International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 5118-5128

https://doi.org/10.70517/ijhsa463439

Analysis of Innovative Application of Artificial Intelligence Technology and Teaching Transformation in English Education in Colleges and Universities

Li Yang[1,*]
[1] Shanxian College, Heze University, Heze, Shandong, 274015, China
Corresponding authors: (e-mail: 15053029358@163.com).

Abstract This study focuses on the innovative application of artificial intelligence technology and teaching transformation in English education in colleges and universities, constructs a three-level theoretical framework (student model, teacher model, domain model) for intelligent computer-assisted instruction ICAI system, and explores the data-driven teaching optimization path by combining the time-series clustering and decision tree algorithms. Based on the real data from the intelligent teaching platform of a university, the study clusters students' learning behaviors through the time-series k-means algorithm KmL, identifies four types of differentiated learning groups, efficient learners (N=625), task-oriented (N=3011), passive participants (N=4276), and passive groups (N=247), and reveals their behavioral characteristics in resource use, interactive participation, and other 10 behavioral characteristics in the dimensions of resource use, interaction participation, etc. The decision tree algorithm was further utilized to mine the academic performance association rules, and found that classroom mastery, listening time and vocabulary were the core factors affecting the performance, such as the Rule 1 confidence level of 61.23%. The study shows that the data-driven ICAI system can realize the dynamic adaptation of teaching strategies, provide technical support for personalized teaching and precise intervention, and promote the transformation of English education in colleges and universities to intelligence and refinement.

Index Terms ICAI system, temporal clustering, decision tree algorithm, college English education, learning behavior

I. Introduction

With the acceleration of the development process of economic globalization, the demand for English talents in globalized communication is also increasing. Colleges and universities, as the training base for talents of various specialties, need to keep abreast of the times in their English teaching, and cultivate more high-quality English talents from the point of view of social demand and national development needs [1]. In order to further improve the quality of English education and teaching in colleges and universities, there is an urgent need for the education field to carry out teaching reform with the help of artificial intelligence and other advanced technologies [2].

Influenced by the employment environment and economic globalization, the demand for English professionals in the social environment and the rapid development of information technology have prompted the English language in colleges and universities to have to seek a breakthrough, reform and innovate the education model and education concept [3]-[5]. In the era of artificial intelligence, teachers need to continuously improve their information technology application ability, master various teaching tools and platforms, and actively explore new teaching methods and means, innovate the teaching mode, and improve the teaching effect and quality [6]-[8]. The application of artificial intelligence technology contributes to the intelligent development of English education and teaching, and makes an important contribution to the intelligent change of teaching resources, teaching mode and teaching tools in the education process [9], [10]. Through the reform practice in these aspects, the development of English teaching in colleges and universities will be more modernized and intelligent, which will help China cultivate more high-quality English talents to meet the needs of social development [11], [12]. In the future, educators should also actively explore the specific application of artificial intelligence and other new technologies in the field of English education and teaching to promote the innovative development of English education [13].

This study focuses on constructing a theoretical framework for the ICAI system of intelligent computer-aided instruction and exploring data-driven teaching optimization paths by combining temporal clustering and decision tree algorithms, aiming to provide precise and dynamic technical support for English education. The article first proposes a three-level theoretical model (student model, teacher model, domain model) of the ICAI system, which realizes the dynamic adaptation of teaching strategies by simulating the teacher's decision-making mechanism and students' cognitive characteristics. The student model constructs a multidimensional learner portrait through historical learning data and real-time assessment. The teacher model relies on the teaching strategy library and intelligent reasoning machine to generate personalized teaching paths. The domain model integrates the subject knowledge base and teaching experience base to form a structured knowledge network. For the temporal data characteristics generated in the teaching process, the temporal k-mean algorithm (KmL) is introduced to analyze the learning trajectory through Euclidean distance and dynamic time regularization (DTW). The algorithm can identify the stage-specific learning pattern differences of the student population, and in order to further explore the teaching decision-making law, the decision tree algorithm is used to construct a classification prediction model. Key features are screened by information entropy and Gini coefficient to generate interpretable rule sets.

II. Research on the construction of personalized teaching model and data analysis method based on ICAI system

II. A. Theoretical design model of ICAI system

The structure of the ICAI system model is shown in Figure 1. A detailed overview of the student, instructor, and domain models follows.

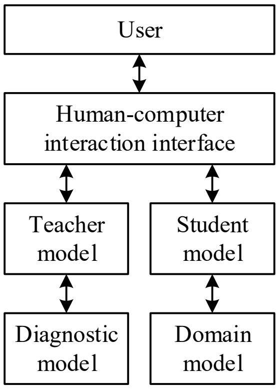

Figure 1: Structure diagram of the ICAI system model

II. A. 1) Student model

The student model is the core of ICAI's implementation of individualized instruction, which is the system's characterization of the student's knowledge level, cognitive ability, motivation to learn, learning style and learning history variation and other information after the student passes the study and test in the process of systematic instruction. Its main role is to provide a judgment basis for the realization of teaching objectives, teaching contents and teaching strategies so that the system can establish suitable individualized teaching according to students' characteristics.

In the construction of the student model, a construction method with learning history inheritance is used. When a student is learning new knowledge, there is an assessment test of his/her previous learning history, and the corresponding learning strategy is given according to the assessment value. If they pass, they can proceed to the next stage of learning; otherwise, repeat learning or supplementary exercises are given. Only those who pass the assessment can move to the next stage of learning.

II. A. 2) Teacher model

Teacher model is the center of the intelligent teaching system for organizing, managing and implementing the whole teaching activities, whose role is to combine the students' goals and actual levels, synthesize the learning results provided by the student module, analyze the current state of the students, make a teaching strategy, choose the most effective teaching methods for the students to carry out the teaching activities, supervise and evaluate the learning effect of the students, and realize individualized instruction.

The teacher model consists of two parts: the intelligent reasoning machine and the teaching strategy library. The teaching strategy inventory puts inference rules, including teaching methods, teaching experience and best path mathematical strategy, which is the basis of the reasoning machine work. The reasoning machine is a light system that coordinates the whole ICAI system and applies the knowledge in the knowledge base based on the current student hindsight in the database and then generates teaching strategies that are individualized for the students based on the teaching strategy library.

II. A. 3) 2.1.3 Domain model

The domain model contains a student knowledge base and a teacher knowledge base. The student knowledge base contains personal information about students, such as learning history, knowledge level, problem solving, study, learning ability, and overall level. This information provides information for the system to make judgments about students, select learning content, and arrange the organization and difficulty of the content. Teacher's knowledge base is a knowledge base consisting of the teacher's knowledge, which mainly includes teaching knowledge in terms of teaching strategy of the content and course structure, such as the key points and difficulties of the textbook, the connection of each part, the methods of introduction and expansion, and the rules of solving problems.

The expression of knowledge is the primary problem to be solved in establishing a knowledge base system. How to formalize and structure domain knowledge so that it can be stored and utilized in a knowledge base is the basic issue in building a domain knowledge base.

II. B.Time-series clustering

After completing the construction of the core model of the ICAI system, how to effectively utilize the massive time-series teaching data generated by the system becomes a key issue. The time-series clustering algorithm, through quantitative analysis of learning trajectories, can reveal the differentiated development patterns of student groups from a dynamic perspective and provide data support for group adaptation of teaching strategies.

The time-series k-means algorithm (KmL) is an algorithm for time-series data clustering, which is initially used in cohort studies in epidemiology. In such studies, the measurement and recording of observed variables often do not occur at a single point in time, but are recorded continuously over a period of time, with the final result treated as a trajectory, and these trajectories are classified and clustered using statistical methods to identify patients of the same type. The ability of the KmL algorithm to handle missing values and its very user-friendly graphical interface in determining the appropriate number of clusters has led to its factory-wide use in trajectory clustering problems, which has been gradually extended from the field of epidemiology to other areas of research.

The traditional $k-$ mean algorithm is a hill-climbing algorithm that assumes that the set of all data points is $\{x_1, x_2, ..., x_n\}$, where $x_i = (x_{i1}, x_{i2}, ..., x_{in})$ are vectors in the real number space $X = R^r$ and r denotes the number of attributes (i.e., the dimension of the data space). The $k-$ mean algorithm divides all the data points into k clusters, each of which is called a cluster, and considers the mean of all the points in the cluster as the center of the cluster, which is where the name $k-$ mean algorithm comes from. At the beginning of the algorithm's operation, k data points are randomly selected from the entire sample set as the initial cluster centroids, i.e., the seed cluster centers. Then, the distances between the rest of the data points in the sample set and each of the seed clustering centers are computed sequentially, and they are assigned to the centers that are closest to them, and the round of iteration ends when all the data points have been assigned. Next, the mean of each cluster is recalculated to obtain k new cluster centroids, and then the next round of data point allocation begins. This process of sample allocation and class center computation is repeated until the established termination conditions are met the time-series $k-$ mean algorithm is derived from an extension of the basic idea of the traditional $k-$ mean algorithm to the time-series data clustering problem. When considering time-series data clustering, the attribute dimensions of the data set are no longer r-dimensional, but 1-dimensional. The sequence of values of a single variable at different time nodes constitutes a trajectory, and the division of these trajectories into clusters is time-series clustering. Specifically, for the dataset S containing n samples, if we focus on the value of the variable Y for each sample at t time nodes, and denote the value of the variable at time node l for the i th sample as y_u, then is called the trajectory of sample i. The purpose of temporal clustering is to partition the entire sample set S into k homogeneous subsets by calculating the distances between the trajectories of each sample.KmL usually uses the Euclidean distance:

$$Dist^E(y_i, y_j) = \sqrt{\sum_{l=1}^{t}(y_{il} - y_{jl})^2} \tag{1}$$

Or Manhattan distance:

$$Dist^M(y_i, y_j) = \sum_{l=1}^{t}| y_{il} - y_{jl} | \tag{2}$$

to measure the distance between sample trajectories. The former is the most intuitively understood and widely used distance measure, while the latter is more robust in the presence of outliers. In addition to this, KmL is able to use measures specialized for time-series data, such as Frechet distance or Dynamic Time Warping (DTW), as well as user-defined methods.

II. C.Decision Tree Algorithm

Based on the group characteristics division obtained by time-series clustering, further classification algorithms are needed to realize the refinement of personalized teaching decisions. The decision tree algorithm, with its strong interpretability and intuitive rule generation, is able to transform the clustering analysis results into specific teaching intervention programs, forming a closed loop from data insight to practical application of the whole chain.

II. C. 1) Concept of decision trees

Decision tree algorithm is a typical classification method, which belongs to supervised learning, and it can realize some kind of prediction function. The data is analyzed and processed using an inductive algorithm to generate a decision tree, which can be used to analyze new data based on the decision tree.

The decision tree model has a tree-like structure and is a supervised learning based on if-then-else rules. Each internal node in the first method represents a test on an attribute, each branch represents the output of a test and can only output one result, and each leaf node represents a category. The decision tree rule results are obtained by training rather than manually formulated. The decision tree is schematically shown in Fig. 2.

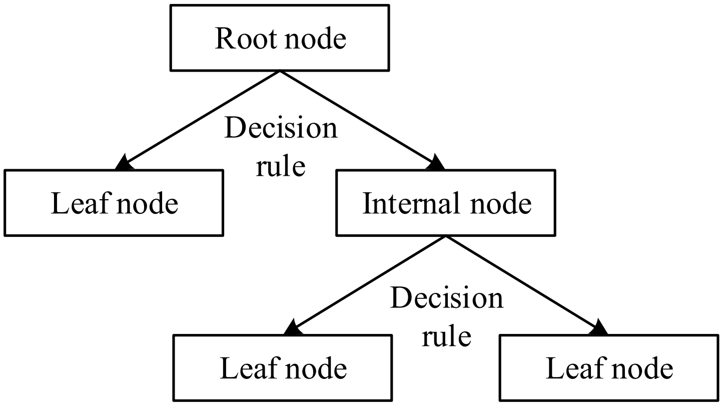

Figure 2: Schematic diagram of Decision Tree

Decision tree is a tree structure similar to flowchart, its advantage is that the construction of the decision tree does not require any domain knowledge and parameter settings, easy to understand, easy to operate, suitable for probing knowledge discovery, and can deal with high-dimensional data, the representation of the form is intuitive and clear, with a good accuracy; disadvantages are that the algorithm for each layer of the tree to scan through all the tuples, and thus is not suitable for dealing with large databases. The disadvantage is that the algorithm has to scan all the tuples at each level of the tree, thus it is not suitable for handling large databases. Typical algorithms for decision trees are ID3, C4.5, CART, etc., and the commonly used data mining tools are SSAS and SPSS.

II. C. 2) Feature selection for decision trees

Feature selection is an important part of the decision tree algorithm construction. The top-down recursive method is used to construct the decision tree, and the basic idea is to use the information entropy as the measure to construct a tree with the fastest decreasing entropy value, and the entropy value at the leaf node is zero.

The “entropy” mentioned above is, in physics, a measure of the degree of chaos. The more organized the system, the lower the entropy; conversely, the more chaotic the system, the higher the entropy. The “information entropy” is the most commonly used measure of the purity of a sample set. The higher the probability of an event occurring, the less information it contains, and the lower the value of “information entropy”.

In the calculation, assume that the proportion of the k th class of samples in the current sample set D is $p_k(k=1,2,...|y|)$, then

$$p_k = \frac{C^k}{D} \tag{3}$$

In this formula, D is the total number of samples and C^k is the number of samples in the k th class. Then the information entropy definition of D can be expressed as:

$$\begin{aligned} Ent(D) &= -\sum_{k=1}^{n}\frac{c^k}{D}\log\frac{c^k}{D} = -\sum_{k=1}^{n} p_k \log_2 p_k \\ &= -p_1\log_2 p_1 - p_2\log_2 p_2 - \cdots - p_n \log_2 p_n \end{aligned} \tag{4}$$

The metric is feature selection, which determines which features are used to make judgments. The more common criteria used in feature selection are information gain, information gain rate and Gini coefficient. In a dataset, since each sample may have more than one attribute, and different attributes play different roles in the classification result. Thus, feature selection can filter out the features that are more relevant to the classification result, that is, the features with stronger classification ability.

III. Feature Mining of English Teaching Behavior in Colleges and Universities Based on Time Series Clustering and Decision Trees

Chapter 2 constructs the theoretical framework and data analysis method of ICAI system, and proposes the technical path of time series clustering and decision tree algorithm. In order to further verify the practical efficacy of the model, Chapter 3 is based on the real data of a university's intelligent teaching platform, from learning behavior feature clustering, teaching session organization mode to performance association rule mining, systematically exploring the teaching optimization strategy driven by AI technology, and realizing the closed-loop validation from theoretical construction to empirical analysis.

III. A. Study design

III. A. 1) Research sample

In this study, we chose the smart teaching platform of a university in the spring semester of 2024 (February 18-July 4) for the behavioral data of the university-level students, with a total of 88,291 research samples, covering all the students of the university at the university level. 84,503 items were retained after removing the noisy data, and the data validity rate was 95.69%.

III. A. 2) Research process

This study is mainly divided into 2 research phases: (1) data collection and preprocessing phase, collecting students' learning behavior logs in the smart teaching platform in the spring of the 2023-2024 academic year, and converting the log data into students' learning behavior data for preprocessing; (2) the phase of students' smart classroom behavior clustering and analysis, adopting the KmL temporal k-mean clustering approach to obtain the different kinds of students' learning behavior performance, and analyze the behavioral characteristics of different student groups in combination with the discipline of English education in colleges and universities.

III. B. Cluster analysis of learning behavior characteristics

Ten dimensions, including downloading resources, watching microclasses, participating in cloud classroom lessons, logging on to the platform, participating in discussions, actively answering questions, participating in teaching activities, submitting assignments, independently cultivating excellence, and participating in online tests, were selected for cluster analysis of students' behavioral characteristics. The full score of each dimension performance evaluation is 10 points.

III. B. 1) Data pre-processing

Before clustering the data, the first step is to standardize the collected data, the purpose of which is to eliminate the differences between the data. From observing the collected student log data, it is easy to find that the disparity of student data between different dimensions is large in magnitude, which is not conducive to the clustering analysis of data mining. Therefore, this study adopts Z-score to convert data of different magnitudes to the same magnitude to improve the comparability of the data. At the same time, in order to improve the efficiency of data processing and reduce the complexity of data, this study used PCA principal component analysis to downscale the 10 dimensional data.

When the clustering cluster value is 4, the error squared and SSE within the cluster obtains the inflection point value, so the optimal clustering cluster value is 4. After obtaining the optimal clustering clusters the data are subjected to cluster analysis with the following rules:

$$P(x) = \frac{D(x)^2}{\sum_{x\in X} D(x)^2} \tag{5}$$

III. B. 2) Cluster analysis

The shortest distance between each sample and the currently existing cluster center is denoted by D(x), and the probability P(x) of each sample point being selected as the next cluster center is calculated first, and then the sample point corresponding to the maximum probability value (or probability distribution) is selected as the next cluster center, and the sample clustering visualization results are shown in Figure 3.

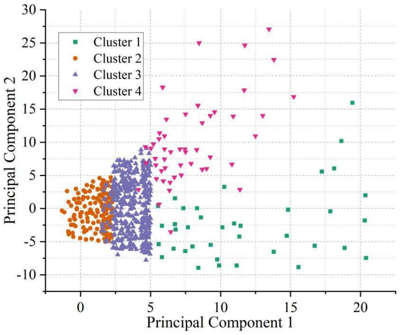

Figure 3: Clustering visualization results

Figure 3 presents the clustering results of all students, with different shades of patterns representing 1 type of student, and the clustering results of this study are in 4 categories. Based on the number of clustering points, the number of students contained in the 4 types of clusters can be judged. The more the clustering points are distributed, the darker the color, the more the number of students included, and vice versa, the less. Through the analysis, it can be seen that cluster 3 has the darkest color and the densest distribution of cluster points, which represents the highest number of students, followed by cluster 2, and finally cluster 1 and cluster 4.

In order to present more clearly the differences in the learning behaviors of the students of the four clustering groups in the Smart Teaching Platform, this study further calculated the mean and standard deviation of the behavioral performance of the different categories of students in each of the 10 data dimensions, and the descriptive statistics of the performance of the clustering groups in each of the 10 dimensions are shown in Table 1.

Table 1: Descriptive statistics of the performance of clustered groups in 10 dimensions

Dimension	Cluster 1 (N=625)		Cluster 2 (N=3011)		Cluster 3 (N=4276)		Cluster 4 (N=247)		Mean value
	M	SD	M	SD	M	SD	M	SD	
Resource download	7.73	1.13	6.74	1.37	4.33	1.47	1.19	3.45	5.38
Watch micro-lessons	9.23	1.2	8.04	1.9	6.41	2.66	1.87	3.51	7.09
Participate in the cloud class	7.07	1.02	5.63	1.83	3.58	2.54	1.02	2.34	4.53
Log in to the platform	8.43	1.31	6.94	1.31	5.44	2.9	2.05	3.15	6.12
Participate in the discussion	9.56	1.19	7.51	1.86	4.78	3.08	2.56	3.35	6.09
Answer questions actively	7.11	1.21	6.07	1.42	4.77	2.31	1.9	2.25	5.34
Participate in teaching activities	8.17	1.36	6.59	1.77	5.46	1.67	2.38	1.55	5.99
Submit the assignment	8.31	1.34	7.16	0.98	5.35	2.98	2.68	3.74	6.16
Self-training for excellence	7.75	0.99	5.94	1.02	3.05	2.87	1.09	2.64	4.42
Participate in the online test	8.29	1.37	7.67	1.61	5.04	1.58	2.49	2.88	6.18

Through the analysis of students' active situation in the smart teaching platform, it can be seen that students' use of resources can reflect their learning input in the smart teaching platform, and the higher the degree of resource utilization, the higher the degree of learning input; the teaching activity module and homework module can reflect the students' initiative in learning; and the test module is a necessary part for students to test their learning effect after completing the stage of learning. Through the above analysis and visualization results, the final four types of learner behavioral characteristics are as follows:

Cluster 1 included 625 students, and this group had the most outstanding performance in all dimensions, especially in the dimensions of "participating in discussions" (9.56), "watching micro-lessons" (M=9.23), and "login platform" M=8.43, with standard deviations of less than 1.5, indicating that their learning behaviors were highly positive and internally consistent. Such students demonstrate a strong level of initiative and engagement, and may be the quintessential "high-performing learner" who is able to make the most of the platform's resources and actively engage in interactions.

Cluster 2 consisted of 3011 students, with medium group performance, with higher scores in task-driven behaviors such as "submitting assignments" (M=7.16) and "participating in online tests" (M=7.67), but weak active behaviors such as "self-cultivation" (M=5.94) and "participating in online classes" (M=5.63). The standard deviation is generally between 1.4 and 1.9, indicating that the behavior pattern is relatively stable, and they belong to the "task-oriented" learner, who relies on external task-driven, but has limited ability to expand independently.

Cluster 3 included 4276 students, as the largest group, their scores in each dimension were generally lower than those of the first two categories, especially in the dimensions of "resource download" (M=4.33) and "self-cultivation" (M=3.05), and the standard deviation was high, such as the standard deviation of "participation in discussion" was 3.08, indicating significant intra-group differences. Such students may be "passive participants", who only complete basic tasks, lack the motivation to actively explore, and need to stimulate their interest in learning.

Cluster 4 included 247 students, with the smallest group size and the most negative behavioral characteristics, and the scores of all dimensions were significantly lower than those of other groups, such as the average value of "resource download" was only 1.19, "watching micro-lessons" M=1.87, and the standard deviation of some dimensions was as high as 3.51, indicating that their behavior fluctuated greatly. These students may face a lack of motivation or barriers to technology use, and are in urgent need of individualized interventions, such as enhancing teacher-student interaction and providing targeted resource recommendations to increase engagement.

The behavioral differences of the four groups reflect the diverse learning needs and provide data support for the refined design of teaching strategies: advanced resources can be provided for high-performing learners, task-oriented learners need to strengthen goal motivation, passive participants need to increase interactive content, and passive groups need to focus on helping to improve the learning experience.

III. C. Cluster analysis of organizational characteristics of teaching sessions

On the basis of identifying the four categories of learning behavior patterns, further clustering and association rule analysis of the teaching sessions of English courses in colleges and universities are carried out to reveal the dynamic matching relationship between instructional design and students' behavioral characteristics, and to clarify the key aspects of course optimization.

After comparing the two characteristics of the data sequence and the proportion of hours of English courses in colleges and universities in the clustering results, when the number of clusters is also set to 4, i.e., the sequence of teaching sessions about English education is clustered into four categories, the characteristic differences between the resulting course categories are the most significant. After the data of English courses related to colleges and universities are processed accordingly, the support degree of each teaching link in various courses related to English can be obtained by analyzing the association rules, and the frequency of jumping between teaching links of English courses in colleges and universities and the adjusted residual value can be obtained by performing the lagged sequence analysis. The study analyzes the organizational characteristics of teaching sessions and the implied teaching ideas of English courses in colleges and universities by observing the organizational characteristics of teaching sessions drawn on the basis of the adjusted residual values.

The frequency of each teaching session in the total sample is as follows: introduction 3875 times, vocabulary learning 3413 times, text learning 5836 times, communicative activities 4457 times, text practice 2134 times, and knowledge summarization 873 times. Text learning and communicative activities were carried out much more frequently than other teaching sessions, while text practice and knowledge summarization had a significant weakness. After KmL clustering, the courses were divided into four unequal categories named Cluster 1, Cluster 2, Cluster 3, and Cluster 4. The study mined the frequent item sets of the courses to obtain the support of teaching sessions of English courses in colleges and universities after clustering as shown in Table 2.

Table 2: The support degree of teaching links in college English courses

	Import	Vocabulary learning	Text study	Communication activities	Text Practice	Knowledge summary
Cluster 1 (N=6128)	0.6372	0.7366	0.6226	0.4252	0.3159	0.2424
Cluster 2 (N=4225)	0.5021	0.4333	0.4604	0.5587	0.4724	0.0427
Cluster 3 (N=3646)	0.2724	0.6592	0.4363	0.3188	0.3073	0.1313
Cluster 4 (N=1355)	0.7588	0.9181	0.8835	0.7617	0.6384	0.3544

According to the results of the cluster analysis in Table 2, the four clustered groups of English courses in colleges and universities show significant differences in the support of teaching sessions. Cluster 1 courses, with 6128 sections, are characterized by vocabulary learning, with a support degree of 0.7366, and introductory sessions, with a support degree of 0.6372 as the core features, followed by text learning with a support degree of 0.6226 and communicative activities with a support degree of 0.4252, but the support degree of text practice only 0.3159 for text practice and 0.2424 for knowledge summarization, indicating that this kind of course focuses on knowledge input and basic activities, but lacks consolidation sessions.

Cluster 2 courses are 4225 sections, highlighting the support of 0.5587 for communicative activities, while vocabulary learning 0.4333, text learning 0.4604 and other basic aspects are weak, and the support of 0.05 for knowledge summarization is less than 0.05, which shows that the teaching mode is mainly interactive practice, but the systematic knowledge grooming is insufficient. Cluster 3 courses are 3646 sections, and the support for vocabulary learning is higher at 0.3159 and 0.2424 for knowledge summarization of this cluster. support is high, 0.6592, but other links are at a low level, such as communicative activities 0.3188, and knowledge summarization is not included, reflecting that its teaching is biased towards the transmission of isolated knowledge points, and the overall coherence is poor. Cluster 4, with N=1355, has a balanced performance and the highest level of support among all the links, with 0.9181 for vocabulary learning, 0.8835 for text learning, 0.7617 for communicative activities, and 0.9181 for knowledge summarization support, and 0.7617 for vocabulary learning, and 0.9181 for vocabulary learning. 0.7617, and 0.3544 for knowledge summarization support are also significantly higher than the other clusters, representing a comprehensive teaching model that takes into account input, practice and reflection.

Overall, English courses in colleges and universities generally emphasize vocabulary learning, text learning and communicative activities, but the weakness of text practice and knowledge summarization exists in all clusters, and only Cluster 4 performs slightly better. Different clusters reflect the differentiation tendency of instructional design, such as Cluster 1 favors basic input, Cluster 2 focuses on interactive practice, Cluster 3 relies on a single link, and Cluster 4 realizes multidimensional integration.

III. D. Implementation and Effectiveness Analysis of Achievement Data Mining

Combining teaching session characteristics and learning behavior patterns, the study mines performance association rules through decision tree algorithms to quantify the impact of variables such as classroom mastery and listening time on academic performance, ultimately forming a complete data-driven pathway from behavioral analysis to instructional intervention.

III. D. 1) Processes and results of excavation runs

Data mining is an important function of this system, in order to dig out the influence of various factors such as students' study habits on their performance, this group issued a total of 8406 questionnaires to 2024 freshmen students in a university, the questionnaire content mainly includes classroom mastery, listening time, reading frequency, vocabulary, etc., and recovered valid questionnaires of 8159 to form a student questionnaire of 8159 items.

In order to facilitate the understanding of the mining algorithm studied in this project, the process of substituting the data into the system for the operation is explained as follows:

Firstly, for computational convenience, the 8159 valid data are numbered as l, 2, …, 8159. Set u={l, 2, …, 8159} and let the attributes R1=gender, R2=classroom mastery, R3=listening time, R4=reading frequency, R5=vocabulary, whether excellent and whether average are the decision attributes d, respectively.

The steps of data mining by the system are as follows:

(1) Firstly, the decision coordination degree of each attribute is calculated

From the calculation, the difference between the coordination degree values between some of the attributes is relatively small, and the information gain needs to be further used to detect the attributes that can be used as split nodes.

(2) The node gain value (D) = 0.2632 is calculated for each attribute

Classroom mastery: InfoR2 (D) = 0.2632, GainR2 = 0.6872

Listening time: InfoR3 (D)=0.2632, GainR3=0.6033

Reading Frequency: InfoR4 (D)=0.2268, GainR4=0.7406

Vocabulary: InfoR5 (D)=0.0712, GainR5=0.8385

(3) Through the data mining algorithm, the decision tree of each part is combined, that is, the following complete decision tree is obtained, and the complete decision tree of whether the grade is excellent or not is shown in Fig. 4, which is drawn based on the results of the system calculations and is not directly generated by the system, indicating

the main factors of whether the grades can reach excellence and other key influencing factors, that is to say, the degree of mastery of the classroom has a direct impact on the grades Whether or not the grade can reach excellent.

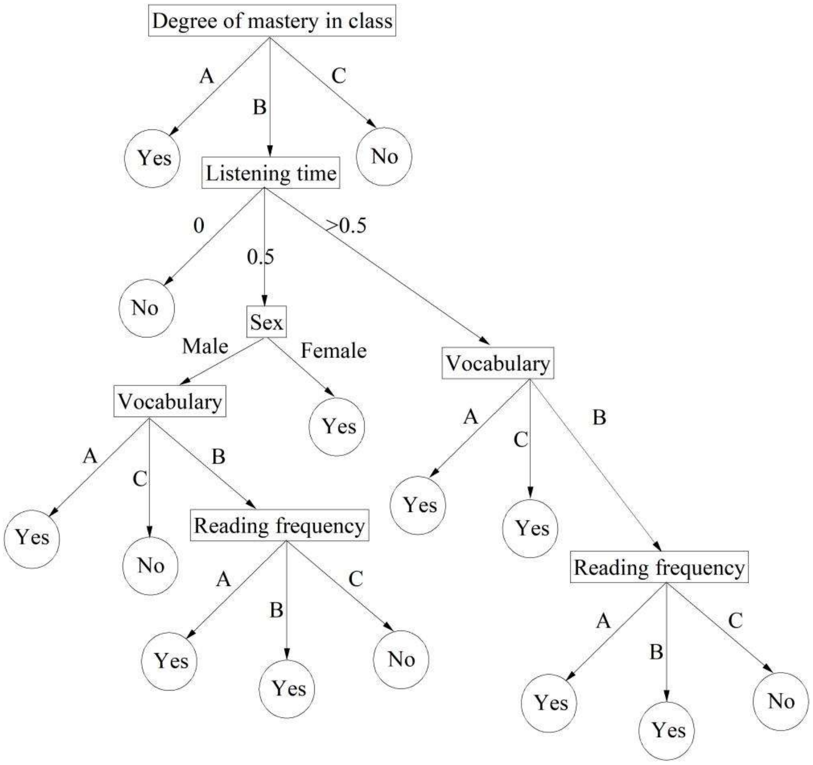

Figure 4: A complete decision tree of whether the grades are excellent or not

(4) Further calculations can be derived from the decision tree of whether the student's grades are general or not, the leaf node is “Yes” means that the grades are general, and the leaf node is “No” means that the grades are not general, and the complete decision tree structure of whether the grades are general or not is shown in Fig. 5. This figure is drawn based on the results of the data table of whether the grades are average or not in the system, not directly generated by the system, indicating the main factors of whether the grades are average or not and other key influencing factors.

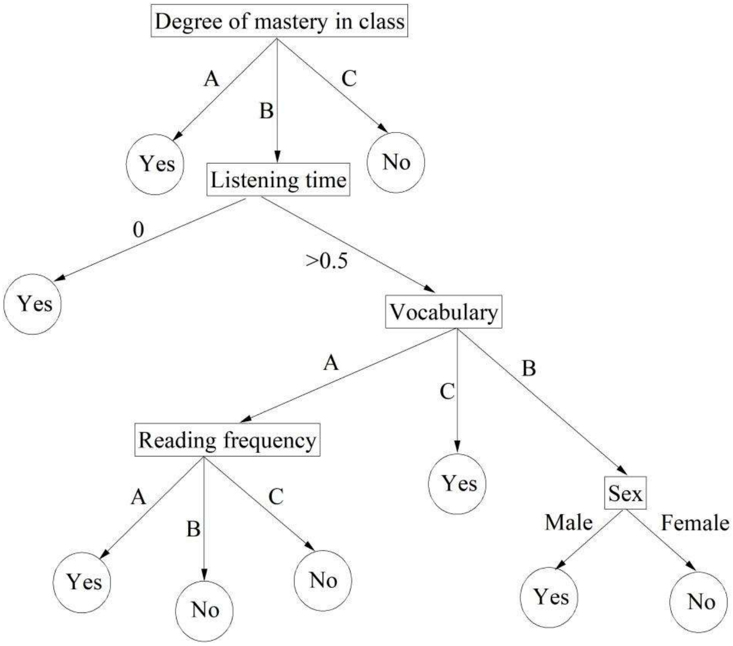

Figure 5: Complete decision tree of general results

III. D. 2) Analysis of outstanding achievements

According to the score mining analysis of the decision tree, the 14 correlation rules for excellent college English scores are as follows.

Rule 1: If class mastery = "very good", then excellent grade = "yes";

Rule 2: If class mastery = "fair" and listening time = ">0.5" and vocabulary = "rich", then excellent grade = "yes";

Rule 3: If class mastery = "fair" and listening time = ">0.5" and vocabulary = "basically sufficient" and reading frequency = "often", then excellent grades = "yes";

Rule 4: If class mastery = "fair" and listening time = ">0.5" and vocabulary = "mostly sufficient" and reading frequency = "occasionally", then excellent grades = "yes";

Rule 5: If class mastery = "fair" and listening time = ">0.5" and vocabulary = "basically sufficient" and reading frequency = "never", then excellent grades = "no";

Rule 6: If class mastery = "fair" and listening time = ">0.5" and vocabulary = "poor", then excellent grades = "no";

Rule 7: If class mastery = "fair" and listening time = "0.5" and gender = "male" and vocabulary = "rich", then excellent grade = "yes";

Rule 8: If class mastery = "fair" and listening time = "0.5" and gender = "male" and vocabulary = "basically sufficient" and reading frequency = "often", then excellent grades = "yes";

Rule 9: If class mastery = "fair" and listening time = "0.5" and gender = "male" and vocabulary = "mostly sufficient" and reading frequency = "occasionally", then excellent grades = "yes";

Rule 10: If class mastery = "fair" and listening time = "0.5" and gender = "female", then excellent grade = "yes".

Table 3 lists the 10 association rules and their support and confidence levels.

Table 3: Regarding the 14 association rules and their support and confidence levels

Rule	Support	Confidence
Rule 1	0.00879	0.6123
Rule 2	0.008822	0.3887
Rule 3	0.007963	0.3062
Rule 4	0.008486	0.4153
Rule 5	0.007903	0.4815
Rule 6	0.004874	0.5348
Rule 7	0.006736	0.4629
Rule 8	0.006109	0.4682
Rule 9	0.007942	0.5702
Rule 10	0.004405	0.2959

Table 3 reveals the key influencing factors of students' excellent performance in college English education. The most supported rules were 0.00879 for Rule 1 and 0.008486 for Rule 4, indicating that "class mastery" = "very good" and "class mastery" = "fair" "listening time>0.5" "Vocabulary" = "Mostly sufficient" "Reading frequency" = "occasionally" is a more common combination of conditions.

In terms of confidence, 0.5702 for Rule 9 and 0.4815 for Rule 5 stood out, indicating that "classroom mastery" = "fair" "listening time" = 0.5 "gender" = "male" "vocabulary" = "mostly sufficient" "reading frequency" = "occasionally" and "classroom mastery" = "fair" "listening time" >0.5 "vocabulary" = "mostly sufficient" "reading frequency" = "never" have high confidence in the prediction of grades. However, some rules, such as Rule 10, have a low support and confidence level of 0.004405 and a confidence level of 0.2959, indicating that the gender "female" has a weaker explanatory power for excellent performance under this rule. Overall, classroom mastery is the core driver, followed by listening time and vocabulary, while reading frequency and gender only play a supporting role under certain conditions.

IV. Conclusion

Based on the smart teaching platform of a university, this study systematically analyzes the student behavioral characteristics, teaching session organization mode and academic performance association rules in English teaching in colleges and universities through time series clustering and decision tree algorithms.

Using the KmL time-series clustering algorithm, students were divided into four differentiated groups: efficient learners (N=625) had outstanding performance in the dimensions of resource downloading with a mean value of 7.73 and participation in discussion M=9.56; task-oriented (N=3011) relied on task-driven, submitted homework dimensions with a mean value of M=7.16, and scored high but were weak in independent cultivation of excellence with a mean value of M=5.94; passive participants (N=4276) were weak in various dimensions such as scoring M=3.05 with significant standard deviation in autonomous meritocracy dimension, such as participation in discussion SD=3.08, which needs to be enhanced with interactive design; the number of negative groups was 247, whose behavioral characteristics were negative, and resource downloading M=1.19, which urgently needs targeted interventions.

The cluster analysis of the organizational characteristics of teaching sessions showed that only Cluster 4 (N=1,355) had balanced support for teaching sessions, with vocabulary learning of 0.9181 and knowledge summarization of 0.3544, whereas the other clusters generally had insufficient consolidation sessions such as knowledge summarization with support <0.25, and needed to optimize the design of teaching in order to enhance the systematicity.

The decision tree algorithm shows that classroom mastery is the core driver, with Rule 1 confidence level 61.23%, listening time Rule 4 confidence level 41.53% and vocabulary Rule 2 support level 0.88% as the key auxiliary variables, and gender and reading frequency only play a limited role in specific conditions, such as Rule 10 confidence level 29.59%.

References

[1] Li, K. (2024). Research on optimization of English teaching in universities under the guidance of applied talent training. Adult Higher Educ, 6.

[2] Huang, L. (2022). An empirical study of integrating information technology in english teaching in artificial intelligence era. Scientific Programming, 2022(1), 6775097.

[3] Wang, Q. (2021). The Research on Methods of Business English Talent Training Based on the Needs of Enterprises. International Journal of Frontiers in Sociology, 3(13).

[4] Jing-hua, Z. H. A. N. G. (2020). A Study of “Foreign Language+” Innovative Talents Training Model Driven by Social Needs. Journal of Literature and Art Studies, 10(4), 321-326.

[5] Wang, C., Chen, X., Yu, T., Liu, Y., & Jing, Y. (2024). Education reform and change driven by digital technology: a bibliometric study from a global perspective. Humanities and Social Sciences Communications, 11(1), 1-17.

[6] Jiang, R., & Zhang, C. (2024). Strategies for Reforming College English Teaching in the Context of Educational Digital Transformation. Curriculum Learning and Exploration, 2(1).

[7] Lu, H. (2024). Professional Development of English Teachers in the Context of Digital Transformation. Journal of International Education and Science Studies Vol, 1(5).

[8] Liang, X., Haiping, L., Liu, J., & Lin, L. (2021). Reform of English interactive teaching mode based on cloud computing artificial intelligence–a practice analysis. Journal of intelligent & fuzzy systems, 40(2), 3617-3629.

[9] Sun, Z., Anbarasan, M., & Praveen Kumar, D. J. C. I. (2021). Design of online intelligent English teaching platform based on artificial intelligence techniques. Computational Intelligence, 37(3), 1166-1180.

[10] Zhang, X., & Chen, L. (2021). College English smart classroom teaching model based on artificial intelligence technology in mobile information systems. Mobile information systems, 2021(1), 5644604.

[11] Jie, Z., & Sunze, Y. (2023). Investigating pedagogical challenges of mobile technology to English teaching. Interactive Learning Environments, 31(5), 2767-2779.

[12] Rohmiyati, Y. (2025). Enhancing English Language Learning Through Artificial Intelligence: Opportunities, Challenges and the Future. DIAJAR: Jurnal Pendidikan dan Pembelajaran, 4(1), 8-16.

[13] Fu, M., Guan, X., Wang, Y., & Chen, Q. (2025). Application of speech recognition algorithm based on interactive artificial intelligence system in English video teaching system. Entertainment Computing, 52, 100859.