Hindawi
Discrete Dynamics in Nature and Society
Volume 2021, Article ID 9963382, 8 pages
https://doi.org/10.1155/2021/9963382

Research Article

Application of an Improved CHI Feature Selection Algorithm

Liang-jing Cai,[1] **Shu Lv**,[1] **and Kai-bo Shi**[2]

[1]*School of Mathematical Sciences, University of Electronic Science and Technology of China, Sichuan, Chengdu 611731, China*
[2]*School of Electronic Information and Electrical Engineering, Chengdu University, Sichuan, Chengdu 610106, China*

Correspondence should be addressed to Shu Lv; lvshu@uestc.edu.cn

Received 11 March 2021; Accepted 4 May 2021; Published 13 May 2021

Academic Editor: Zi-Peng Wang

Text classification is the critical content of machine learning, and it is widely applied in information filtering, sentimental analysis, and text review. It is very important to improve the accuracy of classification results, and this is also the main research purpose of researchers in this field in recent years. Feature selection plays an important role in text classification, which has the functions of eliminating irrelevant features, reducing dimensionality, and improving classification accuracy. So, this paper studies the CHI feature selection algorithm, and the main work and innovations are as follows: firstly, this paper analyzed the CHI algorithm's flaws, determined that the introduction of new parameters will be the improvement direction of the CHI algorithm, and thus proposed a new algorithm based on variance and coefficient of variation. Secondly, experiment to verify the effectiveness of the new algorithm. In terms of language, the experiment in this paper includes two text classification systems, which were Chinese and English. In terms of classifiers, two classifier algorithms were used, which included the KNN classifier and the Naive Bayes classifier. In terms of data types, two distribution types of data were used: balanced datasets and unbalanced datasets. Finally, experiment and result analysis. This paper has conducted 3 comparative experiments and analyzed the results of each experiment. The experimental results obtained are all significantly improved compared to the results before the improvement.

1. Introduction

Text classification is a process of classifying unknown text according to the content of text [1]. It is the most classical field of natural language processing, which is according to the content of the text; using some automatic classification algorithms, the computer divides the texts into predefined categories [2]. With the rapid development of the Internet, information and data exist in a variety of ways, such as images, video, sound, and text. Compared with other forms, texts are applied more widely, because of its faster upload speed, faster download speed, and less use of network resources. Therefore, to quickly and correctly classify the massive text information, text classification technology has emerged. The processing of text categorization can be roughly divided into three stages: text preprocessing, text feature dimension reduction, and classification model construction. Text classification is widely used in filtering spam, analysis of Internet public opinion [3], and clinical diagnosis [4]. At present, the main direction of the text information structure is the vector model, in which the difficulty is mainly the high dimension of text features and the data sparsity [5], so feature dimension reduction plays a vital role in the classification effect. Feature selection method is an effective way to address this issue and play an important role in text classification [6].

At present, many researchers have proposed varieties of text feature selection methods, which include document frequency (DF), information gain (IG), mutual information (MI), expected cross-entropy (ECE), chi-square (CHI), and Gini Index. Many papers have proved that IG and CHI work best for text classification [7, 8]. Because the CHI statistical algorithm has many advantages, such as low complexity, easy understanding, and significant classification effect, people use it widely in actual situations. However, CHI algorithm only considers the influence of the document frequency in the calculation process, but ignores the factor of the feature word frequency; this defect is known as the "low word frequency defect"; many researchers have proposed improved methods for CHI algorithm. This paper also

proposes an improvement in the CHI algorithm based on the variance (Var) and coefficient of variation, which named Var-CV-CHI algorithm in this paper.

The rest of this paper is organized as follows: Section 2 briefly reviews the related work; Section 3 presents the method of improved CHI algorithm; Section 4 shows the experimental results and analyzes the results; Section 5 concludes the paper with some possible work in the future.

2. Related Work

Feature selection aims at removing irrelevant or redundant features for solving a supervised learning problem and reducing the difficulty of learning tasks [9]. It has been widely studied in web document processing (text classification, text retrieval, text recovery, etc.), gene analysis, drug diagnosis, and other fields.

Feature selection has a significant impact on the results of text classification, where it is often used to find the smallest subset of features that maximally increases the performance of the model. An excellent feature selection model can optimize the algorithm and improve the performance of text classification. Facing increasingly diverse forms of data or data flow, how to design better feature selection algorithms to solve the needs of society is a long-term task.

Machine learning divides feature selection into three methods: embedded, wrapper, and filter methods.

In the embedded feature selection method, some machine learning algorithms and models are first used for training to obtain the weight coefficients of each feature, and the features are selected from large to small according to the coefficients. The wrapper method selects or excludes several features at a time based on the objective function (usually the predictive effect score). The filter method scores each feature according to its divergence or correlation, sets the threshold, and selects the features. Filter-based feature selection is an effective solution to improve the performance of classification systems by selecting significant features and discarding the undesirable ones [10]. The features with a score higher than the threshold were selected, or the features with the most significant score in the first k scores were selected. Accurately, calculate the divergence of each feature, remove the features with divergence less than the threshold, and select the features with the most significant first k fractions; calculate the correlation between each feature and the label, remove the features whose correlation is less than the threshold, and select the feature with the most significant first k scores. The classical filter feature selection methods are variance threshold, chi-square, MI, etc. Moreover, chi-square performs best.

Many researchers propose improved methods based on the chi-square feature selection method. The related works are as follows.

Fan et al. [11] introduced word frequency factor, interclass variance, and adaptive scaling factor into CHI, which significantly improved the effect of text classification. Pei [12] introduced factors such as dispersion, concentration, and word frequency into the CHI formula, which improved its classification accuracy on the unevenly distributed corpus. Qiu et al. [13] proposed the variance-based chi-square statistics (Var-CHI) method which has significantly improved the recall and precision. Bahassine et al. [14] proposed an Imp-CHI algorithm, which combined the total number of documents in the corpus with the number of documents belonging to a particular category for Arabic text classification. However, they all did not consider using the variance (Var) of total word frequency in each class and coefficient of variation (CV) within a class to measure the dispersion of feature words in the category.

3. Materials and Methods

3.1. Chi-Square Feature Selection. The CHI algorithm's core idea is hypothesis testing; by observing the deviation between the actual value and the theoretical value, we can judge the correlation between the feature words and the category. In the text classification process, the original hypothesis is that "feature word t is not related to class C_j," and each feature word can get the corresponding CHI value. The larger the CHI value, the more significant the correlation between t and C_j. According to the order of CHI values from large to small, we can select m feature words that have the most significant correlation with class C_j.

The CHI value is defined as the following equation:

$$\chi^2\left(t, C_j\right) = \frac{N \times (AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)}, \quad (1)$$

where N is the total number of documents in the corpus; A is the number of documents that contain feature t and belong to class; B is the number of documents that contain feature t but do not belong to class; C is the number of documents that do not contain feature t but belong to class; D is the number of documents that do not contain feature t and not belong to class.

According to equation (1), each feature word gets a score, and then the first m feature words are selected.

3.2. Improved Chi-Square Algorithm. Yang's research shows that chi-square is one of the best current feature selection methods [8]. Compared with other methods, the CHI method can reduce more vocabulary and eliminate a lot of redundant words and then improve the classification performance. Moreover, with the amount of text gradually increasing, the stability is also considerable. Since $A + C$ is equal to the number of documents belonging to category C_j, $B + D$ is equal to the number of documents not belonging to category C_j, both of which are constant and not related to the feature word t, so does N. Therefore, equation (1) can be written as the following equation:

$$\chi^2\left(t, C_j\right) = \frac{(AD - BC)^2}{(A + B)(C + D)}. \quad (2)$$

In equation (2), A, B, C, and D all come from the statistics of the number of documents containing feature word t_k in each category of the corpus, so the CHI value is only related to whether the feature words appear, but not related

to the frequency of its appearance. This problem leads to incomplete classification information; it is easy to select low-frequency words that are not useful for classification falsely. Especially when the distribution of corpus categories is uneven, equation (2) is more likely to give high scores to low-frequency words, which affects the accuracy of classification results. To correct these shortcomings of the traditional CHI algorithm, this paper proposes an improved algorithm, Var-CV-CHI algorithm. The method is as follows:

(a) In the text classification, we want to select feature words that meet the condition of $AD - BC > 0$ as many as possible. However, according to equation (2), the feature words satisfying $AD - BC < 0$ will be selected at the same time, so the low-frequency words in the category will be selected by mistake; we call it as a negative correlation. Some low-frequency words contain important classification information, so we cannot completely delete them. Therefore, this paper will select feature words according to the positive and negative correlation between feature words and categories and give them different weights to reduce the negative correlation of low-frequency words. The positive correlation and negative correlation are as shown in the following equations:

$$\chi^2\left(t, C_j\right)^+ = \frac{(AD - BC)^2}{(A + B)(C + D)}, \quad AD - BC > 0, \tag{3}$$

$$\chi^2\left(t, C_j\right)^- = \frac{(AD - BC)^2}{(A + B)(C + D)}, \quad AD - BC < 0. \tag{4}$$

Thus, add an adaptive scaling factor α. The equation is improved as the following equation:

$$\chi^2\left(t, C_j\right) = \alpha \times \chi^2\left(t, C_j\right)^+ + (1 - \alpha) \times \chi^2\left(t, C_j\right)^-, \tag{5}$$

($\alpha \in (0.5, 1)$. In the experiment of this paper, $\alpha$is 0.8).

(b) As described above, in the original CHI equation, it does not count t_k's frequency. When multiclass texts are classified, the larger the number of documents containing the feature word t_k and t_k's frequency in category C_j, the better the t_k's classification ability. Therefore, this paper proposes these parameters as the following equations:

$$\beta = \frac{A^2}{(B + 1)}, \tag{6}$$

$$\gamma = \frac{\sum_{i=1}^{n} tf\left(t_k, d_{ij}\right)}{\sum_{j=1}^{r}\sum_{i=1}^{n} tf\left(t_k, d_{ij}\right) + 1}, \tag{7}$$

where $tf(t_k, d_{ij})$ is the number of t_k in the ith file in class C_j, r is the total number of news categories, and n is the total number of documents in a category.

(c) If feature words t_k have good classification ability, they should mostly appear in one category. On the contrary, they seldom or even do not appear in other categories. Therefore, the greater the dispersion of t_k's frequency between categories, that is, the larger the variance between categories, the stronger the classification ability of the feature words t_k. Thus, this paper gives a parameter ξ_1 as the following equation:

$$\xi_1 = \left| \sum_{i=1}^{n} tf\left(t_k, d_{ij}\right) - \frac{\sum_{j=1}^{r}\sum_{i=1}^{n} tf\left(t_k, d_{ij}\right)}{r} \right|^2. \tag{8}$$

$tf(t_k, d_{ij})$'s meaning is the same as equation (7); $\sum_{j=1}^{r}\sum_{i=1}^{n} tf(t_k, d_{ij})/r$ represents the category average word frequency of t_k in the corpus. If $\sum_{i=1}^{n} tf(t_k, d_{ij}) - (\sum_{j=1}^{r}\sum_{j=1}^{n} f(t_k, d_{ij})/r) \geq 0$, it indicates that the word frequency of t_k in the specified category C_j is greater than or equal to the average word frequency of t_k between categories, and it is a meaningful word for C_j; if $\sum_{i=1}^{n} tf(t_k, d_{ij}) - (\sum_{j=1}^{r}\sum_{i=1}^{n} tf(t_k, d_{ij})/r) < 0$, it indicates that the word frequency of t_k in the specified category C_j is less than the average word frequency of t_k between categories, and it is a meaningless word for C_j. According to the mathematical meaning of ξ_1, $\xi_1 \geq 0$ is constantly established; so, if ξ_1 is very large and $\sum_{i=1}^{n} tf(t_k, d_{ij}) - (\sum_{j=1}^{r}\sum_{i=1}^{n} tf(t_k, d_{ij})/r) > 0$, then it shows that t_k has important classification significance to C_j. In order for the formula to be neat and beautiful, command $\sum_{i=1}^{n} tf(t_k, d_{ij}) - (\sum_{j=1}^{r}\sum_{i=1}^{n} tf(t_k, d_{ij})/r) = s$, and adjust parameter ξ_1 to ξ. The value of ξ is shown as the following equation:

$$\xi = \begin{cases} \sum_{i=1}^{n} tf\left(t_k, d_{ij}\right) - \sum_{j=1}^{r}\sum_{i=1}^{n} tf\left(t_k, d_{ij}\right), & s \geq 0, \\ 0, & s \leq 0. \end{cases} \tag{9}$$

(d) Coefficient of variation is a normalization index of the degree of dispersion of a probability distribution, which is defined as the following equation:

$$\mathrm{CV} = \frac{\sigma}{\overline{x}} \times 100\%. \tag{10}$$

In the text classification, feature words are distributed in every text of a category as evenly as possible, which indicates that it has a strong classification ability. The smaller the coefficient of variation, the more uniform the data distribution. To make the improved CHI value proportional to each parameter, we reverse the numerator and denominator of the coefficient of variation and then transform the standard deviation into a variance. Therefore, we introduce the coefficient of variation for chi-square, which is defined as the following equation:

$$\varepsilon = \left(\frac{\mu}{\sigma^2 + 1}\right) = \frac{\sum_{i=1}^{n} tf\left(t_k, d_{ij}\right)/n}{\left|tf\left(t_k, d_{ij}\right) - \left(\sum_{i=1}^{n} tf\left(t_k, d_{ij}\right)/n\right)\right|^2 + 1}. \tag{11}$$

In summary, the feature selection segment should select words that appear concentrated in a specific category and are evenly distributed in that category. At the same time, the more frequently these words are in this specific category, the better their category representativeness. This paper proposes an improved CHI algorithm, which is called Var-CV-CHI algorithm; the calculation method is as the following equation:

$$\chi^2\left(t_k, C_j\right) = \beta \times \gamma \times \xi \times \varepsilon \times \left[\alpha \times \chi^2\left(t_k, C_j\right)^+ + (1 - \alpha) \times \chi^2\left(t_k, C_j\right)^-\right]. \tag{12}$$

4. Results and Discussion

In this experiment, Chinese news corpus and English news corpus are both used. Because of the different language structures of Chinese and English, we construct Chinese text classification system and English text classification system, respectively. The specific experimental settings and results are as follows.

4.1. Data Collection

4.1.1. Chinese News Corpus. The Chinese corpus is from School of Information Management, Sun Yat-sen University. The corpus is a manually labeled corpus with 14 categories and a total of 36865 documents with high accuracy. In order to make the sample size large enough, this paper selects 10 categories with a large amount of text. The experiments use two corpus sets: balanced corpus (Table 1) and unbalanced corpus (Table 2); same 10 categories are selected in total.

4.1.2. English News Corpus. The English corpus is the Reuters-21578 News Corpus. According to the preliminary experiment results, because this corpus is skewed data, the effect of the Var-CV-CHI algorithm on the English balanced corpus is not significantly improved compared with the traditional CHI algorithm, so only the unbalanced corpus is used in this experiment (Table 3).

4.2. Experiment. In this paper, the Var-CV-CHI algorithm is compared with the traditional CHI and other improved algorithms on the balanced corpus or the unbalanced corpus. To exclude the influence of the classifier, we choose to use two kinds of classifiers for experiments, KNN classifier and Naïve Bayes classifier.

Figure 1 shows the experimental process.

In the experiments, the ratio of the train set and test set is 4 : 1. The purpose of CHI feature selection is to select the first m feature words based on the calculated CHI value. According to the size of the dataset, the threshold value of feature words selected from each category is 150 in Chinese corpus and 20 in English corpus. To verify the effectiveness and stability of the Var-CV-CHI algorithm, we set up the following experiments.

Table 1: The number of Chinese balanced corpora.

Category	Number
Culture	1500
Education	1500
Entertainment	1500
Event	1500
Finance	1500
Game	1500
Health	1500
Occultism	1500
Sport	1500
Technology	1500

Table 2: The number of Chinese unbalanced corpora.

Category	Number
Culture	150
Education	610
Entertainment	1050
Event	800
Finance	1200
Game	840
Health	300
Occultism	300
Sport	700
Technology	1500

Table 3: The number of English unbalanced corpora.

Category	Number
Acq	460
Crude	80
Earn	645
Interest	218
Ship	95
Student	105
Trade	300

4.2.1. Experimental in Chinese Corpus

(1) Experiment 1. Tables 4 and 5 show the experimental results of a balanced corpus and unbalanced corpus in the KNN classifier, which contains a precision, recall, and F1-score.

After the contrast experiment, it is found that when $k = 7$, the classification effect on the KNN classifier is the best. Figure 2 shows the F1 value of the classifier when k takes different values.

From Tables 4 and 5, compared with the traditional CHI algorithm, even if the performance of a few categories decline, for example, the F1-value of education and event on the unbalanced corpus drops by 2%, but does not affect the improvement of the overall corpus classification performance. In a word, whether in the balanced corpus or the unbalanced corpus, the Var-CV-CHI algorithm performed

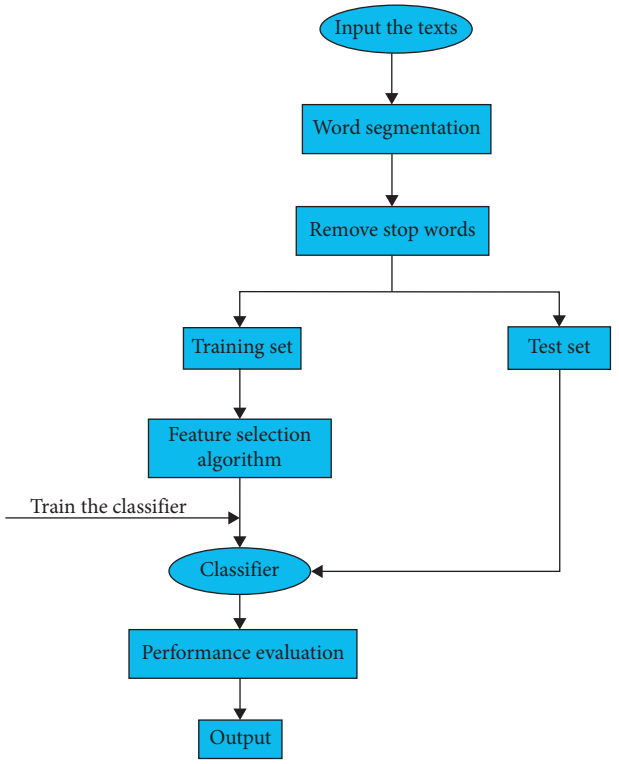

Figure 1: The process of feature selection in text classification.

Table 4: Comparison of classification results between CHI and Var-CV-CHI on the balanced Chinese corpus (KNN).

Category	CHI			Var-CV-CHI		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Culture	0.92	0.95	0.93	0.94	0.94	0.94
Education	0.94	0.93	0.94	0.93	0.92	0.92
Entertainment	0.94	0.88	0.91	0.93	0.87	0.90
Event	0.82	0.82	0.82	0.75	0.85	0.80
Finance	0.82	0.88	0.85	0.87	0.89	0.88
Game	0.98	0.97	0.98	0.98	0.97	0.97
Health	0.92	0.78	0.84	0.95	0.82	0.88
Occultism	0.77	0.98	0.86	0.83	1.00	0.90
Sport	0.98	0.92	0.95	0.98	0.91	0.95
Technology	0.92	0.84	0.88	0.93	0.86	0.89
Macro-avg-P		0.90			0.91	
Macro-avg-R		0.90			0.90	
Macro-avg-F1		0.90			0.90	

well. Macro-avg-P and macro-avg-F1 are 3% and 2%, respectively, on the unbalanced corpus. The traditional CHI algorithm performs better on the balanced corpus, while the Var-CV-CHI algorithm performs well on both the balanced corpus and the unbalanced corpus. Figures 3 and 4 show the comparison of the macroresults of CHI and Var-CV-CHI.

(2) Experiment 2. Tables 6 and 7 show the experimental results of the balanced corpus and unbalanced corpus in the Naïve Bayes classifier.

From Tables 6 and 7, the results on the unbalanced corpus are better than the balanced corpus, in which macro-avg-P, macro-avg-R, and macro-avg-F1 are increased by 8%, 6%, and 9%, respectively; it is a significant improvement compared with traditional CHI algorithm. The experiments show a fact: the classification performance of a balanced corpus is better on the KNN classifier, while an unbalanced corpus is better on the Naïve Bayes classifier. The Var-CV-CHI algorithm is better than the CHI algorithm in any situation and shows excellent performance. Figures 5 and 6 show the comparison of the macroresults of CHI and Var-CV-CHI.

Table 5: Comparison of classification results between CHI and Var-CV-CHI on the unbalanced Chinese corpus (KNN).

Category	CHI			Var-CV-CHI		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Culture	0.97	0.90	0.93	0.93	0.92	0.92
Education	0.93	0.82	0.87	0.93	0.87	0.90
Entertainment	0.92	0.91	0.91	0.91	0.93	0.92
Event	0.79	0.80	0.80	0.79	0.80	0.80
Finance	0.86	0.84	0.85	0.88	0.88	0.88
Game	0.97	0.97	0.97	0.97	0.98	0.97
Health	0.51	0.92	0.66	1.00	0.68	0.81
Occultism	0.64	1.00	0.78	0.47	1.00	0.64
Sport	0.97	0.86	0.91	0.96	0.84	0.90
Technology	0.93	0.79	0.86	0.94	0.87	0.91
Macro-avg-P		0.85			0.88	
Macro-avg-R		0.88			0.88	
Macro-avg-F1		0.85			0.87	

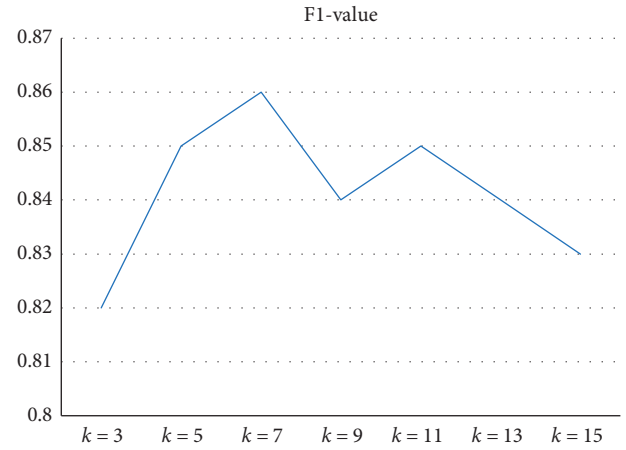

Figure 2: Classification effect under different k values.

4.2.2. Experimental in English Corpus

(1) Experiment 3. Tables 8 and 9 show the experimental results of the English unbalanced corpus using the Naïve Bayes classifier and KNN classifier.

From Tables 8 and 9, it can be seen that the classification results of the Var-CV-CHI are significantly improved in both KNN classifier and Naïve Bayes classifier, and it works best on the KNN classifier, in which macro-avg-P, macro-avg-R, and macro-avg-F1 are increased by 2%, 7%, 5%, respectively. On the Naïve Bayes classifier, the category of

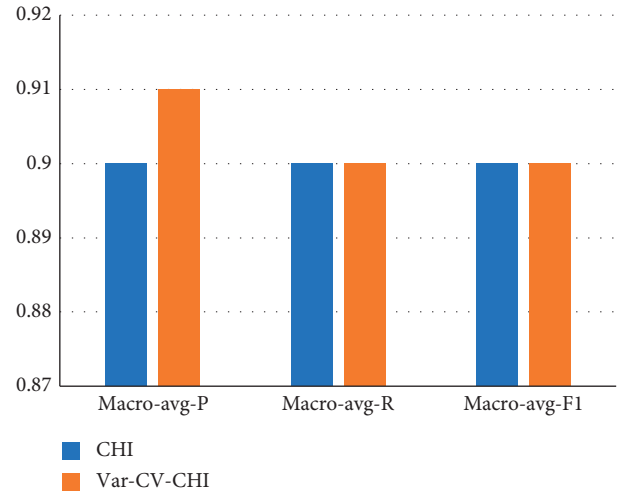

Figure 3: Comparison of macroresults between CHI and Var-CV-CHI on the balanced Chinese corpus (KNN).

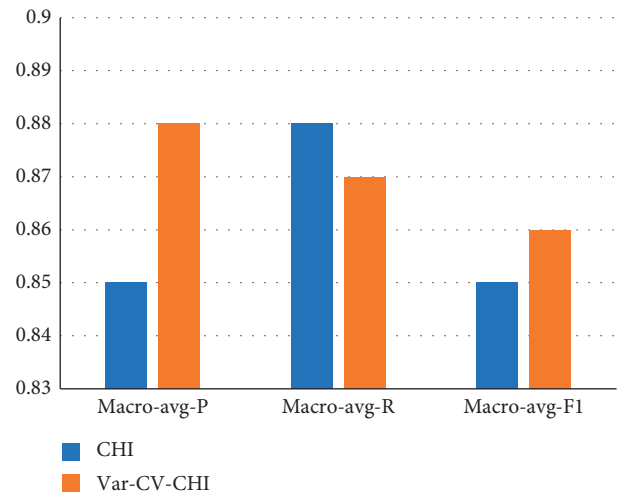

Figure 4: Comparison of macroresults between CHI and Var-CV-CHI on the unbalanced Chinese corpus (KNN).

Table 6: Comparison of classification results between CHI and Var-CV-CHI on the balanced Chinese corpus (Naïve Bayes).

Category	CHI			Var-CV-CHI		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Culture	0.99	0.57	0.72	0.99	0.56	0.72
Education	0.92	0.89	0.91	0.90	0.94	0.92
Entertainment	0.71	0.95	0.82	0.72	0.94	0.82
Event	0.65	0.83	0.73	0.73	0.84	0.78
Finance	0.83	0.73	0.78	0.83	0.82	0.83
Game	0.98	0.94	0.96	0.98	0.94	0.96
Health	0.95	0.91	0.93	0.95	0.93	0.94
Occultism	0.92	0.99	0.96	0.97	1.00	0.98
Sport	0.97	0.97	0.97	0.98	0.97	0.97
Technology	0.88	0.86	0.87	0.89	0.88	0.89
Macro-avg-P		0.88			0.89	
Macro-avg-R		0.86			0.88	
Macro-avg-F1		0.86			0.88	

Table 7: Comparison of classification results between CHI and Var-CV-CHI on the unbalanced Chinese corpus (Naïve Bayes).

Category	CHI			Var-CV-CHI		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Culture	0.96	0.84	0.89	0.96	0.84	0.90
Education	0.75	0.88	0.81	0.92	0.98	0.95
Entertainment	0.94	0.89	0.91	0.93	0.94	0.94
Event	0.68	0.62	0.65	0.73	0.80	0.77
Finance	0.85	0.69	0.76	0.86	0.83	0.85
Game	0.97	0.95	0.96	0.98	0.95	0.96
Health	0.38	1.00	0.55	0.87	1.00	0.93
Occultism	0.92	1.00	0.96	0.95	1.00	0.98
Sport	0.95	0.88	0.92	0.96	0.94	0.95
Technology	0.93	0.83	0.88	0.94	0.90	0.92
Macro-avg-P		0.83			0.91	
Macro-avg-R		0.86			0.92	
Macro-avg-F1		0.83			0.92	

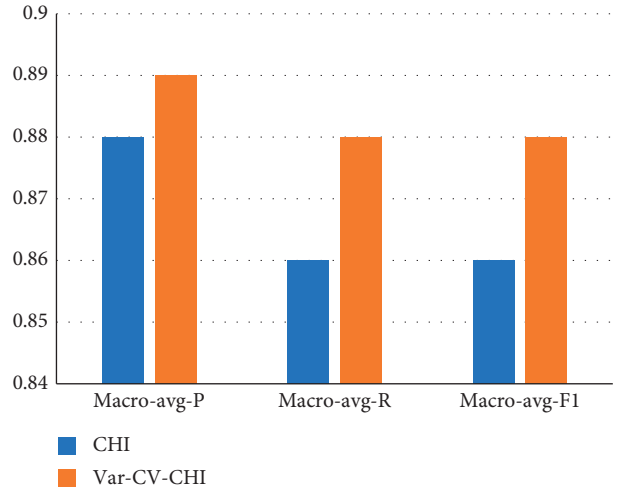

Figure 5: Comparison of macroresults between CHI and Var-CV-CHI on the balanced Chinese corpus (Naïve Bayes).

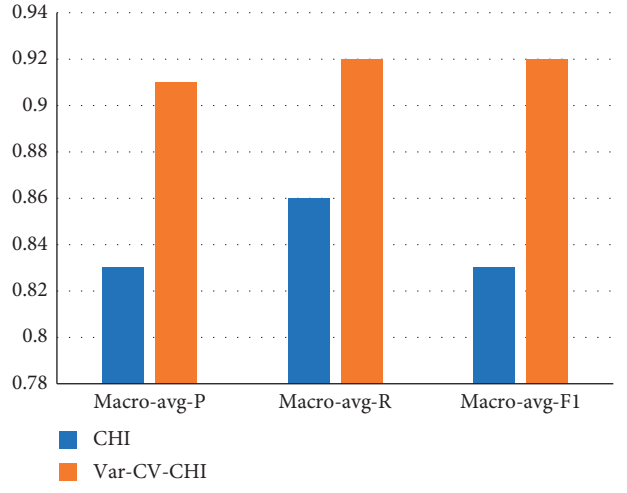

Figure 6: Comparison of macroresults between CHI and Var-CV-CHI on the unbalanced Chinese corpus (Naïve Bayes).

Table 8: Comparison of classification results between CHI and Var-CV-CHI on the unbalanced English corpus (KNN).

Category	CHI			Var-CV-CHI		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Acq	0.73	0.86	0.79	0.73	0.85	0.79
Crude	0.83	0.59	0.69	0.89	0.94	0.91
Earn	0.87	0.89	0.88	0.90	0.87	0.88
Interest	0.90	0.82	0.86	0.93	0.86	0.89
Ship	0.92	0.55	0.69	0.94	0.75	0.83
Student	1.00	1.00	1.00	1.00	0.91	0.95
Trade	0.93	0.89	0.91	0.95	0.90	0.92
Macro-avg-P		0.88			0.90	
Macro-avg-R		0.80			0.87	
Macro-avg-F1		0.83			0.88	

Table 9: Comparison of classification results between CHI and Var-CV-CHI on the unbalanced English corpus (Naïve Bayes).

Category	CHI			Var-CV-CHI		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Acq	0.76	0.84	0.80	0.80	0.84	0.82
Crude	0.78	0.82	0.80	0.84	0.94	0.89
Earn	0.93	0.85	0.89	0.93	0.88	0.90
Interest	0.95	0.91	0.93	0.93	0.93	0.93
Ship	0.85	1.00	0.92	0.92	1.00	0.96
Student	0.73	0.80	0.76	0.88	0.75	0.81
Trade	0.93	0.87	0.89	0.92	0.95	0.94
Macro-avg-P		0.85			0.89	
Macro-avg-R		0.87			0.90	
Macro-avg-F1		0.86			0.89	

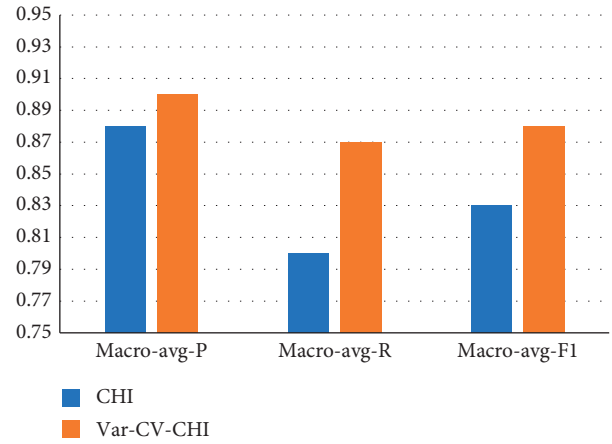

Figure 7: Comparison of macroresults between CHI and Var-CV-CHI on the unbalanced English corpus (KNN).

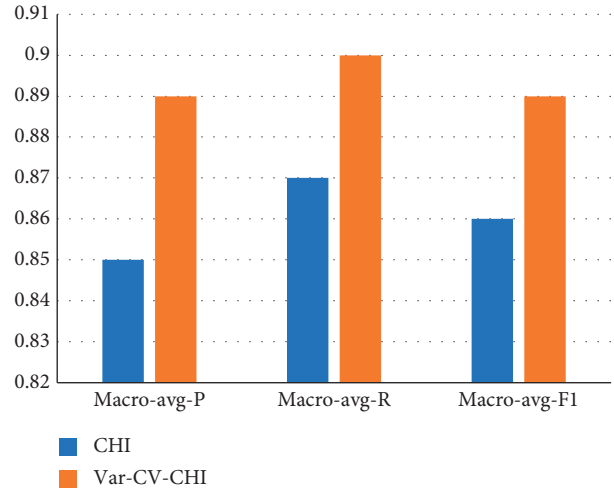

Figure 8: Comparison of macroresults between CHI and Var-CV-CHI on the unbalanced English corpus (Naïve Bayes).

Crude's classification effect of the Var-CV-CHI algorithm is especially improved significantly, and the P, R, and F are increased by 6%, 24%, and 18%, respectively. It can be seen from Table 3 that the number of files in the Crude category is the least among the 7 categories, so it can be seen that in the unbalanced corpus, the Var-CV-CHI's classification effect on the categories with the least number of files is the most obvious. In addition, through multiple experiments, we found that the more the uneven number of categories in the corpus, the better the classification results of the Var-CV-CHI algorithm. Figures 7 and 8 show the comparison of the macro-average results of CHI and Var-CV-CHI.

5. Conclusion

In this paper, through a lot of literature studies, and on this basis, it is determined that the introduction of new parameters will be the direction of CHI algorithm improvement, thus the Var-CV-CHI feature selection algorithm based on variance and coefficient of variation is proposed. Although many of the improved methods of the CHI algorithm are based on the idea of introducing new parameters, the method in this paper is more comprehensive and rigorous than these methods. It combines the differences caused by the different distributions of feature words between categories and within categories and synthesizes the situation of feature words in the document. The classification effect is very excellent. At the same time, compared with the existing works, the experiments in this paper are more abundant. Not only the language factor but also the distributions of classifiers and datasets are considered. Finally, the experiments from multiple angles proved the effectiveness of the Var-CV-CHI algorithm.

Text classification is a complex project, and every segment of it plays a vital role in the classification results. Future work can continue in the following directions:

The Var-CV-CHI algorithm in this paper only considers the distribution of feature words but without considering its semantic information. If the semantic information of feature items can be added to the feature selection algorithm, for

example, the synonyms with the same semantic information can be merged, and the feature words with the largest weight among the words with similar semantic information can be selected. Then, the redundancy of feature space will be reduced, and the performance of feature selection will be improved significantly.

Data Availability

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

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The Chinese corpus of this article was provided by Lu Yonghe from the School of Information Management, Sun Yat-sen University.

References

[1] A. K. Uysal, "On two-stage feature selection methods for text classification," *IEEE Access*, vol. 6, pp. 43233–43251, 2018.

[2] Q. Zhou, M. S. Zhao, and H. U. Min, "Study on feature selection in Chinese text categorization," *Journal of Chinese Information Processing*, vol. 18, no. 3, pp. 17–23, 2004.

[3] Y. Zhao, S. Cheng, X. Yu, and H. Xu, "Chinese public's attention to the COVID-19 epidemic on social media: observational descriptive study," *Journal of Medical Internet Research*, vol. 22, no. 5, 2020.

[4] M. Oleynik, A. Kugic, Z. Kasáč, and M. Kreuzthaler, "Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1247–1254, 2019.

[5] H. Liu, S. U. Zhan, and S. Liu, "Improved CHI text feature selection based on word frequency information," *Computer Engineering and Applications*, vol. 49, no. 22, pp. 110–114, 2013.

[6] X.-T. Wang and X.-Z. Luan, "Bayesian penalized method for streaming feature selection," *IEEE Access*, vol. 7, no. 99, pp. 103815–103822, 2019.

[7] H. Ji, A. H. Tan, and C. L. Tan, "On machine learning methods for Chinese document categorization," *Applied Intelligence*, vol. 18, no. 3, pp. 311–322, 2003.

[8] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, vol. 39, pp. 42–49, New York, NY, USA, August 1999.

[9] B. Li, "Importance weighted feature selection strategy for text classification," in *Proceeding of the 2016 International Conference on Asian Language Processing (IALP)*, Tainan, Taiwan, November 2016.

[10] O. A. M. Salem, F. Liu, Y.-P. P. Chen, and X. Chen, "Ensemble fuzzy feature selection based on relevancy, redundancy, and dependency," *Entropy*, vol. 22, no. 7, 2020.

[11] C. J. Fan, Y. S. Wang, and Y. T. Wang, "An improved CHI text feature selection algorithm," *Computer and Modernization*, vol. 11, pp. 7–11, 2016.

[12] Y. Pei, "Study on improved CHI for feature selection in Chinese text categorization," *Computer Engineering and Applications*, vol. 47, pp. 128–123, 2011.

[13] Y. F. Qiu, W. Wang, and D. Y. Liu, "CHI feature selection method based on variance," *Application Research of Computers*, vol. 29, pp. 1304–1306, 2012.

[14] S. Bahassine, A. Madani, and M. Kissi, "An improved Chi-sqaure feature selection for Arabic text classification using decision tree," in *Proceeding of the International Conference on Intelligent Systems: Theories & Applications*, Mohammedia, Morocco, October 2016.