Research Article

Machine Learning-Based Intelligent Scoring System for English Essays under the Background of Modern Information Technology

Shaoyun Fu and Hongfu Chen

Minjiang University, Fuzhou, Fujian, China

Correspondence should be addressed to Shaoyun Fu; fushaoyun@mju.edu.cn

Received 19 January 2022; Revised 17 February 2022; Accepted 22 February 2022; Published 23 May 2022

Academic Editor: Arpit Bhardwaj

Copyright © 2022 Shaoyun Fu and Hongfu Chen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This work is to reduce the workload of teachers in English teaching and improve the writing level of students, so as to provide a way for students to practice English composition scoring independently and satisfy the needs of college teachers and students for intelligent English composition scoring and intelligently generated comments. In this work, it firstly clarifies the teaching requirements of college English classrooms and expounds the principles and advantages of machine learning technology. Secondly, a three-layer neural network model (NNM) is constructed by using the multilayer perceptron (MLP), combined with the latent Dirichlet allocation (LDA) algorithm. Furthermore, three semantic representation vector technologies, including word vector, paragraph vector, and full-text vector feature, are used to represent the full-text vocabulary of English composition. Then, a model based on the K-nearest neighbors (kNN) algorithm is proposed to generate English composition evaluation, and a final score based on the extreme gradient boosting (XGBoost) model is proposed. Finally, a model dataset is constructed using 800 college students’ English essays for the CET-4 mock test, and the model is tested. The research results show that the semantic representation vector technology proposed can more effectively extract the lexical semantic features of English compositions. The XGBoost model and the kNN algorithm model are used to score and evaluate English compositions, which improves the accuracy of the scores. This makes the management of the entire scoring model more efficient and more accurate. It means that the model proposed is better than the traditional model in terms of evaluation accuracy. This work provides a new direction for the application of artificial intelligence technology in English teaching under the background of modern information technology.

1. Introduction

Nowadays, the Internet industry is developing rapidly, and the continuous improvement of modern information technology drives the progress of human civilization. As an important information carrier, English is the most widely used language in the world. Accordingly, learning English has become a necessary requirement, and English proficiency certificate has become a must for every college student [1].

Modern information technology is a comprehensive computer and telecommunication technology based on modern microelectronic technology, which can collect, process, store, disseminate, and use various sensor signals such as sound, image, text, and numbers [2], and its core is informatics. Modern information technology includes enterprise resource planning (ERP), Global Positioning System (GPS), and radio frequency identification (RFID), so it can learn from RFID knowledge and application of the above aspects. Modern information technology is a technology group with a wide range of contents, including microelectronic technology, optoelectronic technology, communication technology, network technology, sensing technology, control technology, and display technology. Machine learning is a multidisciplinary interdisciplinary major, covering knowledge of probability theory, statistics, approximate theory, and complex algorithms. A technology that uses computers as a tool and is committed to real-time simulation of human learning methods and divides existing content into knowledge structures to effectively improve learning efficiency [3]. Machine learning uses algorithms and models to incrementally improve the ability to accomplish specific tasks. Machine learning builds
2. Construction of the Scoring System for English Essays

2.1. English Teaching in Colleges and Universities. English is a basic compulsory course for college students in China. It mainly teaches English language and application skills and includes multimedia teaching and listening and speaking teaching. Currently, there are three levels of minimum requirement, general requirement, and high requirement on teaching, and the minimum requirement needs 340 hours. Generally, the 340 hours are distributed in four semesters, namely, 90 hours per semester and 5 hours per week on average, including 3 hours for comprehensive understanding and 2 hours for listening and speaking. Nowadays, colleges and universities do not put much attention on English writing, which needs some changes [6].

2.2. Machine Learning. Machine learning focuses on how computers simulate the behavior of humankind and animals. From a statistical standpoint, it establishes a model based on historical data and then uses the model to predict the distribution of data. This requires that test data and training data must be identically distributed. Its basic feature is that it tries to mimic the patterns in which neurons in the brain transmit and process information. Representatively, it is used in computer vision and natural language processing [7]. Obviously, deep learning is strongly related to neural network in machine learning, or we can call deep learning improved neural network algorithms. The main idea of deep learning is to simulate human neurons, and after each neuron receives information and processes it, it can be passed to all adjacent neurons [8, 9].

Artificial neuron is a mathematical model created by imitating the biological neuron. The artificial neuron receives the given signal of the previous neuron, and each given signal will attach a weight. Under the combined action of all weights, this neuron will show a corresponding activation state [10–12], expressed as follows:

\[ f(x) = \sum_{i=1}^{n} x_i w_i, \]  

where \( f(x) \) is the final output, \( x_i \) is the input signal, and \( w_i \) represents the weight corresponding to the input signal, with \( n \) groups in total.

When a neuron receives the input signal, it will give an output. Each neuron has a corresponding threshold. If the sum of the input is greater than the threshold, it enters the activated state. Otherwise, it is in the inhibited state. Artificial neuron has several transfer functions expressed as follows [13, 14]:

(a) Linear function:

\[ f(x) = kx. \]

(b) Slope function:

\[ f(x) = a(x \geq \theta), \]
\[ f(x) = kx (-\theta < x < \theta), \]
\[ f(x) = -a(x \leq \theta). \]

(c) Transition function:

\[ f(x) = a(x \geq \theta), \]
\[ f(x) = \beta(x \leq \theta). \]

The transfer function should be selected according to the specific application range. Linear function can amplify the output signal, nonlinear slope function can prevent the impact of degraded network performance, and S-type function is used for the hidden layer.

2.2.1. MLP (Multilayer Perceptron) Model. The multilayer perceptron model is to obtain the vector of the modeling target, and the basic structure of the multilayer perceptron MLP can be obtained based on the biological neuron model. However, the learning algorithm of the perceptron cannot be directly applied to the parameter learning of the multilayer perceptron model. Superior. Therefore, the original proposed learning scheme is described as follows. Except for the last
neuron, the weights of all other neurons are fixed in advance, and the learning process is just to use the perceptron learning algorithm to learn the weight coefficients of the last neuron. In fact, this is equivalent to transforming the original feature space into a new feature space through the first layer of neurons. Each neuron in the first layer constitutes one dimension of the new space, and then the perceptron learning algorithm is adopted to construct a linear classifier in the new feature space. A neural network training network takes a feature vector as input and passes this vector to the hidden layer. The result is then calculated through the weights and excitation functions and passed to the next layer until it is finally passed to the output layer. The weights, synapses, and neurons of each layer are computed and learned by training the ANN algorithm. Layers of MLP are fully connected, that is, any neuron in the upper layer is connected with all neurons in the next layer [15, 16], and Figure 1 shows its structure.

It is noted that the neural network has three basic elements: weight, bias, and weight of activation function. The strength of connections between neurons is represented by weights, the size of which represents the bias of possibility. Bias is set to correctly classify samples and is an important parameter in the model, that is, to ensure that the output value cannot be activated randomly. Activation function: it functions as a nonlinear mapping, which can limit the output of neurons within a certain range, generally between (−1, 1) and (0, 1). The most commonly used activation function is the sigmoid function, which can map the number of (−∞, +∞) to the range of (0, 1) [17].

2.2.3. Representation of the Word Vector. Word vector is the basic structure of text. A good word vector can make semantically similar words gather together, which facilitates the subsequent text classification, text clustering, and other operations. Here, the word2vec model is used. The core of the model is the three-layer NNM, which predicts the words that will appear in the context according to the current words. Figure 3 is the schematic diagram of the word2vec model [21].

\[ p(w|w_1) = \frac{\text{exp}(v_w^T v_{w_1})}{\sum_{w'} \text{exp}(v_w^T v_{w_1})} \]

\[ \sigma(x) = \frac{1}{1 + \text{exp}(-x)} \]

where \( v_w \) and \( v_{w_1} \) respectively, represent the input and output vectors of the word \( w \), \( W \) represents the total words.
of the context within the search scope, and $n(w, j)$ represents the path from $j$ to the root node and it is calculated as follows.

$$v_\omega' = b + Wv_\omega + U \tanh (d + Hv_\omega), n(w, j) = \text{root}, n(w, L(w)) = w,$$

where $\tanh$ is the activation function of the neuron, and $b, d, W, U, \text{and } H$ are the parameters to be solved. SGD (stochastic gradient descent) [22] and GA (genetic algorithm) [23] are used to seek optimal solutions. The part of speech vector will convert the word sequence $w_i$ into part of speech sequence $\psi_i$. Word feature vector mainly investigates the rationality of word collocation.

GD is a simple but very effective method, which is mostly used for the learning of linear classifiers under convex loss functions such as support vector machines (SVM) and logistic regression. SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. GA originates from computer simulation studies of biological systems. It is a random global search and optimization method developed by imitating the biological evolution mechanism in nature, drawing on Darwin’s theory of evolution and Mendel’s theory of genetics. Its essence is an efficient and parallel global search method, which can automatically acquire and accumulate knowledge about the search space during the search process and adaptively control the search process to obtain the best solution.
2.4. Representation of the Paragraph Vector. Paragraph representation conforms to the calculation principle of the paragraph2vec, and MLP model is used to obtain the vector of modeling target. Paragraph ID is added to paragraph2vec, making all sentences or paragraphs have their unique identity. Figure 4 shows the network model principle [24].

2.5. Representation of the Full Text. LDA is a generative Bayesian probability model, including three parts for words, topics, and documents. Figure 5 shows the calculation process of its parameters.

The parameters in Figure 5 follow the relationship expressed as follows, where $p$ is a function, and $w$, $\alpha$, $\beta$, and $\theta$ are all parameters to be solved [25].

\[
p(w | \alpha, \beta) = \int p(w_n, \theta, \beta) | d\theta,
\]

\[
= \int p(\theta | \alpha) \left( \prod_{n=1}^{N} p(z_n, \beta) p(w_n | z_n, \beta) \right) d\theta
\]

\[
= \frac{\Gamma \left( \sum_{i=1}^{k} \alpha_i \right)}{\prod_{i=1}^{k} \Gamma (\alpha_i)} \int \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \left( \prod_{n=1}^{N} \prod_{i=1}^{V} \prod_{j=1}^{y_i} \left( \theta_i \beta_{ij} \right)^{w_{ij}} \right) d\theta
\]

\[
p(\theta | \alpha) = \frac{\Gamma \left( \sum_{i=1}^{k} \alpha_i \right)}{\prod_{i=1}^{k} \Gamma (\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1}.
\]

We use the LDA model to represent the full text. Table 1 shows the production process of an article and the calculation steps of LDA model [26].

2.6. XGBoost Algorithm Model. XGBoost is an optimized distributed gradient boosting library designed to be efficient, flexible, and portable. XGBoost provides parallel tree boosting to solve many data science issues quickly and accurately. In terms of large-scale data in the industry, the distributed version of XGBoost has extensive portability and supports running in various distributed environments, making it a good solution to the problem of large-scale data in the industry. Equations (8) and (9) are the objective functions of each parameter.

The objective functions of parameters are expressed as follows:

\[
y_i' = \sum_{k=1}^{K} f_k (x_i), f_k \in F,
\]

\[
L = \sum_{i} l(y_i - y_i')^2 + \sum_{k} \left( yT + \frac{1}{2} \lambda w^2 \right),
\]

$f_k (x_i)$ contains two items: to calculate the influence of the number of nodes $T$ of the parameter $\gamma$ on the error and to calculate the influence of node weight $w$ of the parameter $\lambda$ on error. The regularization method is used to avoid overfitting due to too many nodes. $y_i'$ represents the output value of the composition model, $y_i$ represents the standard
value of the composition model, \( f_k(x_i) \) is the tree function, \( x_i \) is the input value of the function, and \( L \) is the objective function of each parameter [27].

### 2.7 kNN Algorithm Model

The kNN classification algorithm is one of the simplest algorithms in data mining classification technology, and its guiding idea is to infer its category from its neighbors. The realization principle of the kNN classification algorithm is explained as follows. In order to judge the category of the unknown samples, the distances between the unknown samples and all the known samples are calculated with the samples of all known categories as the reference, and the \( K \) known samples with the closest distance to the unknown samples are selected. According to the voting rule that the minority obeys the majority, the unknown samples are classified into one category. The kNN classification algorithm only determines the category of the sample to be classified according to the category of the nearest one or several samples, rather than the method of discriminating the category domain to determine the category to which it belongs, so the kNN method is more suitable than other methods for the sample set to be divided with more overlapping or overlapping class domains.

TF-IDF (term frequency—inverse document frequency) is a statistical method used to evaluate the importance of words in articles or in context. Table 2 shows the general idea to score essays [28].

The importance of words is represented by the number of digits but is the opposite of the frequency in the corpus. Search engines often use TF-IDF weight to measure the level of relevance between files and user questions. In addition to TF-IDF, search engines also use ranking based on link analysis to determine the order in which files appear in search results.

The TR (TextRank) algorithm is a graph-based algorithm to extract keywords and rank abstracts. It is improved from Google’s PR (PageRank) algorithm. It can extract keywords using the co-occurrence information (semantic) between words. It can extract key words and key phrases from a given text and extract key sentences from the text.

### 2.8 Test Results of the Scoring System

800 essays were randomly selected from the homework of students in colleges and universities, with a vocabulary of 160–210 words. Two teachers scored each essay separately to obtain the average score. Then, the number of essays in each score segment was counted. 800 essays were divided into 5 groups (160 for each), of which 4 groups were randomly selected as the training set and the remaining 1 group was used as the test set. Five rounds of cross-validation were performed on training and testing, and the evaluation index was recorded each time. Finally, the average of these scores was taken as the final score.

The intelligent scoring model constructed in this work is compared with the traditional gradient boosting decision tree (GBDT) method and the effect of the SVM method. The GBDT method trains the base classifiers in a serial manner, and there are dependencies between the base classifiers. Its basic idea is to stack the base classifiers layer by layer, and it will give higher weights to the samples misclassified by the previous layer of the base classifier when each layer is training. During testing, the final result is obtained according to the weighting of the results of each layer of classifiers. The principle of GBDT is very simple. The sum of the results of all weak classifiers is equal to the predicted value, and then the next weak classifier is to fit the residual of the error function to the predicted value (this residual is the error between the predicted value and the true value). SVM is a binary classification model that maps the feature vector of an instance to some points in the space. The purpose of the SVM is to draw a line that “best” distinguishes the two types of points, so that if new points come later, and the line can also make a good classification. SVM is suitable for small and medium data samples, nonlinear, high-dimensional classification. The classical SVM algorithm only provides two-class classification algorithms, but it is generally necessary to solve

<table>
<thead>
<tr>
<th>Order</th>
<th>Step</th>
<th>Calculation process</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>To determine the distribution of topics and words</td>
<td>LDA is used to calculate the polynomial distribution of feature words and describe the distribution with parameters</td>
</tr>
<tr>
<td>2</td>
<td>To determine the distribution of articles and topics</td>
<td>According to PD (Poisson distribution), the scale of feature words is calculated</td>
</tr>
<tr>
<td>3</td>
<td>To randomly determine the number ( N ) of words in the article</td>
<td>LDA is used to calculate the probability vector of topic distribution</td>
</tr>
<tr>
<td>4</td>
<td>If the number of currently generated terms is less than ( N ), go to step 5; otherwise, step 6 is performed</td>
<td>From the ( m )th essay ((m = 1, 2, \ldots, M)); and ( M ) is the total number of essays, a feature word ( w ) of a topic is extracted. Then, the expectation maximization method is used to estimate the maximum likelihood of parameters, so as to establish the LDA three-layer model.</td>
</tr>
<tr>
<td>5</td>
<td>A topic is generated randomly according to article and topic distribution, and then a word is generated randomly based on topic and word distribution. Next, proceed to step 4.</td>
<td>A topic is generated randomly according to article and topic distribution, and then a word is generated randomly based on topic and word distribution. Next, proceed to step 4.</td>
</tr>
<tr>
<td>6</td>
<td>The article generation is finished</td>
<td>The article generation is finished</td>
</tr>
</tbody>
</table>
multiclass classification in the practical application of data mining. It can be solved by a combination of multiple second-class support vector machines. There are mainly one-to-many combination mode, one-to-one combination mode, and SVM decision tree; then, it is solved by constructing a combination of multiple classifiers. The main principle is to overcome the inherent shortcomings of SVM and combine the advantages of other algorithms to solve the classification accuracy of multiclass problems.

The precision, recall, and F1-score are compared. The F1-score is the harmonic mean of precision and recall, and its calculation formula is

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{9}$$

Mean square error (MSE) is a method to measure the difference between two estimators, as follows. $E$ is the expectation, and $D$ is the variance.

$$\text{MSE}(\theta') = E(\theta' - \theta)^2,$$

$$\text{MSE}(\theta') = E[(\theta' - E(\theta)) + [E(\theta') - \theta])^2, \tag{10}$$

$$\text{MSE}(\theta') = D(\theta') + [E(\theta') - \theta]^2.$$

The Pearson correlation coefficient (PCC) is used to represent the linear correlation between two variables: $X$ and $Y$. In the following equation, $\text{cov}$ represents the covariance, $\sigma$ represents the mean, and $\mu$ represents the standard deviation.

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}. \tag{11}$$
3. Intelligent Scoring System for English Essays

Testing and Analysis

3.1. The Categories and Scores of English Essays in the Experiment. Figure 6 shows the number of essays belonging to various categories, and Figure 7 shows the number of essays in each score range.

Figure 6 shows the five types of online shopping, online learning, the importance of investment, university part-time jobs, and career choice. The number in each category is 172, 147, 170, 158, and 153, respectively. Overall, there numbers do not change largely. Figure 7 shows the number of essays in all score ranges. There are 43 essays in [0, 60], 188 in [60, 70], 275 in [70, 80], 242 in [80, 90], and 52 in [90, 100]. It suggests that there is no significant difference in the choice of essay topics by students, and the proportion of students with scores in the [70, 80] and [80, 90] is close to 250, of which the people in the [70, 80] range most, satisfying normal distribution.

3.2. Performance of Different Scoring Models. Figure 8 shows the MSE (mean square error) and PCC (Pearson correlation coefficient) results of various essay scoring methods.

As shown in Figure 8, the method adopted in this work shows the smallest mean square error, which is only 10.39. After the traditional vector machine SVM method is added to the method proposed, the MSE rises to 18.09; after the GBDT method is added, the MSE rises to 12.80. The smaller the mean square error of the scoring method, the better. The MSE of the traditional scoring method is much higher than that of the scoring method proposed. The proposed method shows the highest PCC, reaching 0.9546; after the method proposed is combined with the traditional method, the PCC decreases. The larger the PCC, the better the performance of the intelligent scoring model, indicating that the method proposed has the highest correlation with the smallest error in the manual teacher scoring results, so the performance of the method is the best. Table 3 shows the results of precision, recall, and F1-score.

According to Table 3, the proposed method combines the TF-IDF method and TR method to obtain the average score of each essay. It has the best accuracy, recall rate, and F1-score. Its F1-score of 80% is much higher than the other two methods, and its average accuracy is over 80%, while that of the other two methods is less than 70%. Among them, the precision, recall rate, and F1-score of the TR + kNN method are below 70%, while the F1-score value of the TF-IDF + kNN method is 65%, the recall rate is 63%, and the precision is 68%, which are lower than the other two methods. Therefore, the effect of the algorithm model proposed is obviously better than that of the mainstream algorithms.

4. Conclusion

In this work, it establishes an intelligent scoring system for English composition under the background of modern information technology. The NNM model is constructed by using MLP technology combined with the LDA algorithm, and three semantic representation vector technologies are used to represent the full-text vocabulary of English composition. Then, the English composition evaluation model based on the kNN algorithm and the final composition scoring model based on XGBoost are proposed. Finally, 800 English compositions are randomly selected from the four-level examination training compositions in colleges and universities, and the established English composition data set is used for training and testing. Finally, the comparative
test results between the method used in this work and the traditional method are obtained. The final practical application results reveal that the results of the English composition intelligent scoring system based on machine learning proposed show the smallest mean square error and the highest accuracy and the highest F-score score compared with the artificial teacher’s scoring results. Therefore, the model constructed shows the best performance.

Although it has basically achieved the original expected research goals, the research work still has the following shortcomings due to my limited academic quality. Firstly, it only analyzes the composition for the CET-4 test in colleges and universities. In the future, it will collect the composition data of primary and secondary schools and build an intelligent English composition scoring system suitable for various teaching tasks. Secondly, only a few college students’ English compositions are collected. In the future, the scale of the data set will be further expanded, data collection will be carried out on the English compositions of college students of different levels, and the number of training sessions will be increased, so as to obtain better training effects.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare no conflicts of interest.

References


