Hindawi
Scientific Programming
Volume 2022, Article ID 7781741, 11 pages
https://doi.org/10.1155/2022/7781741

Research Article

Chinese Language Literature Emotion Analysis Model Based on Unbalanced K-Nearest Neighbor Classification Method

Jiaying Ji

College of Division of Personnel and Party Affairs, NanTong Institute of Technology, NanTong 226002, China

Correspondence should be addressed to Jiaying Ji; jjy198208@163.com

Received 15 November 2021; Revised 15 December 2021; Accepted 16 December 2021; Published 3 March 2022

Academic Editor: Tongguang Ni

The Chinese language is a nation's symbol, the accumulation of a country's magnificent culture, and the pearl precipitated in the long history's washing. The Chinese language is rich and complex, and there are still many topics and issues that merit repeated exchanges and discussions in academic circles. This study proposes a classification method of emotion polarity based on reliability analysis in order to identify the tendency of literary emotion in Chinese language. Support vector machine (SVM), class center, and KNN (K-nearest neighbor) are included in the combined classifier, which effectively improves the accuracy and efficiency of emotion polarity classification. A Chinese literary emotion analysis model based on the method of UKNNC (unbalanced K-nearest neighbor classification) is proposed at the same time by analysing the characteristics of text structure and emotion expression. The experimental results show that, when compared to traditional machine methods, the UKNNC method can analyse text sentiment in fine-grained and multilevel ways, while also improving the accuracy of Chinese literary sentiment analysis.

1. Introduction

The term "Chinese language and literature" specifically refers to a literary work that is created through Chinese editing and then communicated through Chinese works. The analysis of language application and artistic conception features in Chinese language and literature that follows is useful for improving human language expression ability and application. Using the content covered in Chinese language and literature can help students improve their language skills, improve communication with others, expand their thinking abilities, and better taste Chinese language and literature works. In essence, students can only grasp the applied context effectively if they have a broad understanding of Chinese language and literature, which helps to improve students' overall Chinese literacy.

Emotion analysis of Chinese language and literature refers to judging its emotion tendency by analysing and mining information such as emotions, positions, and opinions in the text [1]. It involves many fields such as data mining, information retrieval, and so on, and has a wide range of application value. Data level processing is very important for unbalanced classification [2, 3]. No matter whether the importance of samples is consistent in practical problems, data level processing will have a relatively positive impact on the final classification results. Traditional learning analysis techniques in the field of education frequently focus on analysing structured data but rarely consider unstructured data, making it difficult to accurately identify learners' attitudes, emotions, and psychological states [4, 5]. The support of classification methods is essential for successful classification of imbalanced data sets, and a good classification method is the key to success. When traditional classification methods are used to classify imbalanced data sets, however, test samples from a few classes are frequently misclassified into the majority of classes, making it easy to overlook a few classes and resulting in poor classification performance [6]. Because the complex distribution of unbalanced data is difficult to capture, current technology in this field still has many flaws. As a result, investigating an effective unbalanced data classification technology [7] is extremely important.

K-nearest neighbor (KNN) method has become one of the most famous algorithms in the field of pattern

recognition [8–10] and statistics because of its simple algorithm, easy realization, no need to estimate parameters, and high classification accuracy, and it is also one of the earliest nonparametric algorithms applied to automatic text classification in machine learning [11, 12]. However, KNN method needs to store all the training sample data in the process of calculating the nearest neighbor of each sample to be tested, which leads to a large number of similarity calculations for classification and significantly increases the complexity of classification calculation with the increase of sample data set [13], thus reducing the classification efficiency. Based on this, this study puts forward a Chinese literary sentiment analysis model based on the method of unbalanced K-nearest neighbor classification (UKNNC) for learners' learning experience texts, which classifies the learning experience text information in multiple levels to improve the performance of learners' sentiment analysis and provide more effective support for teaching design and management.

Emotion analysis of Chinese literature belongs to the category of computational linguistics, which involves many research contents such as artificial intelligence, machine learning, information extraction, information retrieval, and data mining. It is closer to the goal of artificial intelligence than traditional technology, and has important research value in theory and application. Therefore, this study innovatively proposes an emotion analysis model of Chinese language literature based on UKNNC method. The difference between this method and the traditional voting strategy of combined classifiers lies in whether to vote on the categories of samples by multiclass classification by analysing the credibility of samples. Experiments show that the classification speed of this method is lower than that of SVM.

2. Related Work

As far as the nature of Chinese language is concerned, it involves a wide range of majors, but as a whole, liberal arts students are the main teaching objects, and some students do not study Chinese language and literature until they enter the university, which also reflects the charm of Chinese language and literature from the side, and makes the analysis of language application and artistic conception get corresponding attention and attention. Literature [14] puts forward Fisher discrimination rate that calculates the distance between positive and negative classes of each feature. The larger the distance, the easier it is to classify the data. Literature [15] puts forward the complexity measurement (CM) as the data set measurement index, and CM calculated the proportion of less than half of the samples in the nearest neighbors of the samples in the data set. The larger the CM, the larger is the cross between positive and negative samples. When IR is used as the data set evaluation index, there is no fixed value range of IR, but the CM range is 0–1, which makes it more comparable. Literature [16] holds that each sample has its neighbor distribution. When the sample is surrounded by samples of the same kind, the classification of the sample is simple, while when the sample is surrounded by samples of different kinds, the classification of the sample will be more difficult. In literature [17], the author used Naive Bayes and maximum entropy method to study the emotion classification of news and commentary corpus. Through experiments, it was found that the accuracy of binary value as feature weight was better. Literature [18] has studied the correlation between the extracted subject and the evaluation words of rhetoric. They introduce the relationship extraction into data mining, take the evaluation words and subjects co-occurring in the same sentence as candidate sets, and apply the maximum entropy model to extract the relationship with various features, such as words, parts of speech, semantics, and positions. The results show that degree adverbs can help improve the performance of subjective relation extraction.

In terms of classification method selection and design, it is primarily based on a thorough examination of the degree of influence of indexes (parameters, structures, and so on) of traditional methods on unbalanced data sets, with corresponding improvements to existing methods or the creation of new classification methods. The original training samples are primarily clustered or blocked in literature [19], and then representative training samples are selected to replace the original training samples. The literature [20] divides samples based on density. The KNN rule of distance weighting is based on the characteristics that the nearest neighbor points near the test sample contribute a lot to classification, whereas the opposite contribution is small, according to literature [21]. Literature [22, 23] proposes a generalised nearest neighbor classification method that assigns different weights to each dimension of classified data. The importance of boundary samples has been improved in the literature [24]. Literature [25] defines the representative function from two factors: the distance between two samples and the included angle and replacing the category attribute function with the defined function when calculating the weight by KNN method. In literature [26, 27], by summarizing the construction of praise and derogatory words, the table of influencing factors, the matching word list, and the clear word list in the field of electronic products, an evaluation system for electronic products was constructed, and the opinion mining of electronic products was realized, with the correct rate reaching 93%.

Emotion classification and extraction, as the basic research of emotion analysis, provide effective support for the application of emotion analysis, but there is still a big gap between the results of classification and extraction and the specific needs of users. Therefore, in order to narrow the gap with users' needs, the application research of emotion information becomes essential.

3. Research Method

This study examines the problem of Chinese literary emotion analysis, adopts the classification method of emotion polarity based on reliability analysis and the model of Chinese literary emotion analysis based on UKNNC method, and makes an in-depth study with various linguistic features, which greatly improves the effect of Chinese literary emotion analysis as a whole.

3.1. Traditional KNN. On the basis of the nearest neighbor method, the K-nearest neighbors (KNN) method was developed. It is one of the most widely used classification methods in the fields of data mining and machine learning, as well as the US Census Bureau's default data preprocessing method. It is now widely used in a variety of fields, including classification, clustering, and regression. This study will focus on how it can improve its classification.

Emotion tendency includes two categories: positive and negative. Therefore, emotion recognition can be regarded as two kinds of classification problems, namely, the recognition of positive and negative meanings. Based on this, KNN algorithm is used to identify the emotion of the text. The algorithm is a simple, effective, and nonparametric method, its essence is a predictive supervision algorithm, and its rules are data samples [28].

KNN's classification idea is very simple, that is, suppose that for a given sample $D = D(d_1, d_2, \ldots, d_l)$ to be classified (where l is the dimension of the sample), the computer trains the training data sets of known categories, finds out the k nearest neighbor samples most similar to D from these training data sets, and then classifies and votes D to determine its category.

Suppose the training set is

$$T = \{X, C\} = \{x_i, c_i\}_{i=1}^{n}, \quad x_i \in R^l. \tag{1}$$

Here, X is the training sample set, C is the category of training sample set, and test set is $S = \{x_j\}_{j=1}^{m}$.

The traditional KNN method mainly finds out the k training samples with the highest similarity with x_j from the training sample set X of known categories for each sample x_j to be tested:

$$M = \{X', C'\} = \{x_d, c_d\}_{d=1}^{k}, \quad M \in T. \tag{2}$$

The similarity of each sample x_i and x_j in the training sample set X is calculated as follows:

$$Sim = (x_i, x_j) = \frac{\sum_{t=1}^{n} w_{it} w_{jt}}{\sqrt{\sum_{t=1}^{n} w_{it}^2 \sqrt{\sum_{t=1}^{n} w_{jt}^2}}}. \tag{3}$$

Here, w_{it}, w_{jt} represent the weight of the t-th feature item in samples x_i and x_j respectively, and the weight calculation formula of x_j belonging to class c_t is as follows:

$$W(x_j, c_t) = \sum_{x_i \in x_j} v(x_i, x_j) \phi(x_i, c_t). \tag{4}$$

Among them, $v(x_i, x_j)$ is the weight function of voting, generally taking 1 or $Sim = (x_i, x_j)$, and the ϕ function is as follows:

$$\phi(x_i, c_t) = \begin{cases} 1, & x_i \in c_t, \\ 0, & x_i \notin c_t. \end{cases} \tag{5}$$

x_j is classified into the category with the largest class weight W. Repeat this process to classify all samples to be tested. Finally, the classified category is classified with its real class label to measure the performance of this classification method.

The original KNN algorithm's classification principle is simple and straightforward to use, but there is no systematic consideration of the impact of sample data set imbalance on classification, that is, each selected nearest neighbor sample's representative degree to its class is treated the same. To solve this problem, an improved KNN algorithm is proposed in this study. This method primarily improves the algorithm in two areas: sample set imbalance between classes and sample set imbalance within classes.

3.2. Text Emotion Information Classification. The classification of text emotion information is primarily used to identify opinions, attitudes, preferences, and other related information expressed in natural language. Emotion classification research focuses on unstructured text information such as forums and blogs that come in a variety of formats and are written in a colloquial style. Emotion classification, unlike traditional topic-based text classification, focuses on subjective and objective classification, as well as emotion polarity classification. It is mainly divided into word-level emotion information classification, text-level emotion information classification, and sentence-level emotion information classification, depending on the granularity of the research.

Level-I sentiment information classification includes subjective and objective text classification and text sentiment polarity classification. The process is shown in Figure 1.

The mainstream method used in the classification of level-level emotion polarity is the classification method based on statistical learning. From the existing related research, it is found that the classification method based on statistics is also the most effective method to solve the classification of emotion polarity at present. In the existing research, scholars have tried to apply various statistical classification methods and linguistic features such as words and parts of speech to the study of emotion polarity classification, and achieved good results [29]. In order to further improve the classification accuracy, this study looks at two aspects: optimising the classification model and selecting effective features.

This study proposes a combined classifier method based on reliability analysis for optimising classification models. This study employs a method based on category attribute analysis for feature selection. Experiments show that the method used in this study can effectively improve classification accuracy without slowing down the classification process.

In the process of emotion polarity classification, through the analysis of emotion information, it is found that not all samples need to be discriminated by combined classifiers, and there is no obvious difference in the classification results of various single classifiers for samples that are easy to distinguish.

From Figure 2, we can see that the hollow circle and square respectively represent different types of sample points, and the circle and square which are far from the

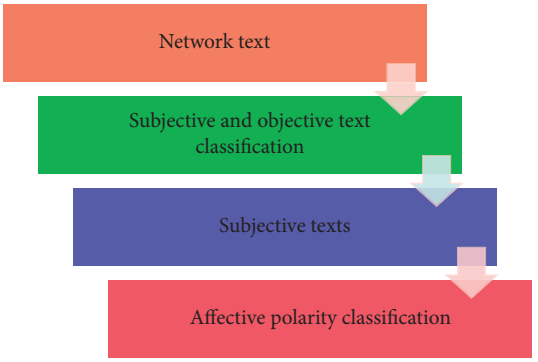

Figure 1: Emotion information classification process.

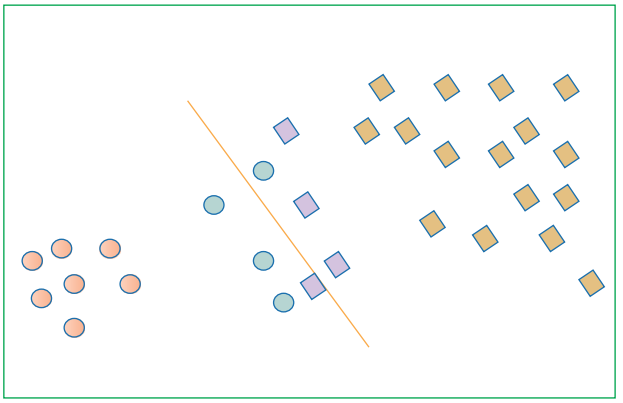

Figure 2: Schematic diagram of binary classification sample distribution.

classification hyperplane respectively represent samples with obvious classification characteristics.

Therefore, if we can find a way to determine which samples belong to easily distinguishable samples and which samples belong to indistinguishable samples, then we can treat the two types of samples with different methods and improve the accuracy and speed of classification. In this section, a classifier fusion strategy based on reliability analysis is proposed to solve the polarity classification problem of these samples.

Assuming that a confidence function $f(x)$ is determined according to the classification principle of the main classifier, and the discrimination threshold $\lambda(\lambda > 0)$ is set, if $f(x) > \lambda$ is used when the main classifier discriminates the samples, it is considered that the discrimination result of the main classifier has high confidence, and the result can be used as the final classification result of the samples.

Otherwise, the classification of the sample to be marked is determined by the method of voting together by the main and auxiliary classifiers. The discrimination process is shown in Figure 3.

We define the credibility function according to the distance between the test sample in the main classifier and the class center, as shown in the following formula:

$$f(x) = \frac{f_1}{f_2}, \tag{6}$$

where f_1 is the distance between the nearest class center and the test sample, and f_2 is the distance between the next nearest class center and the test sample.

It can be analysed that when the cosine similarity is used to calculate the distance between the sample and the class center, if $f_1 > f_2$, it means that the distance between the test sample and the nearest class center is relatively close, and the distance from the next nearest class center is relatively far, and this kind of sample is considered to be relatively easy to distinguish.

If $f_1 < f_2$, it means that the distance between the test sample and the nearest class center and the distance between the next nearest class center may be relatively close, and this kind of sample is considered not easy to distinguish.

In the classification problem, apart from the classifier used, the influence of feature selection on the classification results is also very important. The feature selection algorithm adopted in this chapter is a feature selection algorithm based on category attribute analysis [30], which has achieved good results in traditional text classification. The formal description of feature selection algorithm based on category attribute analysis is as follows:

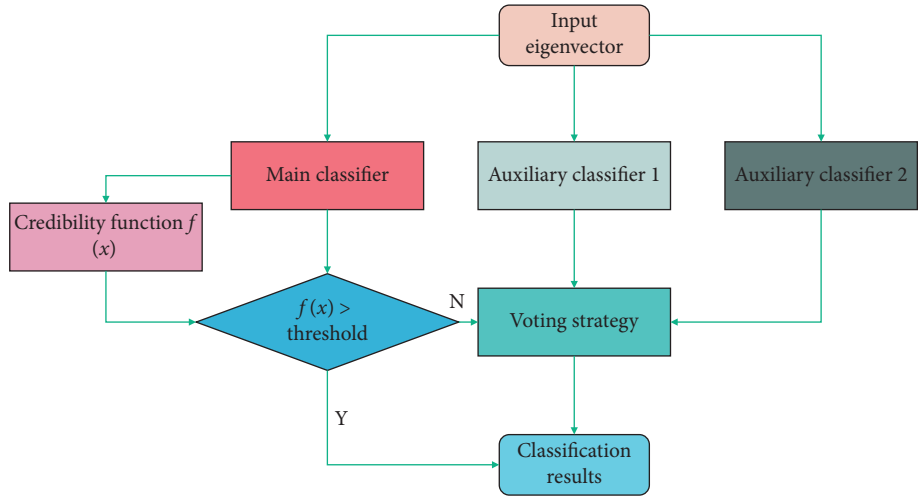

Figure 3: Discriminating process of combined classifier.

Let us assume that $X = (\omega_1, \omega_2, \ldots, \omega_n)$ represents a text, and $C_j = (X_1^j, X_2^j, \ldots, X_{n_i}^j)$ represents a text set with n_i texts and the category mark C_j, $j = 1, 2, \ldots, m$. The following formula represents the intraclass distribution law and interclass distribution law of feature ω_i in text class C_j.

Assuming that the word frequency of feature ω_i in C_j is T_{ij} and the document frequency is d_{ij}, the distribution within the class is as follows:

$$S\omega_{ij} = \frac{f\omega_{ij} \cdot \log\left(d\omega_{ij} + 1\right)}{\sqrt{\sum_{i=1}^{n} \left[f\omega_{ij} \cdot \log\left(d\omega_{ij} + 1\right)\right]^2}}. \tag{7}$$

Here, $f\omega_{ij} = T_{ij}/L_i$, L_i is the number of times ω_i appears in all categories; and $d\omega_{ij} = d_{ij}/D_i$, D_i is the number of all texts where ω_i appears.

Class distribution is recorded as follows:

$$B\omega_i = \left(S\omega_{i1}, S\omega_{i2}, \ldots, S\omega_{im}\right). \tag{8}$$

Generally, when the components of the distribution vector between classes differ greatly, the feature has a strong ability to distinguish between classes. When the components of the interclass distribution are similar or identical, the feature's ability to distinguish between classes will be very low, so the feature's resolution can be expressed as follows:

$$\mathrm{Im}p\left(\omega_i\right) = \sqrt{\frac{\sqrt{\sum_j \left(S\omega_{ij} - \overline{S}\omega_i\right)^2}}{\sum_j S\omega_{ij}}}. \tag{9}$$

In the above formula, $\overline{S}\omega_i = \sum_j S\omega_{ij}/m$.

From the above formula, we can find that the feature selection method based on category attribute analysis pays attention to the distribution of features within and between classes and quantitatively describes this distribution through variance mechanism. The features retained by the algorithm usually have obvious category attributes, so these features have strong representation ability to the corresponding categories, whereas those filtered out features have weak or zero representation ability.

3.3. An Analysis of Chinese Literary Emotion Based on UKNNC. The classification algorithm and the traditional oversampling algorithm are separate. To train the classifier, new samples are first generated and then added to the training set. The test results reflect the quality of samples after the classifier has been trained. However, the advantages of the oversampling algorithm in generating samples are not fully exploited in this framework, and the role of the samples generated in oversampling on the classifier cannot be guaranteed.

This study starts with the goal directly, draws on the idea of confrontation architecture, and proposes the concept of expected classifier, which refers to a new classifier trained by the current classifier under the update of synthesised samples, in order to accurately measure the quality of synthesised samples and judge whether samples can truly improve the performance of classifiers.

Assuming that the classifier used is a single-layer neural network, the influence of the new sample on the performance of the classifier is influenced by the update on the w, b of the classifier, that is, if the w, b calculated by the new sample are in the same direction as the calculation gradient of the original sample, they will reduce the loss function value of the original sample on the new network parameters, and the update of the parameters on the new sample is shown in the following formula:

$$\begin{aligned} w &= w - \eta \frac{\partial \mathrm{loss}}{\partial w} \left(x = x_{\mathrm{new}}\right), \\ b &= b - \eta \frac{\partial \mathrm{loss}}{\partial b} \left(x = x_{\mathrm{new}}\right). \end{aligned} \tag{10}$$

Let the classification cost function use cross entropy, which is defined in the following formula:

$$\begin{aligned} \text{loss}(y, \hat{y}) &= -\sum y \log \hat{y} + (1-y)\log(1-\hat{y}), \\ d\hat{y} &= -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}, \\ z &= w^T x + b, \\ \hat{y} &= s(z), \\ dz &= \hat{y}(1-\hat{y}) \times d\hat{y} = \hat{y} - y, \\ db &= dz = \hat{y} - y, \\ dw &= xdz = (\hat{y} - y)x. \end{aligned} \tag{11}$$

The specific steps of the improved KNN classification are as follows:

Input: training sample set $T = \{X, C\} = \{(x_1, c_1), (x_2, c_2), \ldots, (x_l, c_l)\}$.

Among them, $c_i \in D\{c_1, c_2, \ldots, c_{|C|}\}$ is sample category, and $i = 1, 2, \ldots, l$, sample to be tested x_j.

Output: the class C to which the sample x_j to be tested belongs

(1) Standardization:

$$X_{\text{new}} = \frac{X - \min(X)}{\max(X) - \min(X)}. \tag{12}$$

(2) Calculate the class representation and sample representation of each class.

(3) Calculate the similarity of each sample x_i and x_j in the training sample set X.

(4) Calculate the weight $W(x_j, c_t)$ of the class to which the selected k samples belong, and assign x_j to the class with the largest weight.

(5) Repeat the above steps until all samples to be tested are classified.

UKNNC algorithm uses variational self-encoder as the $P(X)$ model for modeling in oversampling. When training the classifier, because incremental classifier is used, the modification direction of new sample to classifier can be calculated. When the expected classification result of classifier is poor, it means that the modification direction of new sample to classifier is wrong, so the parameters of generator are modified. The flow of UKNNC algorithm is shown in Figure 4.

4. Analysis and Discussion

4.1. Improved KNN Experimental Analysis. In the experiment, the traditional KNN method and the improved KNN method are used to experiment when the nearest neighbor parameter k takes different values, and the results are shown in Figure 5.

When the traditional KNN method is used to classify the experimental data set, the accuracy is the highest when the parameter $k = 50$, while when the improved KNN method is used, the parameter $k = 20$ is the best, so $k = 20$ is finally selected uniformly after comprehensive consideration.

When $k = 20$, the accuracy, recall, and comprehensive classification rate obtained by using the improved classification method on the above data set are shown in Figure 6.

It can be seen from Figure 6 that the improved KNN method proposed in this study has higher accuracy than the existing KNN method on data sets 3, 4, 5, 6, 7, and 8. Especially, on data sets 3, 4, and 8, the original 0 has been improved to a certain accuracy, but only on data set 9, the accuracy has not changed, which is mainly due to the fact that there are few such data sets and the degree of similarity with category 8 is too great.

To summarise, the improved KNN method proposed in this study adopts the idea of weighted calculation of nearest neighbor samples, aiming to solve the problem of equal weight treatment of nearest neighbor samples in the existing KNN method. As a result, each nearest neighbor sample has a specific weight, reducing the impact of unbalanced data sets on classification results and improving classification accuracy.

4.2. Analysis of Polarity Classification Results. In the experiment of Chinese language and literature emotion corpus, documents are represented by vector space model, feature selection algorithm based on category attribute analysis is used, term frequency-inverse document frequency (TF-IDF) value is used as feature item weight, and Chinese word segmentation system uses ICTCLAS3.0 of Chinese Academy of Sciences. The value of k in KNN method is 100.

Figure 7 shows the classification comparison results of the reliability analysis method at $\lambda = 2$, three single classifiers and voting method under the 50% cross-validation method. Among them, for the credibility analysis method, 44.83% of the texts need auxiliary classifiers to participate in voting decisions.

It can be seen from Figure 7 that among the three single classifiers, support vector machine (SVM) has the best classification effect, followed by the classification results of class center and KNN. The F value of reliability analysis is 1.7% higher than that of SVM, which is basically the same as that of voting method.

Chi-square significance test shows that the performance improvement of the method proposed in this study is statistically significant ($p < 0.05$) compared with other methods, which shows that the classification accuracy of the classification fusion strategy based on reliability analysis can exceed the results of SVM model under the condition of setting a reasonable threshold.

In addition, in order to verify the classification speed of reliability analysis method, the classification speed was tested on a computer with CPU frequency of 2.6 GHz and memory capacity of 3.8 GB. The average test result is shown in Figure 8.

Referring to the classification speed values of each single classifier obtained in Figure 8, we can see that the classification speed of the class center method is the fastest, and that

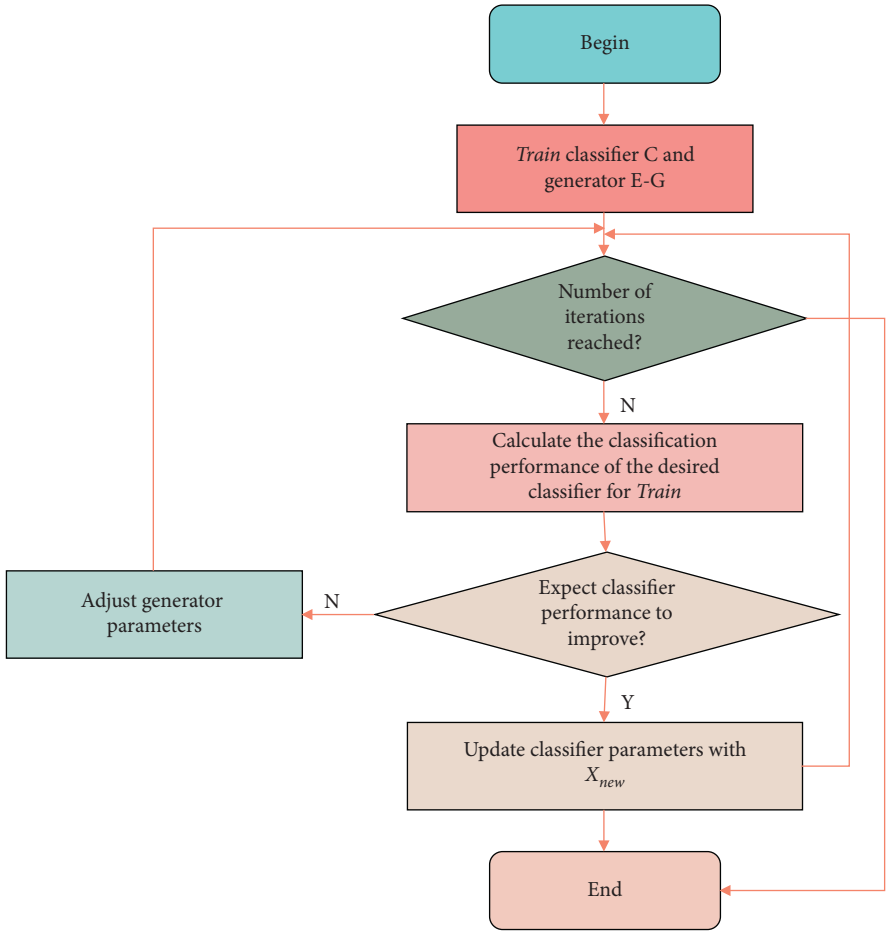

Figure 4: UKNNC algorithm flow chart.

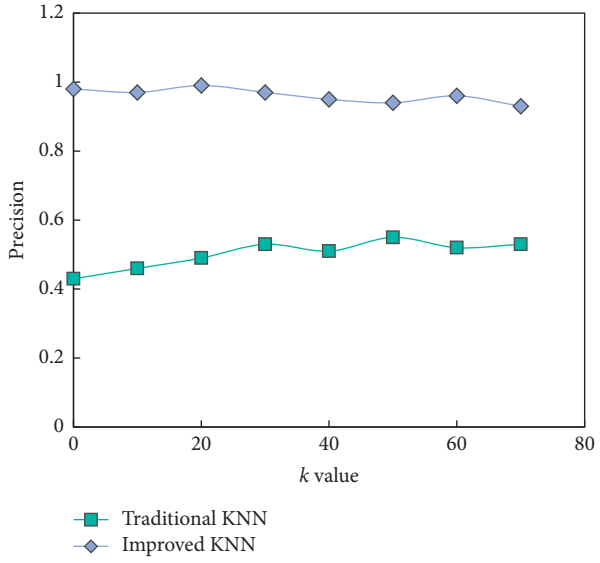

Figure 5: Experimental results of KNN precision before and after improvement.

of the SVM method is the slowest, while the speed of the credibility analysis method is obviously higher than that of SVM, which is between the class center and SVM.

Compared with the voting method, the speed of the credibility analysis method is obviously improved, which shows that the combined classification method based on credibility analysis has an obvious improvement effect compared with SVM and voting method in classification speed under the condition of setting a reasonable threshold.

In the evaluation, two groups of evaluation results of credibility analysis method and single classifier SVM are presented, respectively. The specific evaluation results are shown in Figure 9:

From the evaluation results, it can be seen that the results of reliability analysis method are lower than those of SVM on R-Accuracy and Acc_1000. The reason is that the evaluation does not provide training set, but the test set covers a wide range of fields, which leads to a great difference in the feature distribution between the test set and the training set.

The λ value obtained through the development set cannot well reflect the sample distribution characteristics of the test set, which may lead to the results of reliability analysis method being lower than SVM on some indicators.

To sum up, when the reasonable threshold λ is determined, the credibility analysis method can get better results than SVM method in classification speed and accuracy. At the same time, because the combination classification method based on credibility analysis can ensure faster

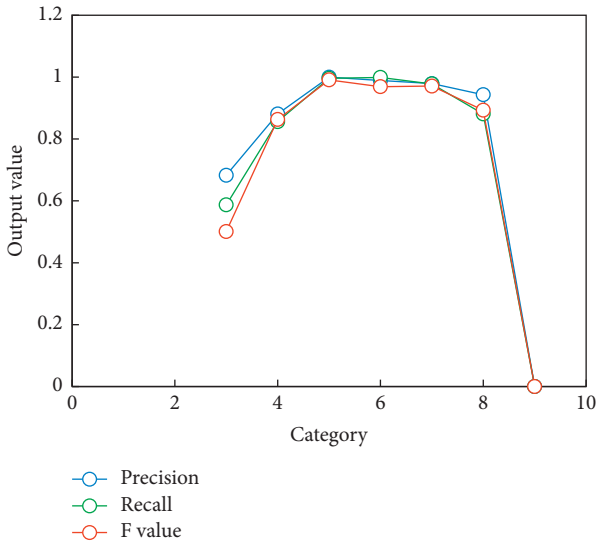

Figure 6: Experimental result comparison data set.

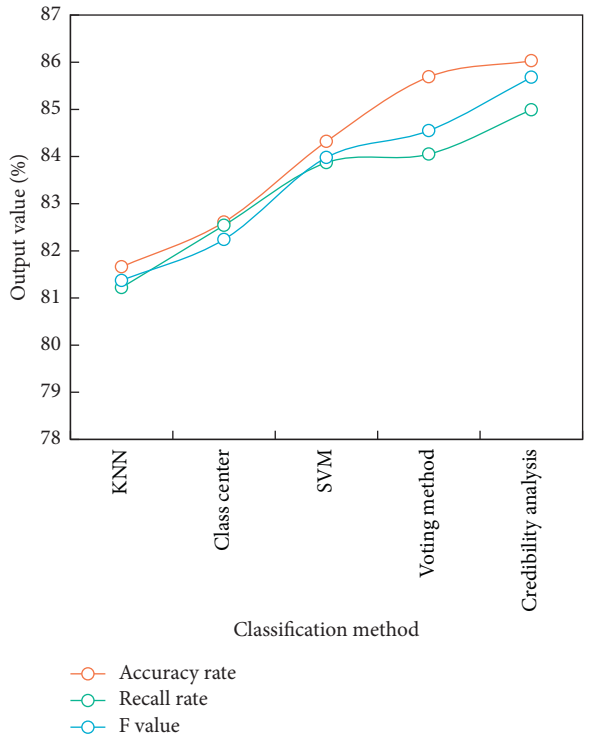

Figure 7: Emotion polarity classification result.

classification speed and better classification accuracy, it is more suitable for emotion classification of large-scale online texts.

4.3. Validity Verification of Unbalanced Classification Algorithm. In this section, in order to verify the rationality of the UKNNC algorithm, the classifier uses a single-layer neural network. The network architecture is consistent with that of the UKNNC algorithm, and it is consistent with the UKNNC algorithm in terms of training times and learning rate.

The VAEOS algorithm is used to generate a few samples that are added to the training set to directly train the LR classifier in order to compare the joint oversampling algorithm with the independent oversampling algorithm. To avoid randomness, the average of 10% cross validation is used in this section as a comparison between the joint oversampling algorithm and the independent oversampling algorithm.

Because it is necessary to make assumptions about the form of data distribution in an oversampling algorithm based on data distribution, this type of algorithm has higher requirements for the form of data distribution, as shown in Figure 10.

Ideally, the oversampling algorithm based on data distribution first reconstructs the probability distribution function of data and then conducts more intensive sampling according to the distribution function. The credibility of sampling results depends on the credibility of modeling. The improvement of classifier performance by the oversampling algorithm based on data distribution depends on whether the current sample meets the hypothesis. Therefore, in UKNNC, when there is no need to set the hypothesis of the true distribution form of samples, the modeling reliability of the algorithm for minority classes will increase, so the generated samples have higher reliability, which shows that the classification performance of minority classes is better.

Figure 11 shows the comparison of the classification performance of CGMOS in the case of Naive Bayes classifier and logistic regression classifier. From the experimental results, it can be seen that due to the different data characteristics, the experimental results are better than other frameworks on the data set that fits the classifier.

However, it synthesizes samples for Bayesian framework classifier, so the classification results on multiple data sets are similar to those of VAEOS algorithm on naive Bayes. All the results of UKNNC algorithm are better than those of VAEOS algorithm under LR classifier, which proves the effectiveness of UKNNC algorithm and joint oversampling algorithm.

Because the data sets in this study are all numerical values, the average classification value of Naive Bayes is slightly lower than LR. In order to obtain a better classification effect, firstly, the classifier should be selected according to the data characteristics; secondly, the oversampling algorithm has a higher probability to improve the classification effect. However, the joint oversampling algorithm based on the characteristics of the classifier has better promotion effect on the classification algorithm in its own framework.

After comparison, all the classification indexes are superior to the emotion classification results of other emotion dictionaries, which not only shows the effectiveness of introducing the expression dictionary to the analysis of Chinese literary emotion but also verifies the application value of the UKNNC method proposed in this study.

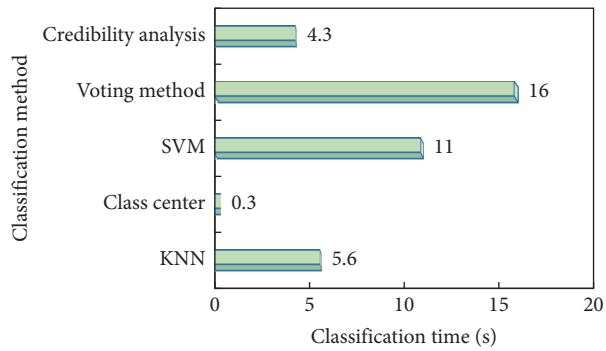

Figure 8: Classification comparison result.

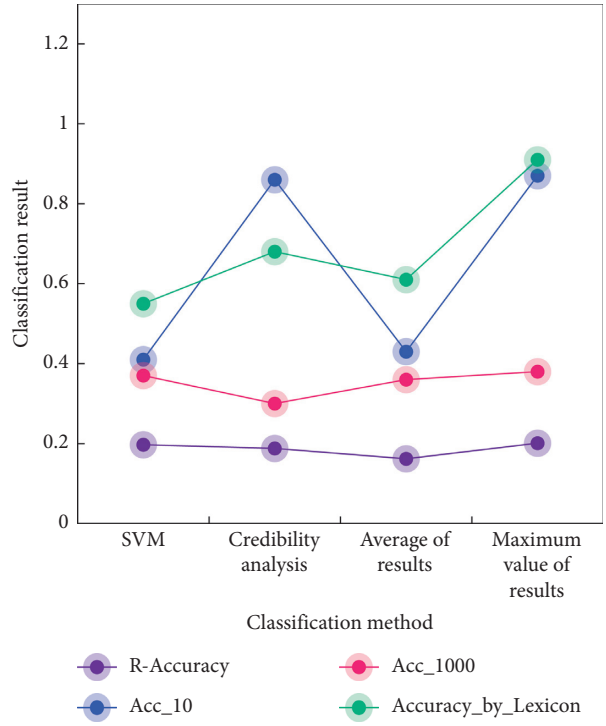

Figure 9: Emotion classification evaluation results.

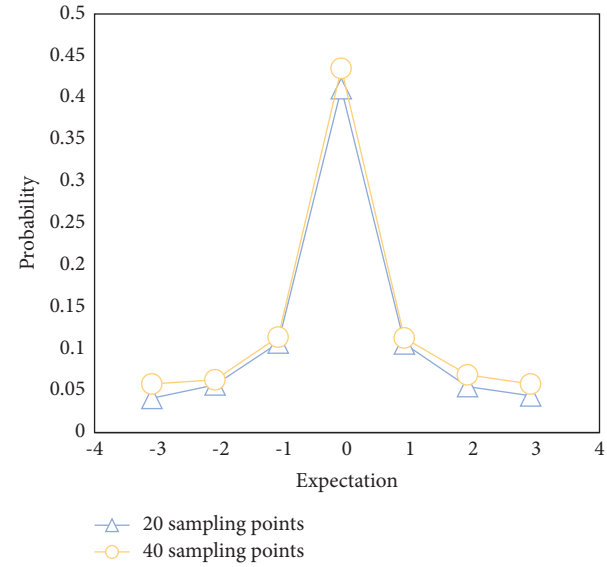

Figure 10: Different sampling points.

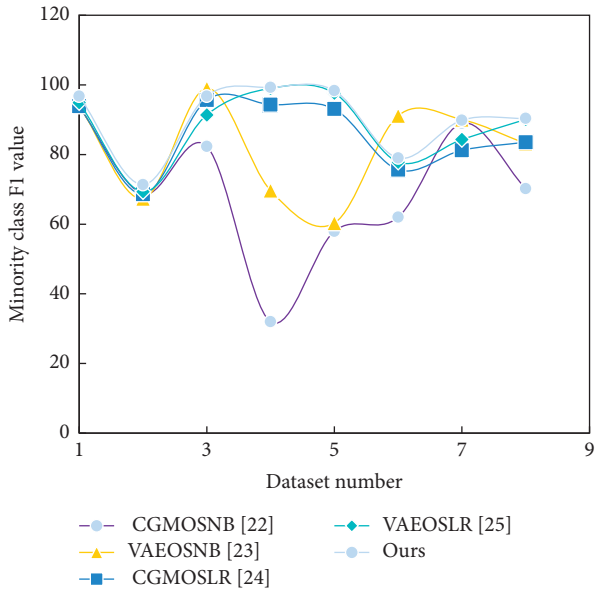

Figure 11: Comparison of oversampling algorithms in different classification frameworks.

5. Conclusion

Chinese development has been shaped by thousands of years of precipitation, resulting in unique aesthetic characteristics that are extremely appealing. We should pay attention to the beauty of temperament when writing poetry. Music is frequently used in the creation of ancient poems. The beauty of images should be considered when describing lyricism. Different images are used to convey specific emotions, and the same images can convey joy and sorrow on multiple levels. The use of reliability analysis to classify emotion polarity is proposed. Experiments show that this voting method based on credibility analysis can outperform SVM in terms of classification accuracy, though its classification speed is slower than SVM's when the appropriate credibility threshold is selected. Simultaneously, a UKNNC-based emotion analysis model of Chinese language literature is proposed, with the emotions of each paragraph in the text obtained through the two-layer model. The experimental results show that, when compared to the traditional machine learning analysis method, the emotion analysis method proposed in this study significantly improves the accuracy of Chinese literary emotion recognition.

Currently, the expected classifier is calculated using synthetic samples and random gradient descent. There is still a lot of analysis and design work to be done on how to use a more accurate gradient descent method and how to avoid prematurely falling into a local optimum in order to stop training.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] C. Huang and J. Chen, "Chinese text classification based on improved K Nearest Neighbor algorithm," *Journal of Shanxi Normal University (Philosophy and Social Sciences edition)*, vol. 48, no. 1, pp. 96–101, 2019.

[2] R. Liu, W. Cai, G. Li, X. Ning, and Y. Jiang, "Hybrid dilated convolution guided feature filtering and enhancement strategy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2021.

[3] M. Gao, W. Cai, and R. Liu, "AGTH-Net: attention-based graph convolution-guided third-order hourglass network for sports video classification," *Journal of Healthcare Engineering*, vol. 2021, Article ID 8517161, 10 pages, 2021.

[4] J. Du and F. Bian, "A privacy-preserving and efficient k-nearest neighbor query and classification scheme based on k-dimensional tree for outsourced data," *IEEE Access*, vol. 8, Article ID 69333, 2020.

[5] N. R. Zhou, X. X. Liu, Y. L. Chen, and N. S. Du, "Quantum K-Nearest-Neighbor image classification algorithm based on K-L transform," *International Journal of Theoretical Physics*, vol. 4, no. 4, pp. 1–16, 2021.

[6] C. Jatmoko and D. Sinaga, "A classification of batik lasem using texture feature ecxtraction based on K-nearest neighbor," *Journal of Applied Intelligent System*, vol. 3, no. 2, pp. 96–107, 2019.

[7] P. Kasemsumran and E. Boonchieng, "EEG-based motor imagery classification using novel adaptive threshold feature extraction and string grammar fuzzy K-nearest neighbor classification," *Computer Journal*, vol. 30, no. 2, pp. 27–40, 2019.

[8] M. Zhao, A. Jha, Q. Liu et al., "Faster Mean-shift: GPU-accelerated clustering for cosine embedding-based cell segmentation and tracking," *Medical Image Analysis*, vol. 71, Article ID 102048, 2021.

[9] Z. Wu and W. Chu, "Sampling strategy analysis of machine learning models for energy consumption prediction," in *Proceedings of the IEEE 9th International Conference on Smart Energy Grid Engineering (SEGE)*, pp. 77–81, IEEE, Oshawa, ON, Canada, 2021 August.

[10] J. Xue, L. Wang, Z. Sun, and C. Xing, "Basic research in diabetic nephropathy health care: a study of the renoprotective mechanism of metformin," *Journal of Medical Systems*, vol. 43, no. 8, pp. 1–13, 2019.

[11] F. Ablayev, M. Ablayev, J. Z. Huang et al., "On quantum methods for machine learning problems part II: quantum classification algorithms," *Big Data Mining and Analysis*, vol. 1, no. 1, 2020.

[12] P. Bhimte, "Review on a privacy-preserving and efficient kNearest neighbor query and classification scheme based on k-dimension tree for outsource data," *International Journal of Innovations in Engineering and Science*, vol. 5, no. 12, pp. 16–19, 2021.

[13] 于. Yu Ting and 杨. Yang Jun, "Point cloud model recognition and classification based on K-nearest neighbor convolutional neural network," *Laser & Optoelectronics Progress*, vol. 57, no. 10, Article ID 101510, 2020.

[14] A. J. Gallego, J. Calvo-Zaragoza, and J. R. Rico-Juan, "Insights into efficient k-nearest neighbor classification with convolutional neural codes," *IEEE Access*, vol. 8, no. 99, p. 1, 2020.

[15] A. Putri, "Analysis of image processing in barcode using the K-nearest neighbor (K-nn) classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, pp. 185–190, 2020.
[16] S. H. Rukmawan, F. R. Aszhari, Z. Rustam, and J. Pandelaki, "Cerebral infarction classification using the K-nearest neighbor and naive Bayes classifier," *Journal of Physics: Conference Series*, vol. 1752, no. 1, Article ID 012045, 2021.
[17] P. Rajarajeswari, "Hyperspectral image classification by using K-nearest neighbor algorithm," *International Journal of Psychosocial Rehabilitation*, vol. 24, no. 5, pp. 5068–5074, 2020.
[18] A. Azhari and F. I. Ammatulloh, "Classification of concentration levels in adult-early phase using brainwave signals by applying K-nearest neighbor," *Signal and Image Processing Letters*, vol. 1, no. 1, pp. 14–24, 2019.
[19] M. Krommyda, A. Rigos, K. Bouklas, and A. Amditis, "An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media," *Informatics*, vol. 8, no. 1, 2021.
[20] M. Piotrowska, G. KOrvel, B. Kostek, and T. Ciszewski, "Machine learning–based analysis of English lateral allophones," *International Journal of Applied Mathematics and Computer Science*, vol. 29, no. 2, pp. 393–405, 2019.
[21] R. Alhalaseh and S. Alasasfeh, "Machine-learning-based emotion recognition system using EEG signals," *Computers*, vol. 9, no. 4, 2020.
[22] N. Darapureddy, N. Karatapu, and T. K. Battula, "Comparative analysis of texture patterns on mammograms for classification," *Traitement du Signal*, vol. 38, no. 2, pp. 379–386, 2021.
[23] J. Shi, Z. Jiang, and H. Zhang, "Few-shot ship classification in optical remote sensing images using nearest neighbor prototype representation," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, no. 99, p. 1, 2021.
[24] S. R. Basha, J. K. Rani, and J. Yadav, "A novel summarization-based approach for feature reduction enhancing text classification accuracy," *Engineering, Technology & Applied Science Research*, vol. 9, no. 6, pp. 5001–5005, 2019.
[25] R. Touati, I. Messaoudi, A. E. Oueslati, I. Messaoudi, and Z. Lachiri, "The Helitron family classification using SVM based on Fourier transform features applied on an unbalanced dataset," *Medical, & Biological Engineering & Computing*, vol. 57, no. 10, pp. 2289–2304, 2019.
[26] C. Li, "Research on the applicability of unbalanced data algorithm based on logistic regression," *Computer Science and Application*, vol. 10, no. 11, pp. 2049–2057, 2020.
[27] Z. Liang, X. Li, and W. Song, "Research on speech emotion recognition algorithm for unbalanced data set," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 5, pp. 1–6, 2020.
[28] J. Jia and W. Qiu, "Research on an ensemble classification algorithm based on differential privacy," *IEEE Access*, vol. 8, no. 99, p. 1, 2020.
[29] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic tenets of classification algorithms K-Nearest-Neighbor, support vector machine, random forest and neural network: a review," *Journal of Data Analysis and Information Processing*, vol. 8, no. 4, pp. 341–357, 2020.
[30] X. Chen and J. Deng, "A robust polarmetric SAR terrain classification based on sparse deep autoencoder model combined with wavelet kernel-based classifier," *IEEE Access*, vol. 8, no. 99, p. 1, 2020.