





Optimization of Computer-aided English Classroom Teaching System Based on Data Mining

Jiling Shang ¹  and Chaohui Liang ² 

¹School of International Education & Europe-Asia Jiaotong, Zhengzhou Railway Vocational and Technical College, Zhengzhou 450000, China, redballons99@163.com

²School of International Education & Europe-Asia Jiaotong, Zhengzhou Railway Vocational and Technical College, Zhengzhou 450000, China, yjfgef@126.com

Corresponding author: Jiling Shang, redballons99@163.com

Abstract. The conventional computer-assisted classroom teaching system has a small number of concurrent users and long information retrieval time. For this reason, we first designed a computer-assisted classroom teaching system based on data mining. Secondly, use the improved decision tree TG-C4.5 algorithms to conduct data mining on the learning behavior data to achieve the goal of predictive classification of performance. Finally, solve the problem that educators who lack data mining knowledge understand the mining results, integrate the algorithm into the education assistance system, realize the visualization of the predictive analysis results, and give personalized prompts to students' learning behaviors, thereby improving the teaching assistance the system makes it personalized.

Keywords: optimization; computer aided systems; teaching system; data mining.

DOI: <https://doi.org/10.14733/cadaps.2021.S4.95-105>

1 INTRODUCTION

In the education system, the old-style pure English classroom teaching model has been strongly impacted and cannot meet the new requirements of the current society for the training of advanced talents. People are paying more and more attention to the innovation of English teaching models. In this environment, the modern English teaching model gradually replaces the old English teaching model [1-2]. The combination of computer network and educational English teaching is a representative modern English teaching mode. This English teaching model uses computers as teaching aids and the Internet as one of the main ways of knowledge dissemination. It has the advantages of sharing English teaching resources, enhancing teacher-student interaction, unrestricted learning space, and teaching students in accordance with their aptitude [3-4].

Paternotte et al. [5] explored learners' communication preferences and the value of preference-based groups, and at the same time studied the grouping of learners according to their

preferences to enhance their learning ability or course completion, which greatly promoted learning new directions for research. Van et al. [6] selected the user-related information of "school online" as the data source, and concluded that the online learning behavior of users has obvious non-uniformity. Li et al. [7] conducted a comprehensive study on the domestic teaching system with the help of the Edutools evaluation system, found the differences between different teaching systems, and proposed optimization strategies for the differences, providing good decisions for the construction of the English classroom teaching system support and advice. Asri et al. [8] proposed the C4.5 in response to the low accuracy problem in the current C4.5 algorithms to construct a predictive model. This algorithm helps to achieve the construction of a predictive model for teaching system performance. Chen et al. [9] selected the 8 courses with the largest number of learners in the teaching system as the research objects, trained on the 11 behavioral data obtained, and constructed the performance prediction model by using multiple linear regression and BP methods to obtain 90% of the prediction accuracy rate has a great early warning effect for teachers to interfere with students' learning. The existing computer-assisted English teaching models generally have problems such as poor information utilization, low intelligence, and low personalization [10]. As a result of database technology development and research, data mining technology can mine potentially useful information in a large amount of information storage, and its essence is a data value-added process. Applying data mining technology to the computer-assisted English classroom teaching model, designing a computer-assisted classroom teaching system based on data mining, improving information utilization, providing teaching resource downloads, network interaction, score management and other services, it is helpful to the "teaching" and "learning" have greater positive significance. Secondly, use the improved decision tree TG-C4.5 algorithms to conduct data mining on the learning behavior data to achieve the goal of predictive classification of performance. Finally, solve the problem of educators who lack data mining knowledge about the understanding of the mining results, integrate the algorithm into the educational auxiliary system, realize the visualization of the predictive analysis results, and provide the basis for teachers' decision-making on teaching early warning and other information.

2 DATA MINING TECHNOLOGY

2.1 Definition

The development of the Internet has caused people to leave a large amount of user behavior information on the Internet, coveting the potential value of these hidden information, domestic Internet companies and other related institutions have begun to conduct data mining research, and they can make correct decisions based on the results obtained by data mining, effectively meet customer needs.

From the definition of data mining, we can conclude that the characteristics of data mining are as follows:

(1) Based on massive data. But it does not mean that a small amount of data cannot be applied to data mining. In practical applications, small amounts of data can also be studied using related algorithms in data mining.

The prominence of massive data in the definition is mainly because: on the one hand, massive data can show universal laws and avoid accidental phenomena. On the other hand, the corresponding knowledge can be obtained by analyzing a small amount of data only manually.

(2) Deep level. This shows that the information obtained through data mining is not simply obtained by observation and reasoning, but by means of statistical analysis, probability theory and other related knowledge, relying on computers to dig deep from the data.

(3) Different fields of value use data mining to obtain information. This information can bring direct or indirect profits to the company or decision-makers. However, some people have a certain misunderstanding of data mining and think that it is just superficially brilliant. There are many

reasons for this misunderstanding. It may be the researchers' insufficient experience, unreliable data sources, and unclear goals. But on the whole, all walks of life have achieved their own success by using data mining, proving the value and effectiveness of data mining.

Data mining is not simply using a computer to process data, but requires users to further process the data based on the information fed back by the computer. The entire processing process involves six stages: business understanding, data understanding, data collection and processing, model building, model evaluation and program implementation. These 6 stages can be adjusted in sequence according to actual needs, as shown in Figure 1.

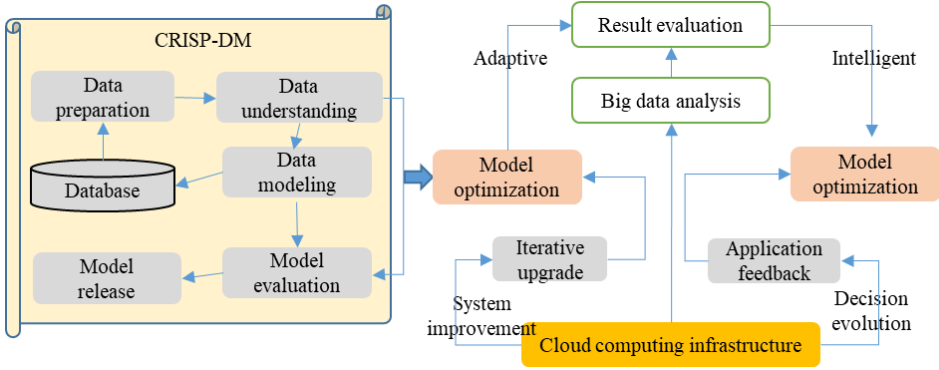


Figure 1: Data mining process.

2.2 Decision Tree Classification Technology

When encountering classification problems in data mining, the decision tree algorithm is often the best choice. Because the classifier constructed by the decision tree algorithm is intuitive, the calculation speed is relatively fast, and the classification accuracy is relatively high, so the decision tree application is more appropriate in the system.

In the process of building a decision tree, it is necessary to make judgments based on the size of the attribute measurement value. There are three commonly used decision tree attribute selection metrics, as shown in Table 1.

Attribute measurement	Algorithm	Disadvantage
Information gain	ID3	The choice of attributes tends to be multi-valued
Information gain rate	C4.5	The gain rate will change with recursion
GINI Index	CART	Poor simulation stability

Table 1: Attribute metrics commonly used in decision trees.

Now we will introduce related attribute metrics, firstly introduce information entropy, secondly introduce information gain and information gain rate formed by information entropy evolution, and finally introduce GINI index.

(1) Information entropy

Information entropy is a measure of the uncertainty of the information source. Suppose S is a sample set, and the set contains s data. There are p different types, corresponding to the set sk, the value range of k is (1,2,..,p). Let mk be named as the number of categories in the sk set. When classifying, the information entropy needs to be calculated, as shown in formula (1).

$$S = -\sum_{i=1}^q a_i \log_2 a_i \quad (1)$$

Where a_i is the probability of belonging to the category s_k . Usually, the choice of logarithm is based on the knowledge of information theory as the logarithm with base 2.

(2) Information gain

Information gain is the change in the amount of information of a given data set before and after classification. Assuming that the value set of attribute X is $\{x_1, x_2, \dots, x_q\}$, according to the different values of X , S can be divided into $\{s_1, s_2, \dots, s_p\}$, where $G(X)$ represents a data set in which the value of attribute X in set S takes the value of a_i . When the X attribute is used as a division, let m_{pq} be the number of s_p belongs to the category of x_q , and its information entropy is shown in formula (2).

$$G(X) = \sum_{i=1}^q m_{pi} S(m_{pi}) \quad (2)$$

By dividing X as an attribute, the information gain obtained is shown in formula (3).

$$G(S, X) = X_1 \dots X_q G(X) \quad (3)$$

(3) Information gain rate

The information gain rate is calculated by dividing the two parts of the information gain and the split information, as shown in formula (4).

$$GR(S, X) = \frac{G(S, X)}{F(X)} \quad (4)$$

The calculation of split information is shown in formula (5).

$$F(X) = -\sum_{i=1}^q \frac{m_k}{m} \log_2 \frac{m_k}{m} \quad (5)$$

(4) GINI index

The GINI index is a measure of the impurity of data division, and its calculation is shown in formula (6).

$$GI = 1 - \sum_{i=1}^q (a_i \log_2 a_i)^2 \quad (6)$$

2.3 The Overall Framework Design of the Teaching System

The computer-assisted teaching system is applied based on the campus local area network, and its information transmission path is mainly through the campus local area network. The computer-assisted teaching system includes hardware environment, teaching resources, teaching activities, and auxiliary teaching system maintenance and management. Through the computer-assisted teaching system design, the user experience is enhanced, and the functions of the computer-assisted teaching system are realized. The overall framework of the system is shown in Figure 2.

The hardware structure in the computer-assisted teaching system combines electronic information technology and data processing, which is a sign of technological development. Therefore, the industry believes that the hardware construction of a computer-aided classroom teaching system based on data mining is not only a systematic project, but also a complicated process. The main task of the teaching interface layer is to provide a human-computer interaction interface for each user. The teaching application layer includes four main modules, namely the information collection module, the information preprocessing module, the personalized analysis module and the information scheduling module. Through these modules to achieve system

personalized functions. The main function of the system resource layer is to store teaching rules, teaching resources, and user information.

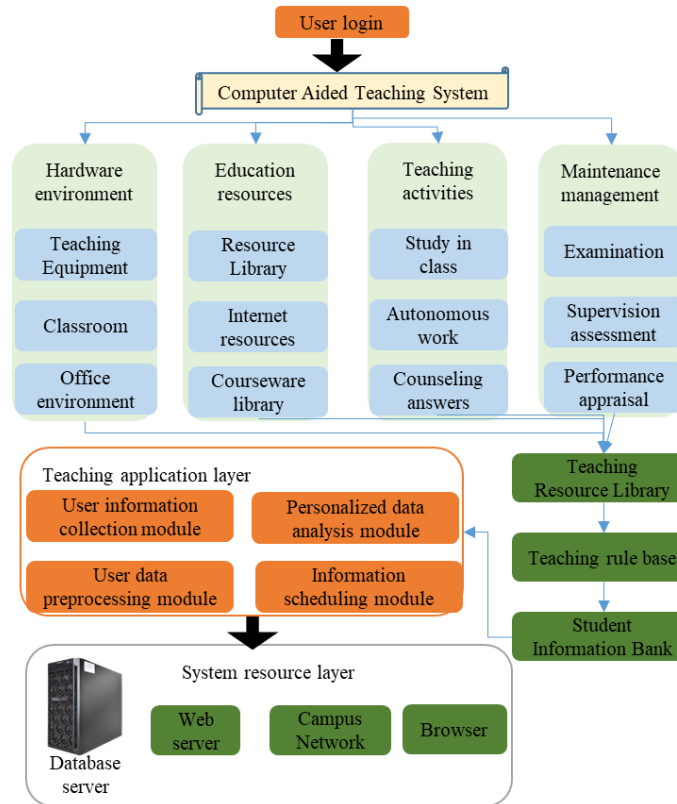


Figure 2: Schematic diagram of the overall system structure.

The hardware structure block diagram contains three parts: data storage and management, OLAP server, front-end tools and applications. The key part of the hardware is the data memory, which is equivalent to the brain of the system. In this design, the data storage part includes SARM memory, CD-ROM and hard disk three parts.

OLAP is also known as online analytical processing server. Its main function is to process, reorganize or integrate the required data, so that analysts can quickly observe information from all aspects from multiple angles and levels, so as to achieve the purpose of in-depth understanding of data and provide multi-dimensional view and analysis of data analysis to realize the sharing of multi-dimensional information. Figure 3 shows the workflow of the OLAP-based information collection module. The information collection module realizes the collection and storage of student information data, supports different source data formats such as OLTP data, OLAP data, and log data, and includes multiple data synchronization methods such as database real-time synchronization and Socket message synchronization. The module adopts a template design method to realize the template and metadata configuration of new data, and complete the unified collection and specification of different information data.

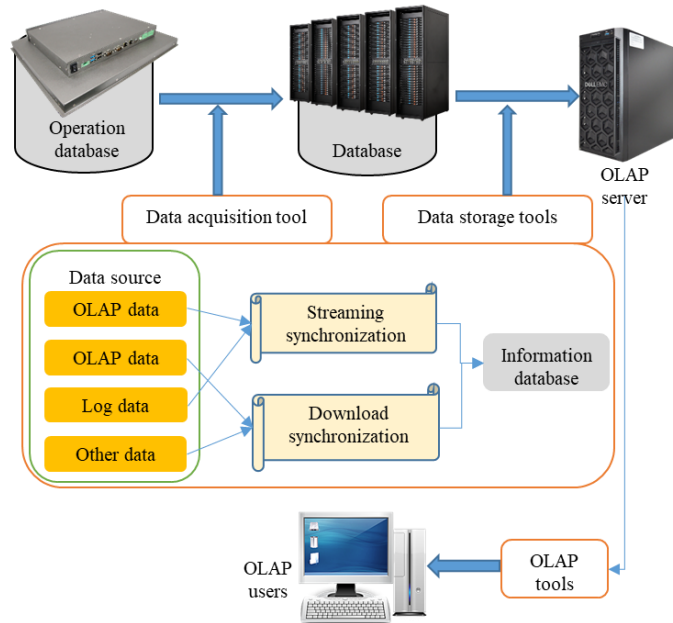


Figure 3: Work flow of OLAP-based information collection module.

2.4 Algorithm Improvement

C4.5 is currently the most widely used decision tree algorithm, but this algorithm also has areas that need to be improved. The specific drawbacks are as follows:

(1) In the process of generating the decision tree, multiple scanning and sorting operations are required for the data set, which easily causes the problem of low classification efficiency. Because the algorithm itself has a large number of logarithmic operations, more logarithmic methods are used. When the actual computer runs the C4.5, it greatly increases the time running cost of the algorithm.

(2) The split attribute calculation used in the process of generating the decision tree by the C4.5 algorithm ignores the influence between different attributes, and only considers the calculation between each attribute and class attribute.

Under the condition of analyzing the shortcomings of the traditional C4.5, regarding the high time complexity of the C4.5, we use Taylor series or McLaughlin's formula to simplify the logarithmic operation into the basic four arithmetic operations, and realize time cost reduction. Regarding the shortcomings of split attribute calculation, the C4.5 optimization algorithm using GINI index is proposed to avoid the influence of correlation between attributes and improve the accuracy of split attributes.

(1) Join Taylor Series

The Taylor series expresses a function as a series of infinite addition, where the coefficient of each added term can be calculated based on the derivative of the function at a certain point.

The function $Z(a)$ has an n -order derivative in a closed interval in the field of c (c is a real number), and an $n+1$ -order derivative in its open interval, then the Taylor series expansion of $Z(a)$, as in the formula (7) Shown.

$$Z(a) = \sum Z^{(n)}(c)(a-c)^n / n! \quad (7)$$

Among them, $n!$ represents the factorial of n , and $Z(n)(c)$ represents the n -th derivative at point c .

$$TC(a) = \sum_{i=1}^q m_{pi} S(m_{pi}) - \sum_{i=1}^q m_{pi} S(m_{pi}) \sum Z^{(n)}(c)(a-c)^n / n! \quad (8)$$

It can be obtained by formula (8) that only using the T-C4.5 algorithm improved by Taylor series, the information gain rate calculation formula avoids the call of the logarithmic operation function, and becomes a simple basic four mathematical operations, saving costs time and improves the accuracy of information gain rate to a certain extent.

(2) Join the GINI index

The GINI index has been used in decision tree classification algorithms, and it can be used to calculate the "purity" of samples in a data set. In practical applications, the size of "purity" measures whether most of the samples in the data set belong to a certain category, in other words: in a data set S , there are most sample data in one of the categories, then its "purity" is great. Conversely, if there are very few samples, the "purity" is small.

Assuming that the data set S is divided into p category subsets, the calculation of the GINI index is shown in formula (9).

$$GI = 1 - \sum_{i=1}^q \left(\frac{TC(a)}{TC} \right)^2 \quad (9)$$

In order to improve the accuracy of attribute selection, we add the mean value of the sum of the GINI index to formula (8), which is then used as the new split information after eliminating the influence between attributes. The improved calculation formula is shown in formula (10).

$$GR(X) = \lambda \bullet TC(a) + (1-\lambda)GI \quad (10)$$

Where λ is the coefficient, which is obtained according to formula (10): if the value of c remains unchanged, if the degree of correlation between attribute X and other conditional attributes is small, the corresponding redundancy is small, and at the same time the larger the mean value of GINI between the attribute X and other conditional attributes, the larger the value of the information gain rate of attribute X . This eliminates the influence of the correlation between conditional attributes on the selection of split attributes, and improves the classification accuracy.

3 ANALYSIS OF RESULTS

3.1 Comparison of Teaching Effects

Taking 124 students from a university as the experimental subjects, they are divided into a control group without using the system of this paper and an experimental group using the system of this paper. Each group has 62 students, and a teacher teaches them.

Fractional segment	0 < score < 6	60 < score < 7	70 < score < 8	80 < score < 9	90 < score < 100
Test group	3	7	14	26	8
Control group	12	23	14	6	2

Table 2: Comparison of the number of people in each score segment between the two groups.

Comparing the final average grades of the control group and the experimental group, the results are shown in Table 2. Describe the data in Table 2 in the form of a bar graph, and the result is shown in Figure 4. It can be obtained from Table 2 and Figure 4 that in the average final exam

score, the number of students in the experimental group failing score segment and the score above 90 are 2 and 16, respectively.

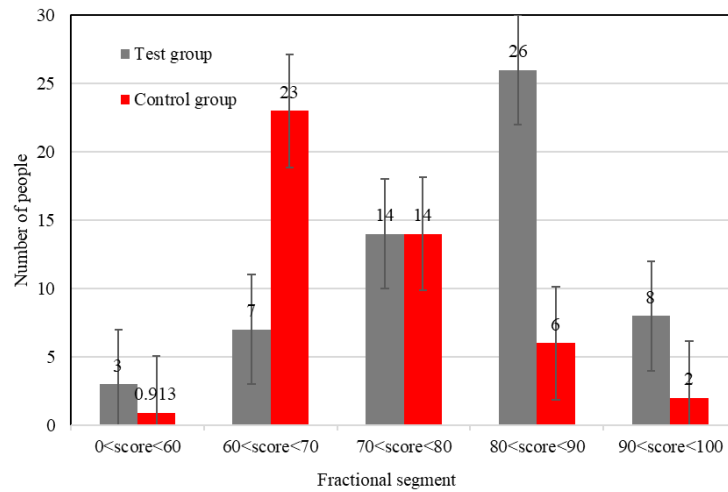


Figure 4: Comparison of the number of people in each score segment between the two groups.

The number of students in the 80~89 score segment is the largest. It accounts for 40.32% of the total number of people. In the control group, there are 13 and 3 people in the failing grade and above 90 grades respectively. The number of students in the 60-69 grade is the largest, accounting for 38.71% of the total. In the paper, the performance of the students who use the text system is significantly higher than that of the non-text system, indicating that the text system can effectively improve the academic performance.

3.2 Experimental Results of Concurrent Users

Using the Simu-works platform, in the above-mentioned experimental environment, the conventional computer-assisted classroom teaching system and the design system of this article are used to test the number of concurrent users, and the experimental results are drawn into a graph. The results are shown in Figure 5(a).

It can be seen from Figure 5(a) that the average number of concurrent users using this system is 131, and the average number of concurrent users in the conventional system is 73. It shows that the system performance in this article is better.

Use the Simu-works platform to test the information retrieval time of the conventional computer-assisted classroom teaching system and the design system of this article, and plot the experimental results into a graph. The results are shown in Figure 5(b). It can be seen from Figure 5(b) that the average information retrieval time of the system designed in this paper is 0.75 s, and the average information retrieval time of the conventional system is 1.47 s. It shows that the retrieval time of this article is shorter.

In summary, in the simulation experiment of the computer-aided classroom teaching system based on data mining, the computer-aided classroom teaching system designed in this paper has a higher number of concurrent users and shorter information retrieval time than conventional teaching systems.

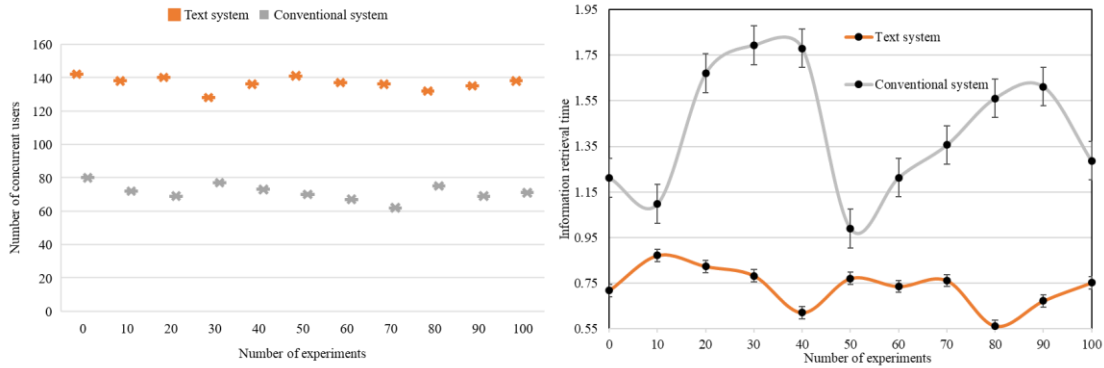


Figure 5: (a) Test result of concurrent user count (b) Test result of information retrieval time

3.3 Comparison of Performance Prediction

This paper uses C4.5, T-C4.5 and TG-C4.5 algorithms to conduct experiments. The experimental results include classification accuracy and time consumption. The detailed experimental results are shown in Figure 6.

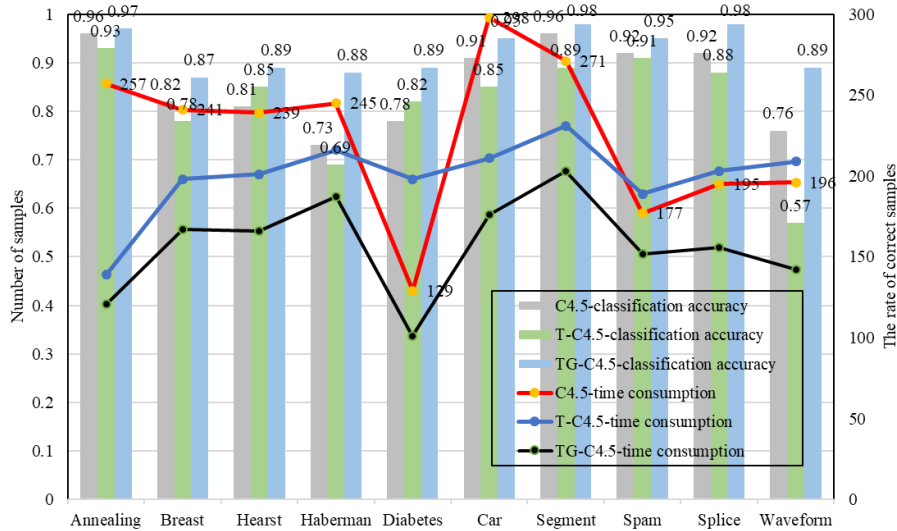


Figure 6: Comparison of C4.5, T-C4.5 and TG-C4.5 algorithm time consumption and classification accuracy.

A comparative analysis of the time consumption of the algorithm in Figure 6 can be obtained: Compared with the traditional C4.5, the introduction of the improved algorithm T-C4.5 and TG-C4.5 of the Taylor series greatly optimizes the operation of the algorithm effectiveness. However, compared with the TG-C4.5, T-C4.5 uses multiple expansions of Taylor series. Compared with T-C4.5, there is a slightly larger amount of calculation, resulting in lower computational efficiency. According to the comparative analysis of the classification accuracy of the algorithm in Figure 6, the improved algorithm is significantly lower than the traditional C4.5 in classification accuracy, which is related to the approximate value of the Taylor series, which makes the accuracy of the information gain rate lower, resulting in a decline in classification accuracy. Compared with the TG-C4.5, the classification accuracy of the TG-C4.5 is improved, but the difference is not very large.

The above experiments have proved the superiority of the TG-C4.5 algorithm, but the information gain calculation formula (10) in the algorithm is affected by the coefficient λ , we need to discuss and analyze to verify the classification accuracy under different values λ . The experimental results are shown in Figure 7.

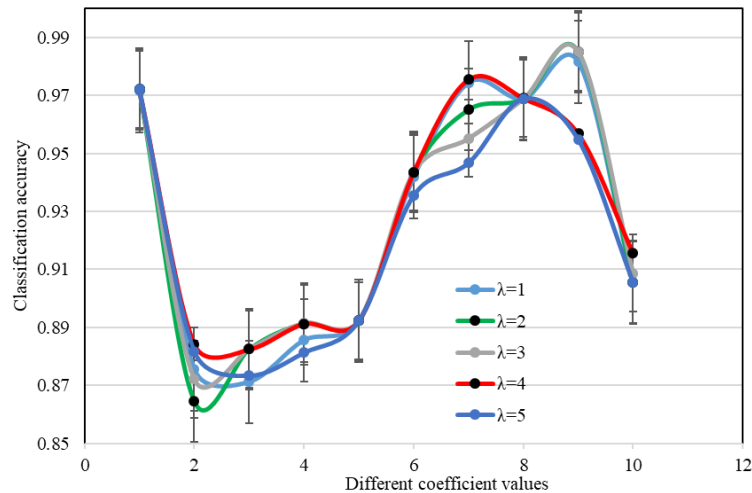


Figure 7: Classification accuracy of different coefficients λ .

According to the results in Figure 7 for statistical findings:

(1) When the values of λ are 1, 2, 3, 4, 5, the number of data sets that obtain the maximum classification accuracy in the corresponding 10 data sets is 1, 5, 6, 8, 2

(2) When the values of λ are 2, 3, 4, 5, compared with the improved TG-C4.5 ($\lambda=1$), the overall classification accuracy of the corresponding 10 data sets is increased by 0.02 %, 0.03%, 0.26%, -1.14%.

It can be obtained that when the value of λ is 4, the classification accuracy of most data sets has a significant improvement effect. Through the experimental results of different values of the coefficient λ , comparing the experimental results, it can be obtained that the TG-C4.5 algorithm has the best classification accuracy of 10 data sets when the value of λ is 4.

Based on the above experiments, we can get: The improved algorithm TG-C4.5 proposed in this paper, which introduces Taylor series and GINI index, has great advantages in classification accuracy; when different coefficients λ , especially $\lambda=4$, the improved TG-C4.5 algorithm has a better classification effect.

4 CONCLUSION

Nowadays, simple classroom teaching can no longer meet the requirements of both teaching and learning in daily teaching. The existence of teaching assistant systems, especially personalized teaching assistant systems, has become an inevitable trend. This article first applies data mining technology to the computer-assisted English classroom teaching model, and designs a computer-assisted classroom teaching system based on data mining. Secondly, use the improved decision tree TG-C4.5 algorithms to conduct data mining on the learning behavior data to achieve the goal of predictive classification of performance. Finally, solve the problem of educators who lack data mining knowledge about the understanding of the mining results, integrate the algorithm into the educational auxiliary system, realize the visualization of the predictive analysis results, and provide

the basis for teachers' decision-making on teaching early warning and other information. However, because the full realization of the individualization of the teaching aid system is a very complex and extensive research content, involving a variety of technologies and fields, there are still many problems that need to be improved by continued efforts. In future research, more data mining techniques will be used to continue to mine the effective information that is not fully utilized in the system, generate more reasonable teaching rules, and make the system more complete.

Jiling Shang, <https://orcid.org/0000-0002-0519-5578>

Chaohui Liang, <https://orcid.org/0000-0001-7072-5128>

REFERENCES

- [1] Xu, Z.; Shi, Y.: Application of constructivist theory in flipped classroom—take college English teaching as a case study, *Theory and Practice in Language Studies*, 8(7), 2018, 880-887. <http://doi.org/10.17507/tpls.0807.21>
- [2] Xu, Y.: Construction of a multiple English teaching mode based on cloud technology, *International Journal of Emerging Technologies in Learning*, 13(08), 2018, 239-253. <http://doi.org/10.3991/ijet.v13i08.9054>
- [3] Zhang, F.: Quality-Improving Strategies of College English Teaching Based on Microlesson and Flipped Classroom, *English Language Teaching*, 10(5), 2017, 243-249. <http://doi.org/10.5539/elt.v10n5p243>
- [4] Wenhong, H.: Study on college English teaching mode multimedia assisted based on computer platform, *International journal of multimedia and ubiquitous engineering*, 11(7), 2016, 351-360. <http://doi.org/10.14257/ijmue.2016.11.7.35>
- [5] Paternotte. E.; Van, S.; Bank, L.: Intercultural communication through the eyes of patients: experiences and preferences, *International Journal of Medical Education*, 8, 2017, 170-175. <http://doi.org/10.5116/ijme.591b.19f9>
- [6] Van, H. -K.; Abbott, K. -M.; Arbogast, A.: A Preference-Based model of care: an integrative theoretical model of the role of preferences in Person-Centered care, *The Gerontologist*, 60(3), 2020, 376-384. <http://doi.org/10.1093/geront/gnz075>
- [7] Li, S.; Gong, W.; Yan, X.: Parameter extraction of photovoltaic models using an improved teaching-learning-based optimization, *Energy Conversion and Management*, 186, 2019, 293-305. <http://doi.org/10.1016/j.enconman.2019.02.048>
- [8] Asri, H.; Mousannif, H.; Al, M. -H.: Using machine learning algorithms for breast cancer risk prediction and diagnosis, *Procedia Computer Science*, 83, 2016, 1064-1069. <http://doi.org/10.1016/j.procs.2016.04.224>
- [9] Chen, L.; Yang, X.; Sun, C.: Feed intake prediction model for group fish using the MEA-BP neural network in intensive aquaculture, *Information Processing in Agriculture*, 7(2), 2020, 261-271. <http://doi.org/10.1016/j.inpa.2019.09.001>
- [10] Sharifi, M.; Rostami, A.; Jafarigohar, M.: Retrospect and prospect of computer assisted English language learning: a meta-analysis of the empirical literature, *Computer Assisted Language Learning*, 31(4), 2018, 413-436. <http://doi.org/10.1080/09588221.2017.1412325>