Student Group Division Algorithm based on Multi-view Attribute Heterogeneous Information Network

Xibin Jia[1*], Zijia Lu[1], Qing Mi[1*], Zhefeng An[2], Xiaoyong Li[3], and Min Hong[4]
[1] Faculty of Information Technology, Beijing University of Technology,
Beijing, 100124, China
[e-mail: jiaxibin@bjut.edu.cn]
[2] Faculty of Humanities and Social Science, Beijing University of Technology,
Beijing, 100124, China
[e-mail: anzhefeng76@163.com]
[3] Information Technology Support Center, Beijing University of Technology,
Beijing, 100124, China
[e-mail: lixiaoyong@bjut.edu.cn]
[4] Department of Computer Software Engineering, Soonchunhyang University, Asan, 31538, South Korea
[e-mail: mhong@sch.ac.kr]
[*]Corresponding authors: Xibin Jia, Qing Mi

Received June 5, 2022; revised October 27, 2022; accepted November 30, 2022; published December 31, 2022

Abstract

The student group division is benefit for universities to do the student management based on the group profile. With the widespread use of student smart cards on campus, especially where students living in campus residence halls, students' daily activities on campus are recorded with information such as smart card swiping time and location. Therefore, it is feasible to depict the students with the daily activity data and accordingly group students based on objective measuring from their campus behavior with some regular student attributions collected in the management system. However, it is challenge in feature representation due to diverse forms of the student data. To effectively and comprehensively represent students' behaviors for further student group division, we proposed to adopt activity data from student smart cards and student attributes as input data with taking account of activity and attribution relationship types from different perspective. Specially, we propose a novel student group division method based on a multi-view student attribute heterogeneous information network (MSA-HIN). The network nodes in our proposed MSA-HIN represent students with their multi-dimensional attribute information. Meanwhile, the edges are constructed to characterize student different relationships, such as co-major, co-occurrence, and co-borrowing books. Based on the MSA-HIN, embedded representations of students are learned and a deep graph cluster algorithm is applied to divide students into groups. Comparative experiments have been done on a real-life campus dataset collected from a university. The experimental results demonstrate that our method can effectively reveal the variability of student attributes and relationships and accordingly achieves the best clustering results for group division.

Keywords: graph deep clustering, heterogeneous information networks, representation learning, student behavior modeling

http://doi.org/10.3837/tiis.2022.12.003 **ISSN : 1976-7277**

1. Introduction

With the increasing improvement of campus informatization construction, the scale of educational big data such as student registration and academic information, student campus smart card records, and library checkout records has exploded. Various types of educational big data can be used by university educators to predict academic performance [1-3], classify academic performance [4], cluster at-risk students [5], and analyze association rules between course participation and student performance [6], etc. Educational data mining can help educators to grasp the multi-dimensional characteristics and changes of students in the educational environment, achieve personalized training programs for students, and improve the quality of their campus life [7].

Discovering the latent similar students and grouping them will facilitate the efficiency of student management. In recent years, it has been considered to make full use of the data of students' behaviors and attributes on campus for more objective and effective student group division. Most research of the existing group division methods is oriented toward community discovery [8,9] and user recommendation [10,11] tasks based on the social network or the heterogeneous network[12-16]. These methods are unsuitable for application to education scenarios directly because educational data mining is characterized by multiple sources, rich types, and rich interactive relationships. Due to the wide variety of big data resources of students on campus, mining students' behavior patterns, analyzing students' multi-dimensional attributes, measuring students' performance similarity, and scientifically dividing student groups remain is still a challenging but valuable research direction in educational data mining. Further work is needed for discriminative representation of student campus behaviors, comprehensive modeling of students based on multi-dimensional attributes, and effective methods for dividing students into different groups. In this paper, we propose a student group division method based on a multi-view student attribute heterogeneous information network for representing students' multi-dimensional attributes and relationships.

The main contributions of our work are as follows:

1) We propose a multi-view student attribute heterogeneous information network (MSA-HIN) to comprehensively represent student in aspects of the campus behavior adopting activity data from student smart cards and student attributes with taking account of multi-dimensional activity and attribution relationship.

2) A deep clustering algorithm is developed on the proposed network MSA-HIN with a multi-view graph auto-encoder. A deep clustering algorithm with a multi-view graph auto-encoder is developed on the proposed network MSA-HIN. The algorithm learns node embeddings by employing the most informative graph views and node content information, which facilitates capturing the shared features of multiple graph views.

3) Experiments are conducted on a real campus dataset, and the experimental analysis of the clustering results is given for demonstrating the effectiveness of proposed student division algorithm based on MSA-HIN.

The rest of the paper is organized as follows. In Section 2, we introduce the work related to group division and deep clustering. In Section 3, we introduce the proposed method for student group division based on the multi-view heterogeneous information network. In Section 4, the experiments and the result analysis are illustrated. Finally, the conclusion is given.

2. Related Work

2.1 Group Division

Group division characterizes the target group by specific technical methods and divides the target group by analyzing the similarity of individual representation. Group division usually includes clustering, classification, and graph partition algorithms. The current group division methods can be divided into three types.

The first type of group division method is performed based on the information related to individuals in the target group. This type of approach focuses only on the target group's existing individual attributes or behavioral information and on how to describe individual characteristics and reasonably represent them comprehensively. For example, Zhang et al. [17] proposed a method using data from two Yelp and Dianping review websites to construct two different aspects of user feature sets. They then applied the Sybil detection algorithm to the two feature sets to divide the target group into Sybil users and real users. Bunic et al. [9] used data from the web-based intelligent teaching system DITUS to construct two types of student features and then used the Euclidean distance-based K-Means algorithm to obtain optimal student group division results. However, this type of approach often ignores the important feature of the social structure among the individuals in the target group.

The second type of group division method is performed based on structural similarity in the social network of the target group. For example, Taheri et al. [18] considered useful information about the structure and properties of the network and proposed a new version of the affinity propagation (AP) method using an adaptive similarity matrix. The community discovery algorithm, as proposed in the classical study by Newman et al. [19], naturally divides the community nodes of the target group into tightly connected subgroups and proposes a measure of the strength of the community discovery structure for assessing the quality of the target group community division, which called modularity. This type of approach focuses only on the social network structure and usually ignores the individual attributes.

The third type of group division method considers all aspects of target group characteristics, including individual attributes and social network structure in the target group.

With the development of graph neural networks [20], more and more studies have applied graph neural network algorithms to group division tasks. The most critical aspect of such problems is learning the comprehensive representation of the targets, measuring the similarity of representation among targets, and then performing clustering algorithms in the target groups. Sankar et al. [21] proposed the GraFRank algorithm based on graph neural networks by specific pattern-specific neighbor aggregators to deal with heterogeneity in pattern homogeneity and learn nonlinear pattern correlations through cross-modal attention, and finally divide the target group into five categories while performing social friend ranking. Zhang et al. [22] proposed a deep node embedding method in attribute interaction graphs, which considers the attributes of edges in the target group graph to dynamically learn the node embedding while exploring the structure of the interaction graph to divide investors into four groups.

In summary, with reference to the good performance of graph neural networks in dealing with the heterogeneity of big data, we discuss a student group division method based on the heterogeneous information network, which can effectively model individual attributes and the social network structure of the target group.

2.2 Deep Clustering

To the best of our knowledge, there is no public dataset on student group segmentation tasks. In this paper, we used real campus dataset from a university and conducted a study using unsupervised clustering methods to reveal the intrinsic connections and behavioral patterns of students on campus. Considering that graph neural networks [20] have become one of the hottest directions in deep learning, we consider to exploit the graph neural network as basis for improvement of clustering algorithms.

Wang et al. [23] propose a deep attentional embedding graph clustering method that uses an attentional network to capture the importance of neighboring nodes to the target node, and encodes the topology and node content in the graph into a compact representation, and trains an inner product decoder to reconstruct the graph structure. In addition, soft labels are generated from the graph embedding itself to supervise the self-training graph clustering process, iteratively refining the clustering results and unifying the graph encoder module and clustering modules are unified under a unified training framework. Bo et al. [24] propose a structured deep clustering network for deep clustering by designing a transfer operator to transfer the representation learned by the deep auto-encoder to the corresponding GCN layer, effectively combining the advantages of an auto-encoder and GCN. Then proposed a dual self-supervised mechanism designed for deep clustering to unify these two different deep neural structures and guide the update of the whole model. Fan et al. [25] propose a one to multi-view auto-encoder clustering framework that learns node embeddings by reconstructing multiple graph views using the most informative graph view and attributes data, which can capture the shared feature representation of multiple graphs well. In addition, a self-training clustering objective is proposed to iteratively improve the clustering results. Good results are achieved by integrating self-training and auto-encoder reconstruction into a unified framework.

In summary, many researchers have worked on combining the powerful representation capabilities of both graph neural networks and heterogeneous information networks to solve deep clustering problems. With reference to the research experience, for the task of student group division in campus management scenarios, this paper uses heterogeneous information networks to model students' behaviors and attributes. Meanwhile, the attributes, behaviors and relationships of students are reasonably measured, and then the student group division problem is solved based on a multi-view deep clustering algorithm.

3. Our Method

3.1 Overall framework of the proposed method

In the university campus scenario, the proposed method is supposed to improve the performance of student group division by considering different attributes of students (e.g., major, hometown, ethnicity, etc.) and different association relationships among them. Due to the lack of well-labeled public datasets for student group division, we use multi-source educational big data and perform unsupervised deep clustering algorithm based on real campus data from a university to discover the intrinsic attributes and behavioral patterns on campus.

The overall framework of the proposed method is shown in **Fig. 1**, which consists of three main parts: the construction of multi-dimensional attribute features of individual students, the construction of multi-view student relationship networks, and the deep clustering module based on the multi-view heterogeneous information network. The general design of the proposed method is as follows.

Firstly, based on students' basic campus information and smart card swipe records, we represent the multi-dimensional attribute features of individual students with low-dimensional vectors. Then, we construct a multi-view student relationship network to comprehensively represent students' different relationships (co-majors, co-occurrence, co-borrowing books). Secondly, we construct a multi-view heterogeneous information network in order to model student attributes and multiple relationships in a complementary and integrated manner. Thirdly, an auto-encoder deep clustering algorithm based on multi-view graphs clusters students by reconstructing multiple network views using a single most informative network view and attribute information.

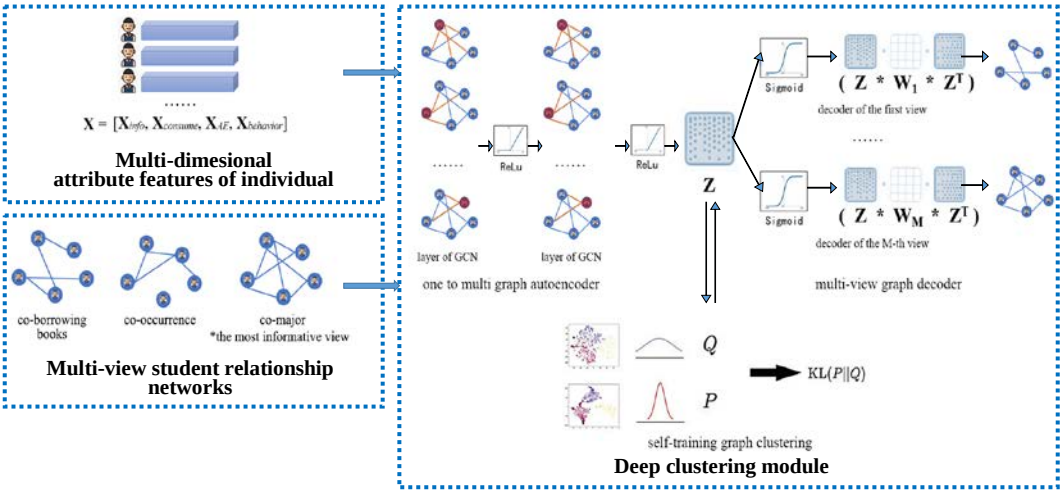


Fig. 1. The framework of the proposed method

3.2 Student behavior pattern learning

From the smart card records, the spatial-temporal information on student activity trajectories can be derived. As shown in **Fig. 2**, <*S*,*T*,*P*> represents a record with the student ID: *S*, the card using time: *T*, the card using position: *P*. It contains the semantic of student behaviors. For example, if value of *P* is cafeteria, a <*S1*,*T1*,*cafeteria*> represents student *S1* having meal at *cafeteria* at time *T1*, which reveals the activity of having a meal. For every activity using the card is recorded, record items in the smart card are highly intensive and repetitive. For example, student may use the smart card to buy the food several times during a short term. Therefore, how to use the spatial-temporal information effectively meanwhile to avoid the negative impact is still a challenge for accurately modeling student behavior patterns from the smart card records. To address this problem, we propose a heterogeneous network-based student behavior representation algorithm. The general framework of proposed network is shown in **Fig. 2**. The data <*S*,*T*,*P*> from student smart cards (SCR) records are generated as input. Then network is constructed calling the students' fine-grained spatial-temporal campus behavior heterogeneous multiple network (FSCB-HMN). The embedding of student behavior patterns is learned under the definition of co-occurrence meta-path and with random walk based meta-path and skip-gram in metapath2vec++. The details of construction of FSCB-HMN and co-occurrence meta-path-based embedding learning of student behavior patterns are illustrated as follows.

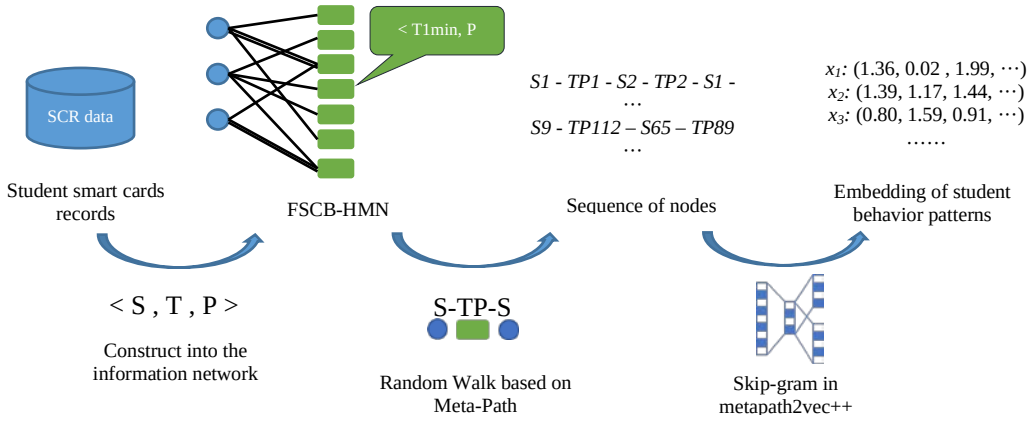


Fig. 2. The framework of the proposed method for student behavior pattern learning

3.2.1 Construction of FSCB-HMN

In order to model student behavior based on the smart card usage records, we propose a student behavior heterogeneous information network. The structure of student behavior heterogeneous information network is shown in **Fig. 3**. In the graph of FSCB-HMN, three kinds of nodes are defined, viz.: student nodes, time nodes, and location nodes. Edges are defined to indicate the relationship of visit or being visited. Meanwhile, student nodes are represented by student ID, time nodes are represented by year, month, day, and hour, and location nodes are represented by the names of locations within the campus. Each record of a student is transferred into three nodes and their associated edges as shown at top line in **Fig. 3**. For example, one type of edge is relationship, where the student node is connected with the time node, i.e. the edge between node student 1 with node 20190101,17 indicates that student 1 does the visit at time 17 on 2019/01/01. The other type of edge is relationship, where the time node is connected with the position node, i.e. the edge between time node 20190101,17 and position node First Canteen indicates that First Canteen is visited at time 17 on 2019/01/01. Then all students in dataset under processed are set in the same way. And nodes among students are connected based the time nodes happening at same time and position nodes happening at same position. In this way, the graph network of FSCB-HMN is constructed.

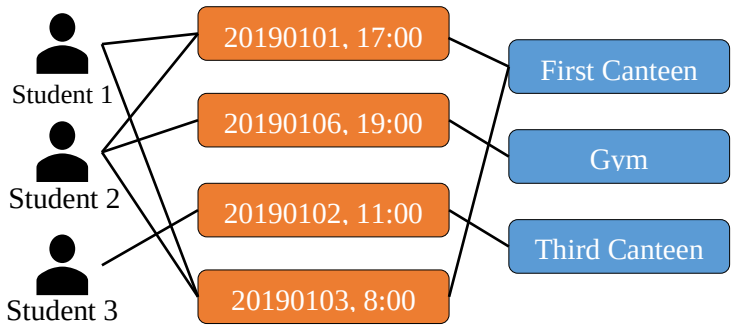


Fig. 3. Structure of student behavior heterogeneous information network

To address the challenge in distinguishable modeling intensive and high regular behaviors of students on campus, we proposed some methods including spatiotemporal dual features nodes, fine-grained temporal division method and multiple edge representation method.

Spatiotemporal dual features nodes: As shown in **Fig. 4**, most smart card records are created during mealtime: 7:00-9:00 am, 11:00-1:00 pm, and 5:00-7:00 pm, and nearly 70% of the smart card records are a combination of <meal time, dining place>. Therefore, students' smart card swipe behavior has the characteristics of time concentration and location density, which will lead to the problem that students' behavior cannot be accurately distinguished. Neither temporal features nor spatial features can independently represent the semantics of students' smart card swipe behavior. Therefore, we merge time nodes and location nodes into one kind of nodes, called spatiotemporal nodes.

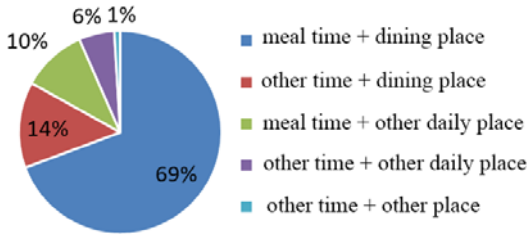


Fig. 4. Number of smart card swipes by students at different times and locations

Fine-grained temporal division method: The temporal information has different division schemes, and different granularities of temporal information will affect the accuracy of modeling. Due to the intensive and repetitive feature of swiping time, swiping behavior cannot be accurately distinguished from each other. Therefore, when modeling student swipe behavior, the model needs to be highly sensitive to temporal fine-grained information. Therefore we propose a fine-grained temporal node segmentation strategy. When representing time information, time is discretely divided according to year, month, day, hour and minute.

Multiple edge representation method: Since single swipe behavior is more contingent than multiple swipe behaviors, when modeling the student smart cards swiping behavior, the multiple swipe behavior semantics should be preserved. Therefore, we propose a multiple-edge representation method, i.e., multiple edges are allowed to appear between two nodes, as shown in **Fig. 5**.

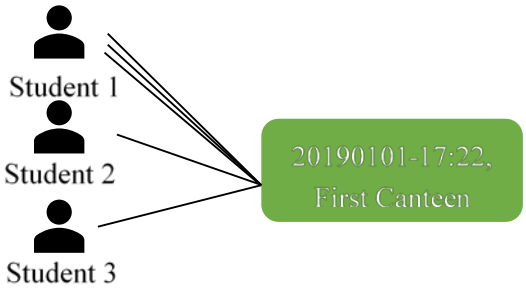


Fig. 5. Schematic Diagram of Multi Edge

Finally, we construct the network structure of student fine-grained Spatiotemporal campus behavioral heterogeneous multiple network (FSCB-HMN), as shown in **Fig. 6**. There are two types of nodes in FSCB-HMN (student node and fine-grained spatiotemporal node) one type of connection (the edge between student node and fine-grained spatiotemporal node is student swipe record connection relationship). There is no edge connection between student nodes and no edge connection between fine-grained spatiotemporal nodes.

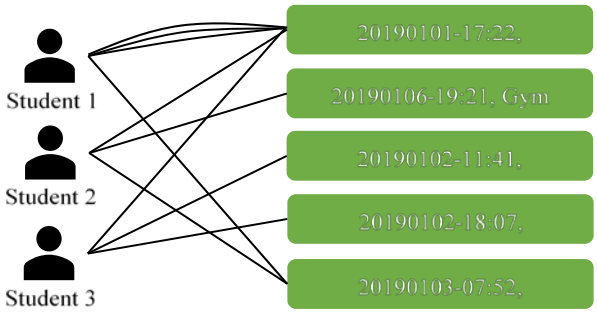


Fig. 6. Structure of FSCB-HMN

3.2.2 Co-occurrence meta-path-based embedding learning of student behavior patterns

Different meta-paths express different semantic relationships. Meta-path-based representation learning of heterogeneous information networks has been widely used in various data mining scenarios as an important approach for information network mining [26-28]. In this section, we first design a meta-path to reveal the co-occurrence relationship between students based on the smart card swipe behavior. Then, we learn the embedding of students' behavioral patterns based on the co-occurrence meta-path.

We propose the student co-occurrence relationship meta-path (*mp*) on FSCB-HMN. *mp* is denoted as: $v_s \xrightarrow{visit} v_{tp} \xrightarrow{visited} v_s$, abbreviated as *S-TP-S*, where v_s denotes a node of student type, v_{tp} denotes a node of fine-grained Spatiotemporal type. The *mp* path denotes a student node visiting a Spatiotemporal node and then visiting another student node, with the semantics that two students have the same smart card record in the same time and location, which is called "co-occurrence".

The student co-occurrence relationship meta-path *mp* can effectively capture the co-occurrence relationship between students and thus more realistically reflect the similarity of behavioral patterns among students. A sequence of nodes is obtained by the random walk algorithm with the guidance of *mp*. Then a low-dimensional vector embedding representation of the student nodes is learned based on the Skip Gram model. The learned embeddings have the feature that the more similar the behavioral patterns of two students are, the closer the metric distance of their embeddings in the vector space.

3.3 Construction of multi-dimensional attribute features of individual students

Considering that the result of student group division is affected by various integrated factors, we extract the multi-dimensional attribute features of individual students to construct student nodes.

In order to comprehensively represent the characteristics of students on campus, we extract many dimensions of student attribute features. The construction of multi-dimensional attribute features of individual students is shown in **Table 1**.

Table 1. Construction of multi-dimensional attribute features of individual students

Attribute features of individual students	Construction
Basic information feature	Gender, major, hometown
Consumption behavior feature	Consumption level and frequency
Lifestyle feature	Regularity of work and rest
Behavioral patterns feature	Behavior patterns based on the smart card records

The basic information features consist of structured data extracted directly from the basic information of the student's academic record. To convert the text corpus in structured data into word vectors, we use the Chinese word vector model [29] trained by the Word2Vec algorithm [30] to obtain the student basic attribute feature representation $\mathbf{X}_{info}$.

The consumption level was taken as the average consumption value of each student, and clustered into three levels of "high consumption", "medium consumption" and "low consumption" by the K-Means algorithm; the consumption frequency was taken as the number of swipes of each student and clustered into three levels of "high frequency", "medium frequency" and "low frequency" by the K-Means algorithm.We use the Chinese word vector model to convert the text into word vectors to obtain $\mathbf{X}_{consume}$, an attribute feature representation of students' consumption behavior.

It is not feasible to analyze the behavioral pattern of students directly on the smart card records, because the records record each student's card swipe in chronological order. We uses the concept of entropy to propose "activity entropy" to quantitatively evaluate the degree of regularity of students' activities, which mainly reflects the regularity of swipe time and location. Based on this, the Spatiotemporal information of students' smart card swiping behavior is represented as a discrete-time series $s(t_i, p_k)$, and the Spatiotemporal information of students is determined by the random variables in the series corresponding to time point t_i and activity location p_k. The probability values of the random variables are calculated to obtain the corresponding entropy values. The activity entropy is defined as Eq. (1):

$$AE(u) = -\sum_{t_i \in T} \sum_{p_k \in P} [P(s(t_i, p_k)) \times \log P(s(t_i, p_k))] \tag{1}$$

Where $P(s(t_i, p_k))$ is the empirical probability distribution of the location p_k of the activity corresponding to student u in time slot t_i. The greater the activity entropy, the more disorderly the students' activities are and the more diffuse the time and location of the smart card swiping bahavior. The lifestyle attribute feature is represented by the student's activity entropy, denoted as $\mathbf{X}_{AE}$.

The behavioral patterns feature is obtained by the method proposed in the section 3.2, denoted as $\mathbf{X}_{behavior}$.

The multi-dimensional attribute features of individual students consist of a fusion of the four aspects of the above-mentioned features, as Eq. (2):

$$\mathbf{X} = g(\mathbf{X}_{info}, \mathbf{X}_{consume}, \mathbf{X}_{AE}, \mathbf{X}_{behavior}) = [\mathbf{X}_{info}, \mathbf{X}_{consume}, \mathbf{X}_{AE}, \mathbf{X}_{behavior}] \tag{2}$$

Where $g(\cdot)$ denotes the fusion method, $[\cdot, \cdot]$ denotes the concatenation of vectors.

3.4 Construction of multi-view student relationship network

Students have various relationships with other students on campus, and these relationships influence student group division on campus. We propose a multi-view student relationship network consisting of several networks, including a co-major student relationship network $G_z = (V, E_z)$ based on basic information, a co-occurrence student relationship network $G_s = (V, E_s)$ based on behavioral pattern embedding, and a co-borrowing books relationship network $G_b = (V, E_b)$. The node sets V in these three networks are all consist of the same student nodes, while the edge sets E_z, E_s, and E_b are constructed from different relationships.

In the campus scenario, the curriculum of the same majors tends to be close, leading to a convergence in the activity habits of students with the same majors on campus, which has a great impact on students' social relationships. Therefore, we construct the co-major student relationship network $G_z = (V, E_z)$ based on the major information. If two students belong to the same major, the student nodes have a connection relationship with each other, otherwise

they are not connected.

Based on the smart card records, the students' behavioral patterns embeddings representation $X_{behavior}$ is obtained using the representation learning algorithm in section 3.2. The most similar student of a certain student is calculated using Annoy algorithm. The most similar student can be called the student's "consumption friend". The student co-occurrence relationship network $G_s = (V, E_s)$ is constructed based on this relationship. If two students are "consumption friends", there is a connection between them, otherwise there is no connection.

Based on students' book borrowing information, students with the same borrowed books are directly connected to construct a co-borrowing books relationship network $G_b = (V, E_b)$.

3.5 Student group division based on multi-view heterogeneous information network

A multi-view student attributes heterogeneous information network (MSA-HIN) is constructed in this section considering three types of student relationships, denoted as $G = (V, E_1, E_2, E_3, \mathbf{X})$, where $V = \{v_i\}^n_{i=1}$ consists of n student nodes, and E_1, E_2, E_3 correspond to the edge sets E_z, E_s, E_b of the co-major relationship network G_z, the co-occurrence relationship network G_s, and the co-borrowing books relationship network G_b. Each student relationship is a network view, and $e^{(m)}_{i,j} \in E_m$ denotes the connection between nodes i and j in the *m-th* view. $x_i \in \mathbf{X}$ denotes the connection associated with each student node v_i attribute value.

Based on the MSA-HIN, we use the O2MAC model to implement the multi-view student group division task. O2MAC model learns node embedding by reconstructing multiple graph views using the most informative graph view and attributes data, which can well capture the shared feature representation of multiple graph views. The model consists of two main components: a multi-view graph auto-encoder and self-training graph clustering.

3.5.1 Multi-view graph auto-encoder

Multi-view graph auto-encoder consists of two components: one to multi-view graph auto-encoder and a multi-view graph decoder.

The most informative graph view $A^* \in \{A(1), A(2), ..., A(M)\}$ and node attribute information X are input to the one to multi-view graph auto-encoder, and the node representation is learned from the most informative view A* by a multilayer graph convolutional network encoder. The most informative graph view can be selected based on the prior knowledge, or modularity metrics. To utilize both the graph structure $A*$ and the structured node attribute $\mathbf{X}$ in a unified framework, the graph convolution network layer extends the convolution operation to the graph data in the spectral domain and learns the hierarchical transformation through the spectral convolution function $f(\mathbf{Z}^{(l)}, \mathbf{A}^* \mid \mathbf{W}^{(l)})$, which is computed in Eq. (3):

$$\mathbf{Z}^{(l+1)} = f\left(\mathbf{Z}^{(l)}, \mathbf{A}^* \middle| \mathbf{W}^{(l)}\right) = \phi\left(\widetilde{\mathbf{D}}^{-1/2}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-1/2}\mathbf{Z}^{(l)}\mathbf{W}^{(l)}\right) \tag{3}$$

where $\phi(\cdot)$ is the activation function, $\mathbf{Z}^{(l)}$ is the representation learned at layer l, $\mathbf{Z}^{(0)} = \mathbf{X} \in \mathbf{R}^{N \times D}$, and $\mathbf{W}^{(l)}$ is the filter parameter matrix learned at layer l, where $\widetilde{\mathbf{A}} = \mathbf{A}^* + \mathbf{I}$, $\widetilde{\mathbf{D}}_{ii} = \sum_j \widetilde{\mathbf{A}}_{ij}$, $\mathbf{I}$ is the unit matrix of $\mathbf{A}^*$.

The multilayer graph convolutional network encoder is two graph convolutional layers, calculated in Eq. (4):

$$\mathbf{Z} = \phi_2(\widetilde{\mathbf{D}}^{-1/2}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-1/2}\phi_1(\widetilde{\mathbf{D}}^{-1/2}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-1/2}\mathbf{X}\mathbf{W}^{(0)})\mathbf{W}^{(l)}) \tag{4}$$

where the activation function $\phi_1(\cdot)$ is the ReLu function and the activation function $\phi_2(\cdot)$ is the linear activation function.

To supervise the encoder to extract shared representations of all views, a multi-view graph decoder is proposed to reconstruct the multi-view data $\widehat{A}^{(1)}$, …, $\widehat{A}^{(M)}$ from the representation **Z**. The decoder consists of *M* view-specific decoders for predicting the existence of a link between two nodes in view *m*. The multi-view connection prediction layer is trained based on graph embedding in Eq. (5):

$$\sum_{m=1}^{M} p(\widehat{\mathbf{A}}^{(m)} \mid \mathbf{Z}, \mathbf{W}_m) = \sum_{m=1}^{M} \text{Sigmoid}(\mathbf{Z}\,\mathbf{W}_m\,\mathbf{Z}^T) \tag{5}$$

where $\boldsymbol{W}_m \in \mathbf{R}^{D\times D}$ is the view-specific weight of view *m*.

For the multi-view graph autoencoder, the sum of reconstruction errors for each graph view is minimized by the reconstruction loss L_r, as Eq. (6):

$$L_r = \sum_{m=1}^{M} L_r^{(m)} = \sum_{m=1}^{M} loss\left(\mathbf{A}^{(m)}, \widehat{\mathbf{A}}^{(m)}\right) \tag{6}$$

where $L_r^{(m)}$ is the reconstruction loss of view *m* and L_r is the reconstruction loss of all views. Due to the structure of the multi-view graph decoder, the gradients of the multi-decoder are propagated through the informative graph encoder during the backward propagation. Therefore, when the forward propagation is processed, the graph encoder will extract the shared representation of all views.

3.5.2 Self-training graph clustering

The aforementioned multi-view graph auto-encoder can encode multi-view graph structure data with attribute information into a compact representation in the low-dimensional embedding space. However, there is a problem: the similarity of distances between nodes is due to maintaining the local structure of the original multi-view graph structure data, and such an embedding representation may not be adaptable to the clustering task. Therefore, the self-training clustering objective module from previous studies is extended to iteratively improve the clustering results, and the self-training clustering objective loss function L_c is calculated in Eq. (7):

$$L_c = \text{KL}(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{7}$$

where $\text{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence between two distributions, *Q* is the distribution of soft labels, and q_{ij} is measured by Student-t to represent the similarity between the embedding h_i of node *i* and the clustering cluster center μ_j, which can be viewed as the soft clustering assignment of each node. q_{ij} is calculated as specified in Eq. (8):

$$q_{ij} = \frac{(1+\| h_i - \mu_j \|^2 / \upsilon)^{-\frac{\upsilon+1}{2}}}{\sum_{j'} (1+\| h_i - \mu_{j'} \|^2 / \upsilon)^{-\frac{\upsilon+1}{2}}} \tag{8}$$

where υ is the degree of freedom of the Student-t distribution. p_{ij} is the target distribution and its calculation is specified in Eq. (9):

$$p_{ij} = \frac{q_{ij}^{\,2} / f_j}{q_{ij'}^{\,2} / f_{j'}} \tag{9}$$

Where f_j is the soft cluster frequency used to normalize the loss contribution of each clustering center to prevent large clusters from distorting the hidden feature space. The target distribution *P* raises *Q* to quadratic, resulting in a denser distribution. The distribution of *Q* can be made denser by minimizing the KL scatter between *Q* and *P*.

Finally, jointly optimizing the learning of the multigraph convolutional encoder as well as the self-training clustering objective, the total loss function is shown in Eq. (10):

$$L = L_r + \lambda L_c \tag{10}$$

where λ is a coefficient controlling the degree of distortion of the embedding space.

4. Experiments

4.1 Dataset

Our experiment uses the data of 759 students in a university in China during the period from February 26, 2018 to July 9, 2018. The dataset is denoted as SCR2018, which mainly contains students' basic information, academic performance information, smart card records, and library book borrowing information.

(1) Basic information: It contains students' basic attributes such as student ID, gender, and major.

(2) Academic performance information: It contains student ID, students' GPA average scores and the number of elective courses.

(3) Smart card records: It contains student ID, card swiping time, card swiping location, and card amount. The experiment uses 261,359 smart card records generated by 759 students. The raw data is saved as the triad format, like <student(s), time(t), location(p)>.

(4) Library book borrowing information: It contains student ID, book borrowing time, book returning time, book title.

4.2 Evaluation metric

Due to the lack of labeling of students, we use the Silhouette Coefficient as the metric to evaluate the clustering results.

Silhouette Coefficient, SC: A clustering evaluation metric that takes both cohesion and separation of clustering results into account, which is calculated by Eq. (11) and (12).

$$SC = \frac{1}{N}\sum_{i=1}^{N} S(i) \tag{11}$$

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{12}$$

where $a(i)$ is the average distance between the representation of sample i and other samples' in its cluster, and $b(i)$ is the minimum of the average distance between the representation of sample i and other samples' in a cluster that does not contain it. $S(i)$ is the silhouette coefficient of sample i. The total silhouette coefficient SC of the clustering result is obtained by averaging the silhouette coefficients of all samples.

4.3 Comparative experiments

The proposed method is compared with the following three algorithms.

(1) DeepWalk-avg [31]: a network representation learning method that combines random walk and Skip Gram models. Since the heterogeneity of the network is ignored, no distinction is made between node semantics and types when performing random walks.

(2) GAE-avg [32]: an unsupervised single-view graph auto-encoder embedding algorithm based on graph convolutional neural networks for learning data representations.

(3) SDCN-avg [24]: an unsupervised single-view graph auto-encoder embedding algorithm for structured deep clustering networks.

Since both GAE and SDCN are single-view homogeneous information network methods, a multi-view down-averaging operation is used to convert multi-view network representation learning to single-view network representation learning when multi-view network clustering is performed.

The comparison experiments of the student group division method based on the multi-view heterogeneous information network with the other three algorithms are conducted on the dataset SCR2018. **Fig. 7** shows the results of the comparison experiments.

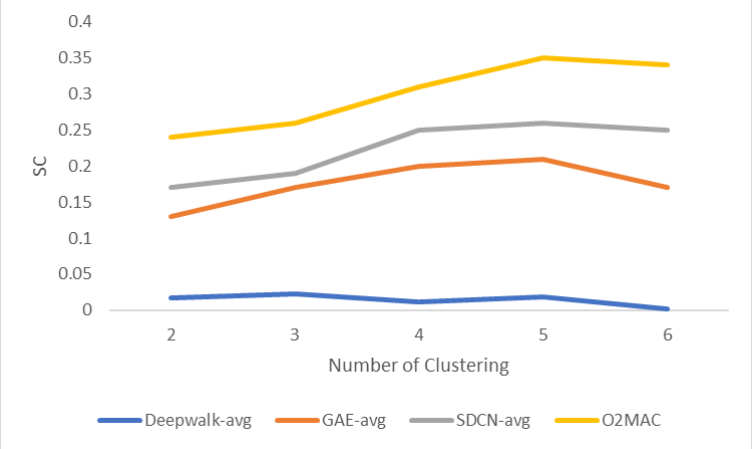


Fig. 7. Comparative experimental results

Analysis of the results leads to the following conclusions.

(1) The O2MAC method has the best clustering effect under the metric. DeepWalk-avg method has the worst clustering effect and can barely distinguish samples from each other. The reason is that DeepWalk does not have the ability to extract sample attribute embedding features. GAE-avg method and SDCN-avg method both can handle node attribute features in the network and have a self-training clustering module, so they have certain clustering ability, but the effect is not as good as O2MAC.

(2) The O2MAC method achieves better results compared with other algorithms because it can fuse multi-perspective information for representation learning; however, the value of its metric is not too high. There are two possible reasons. First, the node attribute features and multiple association relationships between nodes in the multi-perspective heterogeneous information network are artificially defined and extracted, and there is a limitation of the comprehensive description of students' on-campus attributes and multiple relationships description. Second, in the real college scenario, the variability of college students' activities is inherently small, such as the student card swiping behavior that has a greater impact in the division of student groups, and most students tend to go to the cafeteria during meal times, leaving regular and strong data.

The students were finally chosen to be divided into five groups according to the metrics, and the results of the statistical analysis of student behavior in each group are shown in **Table 2**.

Table 2. Behavior statistics of each group

Cluster ID	Percentage of people	Average spending	The standard deviation of average spending	Average Swipe Times	The standard deviation of Average Swipe Times	Average GPA	The standard deviation of GPA	Average AE	The standard deviation of AE
1	7.25%	5.29	11.63	354.03	160.71	3.40	0.49	0.29	0.07
2	8.30%	5.14	3.11	346.98	169.24	3.47	0.4	0.44	0.08
3	6.46%	5.39	3.31	344.34	158.54	3.54	0.44	0.59	0.23
4	57.70%	5.33	5.56	351.71	167.29	3.51	0.45	0.53	0.09
5	20.29%	5.31	9.46	318.86	167.63	3.52	0.33	0.62	0.21

As shown in **Table 2**, the analysis of the statistical characteristics of each group led to the following conclusions.

(1) The greater difference between groups of students is activity entropy, indicating that the proposed method is more influenced by students' swiping behavior characteristics in clustering and has the ability to distinguish differences in behavior characteristics.

(2) The differences in average spending amount, average number of swipes, and GPA among students in each group are small, indicating that the proposed method does not have the ability to discriminate between these characteristics.

The data set of 759 students in this chapter was divided into 5 groups with the following characteristics.

Cluster ID1 group is a small group of students who live a regular life but have fluctuating spending amounts.

Cluster ID2 group is a small group of students who live a regular life but have no other significant characteristics.

Cluster ID3 group is a small group of students who live an irregular life with large fluctuations.

Cluster ID4 group is the majority of students who have an irregular life but with less fluctuation.

Cluster ID5 group is a moderate number of students who have irregular lives with large fluctuations and fluctuating consumer behavior.

Table 3. Statistics of gender, major, and books borrowed by each group

Cluster ID	Male to female ratio	Major Composition	Borrowing books
1	Male 58.18%, Female 41.82%	10 majors involved	72 books involved
2	Male 46.03%, Female 53.97%	9 majors involved	114 books involved
3	Male 51.02%, Female 48.98%	10 majors involved	60 books involved
4	Male 47.72%, Female 52.28%	14 majors involved	183 books involved
5	Male 53.90%, Female 46.10%	14 majors involved	170 books involved

Table 4. Population distribution of each group in each major

Major ID / ClusterID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	14	2	2	2	5	2	4	5	0	0	18	0	0
2	5	1	3	1	0	0	0	1	0	2	45	0	4	1
3	1	2	0	30	1	0	0	1	2	9	1	0	1	1
4	5	21	15	18	18	12	4	8	19	6	17	9	171	115
5	3	7	8	4	5	1	3	4	19	52	8	14	2	24

As shown in **Table 3** and **Table 4**, the analysis of statistics on the ratio of male to female students, the composition of their majors and the books borrowed by each group led to the following conclusions.

(1) The male to female ratio of all students in dataset SCR2018 is 50.2% male and 49.8% female, which is close to 1:1, while the distribution of male to female ratio varies among groups, indicating that the method in this chapter is not random clustering, but clustering based on the extracted comprehensive features, but the influence of male and female gender on the clustering results is not obvious.

(2) There are 14 majors involved in the dataset SCR2018, and the distribution is not balanced. The distribution of specific majors is shown in **Table 4**, in which the number of people in cluster ID1 group from major 2 and major 12 accounts for more than 58%; the number of people in cluster ID2 group from major 11 accounts for more than 71%; the number of people in cluster ID3 group from major 4 accounts for more than 61%; the number of people in cluster ID4 group from major 13 and major 14 accounts for more than 65%; the number of people in cluster ID5 group from each The number of people from major 9 is over 33%. It is not difficult to find that the distribution of majors in different clusters is different, which indicates that the proposed method in this chapter is more influenced by the common major relationship of students when clustering.

5. Conclusions and Future Work

To make full use of students' comprehensive attributes and overcome the problem of student group division under complex relationships, we transform the problem of student group classification based on multi-source campus data into a deep clustering problem based on a multi-view heterogeneous information network. Firstly, we construct multi-dimensional attribute features of individual students based on their basic registration information and smart card records, and establish effective attribute feature representation of students' performance on campus as the attribute information of student nodes in the student multi-view attribute heterogeneous information network. Secondly, we construct different student relationship networks determined by students' multiple association relationships as the multiple graph views in the student multi-view attribute heterogeneous information network. Finally, we evaluate and analyze the reasonableness of the clustering results based on the multi-view graph auto-encoder deep clustering algorithm for deep clustering of students, which verifies the effectiveness of the algorithm and also shows that the proposed method can help provide a picture of student performance in future applications and provide a more objective basis for student management through group division.

Future research should be further carried out in the following aspects to explore the generalization and scalability of the proposed method.

(1) Constrained by the privacy protection of educational big data information, this paper only uses a limited scale of students’ smart card records and student registration information, and future research should further expand the scale of data in time and space to address the problem of limited data volume, expand data sources based on increasing data diversity, add student academic-related data, mental health assessment data, student poverty In the future, we should expand data sources based on data diversity, add data related to students' academic performance, psychological health assessment data, and students' poverty data, and study students' behavioral characteristics in terms of academic performance, psychological status, and economic difficulties.

(2) To address the problems of student behavior modeling and student attribute feature characterization, future research can be oriented toward the improvement of the embedding learning method for heterogeneous information networks to make the network model more powerful in characterization based on more diverse information that can be described by the heterogeneous information network. If all the above data-level attempts can be executed, we can try to overcome the limitations of the meta-path random wandering-based embedding learning algorithm and introduce mechanisms such as long and short-term attention, self-supervision, and reinforcement learning to explore more advanced and robust embedded learning methods for heterogeneous information networks.

Acknowledgment

This work is supported by Beijing Natural Science Foundation (No.4202004). Consulting Project for Major Strategic Decision making for Serving Capital in 2022 from Beijing University of Technology.

References

[1] H. Yao, M. Nie, H. Su, H. Xia, D. Lian, "Predicting academic performance via semi-supervised learning with constructed campus social network," in *Proc. of International Conference on Database Systems for Advanced Applications*, Springer, Cham, pp. 597-609, 2017. Article (CrossRef Link)

[2] Q. Hu, H. Rangwala, "Academic Performance Estimation with Attention-based Graph Convolutional Networks," in *Proc. of the 12th International Educational Data Mining Society*, Montreal, Canada, pp. 69-78, 2019. Article (CrossRef Link)

[3] H. Li, H. Wei, Y. Wang, Y. Song, H. Qu, "Peer-inspired Student Performance Prediction in Interactive Online Question Pools with Graph Neural Network," in *Proc. of the 29th ACM International Conference on Information and Knowledge Management*, Virtual Event, Ireland, pp. 2589–2596, 2020. Article (CrossRef Link)

[4] B. Sekeroglu, K. Dimililer, K. Tuncal, "Student Performance Prediction and Classification Using Machine Learning Algorithms," in *Proc. of the 2019 8th International Conference on Educational and Information Technology*, Cambridge, United Kingdom, pp. 7–11, 2019. Article (CrossRef Link)

[5] J.-L. Hung, M.-C. Wang, S. Wang, M. Abdelrasoul, Y. Li, H. Wu, "Identifying At-Risk Students for Early Interventions—A Time-Series Clustering Approach," *IEEE Transactions on Emerging Topics in Computing*, vol. 5, no. 1, pp. 45-55, 2017. Article (CrossRef Link)

[6] A. Moubayed, M. Injadat, A. Shami, H. Lutfiyya, "Relationship Between Student Engagement and Performance in E-Learning Environment Using Association Rules," in *Proc. of 2018 IEEE World Engineering Education Conference*, Buenos Aires, Argentina, pp. 1-6, 2018. Article (CrossRef Link)

[7] M. Zhou, D. Yang, "Research progress on educational data mining: A survey," *Journal of Software*, vol. 26, no. 11, pp. 3026−3042, 2015. Article (CrossRef Link)

[8] M. McCord, M. Chuah, "Spam detection on twitter using traditional classifiers," in *Proc. of the 8th International Conference on Autonomic and Trusted Computing*, Springer, Berlin, Heidelberg, pp. 175–186, 2011. Article (CrossRef Link)

[9] D. Bunić, I. Jugo, B. Kovačić, "Analysis of clustering algorithms for group discovery in a web-based intelligent tutoring system," in *Proc. of 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia pp. 759-765, 2019. Article (CrossRef Link)

[10] M. Xie, H. Yin, H. Wang, F. Xu, W. Chen, S. Wang, "Learning Graph-based POI Embedding for Location-based Recommendation," in *Proc. of the 25th ACM International on Conference on Information and Knowledge Management*, Indianapolis, Indiana, USA, pp.15–24, 2016. Article (CrossRef Link)
[11] Z. Wang, H. Liu, Y. Du, Z. Wu, X. Zhang, "Unified embedding model over heterogeneous information network for personalized recommendation," in *Proc. of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, pp. 3813-3819, 2019. Article (CrossRef Link)
[12] C. David, B. Cécile, P. François, "Integrating heterogeneous information within a social network for detecting communities," in *Proc. of 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, Niagara Falls, ON, Canada, pp.1453-1454, 2013. Article (CrossRef Link)
[13] Y. Sun, J. Han, "Mining heterogeneous information networks: a structural analysis approach," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 20-28, 2013. Article (CrossRef Link)
[14] C. Shi, Y. Li, J. Zhang, Y. Sun, P. Yu, "A Survey of Heterogeneous Information Network Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17-37, 2015. Article (CrossRef Link)
[15] C. Shi, B. Hu, W.-X. Zhao, P. Yu, "Heterogeneous information network embedding for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 357-370, 2019. Article (CrossRef Link)
[16] J. Gong, S. Wang, J. Wang, et al, "Attentional Graph Convolutional Networks for Knowledge Concept Recommendation in MOOCs in a Heterogeneous View," in *Proc. of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event, China, pp. 79–88, 2020. Article (CrossRef Link)
[17] X. Zhang, H. Zheng, X. Li, S. Du, H. Zhu, "You are where you have been: Sybil detection via geo-location analysis in OSNs," in *Proc. of 2014 IEEE Global Communications Conference*, Austin, TX, USA, pp. 698-703, 2014. Article (CrossRef Link)
[18] T. Sona, B. Asgarali, "Community detection in social networks using affinity propagation with adaptive similarity matrix," *Big Data*, vol. 8, no. 3, pp. 189-202, 2020. Article (CrossRef Link)
[19] M. -E.-J. Newman, M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, pp. 026113, 2004. Article (CrossRef Link)
[20] K. Thomas, W. Max, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv: 1609.02907*, 2016.
[21] A. Sankar, Y. Liu, J. Yu, N. Shah, "Graph Neural Networks for Friend Ranking in Large-scale Social Platforms," in *Proc. of the Web Conference 2021*, Ljubljana, Slovenia, pp. 2535–2546, 2021. Article (CrossRef Link)
[22] Y. Zhang, Y. Xiong, X. Kong, Z. Niu, Y. Zhu, "IGE+: A Framework for Learning Node Embeddings in Interaction Graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 3, pp. 1032-1044, 2021.Article (CrossRef Link)
[23] C. Wang, S. Pan, R. Hu, G. Long, J. Jiang, C. Zhang, "Attributed Graph Clustering: A Deep Attentional Embedding Approach," in *Proc. of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, pp. 3670-3676, 2019. Article (CrossRef Link)
[24] D. Bo, X. Wang, C. Shi, M. Zhu, E. Lu, P. Cui, "Structural Deep Clustering Network," in *Proc. of the Web Conference 2020*, Taipei, Taiwan, China, pp. 1400–1410, 2020. Article (CrossRef Link)
[25] S. Fan, X. Wang, C. Shi, E. Lu, K. Lin, B. Wang, "One2Multi Graph Autoencoder for Multi-view Graph Clustering," in *Proc. of the Web Conference 2020*, Taipei, Taiwan, China, pp. 3070–3076, 2020. Article (CrossRef Link)
[26] X. Kong, P.S. Yu, Y. Ding, D.-J. Wild, "Meta path-based collective classification in heterogeneous information networks," in *Proc. of the 21st ACM International Conference on Information and Knowledge Management*, Maui, Hawaii, USA, pp. 1567-1571, 2012. Article (CrossRef Link)
[27] Y. Dong, N.-V. Chawla, A. Swami, "Metapath2vec: Scalable Representation Learning for Heterogeneous Networks," in *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, pp. 135–144, 2017. Article (CrossRef Link)

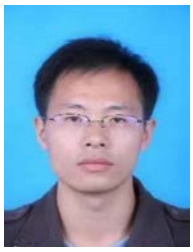
[28] S. Fan, C. Shi, L. Hu, B. Ma, Y. Li, "Metapath-guided heterogeneous graph neural network for intent recommendation," in *Proc. of 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, AK, USA, pp. 2478-2486, 2019. Article (CrossRef Link)
[29] S. Li Shen, Z. Zhao, R. Hu, W. Li, T. Liu, X. Du, "Analogical Reasoning on Chinese Morphological and Semantic Relations," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia, pp. 138-143, 2018. Article (CrossRef Link)
[30] Q. Le, T. Mikolov, "Distributed representations of sentences and documents," in *Proc. of the 31st International Conference on Machine Learning*, Beijing, China, pp.1188–1196, 2014. Article (CrossRef Link)
[31] B. Perozzi, R.-A. Rfou, S. Skena, "Deepwalk: Online learning of social representations," in *Proc. of 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, pp. 701-710, 2014. Article (CrossRef Link)
[32] T.-N. Kipf, M. Welling, "Variational graph auto-encoders," *arXiv preprint arXiv: 1611.07308*, 2016.

Xibin Jia received the B.S. degree in wireless technology from Chongqing University, Chongqing, China in 1991, received the M.S. degree in intelligent instrument from North China Institute of Technology in 1996 and the Ph.D. degree in computer science and technology from Beijing University of Technology, Beijing, China, in 2007. Now, she is a Professor in the Faculty of Information at the Beijing University of Technology (BJUT) in Beijing, China.

Zijia Lu received the B.S. degree in Beijing University of Technology, Beijing, China in 2019. She is currently pursuing a master degree at Beijing University of Technology (BJUT), Beijing, China. Her current research interests include educational data mining and representation learning.

Qing Mi received the B.S. degree in network engineering from Hebei University, Hebei, China in 2009, received the M.S. degree in computer science and technology from the Beijing Institute of Technology in 2012 and the Ph.D. degree in software engineering from the City University of Hong Kong, Hong Kong, China, in 2018. Now,she is a lecturer in the Faculty of Information Technology, Beijing University of Technology, Beijing, China. Her research interests include code readability assessment, data mining and analytics, deep learning and empirical experiments.

Zhefeng An received the Ph.D. degree in psychology from the Beijing Normal University. Now he is a professor in the Faculty of Humanities and Social Sciences, Beijing University of Technology, Beijing, China. His research interests include modern educational technology, psychological measurement and assessment, student psychological development and education.



Xiaoyong Li received the M.S. degree in computer science from the Beijing University of Technology, Beijing, China, in 2009, where he is currently pursuing the Ph.D. degree in computer science with the Beijing Key Laboratory of Multimedia and Intelligent Software Technology. He has been a senior engineer with the Information Technology Support Center, Beijing University of Technology. His research interests include big data analysis and visualization.

Min Hong is a professor of Department of Computer Software Engineering at Soonchunhyang University. He received his B.S. from Soonchunhyang University, M.S. from University of Colorado at Boulder, and Ph.D. received from University of Colorado at Denver and Health Sciences Center in 1995, 2001, and 2005 respectively. His research interests are in computer graphics, physically-based modeling and simulation, bioinformatics, and image processing related applications. Currently he is the Director of Computer Graphics Laboratory.