Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on the Path of Enhancing Teaching Effect of Data Visualization Technology in Civic and Political Education in Colleges and Universities

Zhengwan He[1,] ✉

[1] *Public Foundation College, Anqing Medical College, Anqing, Anhui, 246000, China*

ABSTRACT

The field of education is paying more and more attention to the fundamental task of education by establishing morality, and ideological and political education has become a major project in which all the teaching and learning links cooperate with each other and are accomplished in a concerted manner. This study explores the method of organic integration of ideological and political education and teaching and data visualization technology to enhance the effect of ideological and political teaching. Firstly, the method of portrait construction is introduced, combined with the student behavior dataset, and the student behavior data is preprocessed. Using the user portrait construction method as a hub, a gradient boosting decision tree model was used to predict the students' Civics learning performance. The improved K-prototypes clustering algorithm was used to categorize student groups, which facilitated teachers to develop targeted learning strategies. Finally, group portraits and feature labels are extracted from the students to further help teachers accurately determine the types of student groups and carry out personalized teaching. The classroom teaching model in this paper classifies students into four categories with obvious behavioral characteristics, which increases teachers' understanding of students, and the model not only improves students' academic performance in Civics, but also significantly improves students' level of course Civics and increases students' classroom active response rate by 19.625%. The Civics education data visualization technology proposed in this paper reveals the rules of Civics education and improves teachers' work efficiency.

Keywords: Visualization, Student Portrait, Gradient Boosting Decision Tree, K-prototypes Clustering

1. Introduction

In today's Internet era, data visualization is getting more and more attention, and it is of great significance in the fields of data transparency, decision-making, marketing promotion, scientific research, and national governance [1-4]. In recent years, data visualization technology has gradually

✉ Corresponding author.
E-mail address: hzw@aqmc.edu.cn (Z. He).
Received 25 March 2024; Revised 05 June 2024; Accepted 15 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-317

entered the field of education. In the classroom, data visualization technology presents charts, real-time data, etc., transforming abstract knowledge content into intuitive visual display, helping students better understand and remember [5]. For example, by transforming complex formulas into visual charts, students can better understand their relationships and laws and improve their problem solving ability. In addition, educational data visualization not only helps teachers to understand the overall learning situation of students, teachers can generate personalized learning plans based on students' learning data, targeted tutoring for different students' learning characteristics and needs, and personalized teaching [6-8]. With the implementation of data visualization technology, teachers can monitor students' learning progress and performance changes in real time, understand students' learning situation and problems, and provide timely feedback; at the same time, students and parents can also view students' performance and learning history through the visual interface, which helps to supervise the learning and assessment of teaching [9-11].

Compared with traditional teaching methods, educational data visualization is more attractive and persuasive in the presentation of information, which can make students participate in learning with great interest. Students can analyze, compare and reason more intuitively through the data charts and graphs displayed in the visualization, and develop independent learning and critical thinking skills [12-14]. The application of educational data visualization helps to enhance teachers' understanding of students' learning data and identify students' learning problems in a timely manner, based on which teachers can make targeted teaching plans and adjust teaching strategies to improve academic performance and teaching effectiveness [15-18]. In the practice of data visualization, there is no lack of forms and contents of ideological and political education. Ideological and political education is not only an important way to cultivate students' ideological and moral qualities and national concepts, but also the key to improve the overall quality and social responsibility of universities. Under the development of information technology, it is of great significance to study the teaching effect of data visualization technology applied to ideological and political education.

In this paper, based on the study of student data sets, we use data preprocessing methods to process student behavior data, which makes the student behavior data set more scientific. The gradient boosting decision tree model composed of multiple decision trees is used to predict students' performance, and then the advantages of K-means and K-modes algorithms are combined, and the K-prototypes clustering algorithm is proposed to integrate the two, which adopts different dissimilarity measures for different types of data, so as to get the best effect of dividing the student group. Finally, the user profiles of student groups are constructed, and the behavioral feature labels of different student groups are extracted. Finally, the teaching model designed in this paper is evaluated from three aspects: learning outcome dimensions, course contemplation results, and teacher and student speech behaviors.

2. Student portraits

2.1. Overview of User Profiling

User portrait [19] can accurately abstract and comprehensively display user information, providing data support for analyzing user behavior, consumption habits, etc., and is the basis for analysis and application in the era of big data.

User profiling is a way of describing and classifying users by analyzing their personal information, interests, habits, purchasing behavior and other data. In layman's terms, user profiling is a way of analyzing and mining data to come up with descriptions and portrayals of users, which helps enterprises better understand users and more accurately carry out marketing activities, product promotion and service provision.

User profiling can be divided into two types: static user profiling and dynamic user profiling. Static user profile is based on the basic information and attributes of the user to describe the user, which can

help enterprises to carry out preliminary classification and positioning of the user. Dynamic user profiles, on the other hand, describe users based on their behavioral data and real-time behavior. The information is constantly changing over time, and they can assist enterprises in grasping user interests as well as needs in real time.

Through the construction of user profiles, enterprises can gain a deep insight into user needs and behavioral patterns, so as to carry out targeted product positioning and marketing promotion, and improve the marketing conversion rate. At the same time, user profile helps enterprises to define the target user groups, implement user segmentation, provide them with personalized products and services, optimize user experience and enhance user stickiness.

In practical application, user profiling is usually realized based on data mining, machine learning and big data analysis technologies. Enterprises can collect user data, analyze user behavior and transaction data, build user portrait models, and use algorithms and models to recommend personalized products and services to achieve precision marketing.

To sum up, user profile is an effective tool for in-depth understanding and analysis of users, which can help enterprises better understand user needs and provide more personalized products and services, so as to achieve the purpose of better interaction and communication with users. With the continuous development of science and technology, the application of user profile will be more extensive and in-depth.

2.2. User Profile Construction

User portrait modeling is the process of constructing a user portrait model by analyzing user data and using various data mining and machine learning techniques. The following is a rough portrait modeling process:

1) Data collection: first of all, it is necessary to collect relevant data about users, including their basic information, online behavior, purchasing habits and so on.

2) Data cleaning and pre-processing: After data collection, the original data needs to be cleaned and pre-processed to eliminate abnormal noise data and standardize the data format to ensure data quality and consistency.

3) Feature selection and extraction: after data preprocessing, it is necessary to select and extract features applicable to user profile modeling.

4) Model selection and establishment: after feature selection and extraction, applicable modeling algorithms and models can be selected according to specific needs, such as natural language processing, cluster analysis, deep learning, etc.

5) Model training and validation: after selecting and building the model, it is necessary to use labeled training data to train the model and use validation data to evaluate and tune the model.

6) Model application and update: After model training and validation, the model can be applied to actual user data to classify and recommend users.

7) Model Evaluation and Improvement: After the model is applied and updated, the model effect should be evaluated based on a variety of evaluation indexes (e.g., accuracy, recall, $F1$-value, etc.), and the model should be continuously optimized and improved according to the evaluation results.

The approximate process of user profile construction is shown in Figure 1:

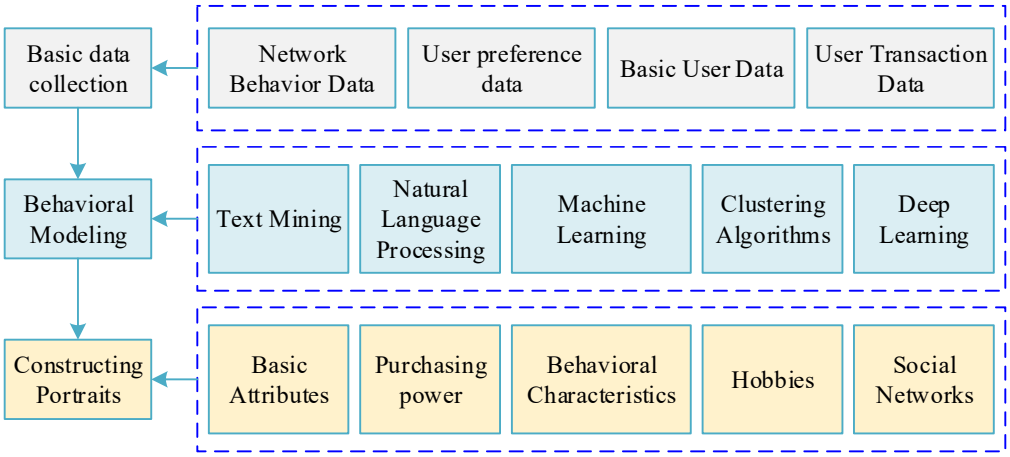

Fig. 1. Portrait building flow chart

User portrait modeling is an iterative process that requires the unified coordination and integration of multiple steps, such as data collection, preprocessing, feature selection and extraction, model selection and building, model training and validation, model application and updating, and model evaluation and improvement. Through continuous iteration and optimization, an accurate and effective user portrait model can be constructed to provide enterprises with accurate user requirements and precise personalized services.

2.3. Student Behavior Data Sources and Preprocessing

The student behavior data used in this study comes from the data generated by undergraduate students in a university during the 2018 to 2020 academic year, including students' basic information, consumption information, book borrowing information, Internet behavior, and grade information.

Data preprocessing refers to a series of processing processes such as cleaning, converting, integrating, normalizing, etc., of raw data before data analysis and modeling. Data preprocessing aims to reduce errors and deviations in the process of data analysis and modeling, and improve the quality and reliability of data.

1) Missing value processing. In the data preprocessing stage, dealing with missing values is a key step in ensuring the accuracy of the data set. The presence of missing values may cause bias in data analysis. In the grade data of this study, the total number of missing values in the usual grades is 5,000, accounting for 11.5% of the overall data volume. Given the large proportion of missing values, this study chose to use the average usual grade of the subject in which it is located to populate it, a method that estimates missing values based on statistical indicators of other data within the same subject.

2) Handling of outliers. Outliers and outliers in the raw data can have a significant impact on statistical analysis and modeling, causing bias in the results. Taking academic grades and regular grades as an example, the red dots in the 1000 randomly selected students' academic grades and regular grades indicate data with more than 100 points or less than 0 points as outliers. The median is used for replacement here, and using median replacement can reduce the impact of these extreme values on the results of data analysis.

3) Data Standardization. Data standardization is the transformation of data through specific rules to achieve a uniform scale and scope of data. This process is especially critical in machine learning and data analytics to eliminate differences in magnitude between different features and ensure that the data is suitable for model training and analysis. In the case of regular grades and GPAs, for example, these two different magnitudes of data may affect analysis and comparison. Through standardization, the magnitude differences can be eliminated, allowing comparisons and analyses to be performed on the same scale. The 1-score standardization method used in this study converts the data to a

distribution with a mean of 0 and a standard deviation of Z by subtracting its mean and dividing by the standard deviation, which preserves the symmetry and shape of the data distribution and is particularly important for some analyses that are sensitive to the shape of the distribution.The Z-Score standardization formula is shown below:

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

Where x is the value of the specific data, μ is the mean and σ is the standard deviation, the formula is shown below:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2}$$

3. Prediction of student performance and population clustering

3.1. Student performance prediction based on GBDT algorithm

GBDT model [20] is fully known as Gradient Boosting Decision Tree.DT stands for Decision Tree and GB stands for Learning Strategy.The algorithm is a composition of multiple decision trees and is an integrated learning algorithm of Boost class. The meaning of integrated learning is that GBDT is using an additive model, i.e., accumulating multiple basis functions and continuously optimizing the error generated during training to achieve the optimal solution, and finally classifying or regressing the numerical analysis results. GBDT practical applications often use regression classification tree as a basis function, each iteration is to learn the residual value of the last round of prediction results, using a serial approach towards the direction of the residuals continue to decrease in the direction of the gradient iteration, the final prediction results are obtained by weighted accumulation of the results of each weak classifier. Given the above requirements, the classification regression trees in training are generally not very deep, commonly between 4 and 6.

The process of GBDT algorithm is as follows:

STEP1: Input training set $\{(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)\}$, number of model iterations M, loss function $L(y_i, r)$, $y_i = \{-1, 1\}$, initialize weak classifier:

$$f_0(x) = \arg_r \min \sum_{i=1}^{n} L(y_i, r) \tag{3}$$

STEP2: For $m = 1, 2, \cdots, M$, perform the following steps:

STEP2.1: For $i = 1, 2, 3, \ldots, n$, respectively, compute the approximate residual values:

$$r_{im} = -\left[\frac{\partial L(y_i f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}} \tag{4}$$

STEP2.2: Fit a categorical regression tree by approximating the residual values r_{im}, giving the leaf node domains R_{jm}, $j = 1, 2, 3, \ldots, J_m$ of the tree.

STEP2.3: Calculate for $j = 1, 2, 3, \ldots, J_m$ respectively:

$$r_{jm} = \arg_r \min \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + r) \tag{5}$$

STEP2.4: Iteratively update the classifier:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} r_{jm} I(x \in R_{jm}) \tag{6}$$

STEP3: Get the final result:

$$\hat{f}(x) = f_M(x) \tag{7}$$

For classification problems it is also necessary to convert $\hat{f}(x)$ to a probability value (positive sample probability):

$$p_i = \frac{1}{1+e^{\left(-\hat{f}(x_i)\right)}} \tag{8}$$

When solving the binary classification problem, the loss function of GBDT can use the negative binomial log-likelihood function to compute the negative gradient as an approximation of the residuals, and the resultant fit is an approximation of the residuals with respect to the classification probability. The expression for the negative binomial log-likelihood function is:

$$L(y,F) = \log\left(1+\exp\left(-2yF\right)\right), y \in \{-1,1\} \tag{9}$$

In the formula:

$$F(x) = \frac{1}{2}\log\left[\frac{\Pr(y=1\mid x)}{\Pr(y=-1\mid x)}\right] \tag{10}$$

The negative gradient approximation residual can be calculated by substituting the above equation into r_{im}:

$$r_{im} = \frac{2y_i}{1+\exp\left(2y_i F_{m-1}(x_i)\right)} \tag{11}$$

GBDT model also has a wide range of applications in the education industry, such as learning behavior research, performance prediction, MOOC research, etc. Therefore, this paper adopts the model to predict the students' performance in Civics class. GDBT has a unique advantage in feature selection, in view of the depth of the tree and the number of tree limitations, in dealing with the massive resources and features of the multi-dimensional, due to the one-hot coding generates a huge amount of resources there will be performance bottlenecks. There will be a performance bottleneck.

3.2. Improved K-prototypes clustering for student segmentation

The K-prototypes algorithm is a clustering partitioning algorithm that combines the K-means and K-modes [21] algorithms, where the K-means algorithm performs dissimilarity measures by calculating Euclidean distances for numerical data, the K-modes algorithm performs dissimilarity measures by calculating Hamming distances for categorical data, and the K-prototypes algorithm is optimized on the basis of both and can deal with mixed data by calculating Euclidean distance for numerical data and Hamming distance for categorical data for dissimilarity measure respectively and assigning different weights to them in order to get the best results.

Assume that dataset X is $\{x_1, x_2, x_i, \ldots, x_n\}$, where each piece of data such as x_i has p features, $a_1, a_2, a_3, \cdots, a_m, a_{m+1}, \cdots, a_p$; where the first m are numerical features and the last $p-m$ are categorical features. The cluster center is assumed to be $C = \{c_1, c_2, c_3, \cdots, c_k\}$, and the dissimilarity function between the cluster center and the sample values is calculated as:

$$d(x_i, c_j) = \beta d_n(x_i, c_j) + \gamma d_c(x_i, c_j) \tag{12}$$

Where $d_n(x_i, c_j)$ is the dissimilarity of numerical features, $d_c(x_i, c_j)$ is the dissimilarity of categorical features, β and γ represent the weights of the dissimilarity of numerical and categorical features, β and $\gamma \in [0,1]$, respectively, and the dissimilarity between the numerical features is calculated using the Euclidean distance, with the following function formula:

$$d_n\left(x_i,c_j\right)=\sqrt{\sum_{u=1}^{m}\left|x_{iu}-c_{ju}\right|^2} \tag{13}$$

For the dissimilarity between category-type features is calculated using the Hamming distance with the following function equation:

$$d_c\left(x_i,c_j\right)=\sum_{u=m+1}^{p}f\left(x_{iu},c_{ju}\right) \tag{14}$$

$$f\left(x_{iu},c_{ju}\right)=\begin{cases}0, x_{iu}\neq c_{ju}\\ 1, x_{iu}=c_{ju}\end{cases} \tag{15}$$

Clustering is also divided into clusters by minimizing the cost function, which is formulated as follows:

$$E\left(X,C\right)=\sum_{i=1}^{n}\sum_{j=1}^{k}U_{ij}dist\left(x_i,c_j\right) \tag{16}$$

Where U is used as the division matrix, the value range of U_{ij} is $\{0,1\}$, which represents whether x_i belongs to the jth cluster or not, if the value is 1, it means it belongs to the cluster, and the value is 0, it means it does not belong to the cluster.

According to the basic principle of K-prototypes algorithm, the process of K-prototypes algorithm is described as follows:

STEP1: Select K initial clustering centers $\{c_1,c_2,c_3,\ldots,c_k\}$ randomly from the experimental dataset $X=\{x_1,x_2,x_i,\ldots,x_n\}$;

STEP2: traverse the dataset, according to the dissimilarity formula to calculate the dissimilarity value of each x_i and c_j in the dataset, according to the dissimilarity value of the smallest division into the specified cluster;

STEP3: After each iteration, update the clustering center, calculate the mean value for numerical features, and calculate the highest frequency value for categorical features;

STEP4: determine whether the value of each cluster changes, if there is a change, repeat step2 to continue iteration until the final number of iterations, if there is no change, the end of clustering.

K-prototypes integrates the respective features of K-means and K-modes algorithms, which has the advantages of simplicity, efficiency and expandability, but there are still some deficiencies in practical use, such as the problem of how to assign the weights of numerical and categorical features. In education data, in addition to the basic information of students, such as name, gender, age, course grades, but also pay special attention to some statistical values, such as the mean, median, and the most value of the student's performance. Then it is obviously inappropriate to calculate the dissimilarity of these statistical values simply as numerical features together with other numerical features. Based on this, this paper from the special characteristics of educational data to improve the K-prototypes algorithm, the traditional dissimilarity formula and the original numerical feature distance plus category-type feature distance deformed into numerical feature distance plus category-type feature distance, plus the distance of statistical numerical features, the formula is as follows:

$$d\left(x_i,c_j\right)=\beta d_n\left(x_i,c_j\right)+\gamma d_c\left(x_i,c_j\right)+\alpha d_m\left(x_i,c_j\right) \tag{17}$$

Where $d_n\left(x_i,c_j\right)$ for the numerical features of the degree of dissimilarity, $d_c\left(x_i,c_j\right)$ for the category-type features of the degree of dissimilarity, $d_m\left(x_i,c_j\right)$ that the statistical features of the degree of dissimilarity, β, γ and α represent the numerical, category-type features and statistical features of the degree of dissimilarity of the weight, about the three weights to determine the value of the simpler is based on the number of their respective features accounted for the ratio of the number of features in the overall calculations, simple but not scientific enough. The essence of

clustering is to distinguish between the sample data with large differences, so in the determination of the weight, can we consider the sample itself attribute differences, so that the characteristics of large differences within the sample to give greater weight, for better clustering. Based on this, a weight determination method that takes into account the differences in the attributes of the sample features is proposed, and the specific steps are as follows:

STEP1: Calculate the degree of variation within each feature $\{b_1, b_2, b_3, \ldots, b_m, b_{m+1}, \ldots, b_o, b_{o+1}, \ldots b_p\}$, i.e., the number of different values of a feature, where the first m are numerical features, $m+1$ to o are categorical features, and $o+1$ to p are statistical features, assuming that for dataset X, a_i denotes the first feature, and b_i denotes the degree of variation within the ith feature;

STEP2: Calculate the mean of the internal variability of the three modular features, M_n, M_c, and M_m, respectively;

$$M_n = \sum_{i=1}^{m} b_i \tag{18}$$

$$M_c = \sum_{i=m+1}^{o} b_i \tag{19}$$

$$M_m = \sum_{i=0+1}^{p} b_i \tag{20}$$

STEP3: Calculate numerical feature, category feature, statistical feature weights β, γ & α, where:

$$\beta = \frac{M_n}{M_n + M_C + M_m} \tag{21}$$

$$\gamma = \frac{M_c}{M_n + M_C + M_m} \tag{22}$$

$$\alpha = \frac{M_m}{M_n + M_C + M_m} \tag{23}$$

STEP4: Calculate the degree of dissimilarity between data samples for clustering division. In this way, this paper is completed to divide the student population.

4. Visualization of student population profiles

4.1. Group portrait construction process

Compared with individual student portraits, the construction of student group portraits is more complex and requires more technical tools. In this section, based on the automatic mining of student data, students are categorized into different groups, and each group of students has its own common characteristics. According to the learning situation of each type of student group, more targeted teaching strategies are designed to improve the personalized learning effect of students. Based on this, this paper proposes a student group portrait method based on clustering integration to better serve students on the premise of mastering a large amount of student data. This method firstly analyzes students' basic data and behavioral data to establish reasonable feature labels of students' group portrait, secondly constructs base clusters using three methods of KMeans, KModes and GMM clustering, and finally realizes the integrated processing of the results of base clusters through the voting method and constructs the students' group portraits by selecting the appropriate number of classes according to the profile coefficients, and at the same time analyzes the main characteristics of

each group, which provides a reference for the subsequent teachers to develop classification teaching tools and strategies to provide reference. Specific implementation steps include: data collection and processing, group portrait feature label construction, base classifier construction, cluster integration, group characterization and student group portrait, Figure 2 shows a brief flow of group user portrait construction.

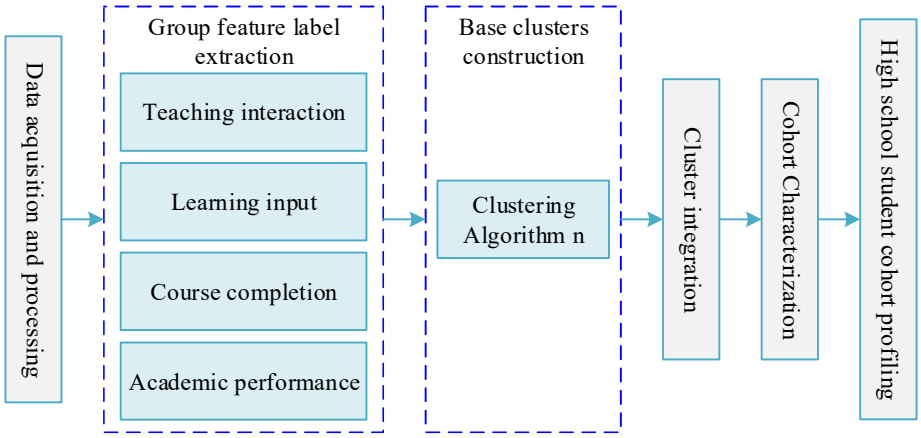

Fig. 2. Construction Fra nework of Group Profile

4.2. Population feature label extraction

Mining and refining students' attribute features from data and labeling the attribute features is a necessary way to build a student group portrait. Students' categorized features contain explicit and implicit features, in which explicit features include students' gender, political profile, ethnicity, whether they participate in evening self-study and place of birth, etc., and implicit features include students' disciplinary cases, total monthly consumption amount, etc. By establishing the label dimensions of students' explicit and implicit features, we use clustering algorithm to divide them into different feature groups and accurately analyze students' learning in various aspects. In clustering algorithms, feature selection directly affects the clustering effect, so the most important features related to students' performance need to be selected.1 Changchun Yang et al. showed that the Random Forest algorithm, as a tool for feature selection for students' portraits, is both robust and well able to deal with abnormal datasets. Thus, in this paper, for each portrait dimension, the Random Forest algorithm is used to select the top three feature importance rankings of student disciplinary infractions $(X1)$, whether they participate in evening self-study $(X2)$, and the total monthly consumption amount $(X3)$ as feature labels for the student population portrait.

5. Analysis of the application of student portraits in civic education

5.1. Cluster Analysis of Student Portraits

5.1.1. Analysis of learning outcome dimensions. The criteria for routinely determining learning outcomes generally use performance indicators, including regular grades and test scores. The author analyzes the learning outcome dimensions with respect to students' usual grades and test scores during their school years.

Students' usual performance is explored through statistical analysis of usual grades. Figure 3 shows the students' usual grades, numbered 1-8 represent the first and second semesters of the four academic years from freshman to senior year, and in general, the students' usual performance is better, and the average of the usual grades of the academic year is around 80 points. College students tend to

perform poorly in the sophomore and junior years, indicating that students are prone to relax their vigilance in the sophomore and junior years, are not active in class, and develop a slacking mentality.

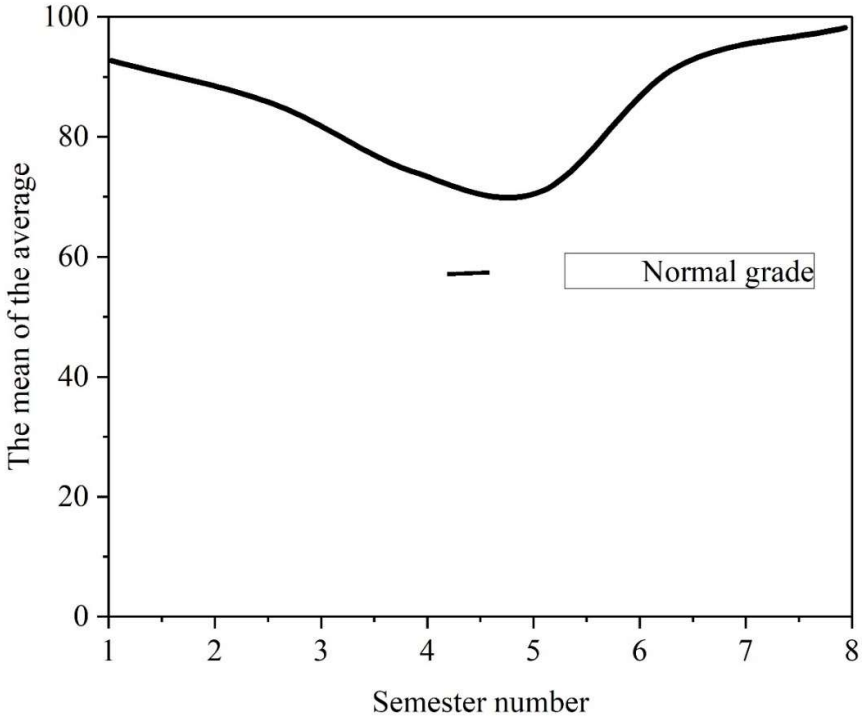

Fig. 3. Students' Usual Results

Based on the analysis of the data in each field, the average of the mapping quiz, midterm grade, final grade, mock exam grade and the total assessment grade of the examination course were selected as the students' learning outcomes. Figure 4 shows the students' examination results in the academic year 2018-2020, and numbers 1-30 are the average values of the mapping quiz, midterm grades, final grades, mock examination grades and total assessment grades of examination classes in the first and second semesters of 2018-2020, respectively, which shows that the students' results are not very stable in general.

Students performed poorly on the touch-tests in the 2018, 2019, and 2020 academic years, but after one semester of study, students' academic performance improved significantly. This indicates that good face-to-face classroom instruction helps to improve students' academic performance.

Through the descriptive statistics and visualization analysis of the students, the study found the following: in terms of demographic dimensions, there is little difference between male and female genders for most of the students, most of the students are members of the Youth League, and the number of first year students is higher, but in the students' age field, there is a big difference in the age of the students, and the students are distributed in the age range of 18-24 years old.

In terms of the behavioral pattern dimension, from the students' attendance results, participation in evening study and the number of breakfasts, it can be seen that students who participate in evening study at school study harder and have a more stable life pattern. From the students' highest single consumption, total monthly consumption amount and total monthly consumption frequency, the students' consumption level varies greatly, and the consumption amount and consumption frequency show different variability.

From the perspective of students' uniform violation rate, late arrival frequency and early departure frequency, the number of disciplinary violations at the beginning and the end of the semester was the highest, and students' late arrival and late departure behaviors were more serious. In terms of learning outcome dimensions, students' test scores were less stable, especially during the touch-tests, and

students performed poorly. Students were less concerned about the results of their regular grades, mainly because students in high school usually perform better and most of their regular grades are good, and perhaps students care more about outcome-based assessment.

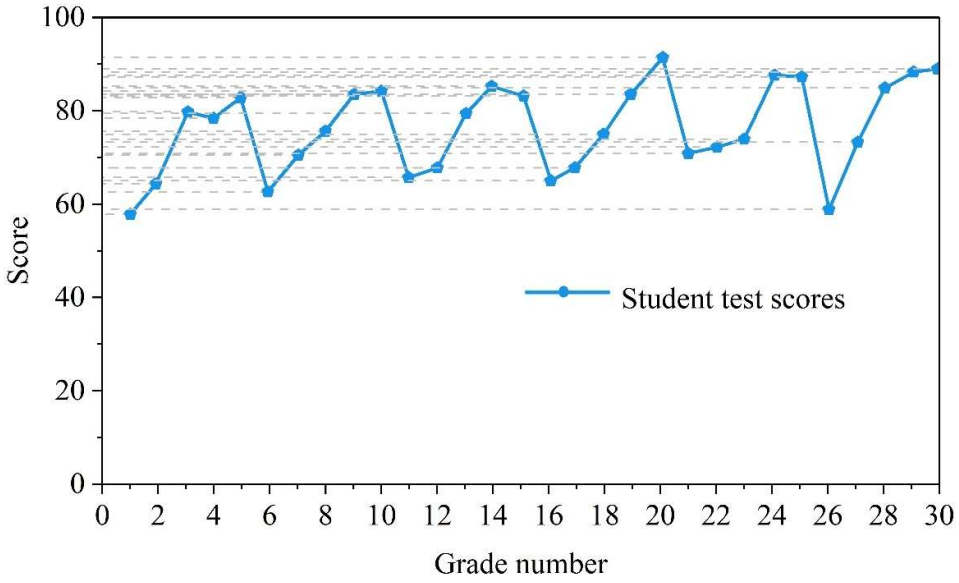

Fig. 4. Students' Examination Results

5.1.2. Student Portrait Group Visualization Analysis. In order to classify the learners of Mit6.00 course, it is necessary to explore the relationship between the four characteristics of course completion, teaching interaction, learning engagement and learning achievement in the dataset. A total of 41,450 pieces of data were extracted to analyze the learning behavior of learners, accounting for about 1/3 of the total, and 7,700 pieces of available data were obtained after data cleaning, and the significance analysis of the four characteristics was carried out. The significance between the four characteristics of "course completion characteristics", "teaching interaction characteristics", "learning engagement characteristics" and "learning characteristics" was less than 0.001, and the significance was obvious. The Pearson correlation between the four traits was greater than 0.6, indicating a strong correlation or a very strong correlation. Therefore, this study can analyze the group portrait from the dimensions of these four characteristics based on the learning behavior data of course completionists.

Due to the inconsistency between the four characteristics of "course completion characteristics", "teaching interaction characteristics", "learning engagement characteristics" and "learning achievement characteristics", this paper uses the Z-Score standardization method to turn the inconsistent data into dimensionless data, that is, the data of different magnitudes are processed to a level, so that the different data are comparable to each other. Table 1 shows some of the normalized Z-Score data.

Table 1. Four examples of the standardized post-standardized data

Interactive characteristics of teaching	Learning input feature	Course completion characteristics	Academic characteristics	Z Interactive characteristics of teaching	Z Learning input feature	Z Course completion characteristics	Z Academic characteristics
7124	63	18	0.84	1.2731	1.67848	1.87424	2.43697

15568	103	17	0.89	3.68925	3.21844	2.06168	2.50621
199	7	1	0.01	-0.7084	-0.67181	-0.93694	-0.57475
1861	23	4	0.01	-0.23288	-0.02343	-0.56212	-0.5402
69	11	2	0.03	-0.7453	-0.46927	-0.74944	-0.57479
7542	70	18	0.91	1.39278	1.881	2.06163	2.5061
293	11	3	0	-0.68092	-0.38812	-0.93696	-0.57478
14414	119	18	0.96	3.35917	3.78578	2.06166	2.71395
1985	29	2	0.04	-0.19769	0.13863	-0.5621	-0.47093
4066	55	15	0.33	0.39803	1.31378	1.31191	0.49834

After data standardization, the data were clustered using the K-prototypes clustering method, and after several attempts, the best clustering effect was achieved with K=4. Tables 2 and 3 show the initial clustering centers and the final clustering centers, in which Zscore1 denotes the instructional interaction features, Zscore2 denotes the learning input features, and Zscore3 denotes the course completion features, Zscore4 denotes learning achievement features, and Table 4 shows the number of cases in each cluster.

According to the four characteristics of "course completion characteristics", "teaching interaction characteristics", "learning engagement characteristics" and "learning achievement characteristics", the learners of Mit6.00 course were clustered into four categories, and the learners in class cluster 1, class cluster 2, class cluster 3 and class cluster 4 were called group A, group B, group C and group D, respectively.

Group A has higher instructional interaction characteristics, indicating that learners in group A watch instructional videos and course interactions more often and more actively, learners in group B are less active in instructional interactions relative to group A, and groups D and C are second most active. In terms of learning engagement characteristics, learners in group A still have the highest learning engagement, learners in group B have relatively less engagement than learners in group A, and group D has a slightly higher engagement compared to group C. However, there is a large gap between the two in terms of engagement compared to group A and B. In terms of course completion, group A is the highest, group B is the second highest, compared to the teaching interaction characteristics, learning engagement characteristics, the gap between the two is not obvious, group D, group C course completion characteristics are not high. In terms of learning achievement characteristics, Group A and Group B are relatively high, and there is no significant difference in achievement between Group A and Group B. Group D and Group C are relatively low.

Table 2. The class are analyzed after the k-means clustering

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Zscore1	11.17473	5.70873	-0.69889	0.7096
Zscore2	0.66546	5.28509	-0.87448	2.32692
Zscore3	1.87424	2.06165	-1.12434	2.06171
Zscore4	2.15999	2.61005	-0.57482	2.71392

Table 3. After the k-means clustering analysis, the final cluster center

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Zscore1	2.95716	1.16119	-0.49703	0.30005
Zscore2	2.62447	1.46188	-0.54016	0.37291
Zscore3	1.96068	1.77523	-0.6067	0.64648
Zscore4	2.18972	1.96433	-0.52108	0.06344

Table 4. Course learners were analyzed by k-means clustering

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Case number	398	891	5210	1201
Proportion/%	5.17	11.57	67.66	15.60

In order to further conduct an in-depth analysis of the clustered four types of learner groups, this study classifies and names different learner groups from four dimensions based on the existing analysis. The learner groups are divided into the following four categories, group A is named as the "hard learner" learner group, group B is named as the "gifted" learner group, group C is named as the "experience" learner group, and group D is named as the "potential" learner group, and the specific explanations of the four learning groups are shown in Table 5.

Table 5. Class 4 types of group types description

Learner group	Group type description
Group A: try to learn bully learners	This kind of learner is excellent in learning, and the degree of input, the degree of teaching and the degree of course is very high. The teaching interaction is so much that it is happy to communicate with the teachers and students, and the dropout rate is very low.
Group B: talented learners	The degree of input degree and teaching interaction is larger than that of the first type of learners, but the difference between the study and the study of the students is very small, and the prediction is that the course is higher, so the learning results are better, and the students can learn better when they are able to invest more time in the course.
Group C: empirical learner	This class is the lowest level of learning, the degree of learning, the degree of teaching, and the degree of course completion. On the other hand, the difficulty of the course is too much for such learners, which leads to boredom and leads to high dropout rates.
Group D: potential learners	This class of learners is generally less engaged in learning, the course is less frequent, the course is less complete, and such learners need to arouse their interest in learning, otherwise they will be more easily lost.

5.2. Analysis of the Results of Curriculum Civics

The results of 100 questionnaire results of the first level dimensions and the overall and pre-test results were analyzed by paired samples t-test, and the results are shown in Table 6.

The t-value of the cultural literacy dimension is -5.624, $p < 0.05$, indicating that there is a significant difference in the results in the cultural literacy dimension and the three rounds of scores are higher than the pre-test scores; the t-value of the value concept dimension is -1.678, $p > 0.05$, indicating that there is no significant difference in the results in the value concept dimension, but the three rounds of scores are slightly higher than the pre-test scores. The t-value for the information perception dimension was -4.068, $p < 0.05$, indicating that there was a significant difference in the results on the information perception dimension and that the three rounds of scores were higher than the pre-test scores. The t-value for the overall was -4.112, $p < 0.05$, indicating that there was a significant difference in the results at the overall level and that the three rounds of scores were higher than the pre-test scores.

After three rounds of action research, in terms of course ideology, the mean of students' cultural literacy increased by about 0.349 compared to the pre-test, the mean of value concepts increased by about 0.105 compared to the pre-test, the mean of information cognition increased by about 0.272 compared to the pre-test, and the overall mean increased by about 0.312 compared to the pre-test. Among them, the students' recognition of value concepts is still the highest, and compared to the pre-test, the three dimensions of endorsement and the overall mean of the questionnaire increased. In terms of standard deviation, the standard deviation of cultural literacy is the smallest, indicating that students' level of cultural literacy varies the least.

Table 6. The matching sample t test of the course and the first order dimension

Pairing	Dimension	Case number	Mean value	Standard deviation	t	Sig.
Pair 1	Pre-survey culture	100	3.477	0.465	-5.664	0.000
	Post-cultural literacy	100	3.826	0.528		
Pair 2	Former value concept	100	3.896	0.482	-1.678	0.095
	Post- value ideas	100	4.001	0.562		
Pair 3	Pre-measurement information cognition	100	3.554	0.467	-4.068	0.000
	Post-information cognition	100	3.826	0.593		
Pair 4	Premeasurement	100	3.684	0.415	-4.112	0.000
	Aftermeasure	100	3.996	0.486		

The paired-sample t-test was conducted to analyze the secondary dimensions of the 150 third-round questionnaire results with the pre-test results, and the test results are shown in Table 7:

The t-values of the six dimensions of rational thinking, digital innovation awareness, integrity and friendliness, national sentiment, information thinking and information ethics and security are all negative and $p < 0.05$, indicating that there are significant differences in the results of these six dimensions, and the third-round scores are significantly higher than the pre-test scores. The $p < 0.05$ but positive t-value for the Ideal Beliefs dimension indicates that there is a significant difference in the results for this dimension, but the three rounds of scores are lower than the pre-test scores. The t-values for the dimensions of scientific spirit, social responsibility and information awareness are negative, but the $p > 0.05$ indicates that there is no significant difference in the results of these three dimensions, but the scores under this paper's pedagogical method are still higher than the pre-test scores.

After the action research in this paper, from the mean value, the more obvious improvement is the rational thinking, digital innovation consciousness, integrity and friendliness, family and national sentiment and information thinking dimensions, all of which are higher than the average of the action research in the pre-test by more than 0.25, and except for the ideal belief dimension, the average of all other dimensions are slightly improved compared with the pre-test.

To summarize, in terms of curriculum ideology, after the action research in this paper, the overall level of students' curriculum ideology has been significantly improved, with cultural literacy and information cognition in the first-level dimensions and the six dimensions in the second-level dimensions, and all the first-level and second-level dimensions, except for the ideal belief dimension in the second-level dimensions, have been improved to a certain extent. This indicates that the action research in this paper has positively affected the students' level of course ideology in general.

Table 7. The t test of the secondary dimensions of the course

Pairing	Dimension	Case number	Mean value	Standard deviation	t	Sig.
Pair 1	Pre-scientific spirit	100	3.948	0.715	-1.554	0.124
	Post-scientific spirit	100	4.087	0.706		
Pair 2	Premeasured thinking	100	3.167	0.578	-7.635	0.000
	Post-rational thinking	100	3.726	0.647		
Pair 3	Digital innovation awareness	100	3.309	0.577	-4.756	0.000
	Digital innovation awareness	100	3.665	0.632		
Pair 4	Honest and friendly	100	3.642	0.556	-3.796	0.000
	The latter measure sincerity and friendliness	100	3.917	0.637		
Pair 5	Former social responsibility	100	4.029	0.635	-0.658	0.516
	Post-urban responsibility	100	4.082	0.677		
Pair 6	The former surveyor's national conditions	100	3.924	0.698	-3.006	0.004
	And then we look at our national conditions	100	4.183	0.736		
Pair 7	Pretest ideal	100	3.987	0.612	-2.063	0.043
	Post-ideal belief	100	3.835	0.647		
Pair 8	Pre-measurement information consciousness	100	3.812	0.678	-1.022	0.306
	Aftersense information consciousness	100	3.906	0.789		
Pair 9	Pre-measurement information thinking	100	3.132	0.746	-4.895	0.000
	Postmeasured information thinking	100	3.644	0.965		
Pair 10	Former information ethics and security	100	3.738	0.623	-2.488	0.017
	Post-mortem information morality and safety	100	3.926	0.664		

5.3. Characterization of Teachers' and Students' Verbal Behavior

The common characteristics of teacher-student verbal behaviors in the traditional classroom and this paper's classroom model based on student portraits are shown in Table 8. Teachers' classroom speech behaviors consist of lecturing, instructing, criticizing, demonstrating, correcting, encouraging and praising students, adopting students' opinions, and asking questions. The ratio of teacher speech in the classroom model and traditional classroom model of Civic Education in this article is 49.55% and 55.21%, respectively. Students' speech behaviors include passive response, active response, active questioning, and peer discussion. The rates of student speech in the classroom model and traditional classroom model of Civic Education in this article were 40.32% and 43.78%, respectively. These values are relatively close to each other. This reflects the balanced behavior of teachers and students throughout the classroom discourse interactions, and both classroom models have good teacher-student interactions. Teachers did not control the classroom discourse and students had ample opportunities for classroom language practice, especially listening and speaking communication exercises.

The data also showed that the direct teacher influence ratio was 43.89% and 33.12% for this text's Civic Education classroom model and traditional classroom model, respectively, while the indirect teacher influence ratio was 16.54% and 12.56% for this text's Civic Education classroom model and traditional classroom model, respectively. The direct influence ratio of teachers in both classroom models is greater than the indirect influence ratio, which indicates that teachers dominate the classroom interaction activities. In ideological and political classroom teaching, the teacher's role should be a designer of the learning environment or an organizer of interactive teaching. His duty is to provide students with as many learning scenarios as possible. Therefore, this common set of data suggests that teachers in both the classroom model of ideological and political education in this text and the traditional classroom model are striving to create an interactive classroom that encourages student participation.

Table 8. The common characteristics of speech behavior of teachers and students

Verbal ratio	This class/%	Tradition hall/%
Teacher speech ratio	49.55	55.21
Student speech ratio	40.32	43.78
Teacher direct influence ratio	43.89	33.12
Teacher indirect influence ratio	16.54	12.56

Some of the teacher-student verbal behavior ratio differences between the two classroom models were highly significant, as shown in Table 9 for specific details. Teacher's indirect influence contains the teacher's behavior of asking open questions and asking closed questions. In this article, the teacher of the classroom model of Civic Education to be more focused on raising open questions, the rate of this behavior is 49.35%. Compared with the traditional classroom model of 11.98%, the difference is very significant. While the rate of teachers to raise closed questions traditional classroom model is 88.56%, the data is significantly higher than the 50.89% of the classroom model of Civic Education in this article. Under the premise that the ratio of teachers' indirect influence behaviors is close, the significant difference between these two sets of data reflects the different learning trajectories of students in the flipped classroom and the conventional classroom. Since the answers to closed questions are fixed, the teacher's role in asking closed questions is not very useful for students' creative language output. Whereas the answers to open questions require active thinking and application of knowledge to arrive at, teachers asking open questions are considered an important pedagogical behavior for training comprehensible output in language learning. This suggests that the classroom model of Civic Education in this article has transformed the teaching process of “knowledge transmission” in class and “knowledge internalization” in class.

Students' active response behaviors include students' free expression of their own ideas and solutions, presentation of the results of self-discussion, and role-playing. The active response rates of students in this Civics Education classroom model and the traditional classroom model are 47.565% and 27.94% respectively, and the passive response rates are 14.35% and 29.56%, which indicate that students in this Civics Education classroom model not only output a large amount of comprehensible language but also have more active thinking.

The data in Table 9 also reflects that the silence rate of 10.17% in this Civics Education classroom model is higher than that of 1.29% in the traditional classroom model. Through classroom observation, it can be seen that there are no disruptive activities that do not contribute to the classroom in both classrooms, and the silent activities mainly come from students' thinking and practicing. This set of data reflects that the teachers of the classroom model of Civic Education in this article pay more attention to providing opportunities for students to think and apply their knowledge for language practice when organizing classroom activities.

Table 9. Differences in the verbal behavior of teachers and students

Verbal ratio	This class/%	Tradition hall/%
The teacher raises the question ratio	49.35	11.98
The teacher raised the problem ratio	50.89	88.56
Students' active response ratio	47.56	27.94
Student passive response ratio	14.36	29.56
Classroom silence ratio	10.17	1.29

In order to further study the characteristics of teachers' and students' speech behaviors in the classroom, the author drew a graph of speech behavior characteristics. The horizontal coordinate of the graph indicates the unit of time per minute, and the vertical coordinate indicates the ratio of teacher's or student's behavior in each minute. The graph is used to describe the behavioral ratios of teachers or students over time. Figure 5 shows the comparison curve of teacher speech characteristics between the classroom model of Civic Education and the traditional classroom model in this article, and Figure 6 shows the comparison curve of student speech characteristics between the classroom model of Civic Education and the traditional classroom model in this article.

In the traditional classroom model, the teacher's verbal behavior appears intensively before 20 minutes. The peak of students' verbal behavior in the traditional classroom mode occurs after 20 minutes. Teachers' verbal behaviors are more than students' verbal behaviors in the first half of the class, which highlights the teaching characteristics of the traditional classroom model of classroom teachers to promote "output" with "input". Combined with classroom observation, this study found that teachers in the traditional classroom model first guided students to become interested in classroom topics through warm-up exercises. Therefore, in the first half of the class, the teacher's discourse activity is more, and the students' response is relatively passive. As the course progressed, the teacher organized the students to have a student-to-student discussion, and the teacher's verbal behavior decreased, and the students' verbal behavior gradually increased.

The teacher verbal behavior of the classroom model of Civic Education in this article was distributed more evenly throughout the classroom. The peak state starts from the first 5 minutes of the classroom and then appears at intervals until the end of the classroom a total of 11 times. The students' verbal behavior of the classroom model of Civic Education in this article appeared in the first peak state from the first 5 minutes of the classroom, and then appeared at intervals, with a total of 10 times. From the figure, it can be seen that the teacher's speech and student's speech in the classroom model of this article's Civics Education are interacting frequently and in a balanced way from the beginning to the end of the classroom, and there is no scene in which the teacher's one-sided lecturing or the student's

speech control the classroom. This text's Civic Education classroom model allows teachers and students to quickly enter into a state of verbal interaction from the beginning of the class, which may be due to the change in students' learning styles. Since the background and main content of the topics covered in the classroom had been conveyed to the students through the learning platform before the class, the diverse exercises organized by the teacher in the classroom enabled the students to respond quickly.

The classroom model of Civic Education in this text has more frequent verbal interactions between teachers and students, and is more closely integrated. The classroom model of Civics education in this article provides real and positive activities and feedback for student-student interaction and teacher-student interaction, which can help students achieve the purpose of learning Civics education knowledge and better promote the mastery of Civics knowledge.

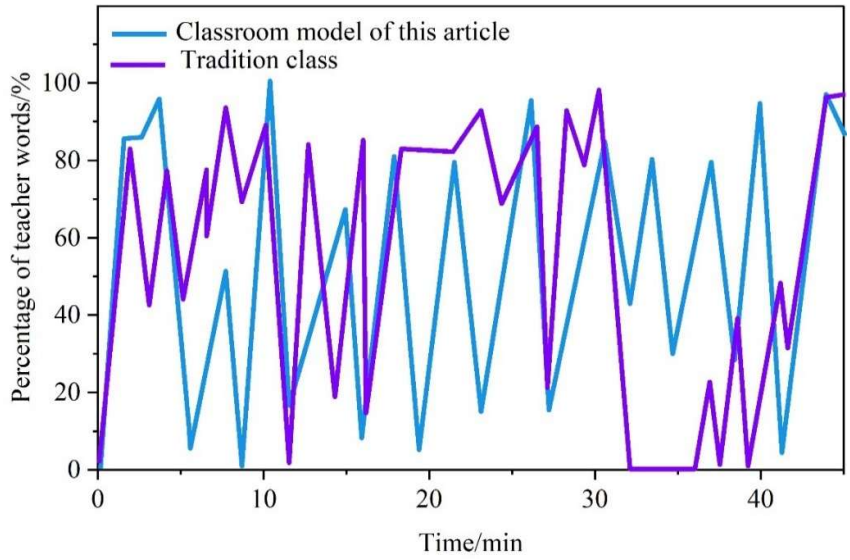

Fig. 5. Teacher's speech characteristic curve

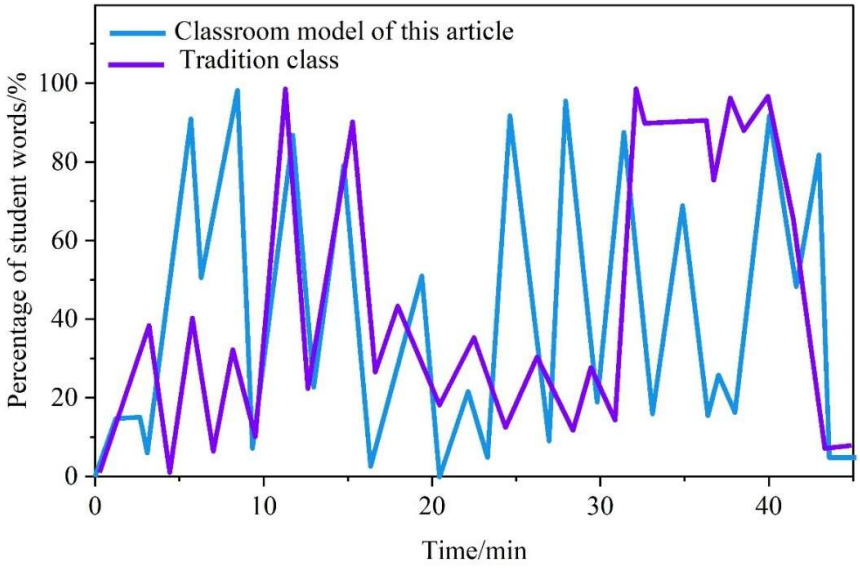

Fig. 6. Student speech characteristic curve

6. Conclusion

This study investigates the teaching practice of integrating the Civics education classroom with student profiling, grade prediction, and group clustering methods, and verifies the feasibility of this paper's instructional design that combines curricular Civics and data visualization techniques.

The overall situation of students' test scores is more unstable. However, its after the teaching mode designed in this paper, the students' academic performance in Civic and Political Education has been significantly improved. This indicates that the classroom teaching mode based on data visualization technology in this paper helps to improve students' academic performance.

The clustering algorithm divides the student groups into four categories with obvious behavioral characteristics, namely the "hard-working" learners, the "talented" learners, the "experiential" learners and the "potential" learners, which provides a basis for teachers' targeted teaching.

In terms of course civics, all of the secondary dimensions related to students and civics, except for the ideal belief dimension, showed significant improvement in the practice process. This shows that the overall level of students' overall course ideology and politics has been significantly improved, and the action of this paper can significantly improve the level of students' course ideology and politics.

Comparing this paper's innovative teaching classroom with the traditional classroom model, the results of the study of teacher-student verbal interaction behaviors in the classroom found that this paper's teaching model stimulated students' thought dispersion and language expression output, and the peak of students' verbal behaviors appeared in the classroom a total of 10 times, and the active response rate of the students in this paper's classroom model increased by 19.625%. It shows that the teaching mode of this paper promotes students' thinking and verbalization in the process of teacher-student interaction.

According to the above conclusions, for the teaching effect of the Civics classroom in colleges and universities under the data visualization technology to improve the teaching and reform, the author has the following suggestions: (1) Teachers should reasonably arrange the visualization classroom and the application of the model according to the course content and objectives to better achieve the teaching effect (2) Teachers through the data visualization technology to strengthen the observation, recording and analysis of students' classroom behavior, so as to improve the teaching method to improve the classroom efficiency.

.

References

[1] Seyi-Lande, O. B., Johnson, E., Adeleke, G. S., Amajuoyi, C. P., & Simpson, B. D. (2024). The role of data visualization in strategic decision making: Case studies from the tech industry. Computer Science & IT Research Journal, 5(6), 1374-1390.

[2] O'Donoghue, S. I., Baldi, B. F., Clark, S. J., Darling, A. E., Hogan, J. M., Kaur, S., ... & Procter, J. B. (2018). Visualization of biomedical data. Annual Review of Biomedical Data Science, 1(1), 275-304.

[3] Ansari, B., Barati, M., & Martin, E. G. (2022). Enhancing the usability and usefulness of open government data: A comprehensive review of the state of open government data visualization research. Government Information Quarterly, 39(1), 101657.

[4] Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., ... & Milic, N. M. (2019). Reveal, don't conceal: transforming data visualization to improve transparency. Circulation, 140(18), 1506-1518.

[5] Wilke, C. O. (2019). Fundamentals of data visualization: a primer on making informative and compelling figures. O'Reilly Media.

[6] Saçak, B., Bozkurt, A., & Wagner, E. (2022). Learning design versus instructional design: A bibliometric

study through data visualization approaches. Education Sciences, 12(11), 752.

[7] Ho, E., & Jeon, M. (2023). Interaction map: A visualization tool for personalized learning based on assessment data. Psych, 5(4), 1140-1155.

[8] Kay, J., Graesser, A. C., & Sinatra, A. M. (2023). DATA VISUALIZATION IN INTELLIGENT TUTORING SYSTEMS SWOTANALYSIS. Design Recommendations for Intelligent Tutoring Systems, 115.

[9] Chiang, J. L., Zhang, Z. H., & Cheng, H. C. (2022, September). Supporting Outcome-Based Education Assessment and Evaluation-Educational Technology Solution of a Data Visualization Framework. In 2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE) (pp. 219-224). IEEE.

[10] Fitri, S., & Lubis, M. (2024). An evaluation on online learning by data visualization: a case study from information technology education program. Jurnal Penelitian Humaniora, 25(2), 67-80.

[11] Makowski, M. B., & Lubienski, S. T. (2023). Classroom data visualization: tracking individuals during group-centered instruction. Educational Researcher, 52(3), 164-169.

[12] Fitri, S., Muhammad, T., Riki, C., & Sampath, S. (2023). Data Visualization for Students' Perception Toward Online and Offline Learning in Information Technology Education Program. International Journal of Quantitative Research and Modeling, 4(3), 178-184.

[13] Mavrikis, M., Geraniou, E., Gutierrez Santos, S., & Poulovassilis, A. (2019). Intelligent analysis and data visualisation for teacher assistance tools: The case of exploratory learning. British Journal of Educational Technology, 50(6), 2920-2942.

[14] Nagaraj, P., & Raja, M. (2024). Development of Inquiry-based Active Learning Pedagogy Approach to Heightening Learners' Critical Thinking Skills in Data Visualization for Analytics Course. Journal of Engineering Education Transformations, 37.

[15] Asamoah, D. (2022). Improving Data Visualization Skills: A Curriculum Design. International Journal of Education and Development Using Information and Communication Technology, 18(1), 213-235.

[16] Ahn, J., Nguyen, H., & Campos, F. (2021). From visible to understandable: Designing for teacher agency in education data visualizations. Contemporary Issues in Technology and Teacher Education, 21(1), 155-186.

[17] Gerela, P., Mishra, P. N., & Vipat, R. (2022). Study on data visualization: It's importance in education sector. International journal of health sciences, 6(S3), 6298-6305.

[18] Shahidan, N. F. A., Ibrahim, A. F., & Jamaluddin, M. N. F. (2024). Data Visualization of Student Academic Performance Analysis. Applied Mathematics and Computational Intelligence (AMCI), 13(4), 49-61.

[19] He Lyuying. (2024). Hotel Loyalty Promotion Mechanism Based on User Portrait. Tourism Management and Technology Economy(3),

[20] Yuxi He,Kaiwei Luo,Han Ni,Wentao Kuang,Liuyi Fu,Shanghui Yi... & Wenting Zha. (2024). Quantitative assessment of Public Health and Social Measures implementation and relaxation on influenza transmission during COVID-19 in China: SEIABR and GBDT models.. Journal of global health05038.

[21] Gavva Surya Teja,Karthik C. S.& Punna Sharath. (2024). Clustering categorical data: Soft rounding k-modes. Information and Computation.