Infrastructure Optimization of CET based on the Decision Tree Sorting Model in the Context of Big Data

Li Ying[1*]

City University of Zhengzhou, Zhengzhou, 452370, China

Corresponding author: Li Ying, liyingsmile123@163.com

Abstract. With the rapid growth of the Internet, we have entered a new era of big data. The uniqueness of the times of big data has led to new changes in the conventional teaching of English in universities and to the dual phenomenon of opportunities and challenges in the reform of English education in universities. To promote the effectiveness of English education quality assessment, the decision tree is used to establish the sorting rules for teaching quality assessment indicators, extract the effective indicators of English education quality assessment, and quantify and refine the attributes of the indicators. According to the samples to be evaluated, the entropy gain of each index attribute is calculated, the entropy gain rate is sorted in descending order, and the root and branch nodes are obtained. Combined with the values of node attributes, the number of branches is obtained, and a complete decision tree for teaching quality assessment is constructed. Statistically, the decision tree sorting model has been widely used in College English Test (CET) practice. In this study, two tree decision algorithms, ID3 and C4.5, were developed and optimized. After the experimental research, the work efficiency of ID3 and C4.5 was increased by 50%, and the accuracy was increased by 5% and 6%, respectively. The proposed method is of great significance in promoting the quality of CET and teachers' work efficiency.

Key words: Decision tree; Large data; CET; Infrastructure Development
DOI: https://doi.org/10.14733/cadaps.2023.S15.309-322

1 INTRODUCTION

Every advancement in science and technology has significantly influenced education and teaching. Relying on large amounts of data, the advent of "Internet +" has changed our way of life and thinking, deeply affecting the education industry. With the rapid growth of the Internet, human social life and economy have entered a newera of informatization, digitization, and globalization, making the importance of English increasingly prominent [20]. English is no longer simply a language and communication tool; it is also a necessary skill. With the rapid growth of IT, the integration of network IT with the university English curriculum has fundamentally changed the essence of the

College English Test (CET). University English teaching in classrooms has changed from conventional classrooms to university English classrooms in a modern network environment. CET in a big data environment brings advantages to classroom teaching but also influences teaching ideas, methods, models, and other aspects [1]. There are incongruities in the teaching reform model at the student, teacher, and environmental levels. The ecological balance of conventional university English classrooms has been disrupted, which affects the implementation of the teaching reform model. Therefore, it is particularly important to analyze and deal with various negative phenomena in college English teaching and seek an ecological approach. Data mining is a method of discovering valuable knowledge from massive amounts of data. Machine learning, as an important research direction, has been widely applied in various data-mining technologies. Data mining has been implemented using various algorithms [7].

The decision tree algorithm is commonly used in data mining. It is a process algorithm that uses tree rules to classify datasets. Generally, it is a data mining algorithm that generates a decision tree model from a sample dataset, which is is used for sorting [11]. Decision tree sorting is the most common sorting method as it has fast calculation speed, and the constructed sorting model can easily be transformed into various rules that are widely used. The decision tree algorithm is one of the most active areas of data mining research. Therefore, this study chose the sort mining method based on the decision tree as the research topic of CET optimization. Septime's classical decision tree sorting algorithm was analyzed, and its advantages and disadvantages were identified [12].

The decision tree algorithm has been an important reference and comparison object in data mining and sorting research for some time. Compared with conventional methods, it has the following advantages: the theory of the decision tree algorithm is clear, the method is simple, and the learning ability is strong; the concepts involved in the algorithm are mature; the algorithm is simple and easy to construct; the sorting rules generated by the algorithm are easy to understand; the readability and interpretability are high; and the algorithm is relatively complete and comprehensive when the training data set is relatively complete. Moreover, the accuracy of the model to correctly predict actual data is relatively high, and the calculation of the algorithm is relatively fast when the dataset structure is relatively fixed [15]. However, the algorithm is not very good at sorting a small number of important attributes due to its infomartion gain calculation characteristics. Only by solving this problem can the tree of decision algorithm be analyzed and promoted to bualgorithmse reasonable and balanced And promote the accuracy of sort prediction. Therefore, the decision tree algorithm will be introduced into the sample structure similarity model to promote the multi-value bias of the original algorithm [4]. In this study, two commonly used decision tree methods, ID3 and C4.5, are promoted. The innovations in this study are as follows:

1. This study combines large amounts of data to build a decision tree that includes comprehensive information and complete sorting. It can more comprehensively understand students' knowledge and identify problems in teaching in a timely and effective manner.

2.The two promoted decision tree algorithms in this study have the advantages of clarity, simplicity, comprehensiveness, fastness, and accuracy, which are helpful in reducing unnecessary consumption and promoting teachers' work efficiency.

3. The proposed method focuses more on teaching practice data and practice between teachers and students, which is helpful for promoting teaching quality. The promotion method proposed in this text has evident effects and is significant for improving CET mode and quality.

2 RELATED WORK

ID3 is a decision tree algorithm with typical significance. This method uses the amount of information to select and segment the samples. Information theory holds that when the amount of attribute information is large, the stability of the system and amount of information are high. Combining information theory and the proposed method, "information increment" is taken as a measure, selecting the attribute with the largest "information increment" in each non-leaf node as the test attribute, classifying it, selecting each value of the test attribute, constructing the corresponding sub node, and then constructing the corresponding discriminative model. Many scholars have improved the multivalued bias in the ID3 algorithm. For example, Tan et al. proposed considering a two-tier node interface when calculating the information gain [11]. Moon et al. proposed a promoted ID3 algorithm based on attribute values; that is, when the maximum number of values in attributes is greater than or equal to two, it will be corrected [8]. Sanz et al. propose introducing a function for the number of attribute values n for correction. Some methods consider only the number of attributes, where the connection between attributes is an important factor [10]. For example, a certain attribute has a strong relationship with the category attributes, and the number of values for that attribute is the largest. In this case, if the information gain is corrected, the decision tree can achieve the optimal effect. The C4.5 algorithm has a complete discriminative classifier system tree, which is its core concept. This method is based on the C4.5 algorithm and incorporates the handling of continuous attributes, attribute value spaces, and other issues, as well as a relatively mature pruning method and rules. Compared with traditional methods based on information growth rate, which is the ratio of information growth to the quantity of segmented information, for each leafless node, the node with the highest amount of information is selected as the detection node. This attribute effectively provides minimum randomness when dividing the samples, such that the expected number of tests required for object sorting is minimized, the amount of information required for classifying the samples in the result division is the smallest, and the final generated decision tree model is simple. Although the C4.5 algorithm allows fewer vacancies in the training sample set and has a certain anti-interference ability, the processing effect is not very obvious. The ID3 and C4.5 algorithms have great advantages in processing small-scale data. They can place all training sample sets into the main memory and have a high execution efficiency. However, in the face of large-scale training sample sets, the efficiency of the algorithm is significantly reduced when all the data cannot be put into the main memory. Therefore, the algorithm needs to be optimized to promote scalability so that it can effectively sort models in the face of large-scale training sample sets. The SLIQ algorithm uses a breadth-first search at the centimeter scale in a binary tree model. Because this algorithm can handle both discrete and continuous attributes, it is a fast and scalable decision tree algorithm that has been widely discussed by many researchers. However, the class table needs to be modifiable at any time when the algorithm is executed; therefore, the class table needs to be stored in memory. However, the size of the class table will increase with an increase in the training sample set size. Therefore, the algorithm requires extremely high memory and cannot remove the limitations of the main memory capacity, resulting in poor practicability and fewer applications of the algorithm. The Sprint algorithm was proposed to effectively solve the problem in which other decision tree algorithms are limited by main memory capacity and are unsuitable for large-scale training sample sets. Because the algorithm runs quickly and allows multiple processors to jointly create a decision tree model, it is an extensible and parallel inductive decision tree algorithm.

3 METHODOLOGY

3.1 Data Mining

To understand the decision tree sorting model, we must begin with data mining. Data mining is the procedure of mining potential or hidden information and knowledge from large amounts of data using relevant algorithms [13]. With the evolution of Web-based techniques, data entered by users through the Web can be automatically saved and continuously collected by sensors. In addition, with the evolution of mobile Internet, the speed of automatic data collection and storage is correspondingly increasing, which makes the amount of data from all over the world constantly expand, and the storage and calculation of data have exceeded the processing capacity of a single computer. This poses a challenge to the implementation of data mining techniques. Data mining involves different application fields and technical methods, and its procedures also vary. According to different mining standards and requirements [6], the data mining procedure can be summarized into five stages: preparation, data preparation and analysis, model training, model validation and evaluation, and online application. The decision tree algorithm is commonly used in data mining and is a procedural algorithm that uses tree rules to classify datasets [2], [3].

3.2 Decision Tree

Figure 1 illustrates a decision tree structure. On this basis, a genetic algorithm is proposed. In the figure, each internal node represents an attribute (N21, N22), and each leaf node, such as L1, L2, and L3, represents a branch of a class.

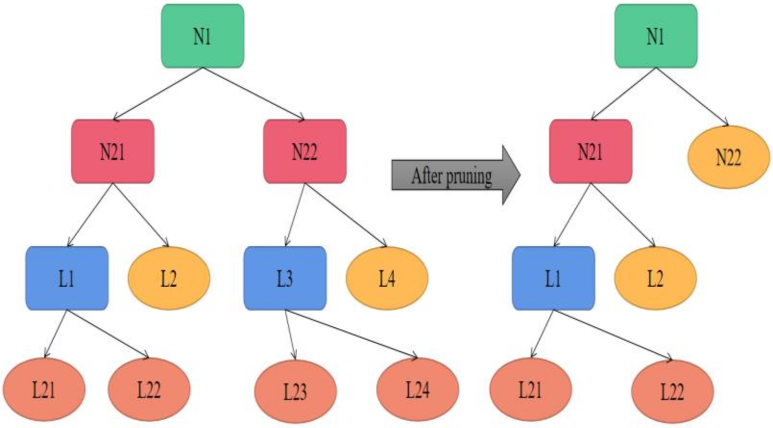


Figure 1: Typical decision tree format.

Sorting by the decision tree is divided into two steps. First, a decision tree is constructed using sample datasets, and the corresponding sorting model is established using the decision tree [20]. This is a procedure for acquiring knowledge from data and refining rules. The second step is to use the established decision tree model to classify the CET practice; that is, the unknown dataset tuples travel through the decision tree from the root node in turn and travel to a leaf node through a certain path to find the class or distribution of the class where the dataset tuples are located. The decision tree model is illustrated in Figure 2.

The ID3 algorithm is based on information theory and uses information entropy and information gain as attribute judgment and selection criteria, as well as greedy search and top-down methods to construct decision trees to realize the induction and classification of data [16]. The ID3 decision tree algorithm has two assumptions based on probability theory and information theory. First, the classification probability of a correct decision tree for any example is the same as the probability of positive and negative examples in S , as shown in Equation (1):

$$Info(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \tag{1}$$

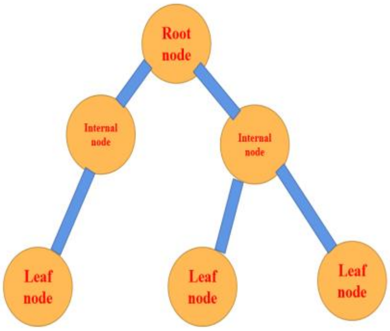


Figure 2: Decision tree model example

If the attribute A is selected as the root node of the ID3 decision tree according to the judgment criteria, then the attribute A is set to have i values $(v_1, v_2, \cdots v_i)$, and it divides S into I subsets $(S_1, S_2, \cdots S_i)$. Assuming that any subset S_1 of S contains p_i positive examples and n_i negative examples, the information of the subset S_i can be obtained as follows:

$$Entroy(p_i, n_i) = -\frac{p_i}{P_i + n_i} \log_2 \frac{p_i}{P_i + n_i} - \frac{n_i}{P_i + n_i} \log_2 \frac{n_i}{P_i + n_i} \tag{2}$$

The information entropy is classified with attribute A as the root-node is $Entroy(A)$, and its calculation formula is as follows:

$$Entroy(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} Entroy(p_i, n_i) \tag{3}$$

Therefore, the formula $Gain(A)$ of the information gain divided by A as the root node can be obtained as follows:

$$Gain(A) = Info(p, n) - Entroy(A) \tag{4}$$

In general, the ID3 algorithm is a very practical decision tree algorithm. Its basic theory is clear and easy to understand, but it also has some shortcomings. The algorithm has the problem of multi-value bias and prefers to select the conditional attribute with more attribute values as the decision attribute; however, in many cases, the attribute with more attribute values is not the optimal attribute [5], [18]. When building a decision tree, each node contains only one feature, which is a single-argument algorithm. The correlation between attributes is not sufficiently strong, and although they are connected by a tree, they are still scattered. Sensitivity to noise data and difficulty in removing noise lead to errors in the selection of eigenvalues. Although the basic theory is clear, its calculation is relatively complicated, and the memory occupancy rate is high, which affects training time and space [9].

We divide the sample training set into three subsets according to academic performance: dominant and difference, and continue to use the promoted C4.5 algorithm to recursively obtain the student achievement tree shown in Figure 3.

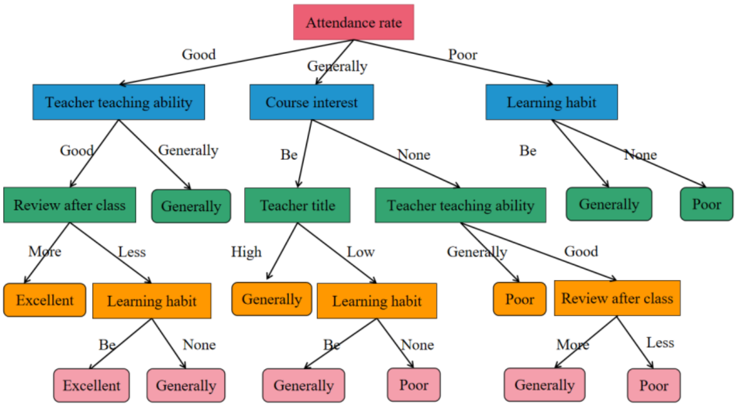


Figure 3: Decision tree constructed by the promoted C4.5 algorithm.

Although the C4.5 algorithm allows fewer vacancies in the training sample set and has a certain anti-jamming ability, the effect of the procedure is not very obvious. Scholars have conducted extensive research on this, and the literature puts forward an effective solution: assigning the most commonly used value of the vacancy attribute in the test training set. However, this method is not very accurate in assigning the attribute value. Another complicated solution strategy was mentioned in the literature: assigning a probability to every possible value of the vacancy attribute. However, this method needs to calculate every possible value of the attribute, which ultimately leads to low efficiency.

The ID3 and C4.5 algorithms have great advantages in processing small-scale data. They can place all training sample sets into the main memory and have high execution efficiency. However, in the case of large-scale training sample sets, the efficiency of the SLIQ algorithm is significantly reduced when all data cannot be stored in the main memory. SLIQ is a fast and scalable decision tree algorithm that has attracted considerable attention. In the SLIQ algorithm, the Gini Index is used instead of information. The Gini index performs better than information and is convenient to calculate:

$$Gini = 1 - \sum_{i=1}^{m} p_i^{2} \tag{5}$$

For a dataset S containing n classes, $Gini(S)$ is defined as , where P_i is the frequency of the jth class data in S . Gini metrics are used to measure the impurity of data partitioning or training tuple sets. If Gini is smaller, the information gain is larger, and the quality of node splitting is better. However, because the class table must be modifiable at any time during the implementation of the algorithm, it needs to be resident in memory. The memory requirement of the algorithm is very high and cannot be removed from the limitation of main memory capacity, resulting in poor practicability and fewer applications of the algorithm. Some scholars have proposed a new method to omit the logarithmic optimization in the algorithm, reduce the amount of calculation, and quickly generate the decision tree by using the Taylor formula, McLaughlin formula, and theory of equivalent infinitesimals. If there are m positive examples and k negative examples and the information quantity of attributes is

$$I(m,k) = -\frac{m}{m+k}\log_2\frac{m}{m+k} - \frac{k}{m+k}\log_2\frac{k}{m+k} \tag{6}$$

then the information entropy is as follows:

$$E(A) = \sum_{i=1}^{v}\frac{m_i + k_i}{m+k}I(m_i, k_i) \tag{7}$$

A portion of the data is extracted from the database as a training set. The training set is a collection of sample vectors, and each attribute corresponds to a column in the training set. After the system enters the training set, it is divided into attribute tables one by one, and all class identifiers are placed into the class table. The leaf field in the class table indicates the tree leaf corresponding to the record.

The SPRINT algorithm was proposed to effectively solve the problem of other tree-based decision algorithms being limited by the main memory capacity and being unsuitable for large-scale training sample sets. Because the algorithm runs fast and allows multiple procedures to create a tree of decision models, it is an extensible and parallel-inductive tree of the decision algorithm. The SPRINT algorithm can handle extremely large training sample sets. The larger the number of data sample sets, the higher the execution efficiency of SPRINT and the better its scalability. However, the SPRINT algorithm has some defects: when the class table of SLIQ is stored in memory, the execution speed of the SPRINT algorithm is slower than that of the SLIQ algorithm, which increases the storage cost because of the use of the attribute table.

3.3 Comparison of Main Decision Tree Algorithms

A performance summary of decision tree algorithms is shown in Table 1.

Characteristic Algorithm	*Structure*	*Test attribute method*	*Attribute handling*
ID3	*Multitree*	*Info gain*	*Discretization*
C4.5	*Multitree*	*Info gain rate*	*Pre sort*
SLIQ	*Binary tree*	*Gini coefficient*	*Pre sort*
SPINT	*Binary tree*	*Gini coefficient*	*Pre sort*

Table 1: Performance summary of decision tree algorithms

The characteristics of various algorithms are compared in Table 2.

Algorithm	*Pruning principle*	*Scalability*	*Parallelism*	*Structure*
ID3	*Sort error*	*Poor*	*Poor*	*Multitree*
C4.5	*Sort error*	*Poor*	*Poor*	*Multitree*
SLIQ	*MDL*	*Very good*	*Very good*	*Binary tree*
SPINT	*MDL*	*Good*	*Good*	*Binary tree*

Table 2: Comparison of characteristics of various algorithms.

4.1 Building a Decision Tree Model for Student Achievement

The optimized ID3 algorithm was used to construct a university English practice teaching effect assessment model and conduct a practice effect assessment test in a university. The results are shown in Figure 4.

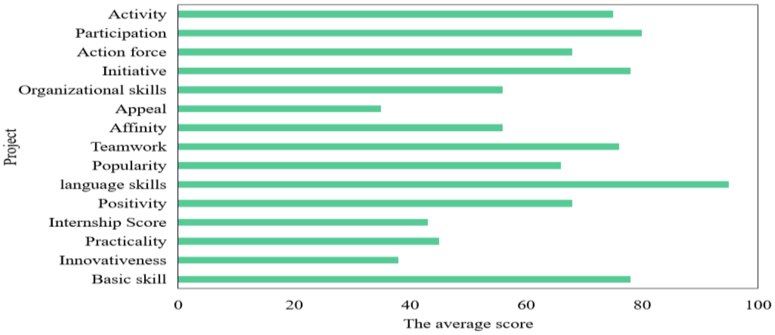


Figure 4: Assessment results of University English Practical Teaching.

4.2 Teaching Quality Assessment of Decision Tree Model and Rule Association Analysis

The assessment indicators of teaching quality were selected from two aspects: the relevant teacher data and related factors of teaching class scheduling and teaching links. The specific indicators are listed in Table 3.

Indicator name	*Index shorthand*	*Indicator description*
Teacher education	*X*	*Undergraduate X1, Master X2, Doctoral X3*
Teacher title	*Z*	*Beginner Z1, Intermediate Z2, Advanced Z3*
Teaching age of Teachers	*N*	*N1 is less than 5 years, N2 is 5-15 years, and N3 is greater than 15 years*
Teaching attitude	*T*	*Good T1, General T2*
Number of teachers evaluated	*P*	*Actual number of students participating in teaching assessment*
Teaching assessment score	*S*	*Poor S1, fail; S2, medium 60-80 points; Good S2, 80-90 points; Excellent S4, above 90 points.*
Multimedia teaching	*D*	*Proficiency (good D1, advantaged D2)*

Course category	*L*	*Major foundation L1, general education L2, major optional L3*
Number of Persons	*R*	*Actual number of course teaching*

Table 3: Specific indicators of teaching scheduling and related factors of teaching links.

The assessment procedure was as follows. First, the data of the teaching quality assessment data sample are sorted according to the indicators in Table 1, and then e (a) and gain (a) are calculated according to Equations (5) and (6). The highest indicator of gain (a) is selected as the root node and the second highest value of gain (a) as the branch node. With this method, a complete decision tree is constructed, and then if-then sort rules are obtained according to the tree structure. Finally, the associated rules are used to verify the consistency of the teaching quality assessment. The specific steps of the decision tree generation and associated rule verification are shown in Figure 5.

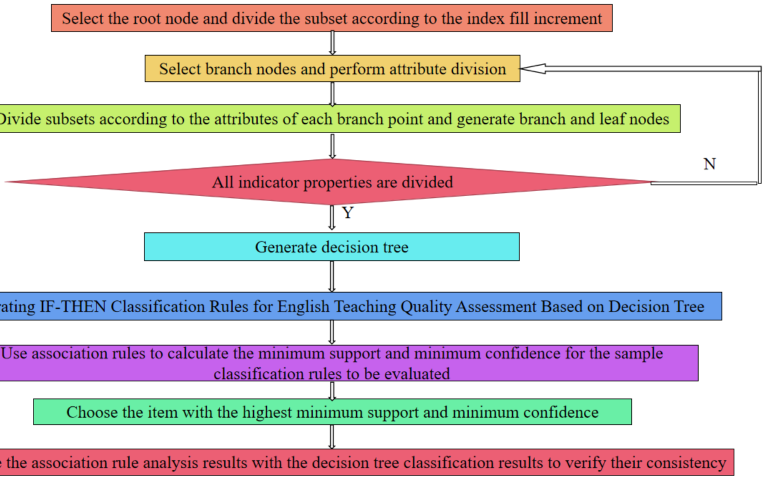

Select the root node and divide the subset according to the index fill increment
Select branch nodes and perform attribute division
Divide subsets according to the attributes of each branch point and generate branch and leaf nodes
N
All indicator properties are divided
Y
Generate decision tree
Generating IF-THEN Classification Rules for English Teaching Quality Assessment Based on Decision Tree
Use association rules to calculate the minimum support and minimum confidence for the sample classification rules to be evaluated
Choose the item with the highest minimum support and minimum confidence
Compare the association rule analysis results with the decision tree classification results to verify their consistency

Figure 5: Specific steps of decision tree generation and associated rule verification.

4.3 ID3 Application of Algorithms in Teaching Quality Assessment of University Teachers

With the continuous expansion of the enrollment scale of universities in the author's country in recent years, society has higher requirements for teachers' teaching levels. Most pre-oral teaching quality assessments adopt the method of assessment by students, which issues a teacher teaching quality assessment card to the students in the mid-term or at the end of term. The teachers are scored according to the assessment items of the assessment card, and then the statistics are passed to the educational affairs management department. Teachers' teaching quality assessment grades are then determined based on the scoring results. The procedure of the ID3 algorithm using selected dataset examples is described below. The examples used here were determined according to the index factors of professional title, political outlook, degree, and age. Table 4 shows whether the teachers' teaching quality is excellent according to the dataset.

Employee number	Job title	Gender	Academic degree	Teaching age	Is it excellent
1001	*Middle-level*	*male*	*bachelor*	*10-20*	*NO*
1002	*Primary*	*female*	*master*	*≦10*	*NO*
1003	*Middle-level*	*male*	*bachelor*	*≧20*	*YES*
1004	*High level*	*female*	*doctor*	*≧20*	*YES*
1005	*Primary*	*male*	*doctor*	*≦10*	*YES*
1006	*Primary*	*male*	*bachelor*	*≦10*	*YES*
1007	*High level*	*female*	*master*	*10-20*	*YES*
1008	*High level*	*female*	*bachelor*	*10-20*	*NO*
1009	*Middle-level*	*male*	*bachelor*	*≦10*	*YES*
1010	*Middle-level*	*male*	*bachelor*	*≦10*	*NO*
1011	*High level*	*female*	*master*	*10-20*	*YES*
1012	*Primary*	*male*	*master*	*10-20*	*YES*

Table 4: Dataset.

A training set S of tuples of labeled classes is given in Table 4, in which the class label attribute "excellent or not" has two different values (i.e. {yes, no}), so there are two different classes (i.e., $n=2$). Let class C correspond to "Yes" and class C_z correspond to "No". Class C has eight tuples, and class C_z has six tuples. First, we use Equation (10) to calculate the expected information required to classify the tuples:

$$I(S_1, S_2, \cdots S_n) = -\frac{8}{14}\log_2\frac{8}{14} - \frac{6}{14}\log_2\frac{6}{14} = 0.985 \tag{10}$$

If the tuples are divided by "job title", the desired information needed to classify the tuples in S is as follows:

$$E(Title) = \frac{5}{14}\times(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) + \frac{5}{14}\times(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) + \frac{4}{14}\times(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}) = 0.925 \tag{11}$$

Therefore, the information gain of this divisionis

$$Gain(Title) = 0.985 - 0.925 = 0.06 \tag{12}$$

Comparing the sorting accuracy of the various algorithms, we can see from the analysis results in Table 5 that the promoted C4.5 algorithm achieves a sorting accuracy of 80.5%, the promoted ID3 algorithm reaches 77.5%, and the sorting accuracies are higher than those of the unprompted ID3 and C4. 5 algorithms.

Algorithm	Classification accuracy
ID3	*72.2%*
promoted ID3	*77.5%*
C4.5	*74.3%*
promoted C4.5	*80.5%*

Table 5: Comparison of sorting accuracy before and after promotion of the two decision tree algorithms.

Table 6 lists the time efficiency of establishing the training data models for various classifiers with a time complexity of seconds. We can see that the advantaged execution time of the promoted ID3 algorithm is reduced to 0.9 seconds, and the advantaged execution time of the C4.5 algorithm is reduced to 0.6 seconds, which promotes the optimization efficiency of the algorithm.

Algorithm	*The advantaged execution time*
ID3	*1.8*
promoted ID3	*0.9*
C4.5	*1.2*
promoted C4.5	*0.6*

Table 6: Correct rates of the algorithm when the number of attributes is continuously increasing.

We can see from the table that although the highest value of each algorithm is almost the same when the number of attributes is the largest, when the number of attributes is slightly less, for example, when there are two or three attributes, the accuracy of the promoted C4.5 algorithm is more stable and shows better robustness, as shown in Figure 6.

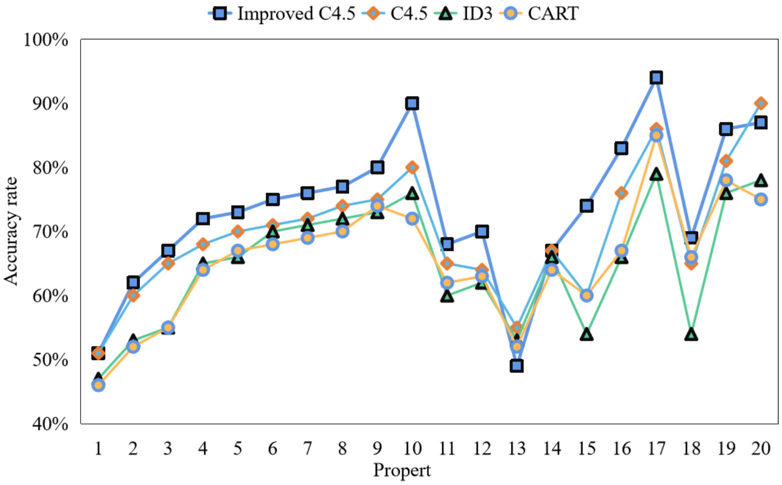


Figure 6: Comparison of algorithm accuracy.

At the same time, in terms of the time efficiency of the algorithm, the promoted C4.5 algorithm takes less time to execute, as shown in Figure 7. We can see from the above experimental results that the recognition accuracy of the promoted C4.5 and ID3 algorithms applied in this study was higher than that of the conventional C4.5 and ID3 algorithms and the other two algorithms. From a technical perspective, this shows that the promoted C4.5 and ID3 algorithms shorten the waiting time of data analysis, promote work efficiency, and ensure the correct rate of sorting. The experimental results are shown in Figure 8. The datasets of university English speaking, reading, audiovisual, and language theory provided in Figure 7 are the original sample datasets, and sorting experiments were conducted on each dataset. In each experiment, one-third of the sample dataset was the training dataset, and this third of the data was randomly selected from the sample dataset each time; the whole sample dataset was the test dataset for calculation. After building the decision tree, the number of decision leaf nodes was counted, the accuracy was calculated, and the advantaged 30 experimental results were calculated. The final experimental data are presented in Figure 9.

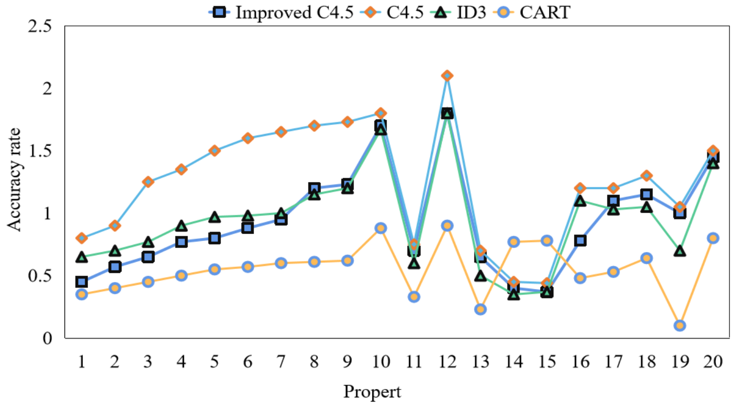
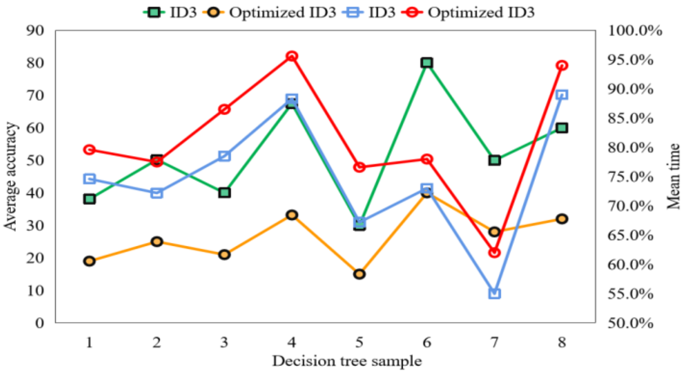


Figure 7: Comparison of algorithm efficiency



Firgue 8: Experimental results of constructing the tree of decision with different samples.

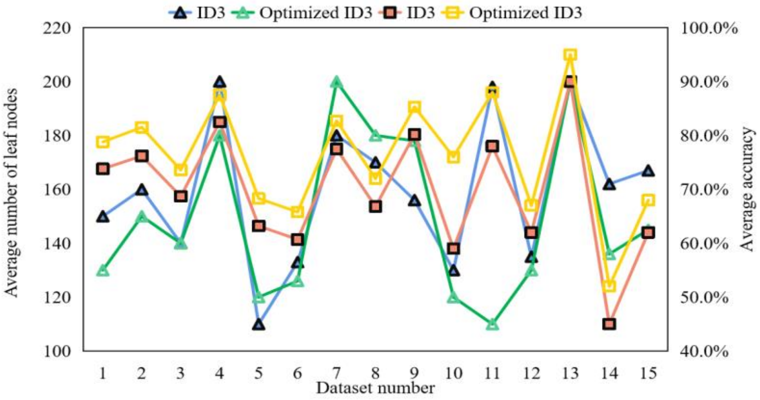


Figure 9: Final experimental data.

The comparison of sorting accuracy before and after promoting is shown in Figure 10.

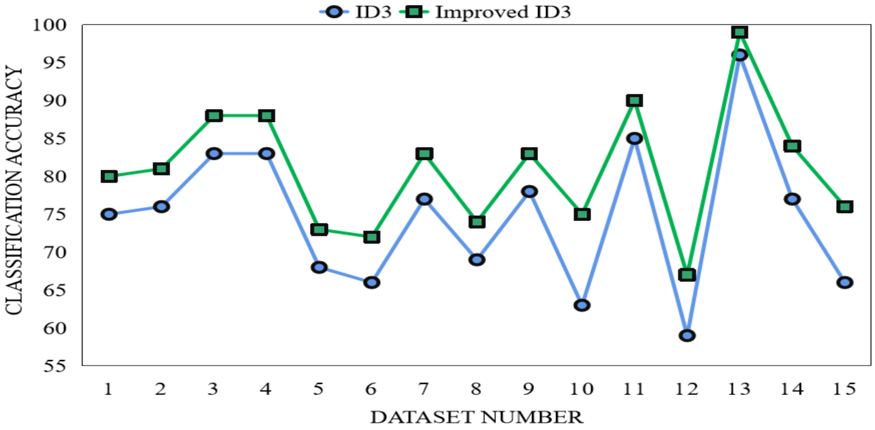


Figure 10: Comparison of sorting accuracy before and after promoting.

4 CONCLUSIONS

In summary, in this era of big data, the traditional teaching mode of college English writing is clearly outdated and needs to be actively changed. Data classification is an important part of data mining, and the decision tree algorithm is an important method for data classification with a mature theoretical basis and a good development platform. The introduction of big data will make English teaching more flexible and diverse. Under such circumstances, English teaching practitioners and planners must change their thinking concepts, adjust their roles, improve their knowledge structure, and update their educational concepts to organically combine big data technology with college English teaching to promote the development of college English teaching in this new era. The analysis and utilization of high-quality classified college English teaching data provides a more accurate and reasonable decision tree algorithm, which not only effectively reduces the invalid consumption of workload but also makes data positioning and analysis more rapid, comprehensive, and accurate. Compared to the original scheme, the improved schemes proposed in this paper improve the work efficiency by 50% and accuracy by 5% and 6%. Therefore, it is important to adopt an improved decision tree method to improve the quality of college English teaching and teachers' work efficiency. Infrastructure development, in the context of this paper, refers to the establishment of a technological framework that supports the analysis and utilization of data in college English teaching. This involves creating efficient systems for data classification, positioning, and analysis using the decision tree algorithm to enhance the accuracy and efficiency of English teaching. By adopting the improved decision tree proposed in this paper, the quality of college English teaching can be improved and teachers' work efficiency can be enhanced.

Li Ying, https://orcid.org/0009-0001-9243-445X

REFERENCES

[1] Arunesh, P.K.A.: Role of Decision Tree Sort in Data Mining, International Journal of Pure and Applied Mathematics, 119(15), 2018, 2533-2543.

[2] Devasenapathy, K.; Duraisamy, S.: Evaluating the Performance of Teaching Assistant Using Decision Tree ID3 Algorithm, International Journal of Computer Applications, 164(7), 2017, 23-27. https://doi.org/10.5120/ijca2017913658

[3] Du; J.: Assessment Method of CET Quality Based on Fuzzy Comprehensive assessment, International English Education Research: English Edition, 65(1), 2018,3.

[4] Guo, B.: Analysis on Influencing Factors of Dance Teaching Effect in Universitys Based on Data Analysis and Decision Tree Model, International Journal of Emerging Technologies in Learning (iJET), 15(9), 2020, 245. https://doi.org/10.3991/ijet.v15i09.14033

[5] Liu, P; Wu C; Tan X.: An Analysis Report of University teaching of English in the classrooms in the Grading Model, Journal of Language Teaching and Research, 8(4), 2017, 768. https://doi.org/10.17507/jltr.0804.17

[6] Lu, J.: Reconstruction of Educational Ecology in CET in info Age, Overseas English, 78(9), 2020, 2.

[7] Mehenni, T.; Moussaoui, A.: Datamining from multiple heterogeneous relational databases using the tree of decision sort, Pattern Recognition Letters, 33(13), 2012, 1768-1775. https://doi.org/10.1016/j.patrec.2012.05.014

[8] Moon Lee, S. K.: Applying of Decision Tree Analysis to Risk Factors Associated with Pressure Ulcers in Long-Term Care Facilities, HEALTHC INFORM RES, 23(1), 2017, 43-52. https://doi.org/10.4258/hir.2017.23.1.43

[9] Rizvi, S; Rienties, B; Khoja, S. A.: The role of demographics in online learning; A the tree of decision based approach, Computers & Education, 137(8), 2019, 32-47. https://doi.org/10.1016/j.compedu.2019.04.001

[10] Sanz, J.; Fernandez, J.; Sola, H.B.; et al.: A decision tree based approach with sampling techniques to predict the survival status of poly-trauma patients, International Journal of Computational Intelligence Systems, 10(1), 2017, 440-455. https://doi.org/10.2991/ijcis.2017.10.1.30

[11] Tan, Q.: Assessment System of CET Based on Large Data, Journal of Physics: Conference Series, 1852(2); 2021, 022014 (6pp). https://doi.org/10.1088/1742-6596/1852/2/022014

[12] Wang, B.; Yang, Y.: CET Assessment System in the times of Large Data, American Chinese and Foreign Languages: English Edition, 15(4); 2017, 5.

[13] Wang, G.: Functional Discourse Analysis and CET, Open Access Library Journal, 9(5), 2022, 9.

[14] Wasil, M.; Sudianto, A. ; Fathurrahman.: Application of the Decision Tree Method to Predict Student Achievement Viewed from Final Semester Values, Journal of Physics Conference Series, 48(5), 2020, 1539. https://doi.org/10.1088/1742-6596/1539/1/012027

[15] Wei, J.X.; Wang, J.; Zhu, Y.X. ; et al.: Conventional Chinese medicine pharmacovigilance in signal detection: The tree of decision-based data sort, BMC Medical Informatics and Decision Making, 18(1), 2018, 19. https://doi.org/10.1186/s12911-018-0599-5

[16] Yi, L.: Optimization of CET Model Based on the Multi-Attribute Decision Method of Random Variables, Dynamic Systems and Applications, 29(5), 2020, 66-69. https://doi.org/10.46719/dsa202029514

[17] Zhang, W.: Research on English score analysis system based on promoted the tree of decision algorithm and fuzzy set, Journal of Intelligent and Fuzzy Systems, 39(4), 2020, 5673-5685. https://doi.org/10.3233/JIFS-189046

[18] Zhang, W.: Research on English score analysis system based on promoted the tree of decision algorithm and fuzzy set, Journal of Intelligent and Fuzzy Systems, 39(4), 2020, 5673-5685. https://doi.org/10.3233/JIFS-189046

[19] Zhao, X.: Application of deep learning algorithm in CET procedure assessment, Behaviour and Info Technique, 133(11), 2019,1-10.

[20] Zhao, H.; Liu, Z.; Yao, X.; Yang, Q.: A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach, Information Processing & Management, 58(5), 2021, 102656. https://doi.org/10.1016/j.ipm.2021.102656