Hindawi
Mobile Information Systems
Volume 2022, Article ID 6306025, 9 pages
https://doi.org/10.1155/2022/6306025

Research Article

Analysis of English Writing Text Features Based on Random Forest and Logistic Regression Classification Algorithm

Chuan Sun[1] **and Bo Luo**[2]

[1]*School of Fundamental Education, Beijing Polytechnic College, Beijing 100042, China*
[2]*School of Electrical and Information Engineering, Beijing Polytechnic College, Beijing 100042, China*

Correspondence should be addressed to Bo Luo; luobo@bgy.edu.cn

Received 9 February 2022; Revised 20 February 2022; Accepted 2 March 2022; Published 19 April 2022

Academic Editor: Chia-Huei Wu

Copyright © 2022 Chuan Sun and Bo Luo. 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.

The characteristics of English writing text in natural scenes are characterized by low character detection rate, difficulty in small character detection, and various character detection categories. In order to improve the classification effect of English-written texts, solve the problem of feature loss and precision reduction of the prediction model based on dimensionality reduction neural network classifier in the analysis process. An improved stacking model combining random forest and logistic regression is proposed to analyze the characteristics of written English texts. The model uses multiple undersampling, trains multiple random forests as primary classifier, and uses logistic regression as secondary classifier. Experimental results show that this model can effectively improve the classification efficiency of unbalanced text classification. While ensuring the main features, the accuracy of prediction is substantially improved. It is proved that the model has high practicability in analyzing the features of English writing texts.

1. Introduction

As the symbol of human civilization and the carrier of information communication, characters exist widely in natural scene images. Compared with other natural scene content in the image, the text in the scene is more logical and general. It can precisely provide high-level semantic information, which is helpful for the analysis and understanding of scene content. The rapid growth of artificial intelligence and deep learning technology provides a new way for the study of end-to-end character detection in natural scenes [1–3].

English text writing is a writing form that uses English language as a tool to express thoughts and emotions. English text writing has basic text types such as description, narration, explanation, and discussion, as well as practical text types such as poetry, prose, and story. In fact, underneath these various text forms, there are many text features. Through the digital construction and analysis of these text features, we can have a clear and rational understanding of English writing which helps to improve our English text writing ability [4, 5].

For text feature detection in natural scenes, domestic and foreign scholars have carried out relevant studies and achieved certain results [6–9]. The core idea of natural scene text feature detection based on traditional methods is to detect artificially extracted features, such as colour levels and regions. Literature [10] proposed a detection algorithm based on the maximum stable extremum region, but this method could not detect the white text region on a black background. Literature [11] proposed a detection algorithm based on Adaboost, but this method has poor robustness in processing low-contrast images. Literature [12] uses FAST corner point extraction, but it is easily affected by lighting changes and shooting angle. The above traditional natural scene text feature detection methods require manual design features; it would extract many poor features and lead to the loss of recall rate and detection accuracy.

With the continuous development of deep learning technology, convolutional neural network (CNN) has made a series of achievements in image classification, object detection, and other fields. Literature [13] proposed a layered detection strategy for text features. Firstly, CNN is used to extract

features, and then, random forest algorithm is used to finely classify text feature candidate regions. However, the subsequent processing of this method is usually complicated and cannot achieve the effect of real-time detection. In order to further improve the speed of target detection, literature [14] proposed an end-to-end target detection algorithm, YOLO (You Only Look Once). The algorithm transforms the target detection problem into a regression problem, to better distinguish the target from the background. However, the size of the input image is fixed, which makes it unable to adapt to the shapes of different objects in the training process. The method in Reference [15] adds orientation information, which enables the SSD (single-shot detector) detector to deal with the feature detection problem of text arranged in any direction, but it is not ideal for text with large spacing. The YOLOv2 algorithm proposed in literature [16] uses Darknet-19 network as the feature extraction network, which greatly simplifies the network structure. At the same time, the accuracy of target detection is improved, but its detection effect on text features in natural scenes is poor, and it is easy to mistakenly identify the background as text features.

The key to the construction and analysis of the digital model of English writing text features lies in its text classification. Text classification is to obtain information from text, analyze and process the information, and dig out more important knowledge. Text classification is divided into feature engineering and classifier. Feature engineering is the process of turning data into information, which is the most time-consuming and labour-intensive, but quite important process [17]. DF (word frequency), CHI (chi square test), IG (information gain), and ECE (expected cross-entropy) are often used as the basis for feature selection [18]. Literature [19] uses TF-IDF to classify text word segmentation backward quantization as a feature of text. Literature [20] improved feature selection algorithm IG and combined it with DF to extract more important features for texting classification and improving the accuracy of text classification. Classifiers turn information into knowledge, the desired result. The classifier algorithm used for text classification is constantly updated, which makes the prediction effect of text classification better and better. Literature [21] uses naive Bayesian as a classifier to classify short texts, which had achieved good results. Literature [22] uses support vector machine to classify short texts and proves its effectiveness.

In the process of feature analysis of English-written texts, in order to improve the classification effect of English-written texts, solve the problem of feature loss and precision decline of the prediction model of dimensionally reduced neural network classifier and adapt to high-dimensional unbalanced characteristics of data. In this paper, an improved stacking model combining random forest and logistic regression was proposed by using TF-IDF feature extraction method. It is used to analyze the characteristics of English writing texts. The experimental results show that the effect of text classification in English writing has been improved.

The main innovations of this paper are as follows:

(1) The processing effect of high-dimensional unbalanced features of data is better

(2) The problem of feature loss and precision decline of the prediction model of dimensionally reduced neural network classifier is solved

(3) Extracting features of English writing text by TF-IDF feature extraction method

This paper consists of four main parts, namely, the introduction in the first section, the related work in the second section, the experiment and analysis in the third section, and the conclusion in the fourth section, with the abstract and reference sections.

2. Related Work

2.1. TF-IDF (Word Frequency-Inverse Document Frequency). TF-IDF is a statistical method. Its calculation formula is TF (word frequency) × IDF (inverse document frequency). It means that the more frequently a word appears in a certain text and the less frequently it appears in all texts, the higher the TFIDF weight of the word is, the more it can represent the text [23].

(1) TF (word frequency) refers to the frequency with which a certain word appears in all texts

$$\mathrm{TF} = \frac{\text{The number of occurrences of } x \text{ in a category}}{\text{The number of all words in that category}}. \quad (1)$$

(2) IDF (inverse document frequency) is the reciprocal of document frequency, indicating that words frequently appearing in each text will have less influence on all texts [24]

$$\mathrm{IDF} = \log\left(\frac{\text{Total number of documents in the corpus}}{\text{Number of documents with word changes} + 1}\right). \quad (2)$$

2.2. Feature Dimension Reduction Algorithm. Feature dimension reduction algorithms are usually divided into two categories: unsupervised algorithms represented by PCA and supervised algorithms represented by linear discriminant analysis (LDA). PCA algorithm is usually applied to prediction models, and its main process includes dataset preprocessing, PCA principal component extraction, and prediction model establishment and prediction [25]. The basic principle of PCA algorithm is to construct a k-order matrix (K represents the number of features after sample dimension reduction) by training the feature vector of covariance matrix of dataset. The k-order matrix is a real diagonal matrix, and the eigenvectors corresponding to different eigenvalues are orthogonal. The k-order matrix is the final result of the algorithm. The PCA algorithm process is as follows: Suppose there is a sample set $X = \{x_1, x_2, \cdots, X_n\}$, the covariance matrix $1/nXX^T$ of the dataset was calculated by subtracting the average value of each feature, and

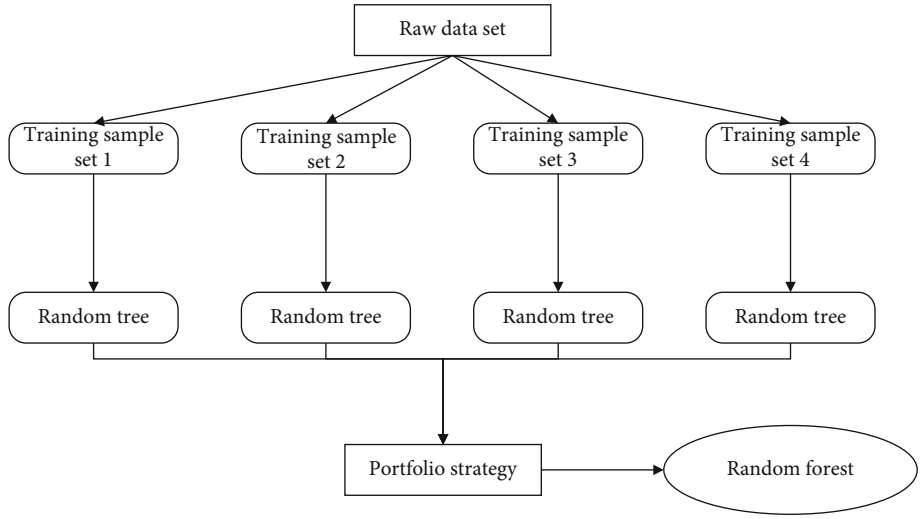

Figure 1: Schematic diagram of random forest.

the eigenvalue decomposition method was used to calculate the eigenvalue and eigenvector of the covariance matrix. Then, the eigenvalues are sorted and top K is selected, and its eigenvectors are, respectively, used as row vectors to form the eigenvector matrix P. Finally, the original dataset is mapped to the space constructed by the new feature vector; that is, $Y = PX$.

2.3. Random Forests. Random forest is an extension of the bagging integration algorithm. Bagging ensemble is constructed by a decision tree-based classifier, and random attribute selection is introduced in the process of ensemble. That is, each attribute is selected to be added into the training process to ensure the diversity of the base learner and improve the final generalization performance of the model [26].

The final decision result of random forest is obtained by the combination of all the decision tree classification results of base classifier. See Figure 1. For the classification problem, the voting method is used to decide, and the classification results of each decision tree are statistically voted, and the minority is subject to the majority. For regression problems, the mean value of decision tree classification results is taken as the result of random forest.

The advantages of random forest are as follows.

(1) It can process high-dimensional data without feature selection, which is time-consuming and labour-intensive

(2) It is easy to parallelize and it is faster

(3) The most important point is that random forest can balance the errors caused by datasets in dealing with unbalanced datasets [27]

2.4. Stacking Integration Algorithm. Stacking is an integrated algorithm that combines several different machine learners together. Different from the integration of voting method, stacking calls the base learner as primary learner and the learner used for combination as secondary learner [28]. The process for stacking is as follows.

(1) Divide dataset D to train primary learners h_1, h_2, h_3...

(2) Use several primary learners trained to predict the test set on D, respectively. All the predicted results are combined as a secondary training set to train the secondary learner

(3) Use each primary learner to predict the dataset that needs to be predicted initially. Then, average all the predicted results and then use the secondary trainer to predict the predicted results after processing to get the result

2.5. Improved Stacking Model Based on Unbalanced Data

(1) Unbalanced data processing methods

In dichotomous tests, one kind of samples concerned, that is, a few kinds of samples, is generally regarded as positive, while the other kind is considered as negative. When the sample number of positive class is much smaller than the sample number of negative class, the data in this case is called unbalanced data.

Sigmoid binary classification algorithm is essentially a discriminant model based on conditional probability. It is usually a dichotomous method with a threshold value of 0.5, a positive sample is greater than 0.5, and a negative sample is less than 0.5 [29]. The sigmoid function formula in multidimensional feature space is as follows:

$$b_\theta(I) = a\left(\theta^N I\right) = \frac{1}{1 + e^{-\theta^N I}}, \tag{3}$$

where θ represents the multidimensional parameter and X is the eigen-space matrix.

For dichotomous problems, the conditional probability function of samples and parameters θ can be expressed as follows:

$$U(j \mid I ; \theta) = (b_\theta(I))^j (1 - b_\theta(I))^{1-j}, \tag{4}$$

where y represents the dichotomous problem output. After the probability function is obtained, the maximum likelihood estimation and logarithm are performed, and the formula is as follows:

$$\rho(\theta) = \log L(\theta) = \sum_{x=1}^{w} j^{(x)} \log b\left(I^{(x)}\right) + \left(1 - j^{(x)}\right) \log\left(1 - b\left(I^{(x)}\right)\right). \tag{5}$$

The derivative of parameter θ is calculated for Equation (5), and the iterative formula of parameter gradient is as follows:

$$\theta_y \coloneqq \theta_y + \alpha\left(j^{(x)} - b_\theta\left(I^{(x)}\right)\right) I_y^{(x)}. \tag{6}$$

Through continuous iteration on the training set, the approximate extreme value of the derivative is obtained, which is called gradient ascent. Finally, the best parameter θ and the available model are obtained.

Unbalanced data is usually sampled to change the data distribution to reduce the degree of data imbalance. Sampling methods include oversampling and undersampling. That is to increase the number of subcategory samples or reduce the number of multicategory samples to increase the influence of positive category features on the classifier. However, if the sample is only oversampled, it will easily lead to overfitting of the model. The generalization ability of the model will be reduced if negative class samples are undersampled [30].

Therefore, this paper is not limited to the data sampling method, but is combined with the sampling method and make improvements on the algorithm level.

2.6. Improved Stacking Model Integrating Random Forest and Logistic Regression. A certain proportion of samples were selected from the negative samples every time, and all positive samples were retained and combined into a training set to train the random forest model in turn. Specific steps include the following: data as much as positive and 5, 10, 16, and 25 times as much as positive samples were randomly selected from the negative class, respectively. A training set was formed with all positive samples, and five random forests were iteratively trained.

With different sampling multiples, classifiers with different parameters can be obtained to ensure the diversity of classifiers. The five random forests are used as primary classifiers. Considering the high-dimensional sparsity of TFIDF, logistic regression classifier was selected as the secondary classifier. Figure 2 shows part of the improved stacking model.

3. Experiment and Analysis

All models in this paper were trained and tested on a computer with core I5-7500 CPU @3.40 GHz and 64 GB memory. The computer system is Windows 10 professional 64-bit. All models are implemented with MATLAB 2020a Deep Learning Toolbox framework.

3.1. The Dataset. This article selects LendingClub lending site (https://www.endingclub.com/info/download-data.action) of the loan customer information as English writing text experiment datasets (Table 1) to 2018 and 2020 loan customer information English text dataset, in which the English-written text contains customers' static information (such as income, work, family, and conditions) and dynamic information (customers' historical credit and others).

In this paper, Loanamnt, Fundedamnt, and other attributes of the dataset are mainly used for the initial prediction of random forest. And home_own, desc, and other attributes are the attributes of the initial prediction of the multilayer perceptron. Due to the large amount of data, data in 2018 and 2020 were selected for analysis in this paper. The sample number of LoanStats_a dataset is 40540, and the number of attributes is 125. The number of samples in LoanStats_b dataset is 40536, and the number of attributes is 56.

3.2. Data Initialization. Firstly, a variety of data preprocessing algorithms are used to remove redundant attributes of data and reduce the number of features. Then, the correlation of each attribute in the dataset is analyzed, the low positive correlation and negative correlation features in the analysis results are removed (that is, redundant features are removed), and the incomplete attribute values related to the results are filled in. After obtaining the available dataset, various algorithms are tested on the dataset. The influence of its performance on the test results before and after feature extraction is investigated, and the obtained data will be used as the theoretical basis for adjusting the structure of the model.

The preprocessing of initial data is completed through the following process:

(1) Use the same number of statistical attributes to determine whether it is the unique attribute and delete the unique attribute in the dataset, such as ID, member ID, and subgrade

(2) Process the nonnumeric data in the dataset, remove the "%" in int_rate, keep only its numeric part, and convert to float. Replace all "fully paid" in the loan status attribute with 1. "Charge doff" is replaced with all zeros. All other fields are replaced with Nan. Replace all "n/A" in the dataset with Nan values. Then, delete the Nan sample in the loan status property. Because loan status is a result set, it is not allowed to have a Nan value

(3) Delete attributes or samples that are Nan in the dataset

(4) Process the incomplete value of the object, int, and float attributes in the data item. If the value in the data item is 0 and Nan is empty, it is the missing item. The missing rate (the percentage of missing

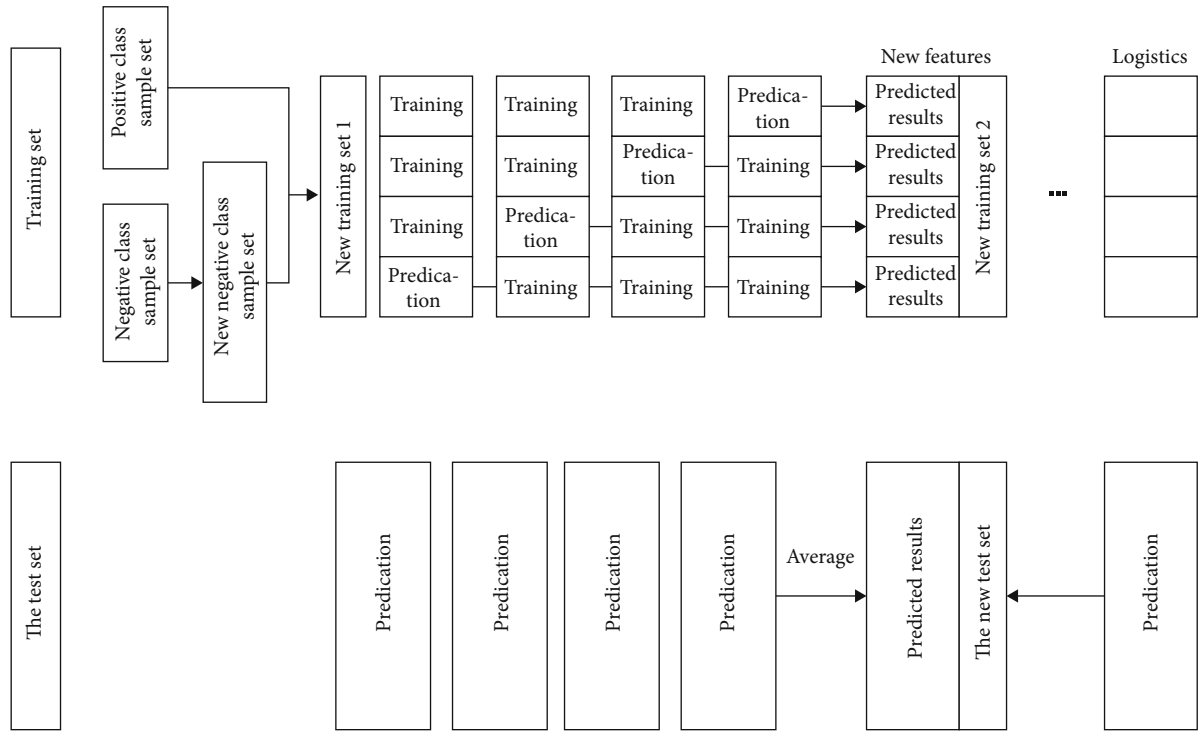

Figure 2: Improved stacking model.

Table 1: English text datasets of loan customer information in 2018 and 2020.

Dataset	Sample number	Attribute number	File format	Statistical year
LoanStats_a	40540	125	Csv	2018
LoanStats_b	40536	56	Csv	2020

items in the whole population) is calculated as follows:

$$\text{Missing}_{\text{pct}} = \frac{L - \text{num}}{L}, \tag{7}$$

where L represents the length of the target attribute, num, which represents the number of target attributes that are 0.

(5) After the above data cleaning, calculate the linear correlation coefficient between the attributes of the dataset, and the formula is as follows:

$$r(I, J) = \frac{\text{Cov}(I, J)}{\sqrt{\text{Var}(I)\text{Var}(J)}}, \tag{8}$$

where $\text{Cov}(X, Y)$ is the covariance of X and Y. $Var\,(X)$ is the variance of X. Var (Y) is the variance of Y. X and Y are attributes of the input. Finally, only $r(X, Y) > 0.7$, $r(X, Y) \in [0, 1]$ is retained. An $n \times n$ dimensional lower triangular correlation coefficient matrix is obtained from Equation (8), where N is the number of dataset attributes.

(6) Fill in missing data. Properties of numeric or custom type fill the mean or 1. Properties of object type are randomly selected in the domain of this property. After processing, two datasets feature 1 CSV and feature 2 CSV can be obtained, whose information is listed in Table 2

The model divides the dataset into two groups (sets S and Z) by attribute. The performance of different algorithms on dataset Z is experimentally observed. The size of the uncertain item set available in the dataset is counted, and the best collocation scheme is selected. After the above work was done, the uncertainties previously on the training set were extracted and trained to update the logistic regression layer parameters of the improved stacking model.

3.3. Feature Extraction. The text information in the dataset of this paper is English text, and the processing of English text includes HTML character conversion, data decoding, removal of Stop Word, removal of punctuation marks, removal of emoticons, separation of words stuck together, and removal of URL.

Text data belongs to unstructured data, and machines are often unable to conduct operation analysis on such data. In general, they need to be converted into structured data that can be analyzed by machines. Therefore, text data features are vectorized [31]. In text classification, word vector

Table 2: Dataset information after pretreatment.

Dataset	Sample number	Attribute number	File format	Statistical year
Feature 1	40540	42	Csv	2018
Feature 2	7800	26	Csv	2020

is a commonly used text representation method. The calculation of entry weight often needs to consider:

(1) The more frequently a word appears in a document, the greater its contribution to text recognition

(2) The fewer times a word appears in all documents, the better it can distinguish between different documents

TFIDF takes both into account. In this paper, all the words in the comments are put into the TFIDF thesaurus, and then, the TFIDF value is calculated as the entry weight, and the text data is converted into word vector, to carry out classifier training [32].

3.4. Experimental Analyses. Before building this model, the performance of MLP, random forest, logistic regression cross-validation (CLF), and logistic regression (LR) algorithms on feature 1 dataset was analyzed separately. Because these results help to better adjust the structure and parameters of the model, finally, one of them was selected to optimize the improved stacking model. In the experimental analysis, PCA algorithm will be added after data pretreatment for comparison. The statistical results show that CLF algorithm and PCA algorithm have a negative effect on the final prediction results of other algorithms. So, consider removing this process in the experiment. The average $F1$ values of the 10 times prediction results of the above algorithms on the training set, test set, and test set with PCA process are listed in Table 3. As can be seen from Table 3, there is no obvious difference between the data of each algorithm. In this paper, the differences between them are shown in the form of graphs (Figure 3). In order to increase the gap between data, only the parts whose $F1$ value exceeds 0.8 are shown in Figure 3.

It is found in the experiment that the two conditional probabilities using logistic regression layer obey Laplace distribution, so penalty is set as $L1$. Parameter C is the reciprocal of regularization coefficient, representing the degree of regularization. The dataset used by the error correction model has a small sample size and large random factors, resulting in a small degree of regularization. Therefore, parameter C should be set to 1. The fitting intercept parameter indicates whether there is a bias, and it is usually set to true. The max number of iterations is set to 100. The larger the value is, the higher the convergence degree of the cost function is. The parameters of MLP layer in the model are set as follows: The hidden layer is set as $100 \times 200 \times 100$, and the activation function is ReLU. The optimizer uses L-BFGS, and the dataset used in the experiment has many attributes. The approximate Hesse matrix needs to be stored in each iteration, which occupies too much storage space and reduces the efficiency of the algorithm. L-BFGS algorithm is an improved algorithm of BFGS algorithm, which only saves the information of the latest m iterations to reduce the storage space of data and improve the efficiency of the algorithm. Use the optimizer to train the weights and biases of the input and output layers.

Table 3: Average predicted $F1$ values of different algorithms on the training set and test set.

Dataset	CLF	LR	Random forest	MLP
Training set	0.9208	0.9208	0.9907	0.9199
Test set	0.9159	0.9159	0.9053	0.9145
Test set (PCA)	0.9161	0.9158	0.9009	0.9009

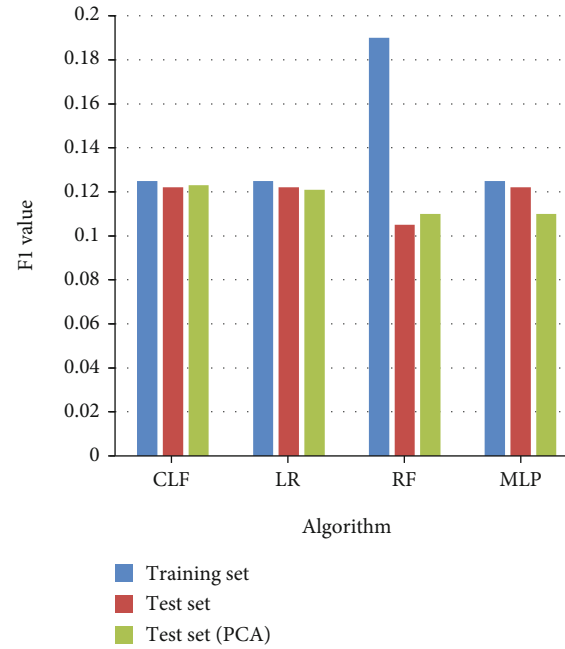

Figure 3: Average prediction $F1$ value of different algorithms.

After the above process, the prerequisite conditions for the realization of the prediction of uncertainty are met. Two comparison models were introduced: Model 1 was used to predict the stacking using the logistic regression layer parameters already trained in the original stacking model. The experimental results show that the prediction results are not ideal, and the accuracy is less than 20%. Model 2 takes all the uncertainties in the training set as the training set of logistic regression layer to train and update the parameters of logistic regression layer. Experimental results show that the logistic regression layer with updated parameters has good prediction results. Figure 4 shows the comparison of error correction results of the above two models.

Figure 5 is the result of the experiment. Experimental dataset feature 2 is selected for prediction. Compared with dataset feature 1, this dataset has fewer data attributes. The pretreatment of the same part of the two datasets adopts

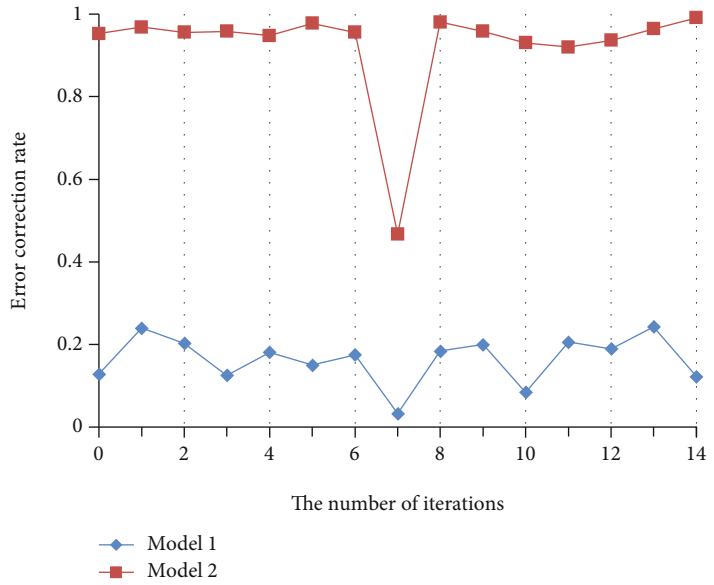

Figure 4: Comparison of error correction results between model 1 and model 2.

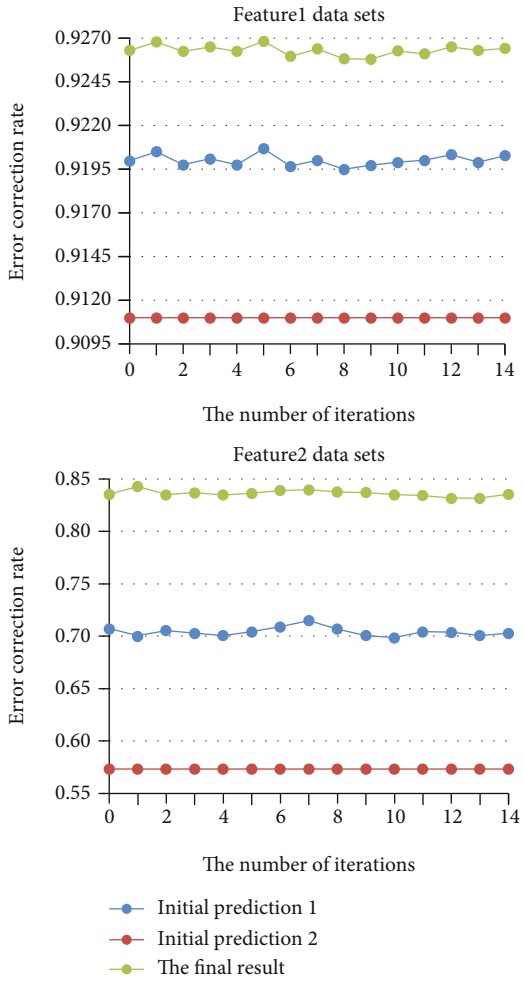

Figure 5: Experimental results of the proposed algorithm on different datasets.

the processing method of feature 1 dataset. Only the attributes used in the experiment of feature 1 dataset are retained. In the experiment using feature 2 dataset, it is found that if only the traditional prediction method is used, the prediction rate will be lower than that in the initial experiment. This is because the reduction of the dimension of the dataset makes the algorithm unable to extract the influence experimental results in a wider range. However, the performance is better in the process of obtaining uncertainties. Many uncertainties can be obtained in each test, which is conducive to correcting many errors and the error correction rate is good. For the wrong prediction results which are not in the uncertain terms, most of them are from the error of the collected data or the factors outside the dataset. It does not belong to the prediction error caused by incomplete feature attribute of dataset. The factors leading to these errors are more complex, and it is more difficult to correct them. Compared with the cost of the whole algorithm, it no longer has a higher predictive value.

After a series of structural adjustment and parameter modification of local algorithm, the final model structure is obtained. The overall experimental results of this model show that the algorithm can improve the final prediction efficiency well in various experimental datasets. In general, compared with nonfeature complete datasets, the number of uncertainties obtained from feature complete datasets is smaller. In actual data mining, there are always more or less incomplete and error in the collected data, which provides conditions for the acquisition of uncertain items. Therefore, it can be proved that the idea of separating out the uncertain term and improving it is feasible (the performance is different in different application fields or datasets). Figure 5 is the result of 15 tests of this model on feature 1 and feature 2 datasets. The initial classification results of the random forest algorithm before the uncertainty are generated are, respectively, represented by initial prediction 1 and initial prediction 2 in Figure 5.

In conclusion, compared with the neural network prediction model based on feature dimensionality reduction, the improved stacking model based on random forest and logistic regression in this paper can analyze practical problems more comprehensively and obtain more reasonable prediction results. An improvement of the model in this paper is that the logistic regression layer of the improved stacking model is no longer used to process the output state of the fully connected layer, but is used to obtain the predicted results of the uncertainties. These uncertainties contain data features that may be ignored by general neural networks. For obvious features in the data, all can be analyzed and processed by the fully connected layer or random forest of the improved stacking model; the classification results can be obtained by simply connecting a softmax or sigmoid classifier at the output location.

4. Conclusion

In order to further improve the recognition and classification of English writing text features without relying on manual feature extraction, in this paper, an improved stacking model based on random forest and logistic regression classification algorithm is proposed to analyze the characteristics of English-written texts. In this model, multiple undersampling is used to train multiple random forests as the primary classifier, and logistic regression is used as the secondary classifier to obtain the prediction results of uncertain terms. English writing text features fully connected layers or random forest computation analysis and processing with improved stacking models. It is classified by a softmax or sigmoid classifier. Experimental results show that this model can effectively improve the classification efficiency of unbalanced English writing text classification. The prediction accuracy is greatly improved while ensuring the main characteristics. It is proved that the model has high practicability in analyzing the features of English writing texts. In the next step, more data on the use of natural scenes in English writing texts will be collected to further study how to improve the feature recognition and classification of natural scene English writing texts.

Data Availability

The labelled datasets used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no competing interests.

Acknowledgments

This study is sponsored by the 2021 National Teaching Reforms on Foreign Language Education in Vocational Colleges (WYJZW-2021-2009).

References

[1] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, May 2019.

[2] S. Qin, A. Bissacco, M. Raptis, Y. Fujii, and Y. Xiao, "Towards unconstrained end-to-end text spotting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4704–4714, Seoul, Korea, 2019.

[3] D. Prasad, A. Gadpal, K. Kapadni, M. Visave, and K. Sultanpure, "CascadeTabNet: an approach for end to end table detection and structure recognition from image-based documents," in *Conference on Computer Vision and Pattern Recognition Workshops*, pp. 572-573, The Washington State Convention Center USA, 2020.

[4] Z. Al-Saadi, "Gender differences in writing: the mediating effect of language proficiency and writing fluency in text quality," *Cogent Education*, vol. 7, no. 1, p. 1770923, 2020.

[5] S. Suhono, M. Zuniati, W. Pratiwi, and U. A. A. Hasyim, "Clarifying Google Translate problems of Indonesia-English translation of abstract scientific writing," *Enterprise Application Integration*, pp. 1–13, 2020.

[6] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, "Scene text detection with supervised pyramid context network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 9038–9045, 2019.

[7] Z. Zhong, L. Sun, and Q. Huo, "Improved localization accuracy by LocNet for Faster R-CNN based text detection in natural scene images," *Pattern Recognition*, vol. 96, article 106986, 2019.

[8] N. Gupta and A. S. Jalal, "A robust model for salient text detection in natural scene images using MSER feature detector and Grabcut," *Multimedia Tools and Applications*, vol. 78, no. 8, pp. 10821–10835, 2019.

[9] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "Textfield: learning a deep direction field for irregular scene text detection," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5566–5579, 2019.

[10] M. Murugesan and S. Thilagamani, "Efficient anomaly detection in surveillance videos based on multi layer perception recurrent neural network," *Microprocessors and Microsystems*, vol. 79, article 103303, 2020.

[11] N. Bhatt and A. Kumar, "A passive islanding detection algorithm based on modal current and adaptive boosting," *Arabian Journal for Science and Engineering*, vol. 45, no. 8, pp. 6791–6801, 2020.

[12] W. Yu, G. Wang, C. Liu, Y. Li, Z. Zhang, and K. Liu, "An algorithm for corner detection based on Contour," *Chinese Automation Congress*, pp. 114–118, 2020.

[13] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE Access*, vol. 7, pp. 15196–15209, 2019.

[14] X. Hou, W. Ao, and F. Xu, "End-to-end automatic ship detection and recognition in high-resolution Gaofen-3 spaceborne SAR images," *IEEE International Geoscience and Remote Sensing Symposium*, pp. 9486–9489, 2019.

[15] Y. Pang, T. Wang, R. M. Anwer, F. S. Khan, and L. Shao, "Efficient featurized image pyramid network for single shot detector," in *Conference on Computer Vision and Pattern Recognition*, pp. 7336–7344, Xi'an, China, 2019.

[16] F. Bi and J. Yang, "Target detection system design and FPGA implementation based on YOLO v2 algorithm," in *3rd International Conference on Imaging, Signal Processing and Communication*, pp. 10–14, Sydney, Australia, 2019.

[17] C. Y. Lee and Y. P. P. Chen, "Prediction of drug adverse events using deep learning in pharmaceutical discovery," *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1884–1901, 2021.

[18] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," *Applied Soft Computing*, vol. 86, article 105836, 2020.

[19] C. Liu and X. Xu, "AMFF: a new attention-based multi-feature fusion method for intention recognition," *Knowledge-Based Systems*, vol. 233, article 107525, 2021.

[20] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 273–292, 2019.

[21] L. Li and W. Li, "Naive Bayesian automatic classification of railway service complaint text based on eigenvalue extraction," *Tehnički Vjesnik*, vol. 26, no. 3, pp. 778–785, 2019.

[22] W. Ahmad, N. Ayub, T. Ali et al., "Towards short term electricity load forecasting using improved support vector machine and extreme learning machine," *Energies*, vol. 13, no. 11, p. 2907, 2020.

[23] T. Zhang and S. S. Ge, "An improved TF-IDF algorithm based on class discriminative strength for text categorization on desensitized data," in *International Conference on Innovation in Artificial Intelligence*, pp. 39–44, Suzhou, China, 2019.

[24] Y. Li and H. Ning, "Multi-feature keyword extraction method based on TF-IDF and Chinese grammar analysis," in *International Conference on Machine Learning and Intelligent Systems Engineering*, pp. 362–365, Dalian, China, 2021.

[25] C. Kim and D. Klabjan, "A simple and fast algorithm for L1-norm kernel PCA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 1842–1855, 2020.

[26] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: an ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[27] B. F. Tanyu, A. Abbaspour, Y. Alimohammadlou, and G. Tecuci, "Landslide susceptibility analyses using random forest, C4.5, and C5.0 with balanced and unbalanced datasets," *Catena*, vol. 203, p. 105355, 2021.

[28] H. Jia, J. Zhao, and W. X. Sun, "Accurate heart disease prediction via improved stacking integration algorithm," *Journal of Imaging Science and Technology*, vol. 65, no. 3, pp. 30408-1–30408-9, 2021.

[29] A. Wanto, I. S. Damanik, I. Gunawan et al., "Levenberg-Marquardt algorithm combined with bipolar sigmoid function to measure open unemployment rate in Indonesia," in *International Conference of Computer, Environment, Agriculture, Social Science, Health Science, Engineering and Technology*, Beijing, China, 2021.

[30] X. Tang, T. Machimura, J. Li, W. Liu, and H. Hong, "A novel optimized repeatedly random undersampling for selecting negative samples: a case study in an SVM-based forest fire susceptibility assessment," *Journal of Environmental Management*, vol. 271, article 111014, 2020.

[31] F. Haque, M. M. H. Manik, and M. M. A. Hashem, "Opinion mining from bangla and phonetic bangla reviews using vectorization methods," in *International Conference on Electrical Information and Communication Technology*, pp. 1–6, Thailand, Bangkok, 2019.

[32] P. Song, C. Geng, and Z. Li, "Research on text classification based on convolutional neural network," in *International Conference on Computer Network, Electronic and Automation*, pp. 229–232, Dalian, China, 2019.