stages and needs, enhancing teaching effectiveness [23]. Additionally, this method helped students better understand and master the TCM knowledge system, promoting the development of TCM education. In clinical applications, the proposed method could provide a reference for TCM clinical decision-making. Through clustering analysis of TCM medical records, physicians could quickly identify cases similar to the patient's condition and formulate more reasonable treatment plans. This study introduced an innovative clustering method for TCM teaching models based on the ResNet50 and K-means++ algorithms. Experiments conducted on TCM teaching image and clinical case datasets demonstrated that the proposed method outperformed traditional approaches in clustering accuracy and stability, which enabled precise classification of various TCM teaching models, providing robust support for personalized and targeted TCM education. This study offered a novel approach for TCM teaching reform and knowledge discovery, contributing to the modernization and informatization of TCM education. Despite the proposed method's excellent performance in experiments, several limitations existed including ResNet50's high data preprocessing requirements and the experience-dependent selection of the K value in K-means++. Future work should explore intelligent K-value selection methods and integration with other DL models to further improve clustering performance and better uncover latent information in TCM teaching data.

Acknowledgements

This research was supported by Industry-University Cooperation and Collaborative Education Project from Ministry of Education (Grant No: 230803924033620) and Educational Scientific Research Project from Shanghai (Grant No.: C2022072).

References

1. Hu L, Zhang M, Hu P, Zhang J, Niu C, Lu X, *et al*. 2024. Dual-channel hypergraph convolutional network for predicting herb-disease associations. Brief Bioinform. 25(2):bbae067.
2. Zeng J, Jia X. 2025. Quantifying compatibility mechanisms in traditional Chinese medicine with interpretable graph neural networks. J Pharm Anal. 15(8):101342.
3. Li Y, Wang N, Huang Y, He S, Bao M, Wen C, *et al*. 2024. CircMYBL1 suppressed acquired resistance to osimertinib in non-small-cell lung cancer. Cancer Genet. 284-285:34-42.
4. Hengrui L. 2023. An example of toxic medicine used in Traditional Chinese Medicine for cancer treatment. J Tradit Chin Med. 43(2):209-210.
5. Liu H. 2020. Effect of traditional medicine on clinical cancer. Biomed J Sci Tech Res. 30(4):23548-23551.
6. Li H, Lu X, Wu Y, Luo J. 2023. Research on a data mining algorithm based on BERTopic for medication rules in Traditional Chinese Medicine prescriptions. Med Adv. 1(4):353-360.
7. Jia L, Li N, Abdelaal TRM, Guo N, Ijsselsteijn M, Unen V, *et al*. 2024. High-dimensional mass cytometry reveals emphysema-associated changes in the pulmonary immune system. Am J Respir Crit Care Med. 210(8):1002-1016.
8. Xie F, Peng Y, Yang W, tang J, Liu J, Li W, *et al*. 2023. Class-imbalance prediction and high-dimensional risk factor identification of adverse reactions of traditional Chinese medicine with centralized monitoring in real-world hospitals. Chin J Exp Tradit Med Formulae. 29(14):114-122.
9. Huang Z, Miao J, Chen J, Zhong Y, Yang S, Ma Y, *et al*. 2022. A traditional Chinese medicine syndrome classification model based on cross-feature generation by convolution neural network: model development and validation. JMIR Med Inform. 10(4):e29290.
10. Siam MHB, Ali MA, Vardaman D 3rd, Acharyya S, Patil M, Tyrrell DJ. 2024. CAFE: An integrated web app for high-dimensional analysis and visualization in spectral flow cytometry. bioRxiv [Preprint]. 12(03):626714.
11. Xia W, Su L, Lai P. 2021. Detection and prediction of active ingredients in fingerprints of traditional Chinese medicine from a high-dimensional perspective. J Chemom. 35(7-8):e3347.
12. Huang YH, Yoon CH, Gandhi A, Hanley T, Castrillon C, Kondo Y, *et al*. 2024. High-dimensional mapping of human CEACAM1 expression on immune cells and association with melanoma drug resistance. Commun Med. 4(1):128.
13. Pan Y, Gao J, Lin J, Ma Y, Hou Z, Lin Y, *et al*. 2023. High-dimensional single-cell analysis unveils distinct immune signatures of peripheral blood in patients with pancreatic ductal adenocarcinoma. Front Endocrinol (Lausanne). 14:1181538.
14. Yang W, Yi D, Zhou XH, Leng Y. 2024. Translational analysis of data science and causal learning in real-world clinical evaluation of traditional Chinese medicine. Sci Tradit Chin Med. 2(1):57-65
15. Lu CX, Zhu H, Xiao MZ, Li X. 2023. Classification of metabolic-associated fatty liver disease subtypes based on TCM clinical phenotype. Gastroenterol Hepatol Res. 5(1):2.
16. Hunewald O, Demczuk A, Longworth J, Ollert M. 2024. CyCadas: Accelerating interactive annotation and analysis of clustered cytometry data. Bioinformatics. 40(10):btae595.
17. Chan A, Jiang W, Blyth E, Yang J, Patrick E. 2021. treekoR: Identifying cellular-to-phenotype associations by elucidating hierarchical relationships in high-dimensional cytometry data. Genome Biol. 22(1):324.
18. Chakladar DD, Roy PP, Chang V. 2024. Integrated spatio-temporal deep clustering (ISTDC) for cognitive workload assessment. Biomed Signal Process Control. 89:105703.
19. Ge P, Liu C, Chan L, Pang Y, Li H, Zhang Q, *et al*. 2022. High-dimensional immune profiling by mass cytometry revealed the circulating immune cell landscape in patients with intracranial aneurysm. Front Immunol. 13:922000.
20. Tang J, Liu Y, Lin K, Li L. 2023. Process bottlenecks identification and its root cause analysis using fusion-based clustering and knowledge graph. Adv Eng Inform. 55:101862.
21. Alvarez-Garcia M, Ibar-Alonso R, Arenas-Parra M. 2024. A comprehensive framework for explainable cluster analysis. Inf Sci. 663:120282.
22. Cao J, Fang J, Meng Z, Liang S. 2024. Knowledge graph embedding: a survey from the perspective of representation spaces. ACM Comput Surv. 56(6):1-42.
23. Zhao Y, Li X. 2023. Deep spectral clustering with regularized linear embedding for hyperspectral image clustering. IEEE Trans Geosci Remote Sens. 61:1-11.