Intelligent Data Mining-Based Method for Efficient English Teaching and Cultural Analysis

Qing Ai, Zhujiang College, South China Agricultural University, China*
Hongyu Guo, Zhejiang Gongshang University, China

ABSTRACT

The emergence of online education helps improve the traditional English teaching quality greatly. However, it only moves the teaching process from offline to online, which does not really change the essence of traditional English teaching. In this work, the authors study an intelligent English teaching method to further improve the quality of English teaching. Specifically, the random forest is firstly used to analyze and excavate the grammatical and syntactic features of the English text. Then, the decision tree-based method is proposed to make a prediction about the English text in terms of its grammar or syntax issues. The evaluation results indicate that the proposed method can effectively improve the accuracy of English grammar or syntax recognition.

KEYWORDS

Data Mining, Decision Tree, English Teaching, Intelligent, Random Forest

1. INTRODUCTION

Since the beginning of the 21st century, the world economy and arms competition have become increasingly fierce, and the demand for English professionals has gradually expanded. In order to meet the needs of rapid economic development, different colleges and universities have implemented a series of reforms in view of the existing English teaching models (Guan 2018). Among them, English grammar and syntactic analysis are the key points in the foundation, and they are the necessary conditions for the establishment and cultivation of compound English talents. With the introduction of the compound talent training mode of “English + news”, colleges and universities gradually begin to actively explore diversified English teaching models, give full play to students’ practical and innovative abilities, and carry out comprehensive practical education around the creative ability of graduates (Mahbub 2021). It plays an important role in cultivating the new ability of English majors.

At present, although the level of English teaching in colleges and universities in China is constantly improving, English, as a public subject, has not been deeply integrated with specific fields and vocational skills, which is reflected in the unfamiliarity of English grammar rules in many fields (Albiladi 2019). French semantic logic is not clear and so on. Therefore, in order to fully meet the needs of social development and educational reform in colleges and universities, we should actively explore an English teaching model dominated by the cultivation of students’ professional ability, focusing on the cultivation of basic English analysis and judgment, and realizing the simultaneous development of students’ English proficiency and vocational skills plays an important role in improving students’ English ability (Mirbabayeva 2020).

DOI: 10.4018/IJMCMC.293745

*Corresponding Author

In the context of the current big data era, the rapid development of technology has also brought opportunities for the development of college English teaching. Especially in the current teaching, the application of some advanced network technology and information processing technology has greatly facilitated English teaching, changed the traditional English teaching mode, made teaching more diversified and more interesting, and significantly improved students' interest in learning (Elyas 2018). With the assistance of big data, school institutions can build an effective English grammar and syntax database, record students' phased English learning, and then organize and analyze them as a basis to constantly optimize teaching ideas. improve the pertinence of English teaching, so as to fully meet the diversified and dynamic needs of students (Hudson 2005). Under the application of big data, we can also strengthen the dynamic monitoring of teaching quality and improve the timeliness of teaching. In addition, the application of information technology can help students understand knowledge and enrich classroom content, so as to improve students' learning efficiency and teaching quality (Guo 2021). Therefore, in order to meet the needs of the development of major industries, colleges and universities at home and abroad gradually combine data mining and other related contents to carry out a wide range of English teaching, including basic grammar, semantics and so on (Lou 2020).

However, big data brings both opportunities and challenges to English teaching. The challenges are mainly reflected in two aspects. First of all, the massive teaching resource in big data's era have weakened the English teaching model based on teaching materials. In the current efficient English teaching system, due to the solidification of the teaching materials used, with the continuous reform and development of teaching, it may have been unable to meet the needs of the current development (Weng 2020). Moreover, a large amount of English teaching resource are usually discrete and there are a lot of repetitive content, which makes it very difficult to obtain effective English knowledge (Bienkowski 2012). Therefore, although big data is rich in resource, students have a negative attitude towards finding effective resource from the vast amount of data on the Internet, so it is difficult to improve their interest in learning. In addition, the teacher-led classroom learning model leads to students' passive acceptance of knowledge and lack of interaction. In the long run, it will not only lead to professional exhaustion of teachers, but also can not bring effective feedback to students' English learning effect (Mihaescu 2021).

Based on the above consideration, in order to improve the quality of English teaching, this paper explores the important writing features of English sentences at the grammatical and syntactic level by using random forest and logical regression machine learning algorithms. Based on this feature, the decision tree model is used to optimize and predict and judge the grammatical and syntactic correctness of sentences, so as to improve the quality of English teaching. The purpose of this paper is to provide a basis for improving English teaching management and improving the quality of English teaching. The main contributions of this paper are as follows:

- By using the random forest theory, he fully analyzes and excavates the grammatical and syntactic features of the text in English teaching, and then extracts the English sentence features at the grammatical and syntactic levels, which can be used for further modeling and analysis of English grammar teaching.
- According to the extracted effective sentence features, an English sentence classification method based on decision tree method is proposed, and an English grammar prediction model is established by using a large amount of English grammar teaching data for learning and classification. and then provide the model and method basis for efficient and automatic English teaching.
- By using a large number of data for statistical analysis and theoretical verification of the proposed model method, the results show that the proposed method can effectively improve the accuracy of English grammar recognition, and then improve students' autonomous learning ability.

The organizational structure of this paper is as follows: the second chapter mainly introduces the related work. The third chapter introduces the model and method proposed in this paper. The fourth chapter introduces the research results of this paper. The fifth chapter is a summary.

2. OVERVIEW OF RELATED TECHNOLOGY AND RESEARCH

Based on the existing research, we divide the related work into two categories, namely, the traditional English teaching research based on data mining and the English teaching research based on intelligent data mining. The following first summarizes the related technologies and current situation of the two kinds of research work, and then carries on the summary and discussion.

2.1 Traditional Data Mining Based English Teaching Work

With the rapid development of information technology, a large amount of data has been accumulated in the daily operation and work in the field of English teaching. With the help of these data, relevant personnel can use artificial or intelligent methods to deal with related work and find many rules and patterns (Dogan 2020). Through the analysis of these data, we can provide important information for the decision-making of English teaching quality improvement and mode reform, which is helpful to improve work efficiency. Data mining technology is an interdisciplinary subject which integrates artificial intelligence technology, visualization technology, database technology, statistics and machine science. It was first proposed in 1989 (Romero 2020). At present, people have different definitions of data mining, among which it is generally accepted that data mining is from a lot of incomplete, large, noisy, fuzzy and random data, through related programs to extract some hidden information and knowledge that people do not know (Kumar 2020). For English teaching, the process of using data mining technology also includes three stages, namely, English text data preparation, English rule feature mining and result expression.

The first is the English text data preparation stage, which mainly covers three sub-steps: text data acquisition, selection, data pre-processing and data conversion. The acquisition of English text data usually adopts the way of web crawler, and the selection is made according to the main goals and tasks of data mining of relevant English institutions. Mining some analytical data from the original and existing data to form the target data of data mining, which lays a good foundation for later data mining (Taleb 2015). Data pre-processing is to make the target data meet the requirements of data mining through the analysis and processing of the target data, which is also the specific function of data mining. In this link, relevant personnel are needed to calculate the missing values of English text data, so as to ensure the standardization of the data format. The second is to mine the grammar rules and rules in the English text, and we need to choose the appropriate mining methods according to the actual goals and requirements, such as decision tree, support vector machine and so on (Zhang 2020, Wen 2020). For example, Zhang W. et al. use the decision tree method to mine English syntactic rules (Zhang 2020), while Wen H. et al. use the support vector machine algorithm. In comparison, the support vector machine method is more suitable for complex English grammar and grammar situations (Wen 2020), while the decision tree method has the advantages of high efficiency and low cost. Finally, there is the expression of the results of English text mining, and the rules to be mined are expressed in a general form.

2.2 Intelligent Data Mining Based English Teaching Work

The traditional methods of machine learning in the process of English teaching based on data mining can be divided into two categories, namely, supervised learning and unsupervised learning. The main difference between the two models lies in whether the data used in the English teaching model not only has the basic characteristics, but also has the corresponding label value to classify it. Generally speaking, there are two main kinds of supervised learning problems, classification and regression. The main algorithms of supervised learning include K-proximity analysis, linear model, naive Bayesian classifier, decision tree, kernel support vector machine, neural network and so on. There are two common types of unsupervised learning: data set transformation and clustering. The unsupervised transformation of data sets is an algorithm for creating new representations of data. Compared with the original representation of data, the new representation may be more easily understood by people or other machine learning algorithms. The clustering algorithm divides the data into different groups, each group contains similar items, and the main clustering algorithms are K-means clustering and agglomeration clustering (Liu 2021). Classification is to label the data according to a certain standard, and then classify them according to the label. Clustering is a process of finding out the reasons for the clustering of data through analysis without a "label" in advance (Zhen 2021).

Many scholars at home and abroad have used intelligent data mining technology to study English teaching and obtained more research results. For example, Xu Z. et al. use stepwise regression and neural network techniques to analyze college students' English learning achievements and their influencing factors (Xu 2019). Pangaribuan T. et al. analyzed a large number of learning data generated in the process of online English learning with the help of statistical analysis and visualization, association rule algorithm and clustering algorithm (Pangaribuan 2018). According to the results of the analysis, some thoughts and suggestions on the supervision and management of web-based English learning process are given. Cui S. et al. applied the multiple linear regression model to predict the academic performance of students in hybrid college English courses based on physical classroom and cloud learning platform, and carried out teaching intervention to improve their learning effect according to the prediction (Cui 2017). Dalton-Puffer C. et al. constructed a classified prediction model of online English learners' academic performance by using the method of nested integrated learning, which provides a reference for the research on the influencing factors and predictive modeling of online English learners' academic performance, and is also helpful to the practice of online learning academic early warning, academic performance prediction and evaluation (Dalton 2011). Eramona F. et al. applied the clustering algorithm to the analysis of English-related course scores and learning information of undergraduate students in online academic education, and realized the subdivision and prediction of adult degree English examination scores (Eramona 2014).

2.3 Discussion

By collating the results of domestic research on the application of data mining technology to English teaching, it is found that many studies have relatively focused on the traditional English teaching model, that is, the use of simple data statistical methods to mine and improve the quality of English teaching. Although the emergence of online education is a great progress in the English teaching model, it only moves the process of data statistical analysis from offline to online, from manual statistics to machine statistics, which has not changed in essence. Although the application of machine statistical analysis is larger and more efficient, it can not flexibly adapt to various situations in English teaching. Therefore, this paper mainly studies intelligent English teaching methods by combining data mining technology and related machine learning technology, so as to provide a basis for improving the quality of English teaching.

3. INTELLIGENT DATA MINING METHOD FOR ENGLISH TEACHING AND CULTURAL ANALYSIS

3.1 Data structure Analysis and Data Pre-Processing in English Teaching

In order to apply intelligent data mining technology to English teaching, it is necessary to analyze the syntactic and grammatical complexity in the process of English teaching. Specifically, syntactic complexity refers to the scope and complexity of different forms of language output, which is mainly used to evaluate language proficiency, describe language ability and measure language development. It is an important index to evaluate learners' language development. At present, the research on syntactic complexity at home and abroad mainly focuses on: (1) horizontal research, that is, by comparing English data representing different English learning levels or writing quality, explore the measurement indicators of syntactic complexity that can effectively distinguish different English learning levels or writing quality; (2) Longitudinal study, that is, by analyzing the multiple writing corpus of the same group of learners at different time points, to explore the development model of syntactic complexity. However, researchers pay less attention to the correlation between the measured syntactic complexity of English learners' writing texts and their writing scores.

In order to solve the bottleneck of data analysis in the study of grammatical and syntactic complexity in English teaching, in order to help researchers carry out a more effective study of English syntactic complexity, five major categories (14 subcategories) are used to analyze the syntactic complexity of written English texts, as shown in Table 1. Specifically, it includes: 1) English sentence unit length (Mean Length of Sentence (MLS), Mean Length of T-unit (MLT), Mean Length of Clause (MLC)); 2) sentence complexity (Clause per Sentence (CS)); 3) subordinate sentence usage (Clause per T-unit (CT), Clause per T-unit Ratio (CTR), Dependent Clause per Clause (DCC), Dependent Clause per T-unit (DCT)). 4) the amount of coordinate structure (Coordinate Phrase per Clause (CPC), Coordinate Phrase per T-unit (CPT), T-unit per Sentence (TS)), and 5) specific phrase structures (Complex Nominal per Clause (CNC), Complex Nominal per T-unit (CNT), Verb Phrase per T-unit (VPT)).

Now, using the web crawlers and other technical means to obtain a large number of English teaching data from open source websites, however, these data need to be pre-processed before they are used. Data pre-processing is the preparation work before data mining, which aims to provide standardized and targeted data for data mining, reduce the data processing capacity of data mining algorithms, improve mining efficiency, and finally improve the accuracy of the model. In this paper, the pre-processing operation is mainly concerned with the standardization and unity of the data format. The complexity of English syntax is different, which may lead to a large gap in the range of corresponding eigenvalues, and the lack of relevant pre-processing will affect the final result. Therefore, in this paper, the complexity data of English text is scaled by a certain proportion, and the maximum-minimum normalization method is used to map the value to [0, 1] interval, that is, given the complexity value *x* of different English syntactics. then the corresponding normalization pre-processing process is defined as follows:

$$x_{new} = \frac{x - x_{\min}}{x_{\max} - x} \tag{1}$$

where $x_{\min}$ is the minimum value; $x_{\max}$ is the maximum value; x_{new} is the new value of *x* after the operation of normalization.

3.2 Model Establishment

Based on the English sentence data obtained by pre-processing, this paper mainly uses random forest to construct English sentence data classifier. For each decision tree in the random forest model, the training set they use is sampled back from the total training set, which means that some samples in the total training set may appear in the training set of a tree many times. or it may never appear in the training set of a tree. When training the nodes of each tree, the features used are randomly extracted from all the features according to a certain proportion. Then, assuming that the overall feature of the English text is *M*, we can set such proportion as $\sqrt{M}$.

The overall training process of this model can be described as follows. First of all, we calculate the overall English text data satisfying the format and denote it by *S*. Then, the testing data set and the feature dimension can be denoted by *T* and *F* respectively. After that, according to the model that need to be trained, we should also decide the training parameters which include the number of decision trees (denoted by *t*) that consist of the random forest, the length of each tree (denoted by *d*), and the number of features of each node of the tree (denoted by *f*). Then, for any decision tree, we define the information entropy for it, as follows:

$$Ent(S) = -\sum_{k=1}^{f} p_k \log_2 p_k \tag{2}$$

where p_k is the rate between the *k*-th type data sample and the overall data set S. Then, the objective is to minimize $Ent(S)$, since the lower the value of $Ent(S)$, the better the training performance.

Then, extracting the training set S(i) from S as the training samples for the root node which is the starting node for the overall training process. However, we should be aware that the training process need to be terminated by certain condition which is defined as follows:

$$Gain(S) = Ent(S) - \sum_{i=1}^{F} \frac{|S^i|}{|S|} Ent(S^i) \tag{3}$$

That is, if the value of the gain value $Gain(S)$ cannot be reduced anymore, the training process for this node will be terminated. After that, the current node will be regarded as the leaf node. Despite this, if the termination condition is not reached, we randomly select the *f* dimensional features from the overall *F* dimensional features without putting them back. Leveraging the *f* dimensional features, we then search for the feature (assume it is indexed by *k*) with the best classification results and its threshold denoted by *threshold*. Then, the samples on the current node with their *k*-th feature smaller than *threshold* will be separated to the left node, while the rest will be separated to the right node. Now, repeating the process until all the nodes are trained or marked as leaf node. Once all the nodes are trained, then the tree is established.

3.3 Prediction Based on Decision Tree

Now, we have well trained the model, based on which we can next use it to prediction the syntax issue of any English text to improve the English teaching quality. Given any English text with fixed structure, we first pick one decision tree from the random forest, and then the searching process starts from the root node of this decision tree. For the current node, if the value of *threshold* of this node is larger than the feature value of this English text, we go into the left node. Otherwise, we go into the right node. Repeating the above searching process until the leaf node is reached. Then, a result of this tree is output, that is, this English text has grammar or syntax issue or not, which can be defined by the following binary variable.

$$z_i = \begin{cases} 0, \text{ this sentence has no grammar issue} \\ 1, \text{ otherwise} \end{cases} \tag{4}$$

where $i > 0$ indicates the number of the decision tree and we should run the above process through each tree. Assuming that there are *N* decision trees in the random forest, then the prediction results can be described as follows:

$$\{z_1, z_2, z_3, \cdots, z_N\} \tag{5}$$

To make it easy-to-be-understood, we generate a probability for English syntax or grammar issue prediction, that is,

$$p = \frac{\sum_{i=1}^{|N|} z_i}{|N|} \tag{6}$$

where $p \in [0,1]$ is the probability that the English text has grammar or syntax issue. Then, if

$$p > 1 - p \tag{7}$$

is true, we decide that this English text has grammar or syntax issue. Otherwise we judge there is no grammar or syntax issue.

The overall working process is shown in Fig. 1, as follows:

4. EXPERIMENTAL RESULTS

4.1 Setup

The key part of We use the crawler technology to obtain tremendous amount of English text data from online websites such as VIPKID and PALFISH. In order to make the model working efficiently, we select 100000 English text data, among which 80% of them are used as the training set and the rest 20% of them are used as the testing set.

To better test the result of the proposed method, we adopt three benchmarks as follows:

- Benchmark1 (Guo 2021): it uses the traditional statistical method to dig the regular from tremendous amount of English text data, such that we refer it as TSM in the comparison.
- Benchmark2 (Lou 2020): it uses the K-nearest neighbor method to explore the relationship between different English texts, such that we refer it as KNM in the comparison.
- Benchmark3 (Xu 2019): it relies on using the supported vector machine method to classify the categories of different English texts, such that we refer it as SVM in the comparison.

For convenience, the proposed method is referred to as Decision Tree based Method (DTM). Then, TSM, KNM, SVM and DTM are evaluated over the following metrics of English syntax statistical result, time overhead and prediction accuracy.

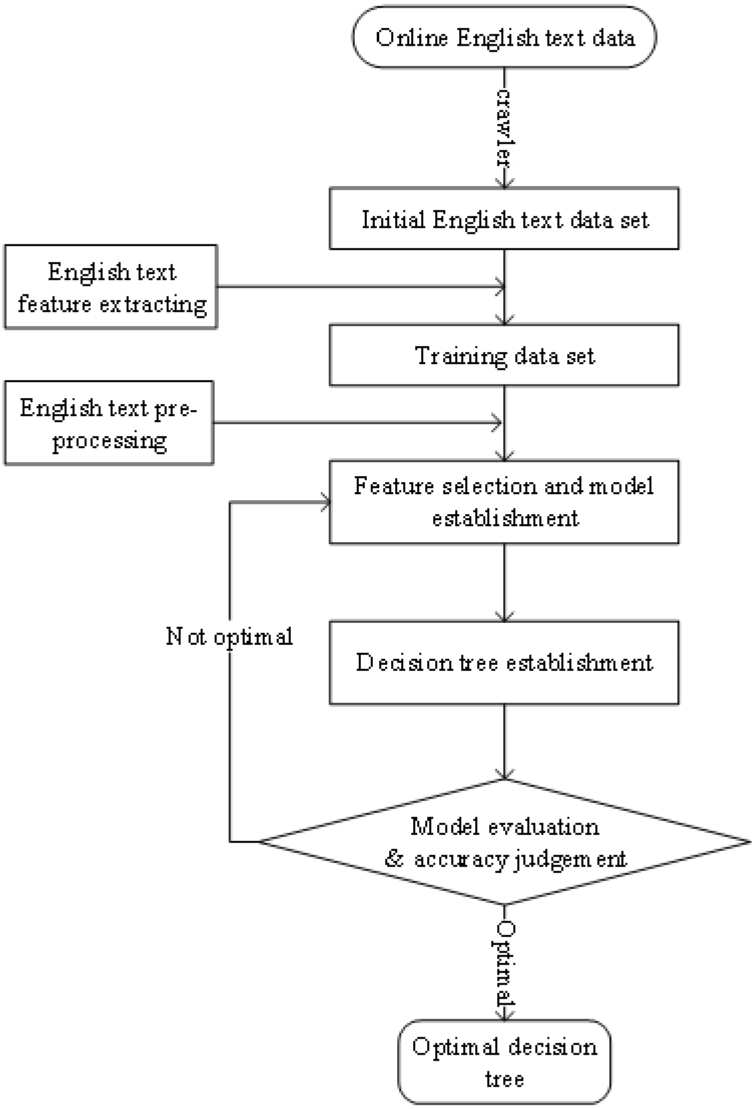
4.2 Results

4.2.1 English Syntax Statistical Result

As explained, we finally obtain 100000 English texts for training the English syntax model. In particular, the quality of such data is vital important for the prediction result. Hence, we first present the statistical results of the data set. Since the English text is classified into 14 kinds, the statistical results are summarized accordingly and shown in Table 2.

In particular, by observing the results, we can conclude that the obtained data can satisfy the requirements of the proposed method. On one hand, the diversity of the data can be guaranteed, since we have collected the data from 14 kinds of English texts and it is easy to note that the number of

Figure 1. The flow chart of the overall working process


Online English text data
crawler
Initial English text data set
English text feature extracting
Training data set
English text pre-processing
Feature selection and model establishment
Not optimal
Decision tree establishment
Model evaluation & accuracy judgement
Optimal
Optimal decision tree

each kind of English text exceeds zero. On the other hand, let us observe the distribution of these English text data, several phenomena can be found, that is: 1) the format of MLS of the English text has the largest mean value, while CPC has the smallest mean value; 2) MLT has the largest Max-value while CTT has the smallest Max-value; 3) the min value of MLS is the largest and that of DCC is the smallest; 4) the standard deviation of MLS is the largest and that of DCC is the smallest. This is reasonable, since MLS structure in English text is the most commonly used structure of the English text, while DCC is the not the general one.

4.2.2 Time Overhead

Now, we have the well prepared data for training the English syntax model. Next, we should start the training model. In this regard, the time overhead becomes a critical factor to evaluate the effectiveness

Table 2. Statistical results of the 100000 English texts

	Mean value	Max value	Min value	Standard deviation
MLS	35.12	391	5.35	39.66
MLT	27.21	392	6.55	33.25
MLC	12.38	115	1.785	9.14
CS	3.5	15	0.8	1.78
VPT	4.84	15	0.98	1.35
CT	2.12	9	0.78	0.95
DCC	1.41	2.441	0.02	0.1333
DCT	1.99	6.78	0.401	0.789
TS	1.07	5.12	0.048	0.31
CTT	0.556	1.1	0.037	0.22
CPT	0.489	4.2	0.02	0.364
CPC	0.255	2.57	0.02	0.178
CNT	3.52	15	0.24	1.398
CNC	1.198	5	0.25	0.385

of the proposed method. This metric is evaluated against the number of English texts and shown in Fig. 2.

Overall, by observing the training time achieved by the four methods, we can discover two obvious phenomena that: 1) DTM has the smallest time overhead and KNM has the second smallest time overhead, that is, given the same amount of training data, DTM spends the least time to finish training the corresponding model; 2)TSM has the highest time overhead, while SVM is the second highest. This is reasonable, because TSM relies on statistical technology which is highly proportional to the scale of the training data. As for DTM, despite the fact that its time overhead also increases with the increasing of the number of English syntax text, it increases slowly, due to the fact that the simple training structure of the decision tree. On the contrary, the training network structure of the supported vector machine is relatively complex for the purpose of achieving high prediction accuracy.

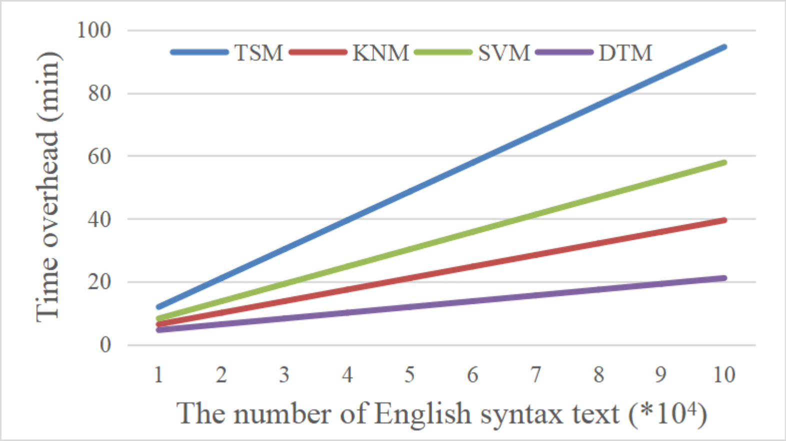
Another deep condition can be seen is that the time overhead of these four methods increases linearly when the number of the English syntax text is increasing, which is reasonable. On one hand, dealing with a lot of data requires time, which supports the situation that the time overhead increases with the increasing of the scale of data. On the other hand, the training model of these methods are fixed, that is, the network structure (i.e., how many nodes and how many links) and the training parameters used for model training are all known, such that the the linear relationship between them is fixed. Therefore, the increasing trend is also linear as shown in Fig. 2.

4.2.3 Prediction Accuracy

As for the prediction accuracy results, we first show the Root Mean Squared Error (RMSE) of the four methods, since RMSE are one critical factor for evaluating the prediction model, and the calculation is shown as follows.

- $RMSE = \sqrt{\frac{1}{|S|}\sum_{i=1}^{|S|}(z^i - \hat{z}^i)^2}$

Figure 2. The results of time overhead against the number of English syntax texts



The corresponding results are presented in Fig. 3. In particular, we can see that DTM achieves the smallest value of RMSE. RMSE indicates the error rate between the actual value and the prediction value. For any prediction model, we want to reduce the difference between the prediction result and the actual result, such that the smaller the value of RMSE, the better. Taking this into consideration, DTM performs the best in terms of the prediction model. SVM achieves the second smallest value of RMSE, which is relatively close to that of DTM. Although they has similar RMSE, their training processes differ a lot.

Based on the above value of RMSE, we now present the results of the prediction accuracy. Apparently, we can see that the results in Fig. 4 show the similar trend as in Fig. 3. That is because the prediction accuracy is highly proportional to the performance of RMSE. On one hand, the method with smaller value of RMSE will achieve higher prediction accuracy. For example, DTM has the smallest value of RMSE (about 0.69 on aveargе) and it achieves the highest prediction accuracy (about 95.4%). On the other hand, any two methods with similar RMSE will achieve similar prediction accuracy. For example, the prediction accuracies of DTM and SVM are close, that is, 95.4% and 93.2%. Moreover, TSM and KNM have lower prediction accuracy, that is, 85.8% and 87.1% respectively.

4.2.1 Result Over Different Platforms

Lastly, the proposed method main relies on using the random forest and the decision tree. Considering the fact that there are a lot of platforms that have already implemented such methods. We then leverage different platforms that has embedded the random forest and decision tree methods, to evaluate the efficiency of DTM on different platforms. In order to comprehensively show the performance, four platforms are selected, that is, the scikit-learn, the Spark MLllib, the DolphinDB, and the XGBoost over the indicators of training speed, memory occuracy and the prediction accuracy. The corresponding results are shown in Table 3. Apparently, although these platforms use the same method DTM, their efficiencies are totally different. For example, XGBoost only demands 0.48GB memory, while Spark MLlib demands 134.6GB memory. On the contrary, XGBoost needs a long training time, that is, 103.8 minutes, while Spark MLlib only needs 20.1 minutes. Nevertheless, their prediction accuracy are almost the same and this is reasonable, because the algorithm essence of these platform are the same.

Figure 3. The results of RMSE

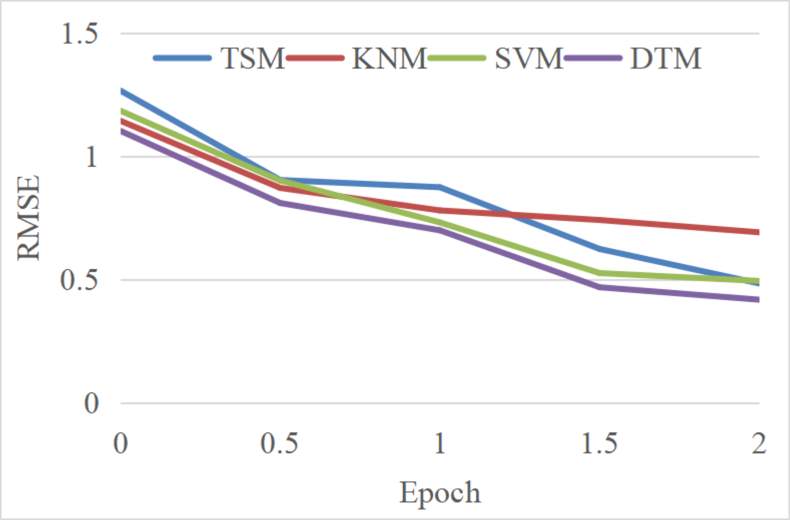
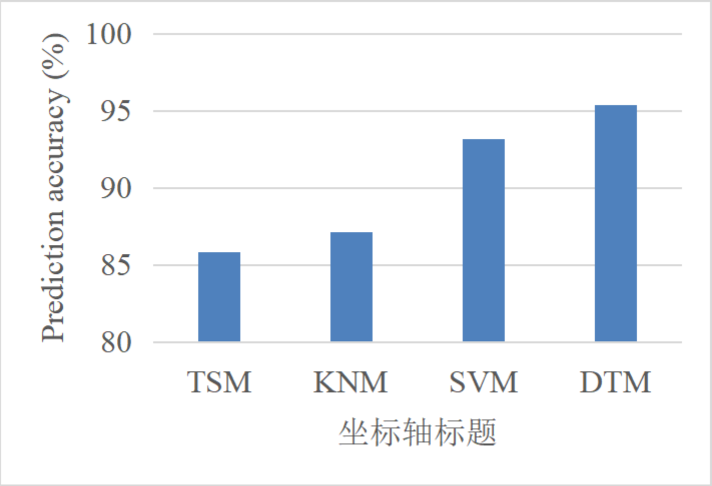


Figure 4. The results of prediction accuracy



5. CONCLUSION

In this work, we study the intelligent English teaching method to further improve the quality of English teaching. Specifically, it consists of several parts. The first part is to implement the data pre-processing, which makes sure that the English syntax texts are in the same format. The second part uses the random forest to establish a training model. The last part relies on using the decision tree to make a prediction about the English text in terms of its grammar or syntax issues. The evaluation results indicate that the proposed method can effectively improve the accuracy of English grammar or syntax recognition and outperforms the other benchmarks greatly.

Table 3. Results of DTM over different platforms

Platform	English syntax text data	Time (min)	Memory (GB)	Accuracy
scikit-learn	100000	78.3	1.1	0.949
Spark MLllib	100000	20.1	134.6	0.956
DolphinDB	100000	56.9	3.3	0.948
XGBoost	100000	103.8	0.48	0.952

ACKNOWLEDGMENT

This research is supported by research and Practice of “online and Offline” mixed teaching based on the cultivation of students’ innovation ability under “One Belt and One Road” -- Taking core Business English courses as an example(1070XJ0520153).

REFERENCES

Albiladi, W. S., & Alshareef, K. K. (2019). Blended learning in English teaching and learning: A review of the current literature. *Journal of Language Teaching and Research*, *10*(2), 232–238. doi:10.17507/jltr.1002.03

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief.* Office of Educational Technology, US Department of Education.

Cui, S. (2017). Research on across-cultural communication in college English teaching based on cloud platform. *Journal of Computational and Theoretical Nanoscience*, *14*(1), 89–93. doi:10.1166/jctn.2017.6130

Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles? *Annual Review of Applied Linguistics*, *31*, 182–204. doi:10.1017/S0267190511000092

Dogan, A., & Birant, D. (2020). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, 114060.

Elyas, T., & Picard, M. (2018). A brief history of English and English teaching in Saudi Arabia. In *English as a foreign language in Saudi Arabia* (pp. 70–84). Routledge. doi:10.4324/9781315688466-3

Eramona, F., & Al-Hafizh, M. (2014). Using The Clustering Technique in Teaching Writing a Descriptive Text to Junior High School Students. *Journal of English Language Teaching*, *2*(2), 73–81.

Guan, N., Song, J., & Li, D. (2018). On the advantages of computer multimedia-aided English teaching. *Procedia Computer Science*, *131*, 727–732. doi:10.1016/j.procs.2018.04.317

Guo, X. (2021, February). Research on the application of data mining in the analysis of college English teaching quality. *Journal of Physics: Conference Series*, *1744*(4), 042024. doi:10.1088/1742-6596/1744/4/042024

Hudson, R., & Walmsley, J. (2005). The English patient: English grammar and teaching in the twentieth century. *Journal of Linguistics*, *41*(3), 593–622. doi:10.1017/S0022226705003464

Kumar, T. S. (2020). Data mining based marketing decision support system using hybrid machine learning algorithm. *Journal of Artificial Intelligence*, *2*(03), 185–193.

Liu, H., Chen, R., Cao, S., & Lv, H. (2021). Evaluation of College English Teaching Quality Based on Grey Clustering Analysis. *International Journal of Emerging Technologies in Learning*, *16*(2), 173–187. doi:10.3991/ijet.v16i02.19727

Lou, H. (2020). Design of college english process evaluation system based on data mining technology and Internet of Things. *International Journal of Data Warehousing and Mining*, *16*(2), 18–33. doi:10.4018/IJDWM.2020040102

Mahbub, M. (2021). *English teaching in vocational high school: A need analysis*. Academic Press.

Mihaescu, M. C., & Popescu, P. S. (2021). Review on publicly available datasets for educational data mining. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *11*(3), e1403. doi:10.1002/widm.1403

Mirbabayeva, S. (2020). Innovative Approaches to English Teaching of Pre-School and Primary School Learners. *Academic Research in Educational Sciences,* (3).

Pangaribuan, T., & Manik, S. (2018). The Effect of Buzz Group Technique and Clustering Technique in Teaching Writing at the First Class of SMA HKBP I Tarutung. *English Language Teaching*, *11*(1), 164–178. doi:10.5539/elt.v11n1p164

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *10*(3), e1355. doi:10.1002/widm.1355

Taleb, I., Dssouli, R., & Serhani, M. A. (2015, June). Big data pre-processing: A quality framework. In *2015 IEEE international congress on big data* (pp. 191-198). IEEE.

Wen, H. (2020). Intelligent English translation mobile platform and recognition system based on support vector machine. *Journal of Intelligent & Fuzzy Systems*, *38*(6), 7095–7106. doi:10.3233/JIFS-179788

Weng, S. S., Liu, Y., Dai, J., & Chuang, Y. C. (2020). A novel improvement strategy of competency for education for sustainable development (ESD) of university teachers based on data mining. *Sustainability*, *12*(7), 2679. doi:10.3390/su12072679

Xu, Z., Qiu, J., Yang, B., Huang, P., Cai, L., Chen, L., Hou, M., Ji, M., & Wu, G. (2019). Evaluation of factors influencing the guide to read biomedical English literature course for Chinese new medical postgraduates—A multiple regression analysis. *BMC Medical Education*, *19*(1), 1–7. doi:10.1186/s12909-019-1731-7

Zhang, W. (2020). (Preprint). Research on English score analysis system based on improved decision tree algorithm and fuzzy set. *Journal of Intelligent & Fuzzy Systems*, 1–13.

Zhen, C. (2021). Using Big Data Fuzzy K-Means Clustering and Information Fusion Algorithm in English Teaching Ability Evaluation. *Complexity*, *2021*, 1–9. doi:10.1155/2021/5554444