2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) | 979-8-3503-7507-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/MLISE62164.2024.10674225

Research on Student Performance Prediction Based on Clustered Graph Neural Networks

*Xiaochen Lai
School of Software, Dalian University of Technology, Dalian, China
* Corresponding author: laixiaochen@dlut.edu.cn

Sheng Zhao
School of Software, Dalian University of Technology, Dalian, China
18742023541@163.com

Zheng Zhang
School of Software, Dalian University of Technology, Dalian, China
zhangzheng@mail.dlut.edu.cn

Xiaodi Pan
School of Software, Dalian University of Technology, Dalian, China
panxiaodi@mail.dlut.edu.cn

***Abstract*—The advent of the big data era has brought about profound changes in modern education, making educational data mining an important field within the realm of data analysis. Predicting students' academic performance is one of the crucial topics in this field. This paper proposes a student performance prediction method: Cluster-based Graph Neural Network Prediction (CGNNP). By clustering students' routine and final exam scores and using the clustering results as category labels, the method employs graph neural networks for training and testing the data. Through analyzing students' background information and learning behavior data, this method effectively predicts students' performance. Experimental results demonstrate a significant improvement in prediction accuracy compared to traditional machine learning methods, better reflecting students' learning situations and providing educators with more accurate decision-making support.**

Keywords-component; clustering, graph neural networks, educational data mining.

I. INTRODUCTION

Data mining can discover hidden information from large amounts of data and is widely used in fields such as image processing and speech recognition, educational data mining (EDM) is an important research direction within data mining. Predicting students' academic performance poses a formidable challenge, and EDM provides technical support for such predictions. It quantifies, analyzes, and models the massive data generated during the teaching process through the use of data mining, machine learning, and statistical methods. This enables the detection and regulation of the teaching process, thereby enhancing teaching effectiveness. Data mining technology has been applied in various aspects of educational data processing, including dropout prediction, academic data analysis, and analysis of academic behaviors [1].

Big data technology is gradually being widely applied in the field of education, many scholars have pointed out the extensive application of data mining techniques in the field of education. Javier [2] pointed out that educational data mining combines data mining techniques with educational data and summarized the most commonly used data mining methods, including factor analysis, regression, and correlation mining. Romero et al. [3] believe that the goal of educational data mining is to better understand students' learning patterns. It is also considered a method for exploring unique data types in the educational environment, defined as the application of data mining techniques.

Currently, the theoretical research on predicting student performance primarily revolves around traditional machine learning algorithms such as Random Forest, Support Vector Machine, and K-Nearest Neighbors. Some scholars apply clustering and other data mining techniques to the analysis of learning performance. Karthik et al. [4] developed a new method called the Mixed Educational Data Mining Model to analyze student performance effectively, aiming to enhance the quality of education. Jun [5] pointed out the application of multimodal information analysis in learning early warning models, aiming to further improve the learning early warning models through multimodal information analysis. Lin et al. [6] proposed a deep learning-based continuous facial emotion pattern recognition method to analyze students' learning emotions. By combining Convolutional Neural Networks and Long Short-Term Memory Networks for deep learning, they identify and analyze students' continuous facial emotions and predict learning emotions. Zhang [7] introduced the Analytic Hierarchy Process online learning warning model, assigning unequal weights to different evaluation indicators to ensure fairness in evaluation.

However, with the development of intelligent technologies and their widespread application in the field of education, educational data is experiencing rapid growth. In addition, in both research and practical applications of artificial intelligence, data structured as graphs plays a significant role.

In recent years, Graph Neural Networks (GNNs) have made significant strides across various domains. The extension of graph neural networks into the realm of educational data mining is gaining increasing attention. This expansion is driven by the recognition that graph structures can effectively model relationships within educational data, providing a promising avenue for improving the accuracy and effectiveness of predicting student performance.

This study primarily focuses on addressing the research problem from a clustering perspective to obtain insights into students' academic performance. By clustering the regular assessment scores and final exam scores, the clustering results are utilized as category labels. A prediction method is proposed that combines K-Means clustering and graph neural networks(CGNNP). The K-Means algorithm is employed to cluster the scores, partitioning the data into training, validation, and testing sets. Subsequently, graph neural networks are utilized for training.

The rest of the paper is organized as follows: Chapter 2 is the relevant theoretical foundation, Chapter 3 describes the research methods of this paper, Chapter 4 is dedicated to experimental performance, and Chapter 5 serves as the conclusion summarizing the entire paper.

II. THEORETICAL FOUNDATIONS

A. Clustering

Clustering is an unsupervised learning method aimed at partitioning the objects in a dataset into several clusters, where objects within the same cluster are highly similar to each other while being dissimilar to objects in different clusters. K-Means is a clustering algorithm known for its simplicity, efficiency, and wide applicability. It is a highly representative clustering algorithm, and its implementation process is as follows:

1) Input Data and Number of Clusters (K): Input dataset D and the desired number of clusters K.

2) Initialization: Randomly select K data points as initial cluster centers.

3) Data Point Assignment: For each data point, calculate its distance to each cluster center and assign it to the cluster with the closest center.

4) Cluster Center Update: For each cluster, compute the average of all data points within the cluster, and set this average as the new cluster center.

5) Iteration: Repeat the data point assignment and cluster center update steps until a termination condition is met, such as the cluster centers no longer undergoing significant changes.

6) Output Clustering Results: Output the final clustering results.

Let X be a dataset containing n student samples with course performance data: X={x_1, x_2, x_3, ..., x_n}, where the i-th student's course grades are represented as x_i={x_{i1}, x_{i2}, x_{i3}, ..., x_{ip}}, and p is the total number of grades. The Euclidean distance between any two students is defined as:

$$d(x_i,x_j)=\sqrt{\sum_{r=1}^{p}\left(x_{ir}-x_{jr}\right)^2} \qquad (1)$$

B. Graph Neural Network

Graph Neural Networks (GNNs) constitute a category of deep learning models specifically designed for learning and inference on graph-structured data. Their outstanding performance has garnered considerable attention and thorough exploration from scholars. Graph-structured data consists of nodes and edges, where nodes represent entities, and edges represent relationships between entities. Mainstream algorithms within the realm of Graph Neural Networks include Graph Convolutional Networks (GCNs), Graph Autoencoders, Graph Generative Networks, Graph Recurrent Networks, and Graph Attention Networks.

A graph consists of two main components: nodes and edges, typically denoted as G=(V, E), where V={v_1, v_2, v_3…v_n} represents the set of nodes and E={e_1, e_2, e_3…e_n} represents the set of edges. The graph neural network (GNN) is a type of neural network based on graph data structure, and it can be classified into spectral domain graph neural networks and spatial domain graph neural networks based on convolution operations. Spectral domain methods primarily rely on the spectral decomposition of the graph's Laplacian matrix. Graph Convolutional Network (GCN) is a typical representative of this approach. In its message-passing process, it uses the normalized form of the Laplacian matrix and transmits information through spectral filtering. This method allows for learning node representations in the frequency domain. Spatial domain methods, on the other hand, focus on the direct relationships between nodes and their neighbors. GraphSAGE is an example of a spatial domain method. It aggregates information by sampling neighboring nodes and processing them. This approach emphasizes local structures between nodes and direct neighbor relationships without involving spectral decomposition.

The message passing process of graph neural networks is one of its core operations, used to transmit and update node information in graph structured data. For graph G, each node v has an initial feature representation $h_v^{(0)}$. For node v, the message passing process involves information transmission and aggregation between it and its neighboring node u. The usual formula for message passing is:

$$m_{u\to v}=M(h_u^{(l)},h_v^{(l-1)}) \qquad (2)$$

Among them, $m_{u\to v}$ represents the message from node u to node v, M represents the message passing function, $h_u^{(l)}$ is the representation of node u in the l-th layer, and $h_v^{(l-1)}$ is the representation of node v in the l-1st layer. After receiving messages from neighboring nodes, node v performs message aggregation operations. Typically, aggregation operations can be represented as:

$$a_v=\mathrm{AGG}(\{m_{u\to v},\forall u\in N(v)\}) \qquad (3)$$

Among them, a_v is the aggregation information of node v, and AGG is the aggregation function of neighbor information. Update node representation using aggregated information:

$$h_v^{(l)}=U(h_v^{(l-1)},a_v) \qquad (4)$$

Among them, U is the update function, and $h_v^{(l-1)}$ is the representation of node v in the l-1 layer. The above steps can be repeated multiple times on each layer to form a multi-layered message passing and node representation update. The message passing diagram is shown in Figure 1.

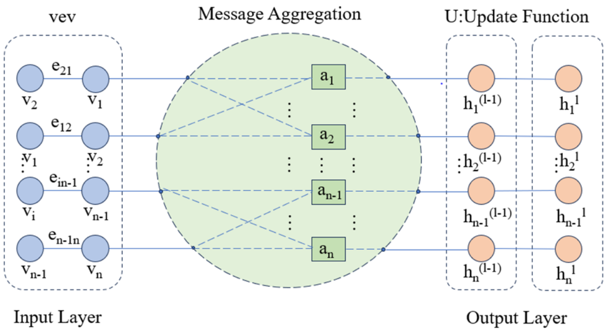

Figure 1. The message passing diagram

III. METHODOLOGY

A. Data Processing

In the process of predicting student performance, data preprocessing is a crucial task. Data preprocessing refers to a series of necessary steps, including cleaning, integration, transformation, and reduction of raw data before data mining. These steps aim to achieve the minimum standards and requirements for algorithms to effectively study the data.

Typically, stored data may have various issues such as missing values, outliers, and duplicate entries. Data cleaning involves efforts to fill in missing values, detect outliers, and remove duplicates. There are several methods for filling missing values in attributes, such as using global constants, means, medians, or modes. This paper adheres to the following principles: deleting samples of students with missing scores in more than two courses. If there are still students with missing scores, the average score for the course is used to fill in the missing values. It is noted that in the dataset, scores of 0 in a course indicate that the student was absent from the exam.

B. Student Grade Aggregation

To ensure teaching quality, student analysis should not be solely based on final exam scores, attention should also be paid to their day-to-day learning. In the dataset used in this study, each course has multiple regular assessment scores. Predicting and analyzing these scores comprehensively can more accurately reflect student learning situations and predict future academic performance, thus contributing to the improvement of overall teaching quality. Therefore, K-Means clustering is employed to effectively integrate routine assessment scores with final exam scores. The routine scores for each course consist of multiple components. Based on the clustering results, students' grades are categorized into different levels, serving as input data for the graph neural network.

In order to determine the optimal number of clusters (K) using the elbow method, one typically plots the within-cluster sum of squares (WCSS) for different values of K and identifies the point where the rate of decrease in WCSS slows down, resembling an elbow on the graph. This point indicates that increasing the number of clusters beyond this value does not significantly reduce the error. The K value corresponding to the elbow point is then chosen as the final number of clusters. The within-cluster sum of squares (WCSS) is defined as follows:

$$\mathrm{WCSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \| X_{ij} - c_i \|^2 \tag{5}$$

Among them, k represents the number of clusters, n_i is the number of data points in the i-th cluster, X_i is the collection of data points in the i-th cluster, and c_i is the center of the i-th cluster.

C. Graph Neural Network Training

Real-world student datasets are typically not in the form of graph-structured data, necessitating the construction of an undirected graph based on the student dataset. To achieve this objective, referencing the work of Chio et al. [8] and other relevant analyses, the construction of graphs in graph neural networks is often based on computing the similarity between samples. In this study, to quantify the correlation between students, the Pearson correlation coefficient is employed for measuring similarity.

The Pearson correlation coefficient is a statistical measure used to assess the strength and direction of a linear relationship between two variables. It quantifies the degree of the linear relationship between two variables, with values ranging from -1 to 1. It is usually represented as r, the specific formula for calculation is as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{6}$$

Where X_i and Y_i are the i-th observed values of two variables, $\bar{X}$ and $\bar{Y}$ are the average values of the two variables, and n is the number of observed values.

With students serving as nodes, the Pearson correlation coefficient is utilized to gauge the interrelation among nodes. Nodes exhibiting a Pearson correlation coefficient exceeding 0.8 are designated as strongly correlated entities, subsequently interlinked to compose an undirected graph.

GPcSAGE is a graph neural network framework based on Pearson correlation coefficient random walk and importance aggregation. The model samples a fixed number of neighboring nodes layer by layer in an outward manner through random walk, then aggregates neighboring nodes layer by layer in an inward manner, and finally utilizes the obtained node embeddings for prediction. By performing random walks on the graph, hidden structures and patterns within the graph can be effectively discovered. Meanwhile, combining importance aggregation can accelerate training and inference, reduce computational resource consumption, and improve model generalization ability. To prevent the issue of neighborhood explosion, the model limits the number of neighbors per hop. Additionally, during the aggregation process, it introduces attention mechanisms to better focus on important neighbor node information.

CGNNP first clusters the cleaned data, then uses the clustering results as category labels, constructs an undirected graph, and finally the GPcSAGE model is used for training. The process is shown in Figure 2:

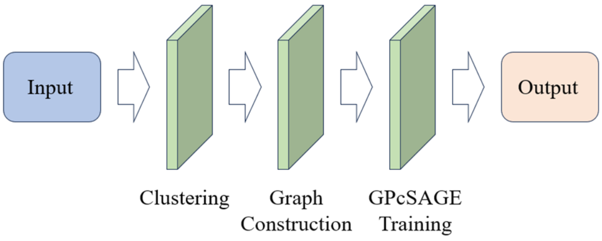

Figure 2. Flowchart of CGNNP

IV. EXPERIMENTS

A. *Determination of Clustering Parameters*

To accurately determine the optimal number of clusters (K), the elbow method is employed to identify the point where the rate of decrease in error slows down. This point, known as the "elbow point," is selected as the final value for K. In the graph below (Figure 3), the optimal number of clusters is determined to be 3 based on the elbow method.

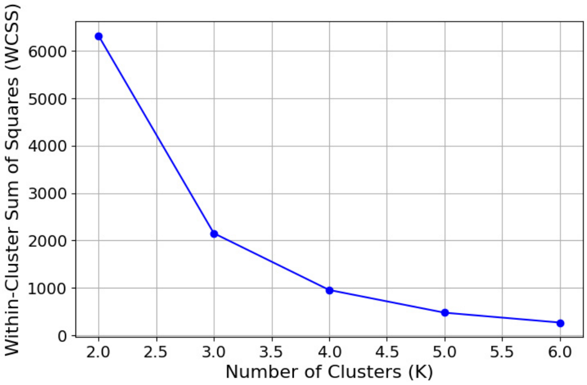

Figure 3. Sum of squared errors within clusters

B. *Experimental Results and Analysis*

After clustering, construct an undirected graph and train on a graph neural network. The dataset is split into three parts: training (60%), validation (20%), and testing (20%). The experiment compares the grades data before and after clustering, and the accuracy results for various models, including traditional neural network models, are presented in Table 1.

Table 1 THE Accuracy of Student Dataset Prediction

Model	MLP	KNN	RF	SVM	GPcSAGE
UnCGNNP	76.79	77.64	74.79	77.04	88.48
CGNNP	79.74	80.16	79.56	80.61	90.33

From the above table, it can be observed that clustering the grades before prediction improves the accuracy of predictions. When using traditional methods like MLP, KNN, and SVM, the accuracy increases by 2.95%, 2.52%, and 3.57%, respectively, compared to using non-clustered data. The RF method shows the highest accuracy improvement of 4.77%. Additionally, GPcSAGE exhibits a significant improvement in prediction accuracy compared to traditional methods. The experimental results indicate that the clustering-graph neural network prediction model can enhance classification performance, making it an ideal optimization method.

V. CONCLUSIONS

This paper proposes a learning performance prediction method: CGNNP, which combines students' regular scores and final scores through clustering, takes the clustering results as category labels, and applies graph neural networks for prediction. Presently, the domain of student performance prognostication predominantly fixates on terminal grades, neglecting the performance of students over time. The amalgamation of these two metrics can furnish a more impartial assessment of students. Comparative experiments have verified the effectiveness of this model in predicting scores, with more ideal prediction accuracy. However, this paper also has shortcomings: the initial clustering centers are randomly determined, which may have a certain impact on the clustering results; using Euclidean distance for distance measurement may not be the optimal choice. In future work, further exploration and optimization of clustering methods, such as density-based clustering methods, can be conducted to study how to choose appropriate density parameters and distance measurement methods to improve the accuracy and stability of clustering results.

REFERENCES

[1] D. Buenaño-Fernandez, W. Villegas-Ch, and S. Luján-Mora, "The use of tools of data mining to decision making in engineering educationA systematic mapping study," Comput. Appl. Eng. Educ., Review vol. 27, no. 3, pp. 744-758, May 2019, doi: 10.1002/cae.22100.

[2] J. Bravo-Agapito, C. F. Bonilla, and I. Seoane, "Data mining in foreign language learning," Wiley Interdiscip. Rev.-Data Mining Knowl. Discov., Review vol. 10, no. 1, p. 16, Jan-Feb 2020, Art no. e1287, doi: 10.1002/widm.1287.

[3] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdiscip. Rev.-Data Mining Knowl. Discov., Review vol. 10, no. 3, p. 21, May 2020, Art no. e1355, doi: 10.1002/widm.1355.

[4] V. G. Karthikeyan, P. Thangaraj, and S. Karthik, "Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation," Soft Comput., Article vol. 24, no. 24, pp. 18477-18487, Dec 2020, doi: 10.1007/s00500-020-05075-4.

[5] G. Jun, "Application of Multimodal Information Analysis in Learning Early Warning Model," Procedia Computer Science, vol. 228, pp. 729-735, 2023.

[6] S. Y. Lin, C. M. Wu, S. L. Chen, T. L. Lin, and Y. W. Tseng, "Continuous Facial Emotion Recognition Method Based on Deep Learning of Academic Emotions," Sens. Mater., Article vol. 32, no. 10, pp. 3243-3259, 2020, doi: 10.18494/sam.2020.2863.

[7] T. Zhang, W. X. Xiao, and P. Hu, "Design of Online Learning Early Warning Model Based on Artificial Intelligence," Wirel. Commun. Mob. Comput., Article vol. 2022, p. 11, May 2022, Art no. 3973665, doi: 10.1155/2022/3973665.

[8] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," Advances in neural information processing systems, vol. 29, 2016.