Online-Journals.org

iJET International Journal of Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 18 No. 21 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijet.v18i21.44687

PAPER

A Neural Networks Based Model to Predict the Interest of College Students in Sports Activities

Xiuzu Xiong(✉)

Physical Education Department, Guizhou University of Commerce, Guiyang, China

201720625@gzcc.edu.cn

ABSTRACT

The widespread application of big data technology in various fields, including research in education and sports in colleges, has also been deeply influenced. College students are the future strength of a country, and their habits and interests in sports activities have profound significance for their physical and mental health, teamwork, and outlook on life. However, traditional research methods, such as questionnaire surveys, observations, or interviews, have obvious limitations when dealing with large amounts of complex high-dimensional data. This study aimed to extract the interesting features of college students regarding sports activities using graph neural network (GNN) technology. Then, the labels of those interest features were further predicted, and a feature matrix was constructed. Finally, the K-means clustering method was used to achieve accurate feature clustering. This study presents a novel idea and approach for physical education and event planning in colleges, offering both practical value and theoretical significance.

KEYWORDS

big data, college students, sports activities, interest features, graph neural network (GNN), K-means clustering

1 INTRODUCTION

With the rapid development of big data technology and the continuous deepening of social informatization, a large amount of data has been collected, stored, and analyzed. In the field of education and sports, the data on sports activities among college students has become a valuable resource for educational research and decision-making [1–4]. College students are the future of a country, and their habits and interests in sports activities have a profound impact on physical and mental health, outlook on life, teamwork, and other aspects [5, 6]. Therefore, conducting comprehensive research on the interests and preferences of college students regarding sports activities, as well as their clustering patterns, holds significant practical and theoretical value [7–11].

Xiong, X. (2023). A Neural Networks Based Model to Predict the Interest of College Students in Sports Activities. *International Journal of Emerging Technologies in Learning (iJET)*, 18(21), pp. 113–128. https://doi.org/10.3991/ijet.v18i21.44687

Article submitted 2023-08-03. Revision uploaded 2023-09-07. Final acceptance 2023-09-07.

The research on the interests and preferences of college students regarding sports activities helps colleges and other educational institutions gain a better understanding of the sports needs and interests of these students. This, in turn, enables them to provide more accurate and personalized sports course designs and sports activity planning [12, 13]. In addition, understanding their interests in sports activities also provides guidance for their health and development and promotes the enhancement of their physical fitness and their long-term passion for sports [14–17]. Such research also provides data support for educational decision-makers, making the allocation of educational resources more reasonable and efficient.

Although research on college students' interests in sports activities has been ongoing for some time, traditional research methods often rely on questionnaire surveys, observations, or interviews. However, these methods may encounter difficulties when dealing with large amounts of data, high dimensions, and increased complexity [18, 19]. Moreover, traditional methods struggle to accurately capture and analyze subtle individual differences and deeper feature associations. In addition, the subjectivity of these methods may also lead to bias and instability in the results [20, 21].

This study focused on extracting the features of interest among college students regarding sports activities using a graph neural network (GNN). Through deep learning of the nodes and edges in the network, the internal structure and correlation of their interesting features were revealed. Then the labels of those interest features were further predicted, and an interest feature matrix was constructed, laying the foundation for subsequent analysis. Finally, using the K-means clustering method, we were able to achieve precise clustering of the interested features. This approach aims to provide more scientific, systematic, and targeted suggestions for planning physical education and sports activities for college students. This study not only promotes theoretical research on college students' interests in sports activities but also brings great value and enlightenment to practice.

2 INTEREST FEATURE EXTRACTION OF COLLEGE STUDENTS REGARDING SPORTS ACTIVITIES

2.1 Constructing the heterogeneous graph of sports activities

Before extracting interest features of college students concerning sports activities using GNN, the construction of a heterogeneous graph played a crucial role in ensuring the accuracy of feature extraction. The graph contained various types of nodes and edges, such as students, sports events, facilities, and time. This is a comprehensive representation of sports activity data for college students. This multidimensional information representation helps capture and reflect the complexity and diversity of those sports activities, providing rich input for subsequent feature extraction. The graph structure reveals the relationships between different types of nodes, such as students and their favorite sports events, commonly used facilities, etc. These relationships are crucial for understanding the interests and preferences of the students. GNN utilized these correlations for deep learning and extracted representative features. Figure 1 shows the framework of the GNN model.

Under the assumption that the time range is determined, let Z represent the sports activity data within the specified time range. Let $z_{i_u} = \{(q_e, q_a, \ldots, q_y) \mid e, a, \ldots, y \in [1, b]\}$ be the sports activity sequence of college student i_u, where $z_{i_u} \in Z$. Compared to a homogeneous graph, the heterogeneous graph represents more complex patterns.

For example, a student may be interested in both basketball and volleyball, but they may prefer playing basketball indoors and volleyball outdoors. This complex interest pattern was accurately represented by the heterogeneous graph, which provided more learning information for graph neural network.

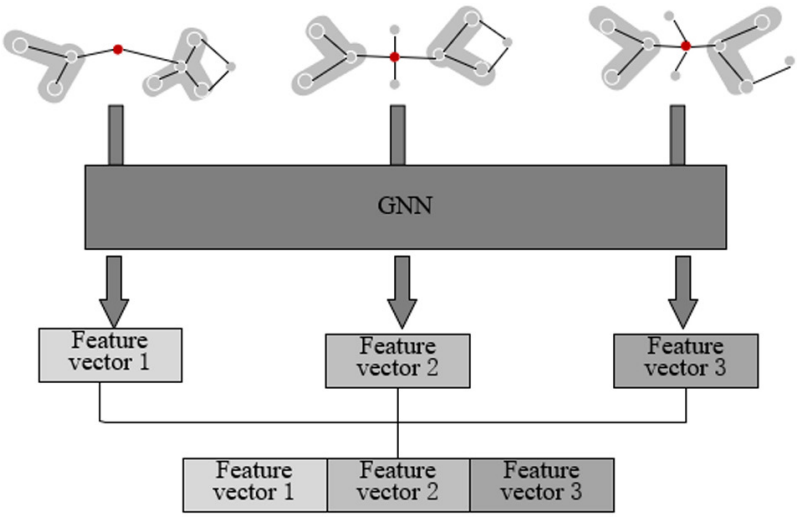

Fig. 1. Framework of GNN model

For college sports activities, the construction steps of the heterogeneous graph are introduced in detail as follows:

First, the types of nodes in the network were defined, such as students, sports events (e.g., basketball, badminton, etc.), and learning resources (e.g., tutorials, training videos, etc.). Let $b_{(iu)} = \{(q_e, q_a, \ldots, q_y) \mid e, a, \ldots, y \in [1, b]\}$ be the neighbor node set of the college student i_u. Based on the sequence of sports activities participated, edges were further established to connect different nodes. For example, when a student watches a basketball tutorial, a connection is established between the student and the tutorial. In the heterogeneous graph H_j of sports activities, there are $2J$ hop neighbor relationships between any two college students i_u and i_k. Finally, the weights of edges were calculated based on factors such as the frequency and duration of interaction between students and learning resources. This calculation reflects their preference or emphasis on a particular resource.

Furthermore, the initial feature vector $x_{iu}^0 = (x_1^0, x_2^0, \ldots, x_h^0)$ was allocated or calculated for each node. For students, this included historical records and preferences of their sports activities. For learning resources, this included their types and difficulty levels. Figure 2 displays the initial characteristics of college students. Appropriate embedding technology was used to represent the aforementioned features as continuous vector forms, which aided GNN in better learning and extracting features. Let $e_{qu} = (e_1, e_2, \ldots, e_g)$ be the initial feature vector of the processed learning resources. To ensure the stability of network training, the feature vectors of all nodes were normalized. This normalization process ensures that the mean value of the feature vectors is 0 and their variance is 1.

It should be noted that the integrity and accuracy of data must be ensured when establishing the heterogeneous information network to avoid information bias caused by data loss or error. Considering that varying interactions may have different levels of importance, it is important to allocate weights to edges based on actual situations when calculating the overall weights. When embedding the initial feature

vector, the features related to the research purpose should be selected, and their dimensions should be paid attention to avoid the curse of dimensionality.

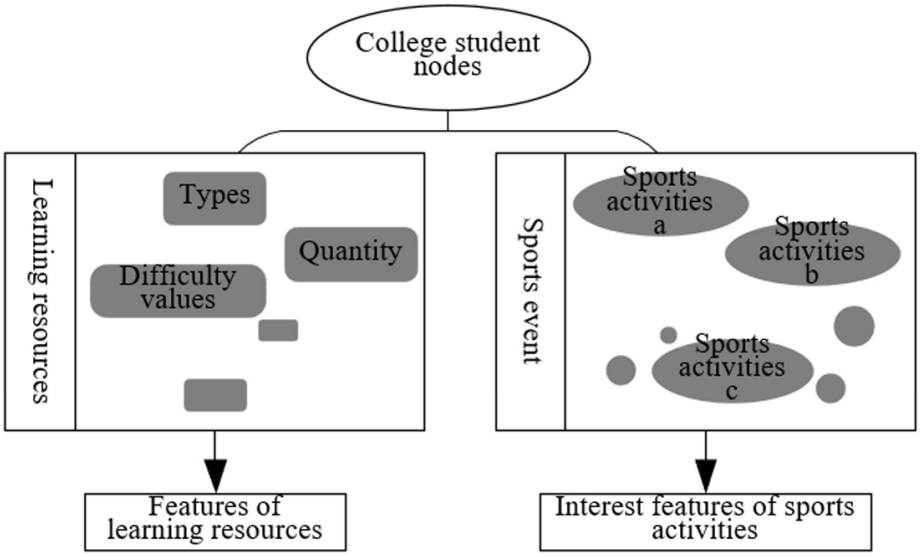

Fig. 2. Initial characteristics of college students

2.2 Feature embedding and fusion

In the realm of sports activities, there exist intricate interactive patterns involving students, sports events, and learning resources. Meta-paths help capture complex relational patterns and transform them into useful feature representations. Meanwhile, students' interests in sports activities may be influenced by various factors, such as their previous activity history, interactions with other students, and utilization of learning resources. Meta-paths provide a method for fusing features through multiple paths, thereby obtaining a more comprehensive feature representation. Therefore, based on the meta-path method, this study provides an efficient feature embedding and fusion mechanism for the constructed heterogeneous graph of sports activities. The heterogeneous graph H_J was decomposed based on meta-paths, resulting in the extraction of heterogeneous sub-graphs representing various sports activities. Let H_{Jo} be the sub-graph structure extracted using the meta-path o. The specific methods and steps are described in detail below.

The sequences of sports activities participated in by college students were first collected. An alignment algorithm was used to compare the sequences of various college students, thereby determining the similarities and differences in their activities. Based on the alignment results, sequences with high similarity were considered similar, thus defining the neighborhood structure. For the college student i_u and the neighbor i_k, who are connected by the meta-path LOj with the maximum number of hops connected by the meta-path L_{Oj}, their sports activity sequences were represented by z_{i_u} and z_{i_k}, respectively. These sequences were aligned and compared, starting from $z_{i_u}[0]$ and $z_{i_k}[0]$. Let μ_1 represent the weight of performing an insertion operation. If F insertion operations were performed during the conversion from z_{i_u} to z_{i_k}, the insertion loss M_1 was calculated using the following equation.

$$M_1 = \mu_1 \cdot F \tag{1}$$

Let μ_2 be the weight of performing a delete operation. If U delete operations were performed during the process of converting from z_{i_u} to z_{i_k}, the delete loss M_2 was calculated using the following equation:

$$M_2 = \mu_2 \cdot U \tag{2}$$

The operational loss, M, during the alignment process of sports activity sequences was calculated using the following equation:

$$M = M_1 + M_2 \tag{3}$$

The heterogeneous subgraphs display the nodes of college students and their neighboring nodes. Neighbor nodes can be other college student nodes, sports event nodes, or learning resource nodes. They were selected based on the alignment results of the sports activity sequences in the previous step. For instance, neighboring nodes were identified by comparing the interests in sports activities of other college students with similar sequences to the target college students. The similarity of interests between the sports activity sequences z_{i_u} and z_{i_k} was calculated as follows:

$$SIM(i_u, i_k) = 1 - \frac{1}{\left|z_{i_u}\right| + \left|z_{i_k}\right|} \tag{4}$$

When college students actively participated in sports activities, they became more susceptible to the influence of other college students who shared similar interests in sports. This resulted in:

$$SIM(i_u, i_k) = \begin{cases} SIM(i_u, i_k), \ if \ SIM(i_u, i_k) \geq \omega \\ 0, \ else \end{cases} \tag{5}$$

Let ω represent the threshold for interest similarity in sports activity sequences, and $B_{(iu)}$ denote the neighborhood space of the neighboring college node of i_u under the corresponding meta-path L_o. The transfer matrix presented below was constructed to assess the similarity of college students in sports activities.

$$N(i_u, i_k) = \begin{cases} 1, \ if \ SIM(i_u, i_k) \exists \wedge S_{PA} = 1 \\ 0, \ else \end{cases} \tag{6}$$

After constructing the neighborhood structure of college student nodes according to the above steps, features were further integrated based on meta-paths. According to the research requirements and the structure of the heterogeneous graph, several meta-paths were initially predefined. Examples of these meta-paths include “college student-sports event-college student” or “college student-sports event-learning resource-college student.” The selection of meta-paths reflected the specific relationship types that were expected to be captured from the graph. To enhance the speed of feature fusion, it is recommended to apply the following linear transformation to the initial features of all nodes. This will ensure that the transformed feature vectors have the same dimension. Let x_{iu}^0 and e_{qu}^0 represent the initial feature vectors of college students, sports events, or learning resources in H_J. Q_i and Q_q represent the

parameter weight matrices. g_{iu}^0 and g_{qu}^0 denote the feature vectors after linear transformation. Therefore, the following equations hold true:

$$\begin{cases} g_{i_u}^0 = Q_i \cdot e_{i_u}^0 \\ g_{q_u}^0 = Q_q \cdot e_{q_u}^0 \end{cases} \tag{7}$$

For each target college student node, random sampling was performed based on predefined meta-paths, resulting in multiple actual path instances. These instances reflected the actual interactions between the target college students and their neighborhood. Let $o(i_u,i_k)$ be the nodes i_u and i_k connected by the meta-path instance o under the meta-path L_o; $\{l^{o(iu,ik)}\} = o(i_u,i_k) - (i_u,i_k)$ be the intermediate nodes in $o(i_u,i_k)$; $g_{o(iu,ik)}$ be a single vector projected on i_u; d_ϕ be the mean aggregation operation. All node features in H_J were converted into $g_{o(iu,ik)}$:

$$g_{o(i_u-i_k)} = d_\varphi\left(g_{i_u}^0, g_{i_k}^0, \left\{g_y^0 \middle| \forall y \in \{l^{o(i_u-i_k)}\}\right\}\right) \tag{8}$$

For each path instance, the features of all nodes involved were extracted and then fused into a unified representation using the fusion method. This method takes into account not only the node features themselves but also their contexts in specific relationships. Let β_o represent the attention parameter of the meta-path instance o; $\|$ represent the vector connection operator. $\beta_{iu,ik}^o$ represent the attention weight of the normalized meta-path instance o. The equation below was used to calculate the normalized weight of each meta-path instance within the neighborhood in H_J for the node iu. It involved calculating the weighted sum of meta-paths in the sub-graph corresponding to L_o in order to achieve the weighted fusion of features between meta-paths. The activation function $\delta(\cdot)$ outputs the feature vectors of sports activity interests of i_u in the subgraph.

$$r_{i_u,i_k}^o = \text{LeakyReLU}\left(\beta_o^Y \cdot \left[g_{i_u}^0 \middle\| g_{o(i_u,i_{7k})}\right]\right) \tag{9}$$

$$\beta_{i_u,i_k}^o = \frac{\exp\left(r_{i_u,i_k}^o\right)}{\sum_{i_k \in N(i_u)} \exp\left(r_{i_u,i_k}^o\right)} \tag{10}$$

$$x_{i_u}^{L_o} = \delta\left(\sum_{o,i_k \in N(i_u)} \beta_{i_u,i_k}^o \cdot g_{o(i_u,i_k)}\right) \tag{11}$$

Finally, the fused features of all path instances were further fused, which obtained the final feature representations of the target college student nodes on the current heterogeneous subgraph.

$$x_{i_u} = \text{CONCAT}(x_{i_u}^{L_1}, ..., x_{i_u}^{L_o}) \tag{12}$$

3 PREDICTING INTEREST LABELS OF COLLEGE STUDENTS CONCERNING SPORTS ACTIVITIES AND CONSTRUCTING AN INTEREST MATRIX

This study predicted the interest labels of college students concerning sports activities based on multi-layer perceptron (MLP), as shown in Figure 3. As a feedforward

neural network, the model is used to perform various tasks, including classification and regression. The model includes an input layer, a hidden layer, and an output layer. The interesting features, such as those extracted based on GNN, of college students' sports activities were fed into the input layer of MLP. These features include, but are not limited to, the frequency of activities in which students participate, the learning resources they use, and their positions in social networks. The input data was processed using one or more hidden layers. Each hidden layer is composed of several neurons, which are connected to all neurons in the previous layer. Each neuron within the hidden layer has a weight, a bias, and an activation function. The output of the last hidden layer is passed to the output layer. The number of neurons in the output layer was equal to the number of target labels related to sports activities that needed to be predicted. The softmax activation function is usually used to ensure that the output value represents the probability distribution.

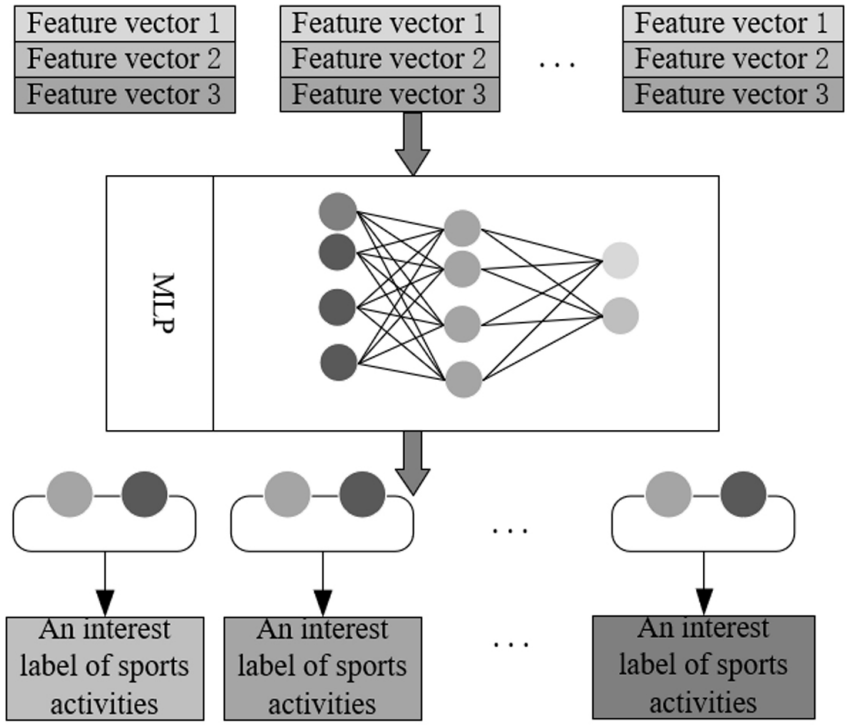

Fig. 3. The prediction principle of college students' interest labels regarding sports activities

To assess the disparities between predicted and actual labels, the model employed the loss function to quantify prediction errors. Let $T = (t_{i1}, t_{i2}, t_{i3}, ...)$ be the prediction results of sports activity labels. $|I|$ represents the number of college students in H_J. t_{iu} represents the true interest label for the college student i_u in sports activities The prediction results output by the model, i.e., the interest labels for the college student i_u in sports activities, are denoted as. The loss function can be represented as:

$$LOSS = -\sum_{u=1}^{|I|} t_{i_u} \log \hat{t}_{i_u} \tag{13}$$

This study constructed an interest matrix based on the known prediction results of interest features and labels of college students regarding sports activities. The matrix is typically represented as a two-dimensional structure, with rows representing college students, columns representing various sports activities, and the value of each cell indicating the student's level of interest in the sports activity. The interest

values were directly derived from the prediction results of interest features and labels of college students regarding sports activities.

A zero matrix of the corresponding size was first initialized based on the number of college students and the types of sports activities. If the interest features were continuous (e.g., feature values based on GNN), these values were directly converted into matrix values. For example, when it comes to feature values that may represent the frequency or duration of students participating in a certain sports activity, they are directly utilized as values in the interest matrix. If MLP or other machine learning models were used to predict interest labels, the interest matrix would be filled based on the prediction results. For instance, if the prediction results indicate that student A's interest in basketball is 0.8, the value corresponding to the basketball column in the matrix was set to 0.8. To ensure that the matrix values remain within a consistent range, they were standardized to range between 0 and 1.

4 INTEREST CLUSTERING OF COLLEGE STUDENTS REGARDING SPORTS ACTIVITIES

Based on the known interest matrix of college students regarding sports activities, this study further clustered their interest features using K-means clustering. Figure 4 shows the clustering algorithm process. Due to the sensitivity of the K-means clustering algorithm to the scale of features, the data in the matrix was standardized. This ensured that the interest level in each sports activity was on the same scale. The silhouette coefficient was used to determine the optimal K value, which represents the number of clusters. This is because different K values can yield different clustering results, and selecting the appropriate K value is crucial for achieving accurate results.

For each student in the matrix, their distance to all K cluster centers was calculated. Each student was assigned to their nearest cluster center, which formed K clusters. For each cluster, the mean value of the interest vectors of all students within it was calculated and then used as the new cluster center. Let $f_{s,vu}$ represent the interest of college student s in the u-th sports event or learning resource type v_u; $f_{n,vu}$ represent the interest of college student n in the u-th type v_u; z represent the number of sports events or learning resource types.

$$SIM(s,n) = \frac{\sum_{u=1}^{z} f_{s,v_u} * f_{n,v_u}}{\sqrt{\sum_{u=1}^{z}\left(f_{s,v_u}\right)^2} * \sqrt{\sum_{u=1}^{z}\left(f_{n,v_u}\right)^2}} \tag{14}$$

The steps of clustering allocation and cluster center recalculation were repeated until the cluster centers no longer underwent significant changes or reached the preset number of iterations. Finally, the silhouette coefficient was used to evaluate the clustering effect and analyze each cluster, thereby identifying the interests of most students in common sports activities in that cluster. Based on these common points, descriptive labels or names were assigned to each cluster, such as "basketball lover," "fitness lover," etc. Finally, customized suggestions for sports activities, resources, or activities were provided to students based on the clustering results, which provided insights into their interests in sports activities for colleges or educational institutions, thereby helping them make decisions and allocate resources.

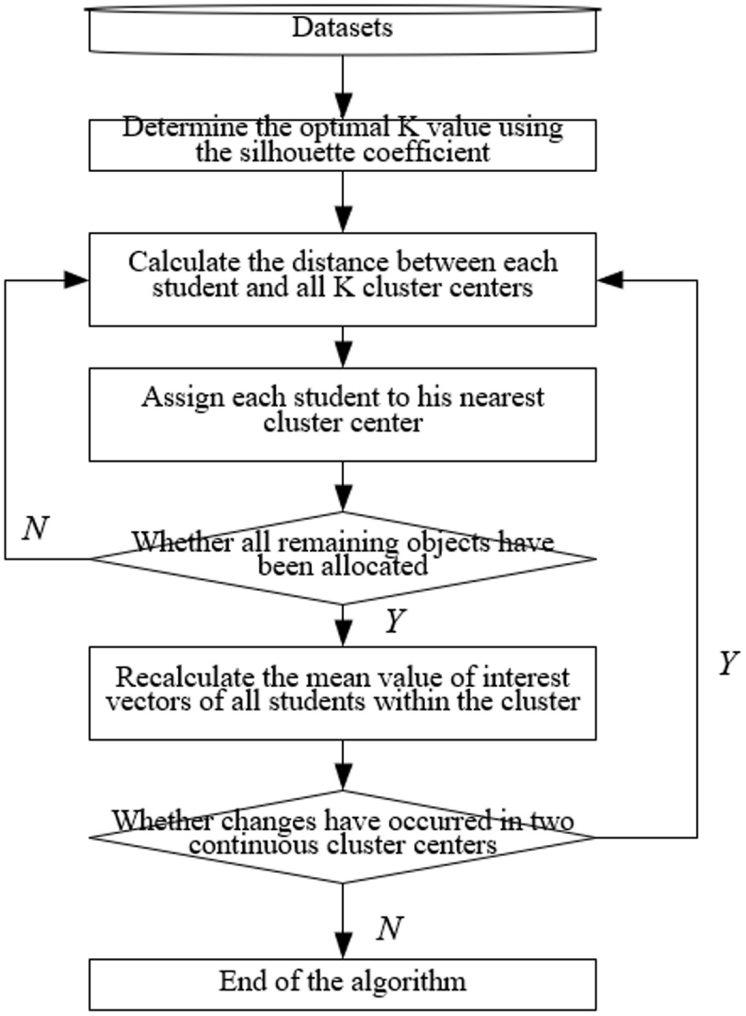

Fig. 4. Interest clustering algorithm process of college students concerning sports activities

5 EXPERIMENTAL RESULTS AND ANALYSIS

To conduct the experiment smoothly, this study constructed three datasets, namely, namely, a sports event dataset, a learning resource dataset, and a comprehensive dataset. The student ID in the datasets is the unique identification for each student. The participated sports event dataset mainly includes names (e.g., "Basketball League," "Table Tennis Training Class," and "Running Club") of sports events, duration of participation, scores and feedback of the events, and other information, which provide intuitive data for the participation and evaluation of students in various sports events, helping analyze their interests in sports activities. The participated learning resource dataset mainly includes names (e.g., "Basketball Skills Tutorial," "Yoga Beginners' Guide," and "Marathon Training Plan") of resources, dates of access, learning duration, feedback on resources, and other information that reflects the participation and evaluation of students in online sports learning resources, helping analyze their online learning habits and preferences. The comprehensive dataset combines the content of the above two datasets.

Table 1. F1-score comparison of different models

Datasets	CNN (%)	RNN (%)	LSTM (%)	The Proposed Model in this Study (%)
Participated Sports Event Dataset	51.29	54.22	55.26	65.62
Participated Learning Resource Dataset	77.36	80.18	81.14	88.33
Comprehensive dataset	84.26	84.55	88.26	92.15

The following comparative analysis can be made based on different models on different datasets in Table 1. In terms of dataset performance, the F1-score of all models achieved an F1-score above 50% on the participated sports event dataset. However, the overall F1-score is relatively low, which may be attributed to the characteristics of the dataset, such as its small size and low data quality. All models have relatively better performance on the participating learning resource dataset, indicating that the dataset may provide more useful information for the models to extract features. The comprehensive dataset has the highest F1-score, which indicates that the data combining sports events with learning resources provides more comprehensive information for the models, thereby improving the prediction accuracy. In terms of model performance, the model proposed in this study has the highest F1-score on all datasets, indicating that the proposed model has higher feature extraction ability and prediction accuracy. Long short-term memory (LSTM) generally performs better than recurrent neural networks (RNN) and convolutional neural networks (CNN). This may be due to LSTM's ability to better capture long-term dependency information, making it more suitable for sequence data. RNN performs slightly better than CNN, especially on the sports event and learning resource datasets, maybe because RNN is specifically designed for sequence data while CNN is more suitable for images and other data. Therefore, the proposed model has the best performance in extracting the interest features of college students concerning sports activities, and its F1-score significantly surpasses that of other models on both single and comprehensive datasets. The performance of models is also influenced by the comprehensiveness and quality of the data. The comprehensive dataset is beneficial for feature extraction in all models because it provides information from multiple perspectives. But the proposed model utilizes the information to the maximum.

The following comparative analysis can be made based on the F1-score of different models on various datasets, as shown in Figure 5. In terms of dataset performance, all models achieved an F1-score above 70% on the participating sports event dataset. However, compared with the participated learning resource dataset and the comprehensive dataset, the F1-score of other models is relatively high, excluding the proposed model, possibly because the information on sports activities contained in the participated sports event dataset is relatively more direct for predicting interest labels. The proposed model has excellent performance on the participating learning resource dataset. However, the performance of other models is relatively poor, which may indicate that the information on learning resources in the dataset poses some challenges in predicting interest labels. The overall performance of the comprehensive dataset is between the above two datasets, but the proposed model still exhibits significant advantages.

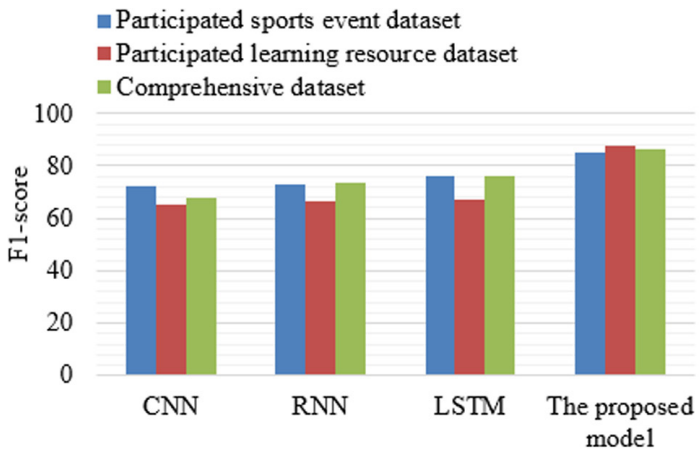

Fig. 5. F1-score of different models on different datasets

In terms of model performance, the F1-score of the proposed model is the highest on all datasets, indicating that the proposed model has significant advantages in predicting the interest labels of college students concerning sports activities. LSTM performs better than RNN and CNN on all datasets, which is consistent with previous analyses. This indicates that LSTM has certain advantages in processing sequence data. The performance of RNN and CNN is relatively similar, especially on the participating learning resource dataset. However, RNN performs slightly better than CNN on both the sports event dataset and the comprehensive dataset. Therefore, the proposed model has significant advantages in predicting the interest labels of college students regarding sports activities, and its F1-score is significantly higher than that of other models on various datasets. At the same time, the comprehensiveness and quality of the data also have a significant impact on the performance of models, with the multi-source information in the comprehensive dataset providing richer contexts for all models, thereby improving the prediction accuracy.

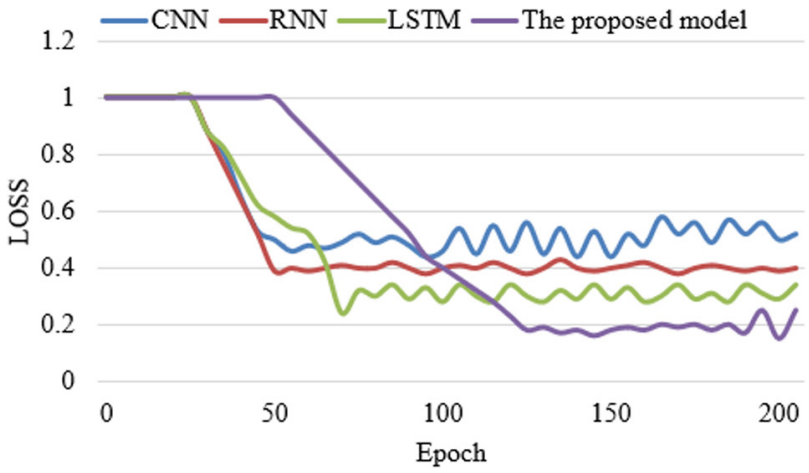

Fig. 6. Loss curve comparison of different models on the comprehensive dataset

It can be observed from Figure 6 that the loss values of different models vary on the comprehensive dataset. The loss value of all models is 1 at the beginning

of training, which means that all models have the same initialized state. The loss value of the proposed model remains at 1 in the first 10 epochs, which means that the model has a slow learning speed in the early stages. However, the loss value begins to rapidly decrease from the 10th epoch and continues to maintain this downward trend in subsequent epochs. The loss values of CNN and RNN remain at 1 in the first six epochs and then begin to decrease, but the rate of decline is slow. The loss value of LSTM also remains at 1 in the first six epochs, and the subsequent downward trend is similar to that of CNN and RNN. However, its decline rate is slightly faster.

The loss value of the proposed model gradually decreases and is relatively stable during training. The loss value of CNN slightly increases in some epochs, indicating a certain degree of instability. The loss value of RNN decreases slowly and fluctuates multiple times during training, indicating its instability. The loss value of LSTM decreases rapidly in the early epochs but fluctuates multiple times in the later epochs, indicating its instability. At 200 epochs, the proposed model has the smallest loss value, reaching 0.15, which is much lower than that of other models. The loss values of LSTM, CNN, and RNN are 0.29, 0.5, and 0.39, respectively. Although LSTM performs the best, it is still inferior to the proposed model. Therefore, the proposed model has significantly better performance than other models in terms of loss value on the comprehensive dataset, with fast learning speed and stable loss values, which indicates that the proposed model better captures information in the comprehensive dataset and more accurately predicts the interest labels of college students concerning sports activities. LSTM has the second-best performance but shows certain instability in the later stages of training. The loss values of CNN and RNN decrease slowly and exhibit instability, and their fitting effect on the comprehensive dataset is not as good as that of the other two models.

It can be observed from Table 2 that the area under curve (AUC) values of different models vary on different datasets. The AUC value of the proposed model is 0.9415 on the participating sports event dataset, which is significantly higher than that of other models. This indicates that the proposed model performs the best in predicting the interest labels of college students regarding sports activities. The proposed model achieved the highest AUC value of 0.8426 on the participating learning resource dataset, indicating superior prediction performance. The AUC value of LSTM is 0.8316, which is second only to the proposed model. The AUC values of CNN and RNN are very close, with values of 0.8218 and 0.8215, respectively. However, both are lower than the AUC values of the proposed model and LSTM. The AUC value of the proposed model on the comprehensive dataset is 0.7125, which is the highest among all models. This indicates that the proposed model has the best prediction performance on the dataset. The AUC value of LSTM is 0.6851, which is lower than that of the proposed model but higher than that of CNN and RNN. The AUC values of CNN and RNN are identical, both measuring 0.6748. This suggests that their performance on this dataset is comparable.

Table 2. AUC value comparison of different models

Datasets	CNN	RNN	LSTM	The Proposed Model
Participated sports event dataset	0.7326	0.7326	0.7426	0.9415
Participated learning resource dataset	0.8218	0.8215	0.8316	0.8426
Comprehensive dataset	0.6748	0.6748	0.6851	0.7125

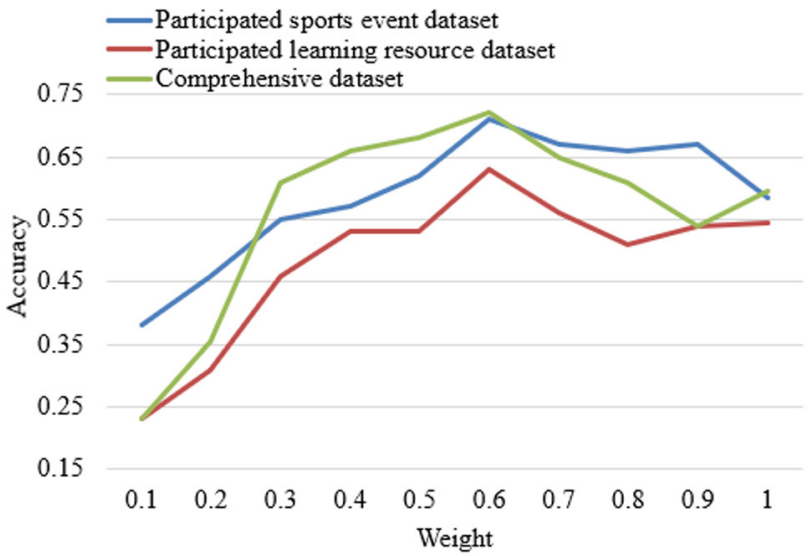

Fig. 7. Impact of different normalized weights on clustering accuracy

Figure 7 shows the impact of different normalized weights on the interest clustering accuracy of college students concerning sports activities. The accuracy of the sports event dataset increases with the initial increase in weight and reaches its peak of 0.71 when the weight is 0.6. The accuracy begins to decrease when the weight increases to 0.7 but stabilizes when the weights are 0.8 and 0.9. Finally, it decreases further to 0.585 when the weight is 1. The accuracy of the dataset for learning resources first increases with the increase in weight and reaches its peak of 0.63 when the weight is 0.6. Similar to the participating sports events, the accuracy of the dataset of learning resources decreases as the weight increases to 0.7. However, it slightly improves when the weight is 0.9 and ultimately reaches 0.545 when the weight is 1. The accuracy of the comprehensive dataset shows an upward trend with an increase in weight and reaches its peak of 0.72 when the weight is 0.6. Then the accuracy begins to decrease as the weight gradually increases and continues to decline to 0.595 when the weight is 1.

The accuracy of the three datasets is highest when the normalized weight is 0.6. This suggests that 0.6 is the optimal normalized weight for accurately clustering college students' interests in sports activities. On the sports event and learning resource datasets, as the weight increases, the clustering accuracy initially increases and reaches its peak before declining. This suggests that surpassing a certain weight threshold on these datasets may result in excessive normalization of information, ultimately reducing the clustering accuracy. The trend of accuracy on the comprehensive dataset is similar to that of the first two datasets, but the highest accuracy is achieved when the weight is set to 0.6. Overall, selecting the appropriate normalized weight is crucial for improving the accuracy of interest clustering among college students in relation to sports activities. The weight of 0.6 yields the optimal clustering results for these three datasets.

6 CONCLUSION

This study focused on extracting and clustering the interests of college students regarding sports activities. Specifically, it aimed to accurately extract the interest

features using a heterogeneous graph structure based on GNN. The sports activity sequences of college students were used to establish the structure of a heterogeneous information network, and the initial feature vector was embedded. Meta-paths were utilized for embedding features and fusing those sequences, thereby capturing the interests of college students in sports activities more accurately. This study utilized a prediction model based on MLP (Multi-Layer Perception) to forecast the interest labels of college students regarding sports activities. An interest matrix was constructed based on the known prediction results of the interest features and labels of the students. Finally, the interested features were clustered using the K-means clustering method.

The experimental results showed that the proposed model outperformed other models on all three datasets, CNN, RNN, and LSTM, in both feature extraction and interest label prediction, as evidenced by higher F1-scores. The proposed model demonstrated its superiority by showing relatively stable and low loss on the comprehensive dataset when compared to other models. In the AUC value comparison, the proposed model performed well on all three datasets and particularly showed significant improvement on the sports event and learning resource datasets. When analyzing the impact of various normalized weights on clustering accuracy, it was discovered that the accuracy reached its optimal level at a normalized weight of 0.6.

This study proposes a feature extraction and clustering method based on a heterogeneous graph neural network, targeting at interests of college students in sports activities. Detailed experimental verification demonstrated that this method was superior in extracting the interesting features of college students regarding sports activities and outperformed common CNN, RNN, and LSTM models. The proposed model exhibited robust stability and accuracy when applied to the participating sports event dataset, the participating learning resource dataset, and the comprehensive dataset. This model serves as a valuable tool for studying college students' interests in sports activities, and it establishes a solid foundation for future research that can delve into these interests in a more detailed and comprehensive manner. This research can include personalized recommendations and health interventions.

7 REFERENCES

[1] B. Yang, "Learning motivations and learning behaviors of sports majors based on big data," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 23, pp. 86–97, 2021. https://doi.org/10.3991/ijet.v16i23.27823

[2] M. Tang and D. Wei, "Cultivation strategies for scientific research and innovation ability of college students majoring in sports," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 6, pp. 154–166, 2021. https://doi.org/10.3991/ijet.v16i06.21093

[3] A. Komaini, Hermanzoni, S. G. Handayani, M. S. Rifki, Y. Kiram, and N. Ayubi, "Design of children's motor training tools using sensor-based agility components in physical education learning," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 5, pp. 207–215, 2022. https://doi.org/10.3991/ijim.v16i05.29731

[4] H. Syahrastani, A. Hidayat, Komaini, A. Gemaini, and Zulbahri, "Smart application for smart learning: How the influence of the factors on student swimming learning outcomes in sports education," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 17, pp. 116–129, 2022. https://doi.org/10.3991/ijim.v16i17.34365

[5] A. O. Ajlouni, W. J. AlKasasbeh, A. Al-Shara'h, and A. Ibrahim, "The impact of mobile application-assisted instruction on intrinsic motivation and sports nutrition knowledge: The case of blended learning," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 11, pp. 251–272, 2023. https://doi.org/10.3991/ijet.v18i11.38637

[6] B. Chen, "RETRACTED: Research on orienteering teaching of college sports based on computer multimedia technology," *Journal of Physics: Conference Series*, vol. 1992, no. 3, 2021. https://doi.org/10.1088/1742-6596/1992/3/032007

[7] J. Alnuaimi, A. Al-Za'abi, I. A. Yousef, M. Belghali, S. M. Liftawi, Z. F. Shraim, and E. A. M. Tayih, "Effect of a health-based-physical activity intervention on university students' physically active behaviors and perception," *International Journal of Sustainable Development and Planning*, vol. 18, no. 5, pp. 1451–1456, 2023. https://doi.org/10.18280/ijsdp.180515

[8] J. Kim et al., "Activities of the creative robot contest for decommissioning at national institute of technology (NIT), Tsuruoka college," *Journal of Robotics and Mechatronics*, vol. 34, no. 3, pp. 527–536, 2022. https://doi.org/10.20965/jrm.2022.p0527

[9] Y. Li, "Study on physical fitness promotion and sports intervention measures and effect of college students based on statistical analysis," *Boletin Tecnico/Technical Bulletin*, vol. 55, no. 15, pp. 306–312, 2017.

[10] Y. Ren, "Research on the network marketing channel strategy in sports life: An empirical analysis based on college students," *International Journal of Security and its Applications*, vol. 10, no. 6, pp. 65–74, 2016. https://doi.org/10.14257/ijsia.2016.10.6.08

[11] M. Bhatia, P. Manani, A. Garg, S. Bhatia, and R. Adlakha, "Mapping mindset about gamification: Teaching learning perspective in UAE education system and Indian education system," *Revue d'Intelligence Artificielle*, vol. 37, no. 1, pp. 47–52, 2023. https://doi.org/10.18280/ria.370107

[12] J. Lin, Y. Jiang, and L. Wang, "Research on impact of network sports information on college students' participation in energy physical exercise in Jiangxi Province," *Energy Education Science and Technology Part A: Energy Science and Research*, vol. 31, no. 1, pp. 171–174, 2013.

[13] Y. Shi and S. Chen, "Development and teaching application of new intelligent courseware in 'Sports Economics'," *International Journal of Emerging Technologies in Learning*, vol. 11, no. 5, pp. 51–55, 2016. https://doi.org/10.3991/ijet.v11i05.5694

[14] L. Massi et al., "Engineering and computer science community college transfers and native freshmen students: Relationships among participation in extra-curricular and co-curricular activities, connecting to the university campus, and academic success," *Frontiers in Education Conference Proceedings*, Seattle, USA, pp. 1–6, 2012. https://doi.org/10.1109/FIE.2012.6462276

[15] K. E. Bigelow, "Reflections of college students promoting engineering through biomechanical outreach activities indicate dual benefits," in *Proceedings of the 117th Annual American Society for Engineering Education Conference and Exposition*, 2010.

[16] M. Ford, "Engineering MythBusters brings engineering principles to kids," *ASEE Annual Conference and Exposition*, Pittsburgh, Pennsylvania, 2008.

[17] W. Nan and W. Lu, "Construction of university student sports activity tendency model based on fuzzy optimization algorithm," in *11th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, Changsha, China, pp. 27–30, 2018. https://doi.org/10.1109/ICICTA.2018.00014

[18] H. Yao, Y. Wu, and X. Wu, "A research of the impact of the network sports information on college students' sports cognition," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 4, no. 21, pp. 4333–4338, 2012.

[19] A. Kannan, R. Priyadharshini, and P. Sinduja, "The influence of online gaming on students' attitudes, behaviors, and academic performance," in *International Conference on Business Analytics for Technology and Security (ICBATS)*, United Arab Emirates, 2023, pp. 1–5. https://doi.org/10.1109/ICBATS57792.2023.10111396

[20] J. Li, X. Li, and Y. Yang, "Effect of power walking on the body shape of college students who sit long and exercise little," in *International Conference on Wearables, Sports and Lifestyle Management (WSLM)*, Kunming, China, 2022, pp. 8–11. https://doi.org/10.1109/WSLM54683.2022.00007

[21] R. Jiang, "Research on innovative practice links of physical education course based on computer big data," *Journal of Physics: Conference Series*, vol. 1865, no. 4, 2021. https://doi.org/10.1088/1742-6596/1865/4/042133

8 AUTHOR

Xiuzu Xiong, obtained a Master's degree from Guizhou Normal University. Currently, she works in the Physical Education Department of Guizhou University of Commerce. Her main research directions include physical education and training (E-mail: 201720625@gzcc.edu.cn; ORCID: https://orcid.org/0009-0007-6018-1930).