Text Classification Based on Improved Information Gain Algorithm and Convolutional Neural Network

Mengjie Dong(✉), Huahu Xu, and Qingguo Xu

School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
1151166299@qq.com, huahuxu@163.com, qgxu@t.shu.edu.cn

Abstract. Feature selection is an important step. It aims to filter some irrelevant features, improve the classifier speed and also reduce the interference during text classification process. Information gain (IG) feature selection algorithm is one of the most effective feature selection algorithms. But it is easy to filter out the characteristic words which have a low IG score but have a strong ability of text type identification. Because IG algorithm only considers the number of documents of feature items in each category. Aiming at this defect, we propose an improved information gain algorithm by introducing three parameters: intra-class word frequency, inter-class separation degree and intra-class dispersion degree. Then, the improved IG algorithm is used for feature selection, and important feature words with high IG value are selected according to the threshold value. Final, the important feature words in the text are expressed as two-dimensional word vectors and input into Convolutional Neural Network (CNN) to train and classify them. Therefore, a text classification model based on improved information gain and convolutional neural network is proposed and abbreviated as "I-CNN". Through experiments, we achieve good experimental results in THUCNews Chinese text classification corpus. Experimental results prove that the improved IG algorithm is better than the traditional feature selection algorithm.

Keywords: Text classification · Feature selection · Information gain · Convolutional Neural Network

1 Introduction

With the rapid development of Internet and cloud computing technology, the scale of data grows exponentially. There is a lot of important information hidden behind massive data. Facing the massive data, how to extract the key and effective information is the current research hotspot [1]. At present, most of the mainstream text representation is based on the VSM. But usually, the dimension of the vector space is very high, which can reach 10^5 for Chinese corpus. And a text usually contains about 10^3 words. It can be seen that the original vector space has the shortcomings of high dimensionality and sparseness, which will seriously affect the classification accuracy of common classifiers. Feature selection can solve the high dimensionality of text

H. Gao et al. (Eds.): TridentCom 2019, LNICST 309, pp. 184–198, 2020.
https://doi.org/10.1007/978-3-030-43215-7_13

representation and select a group of features from the feature set. Meanwhile, it can best express the meaning of the text without losing important information items.

Common feature selection algorithms include Document Frequency (DF), Information Gain (IG), Chi-square Test (CHI) and Mutual Information (MI). However, these algorithms perform in Chinese text not as well as in English text classification, because Chinese text has a higher feature space dimension and word correlation compared with English text. Research shows that theoretical information gain is the best feature selection method [2]. However, owing to the shortage of considering word frequency information of documents in class, the classification effect is not ideal. Literature [3] proposed that characteristics of different categories of data sets should be selected first. Then, different categories characteristics would be optimized and merged. Final, through the appearance of merged features, IG weights should be introduced. Literature [4] uses information entropy and information gain rate respectively as the heuristic information, and proposes an attribute reduction algorithm based on ACO. But it is easy for both to add redundant attributes to the reduced set as selected attributes. Ming [5] introduced the equalization ratio and intra-class word frequency position parameters. His algorithm solved the weakening classification and selection defect of local features problems caused by the traditional IG algorithm with ignoring word frequency distribution.

With the development of deep learning, Convolutional Neural Networks (CNN) have achieved great success in the field of image recognition. Kim [6] proposed that applying CNN to the field of text classification not only improves the classification efficiency but also makes the classification effect better than the traditional classification model. The greatest feature of CNN is that it can automatically extract features and share weights, while the traditional machine learning model with huge parameters needs to manually extract features. Using CNN as the classification model will greatly reduce the training time and improve the classification effect.

Based on the study of traditional IG algorithm and existing improved algorithm, a new improved IG algorithm is proposed in this paper. First, the improved IG algorithm should take into account the word frequency information to select words with high IG value and frequency. These words will be used as the important feature words in the classification. Second, representative feature words should be concentrated in a certain category. For example, “machine learning”, “artificial intelligence” and other words that clearly represent IT texts should be concentrated in the IT category instead of sports, entertainment or other categories. This is called “inter-class separation degree”. Furthermore, representative feature words such as “machine learning” should be evenly distributed in the IT class instead of only appearing in a few documents. This is called “intra-class dispersion degree”. From here we see that, the representative feature words should have a larger degree of separation between classes and a smaller degree of dispersion within classes. Therefore, this paper introduces three parameters: word frequency information of feature items, inter-class separation degree and intra-class dispersion degree. After the feature extraction of the improved IG algorithm, the text is represented as a two-dimensional matrix which is similar to pictures and input into CNN for training. This model (I-CNN for short) can better extract important feature items and improve the classification effect.

The rest of this article is arranged as follows. In Sect. 2, we will describe the information gain algorithm and CNN classification model briefly. Section 3 proposes the improved information gain algorithm. Section 4 introduces the training steps of the proposed model. Section 5 is the experiment and result analysis. Final, there is a summary in Sect. 6.

2 Related Work

2.1 Information Gain Algorithm

Information gain [7] refers to the difference of information entropy, that is, the difference of information entropy whether each feature item appears in the text or not. The larger the information gain, the more important the feature item is in the text. IG algorithm is shown in formula (1):

$$IG(t) = H(c) - H(c|t) = -\sum_{j=1}^{m} p(c_j) \times \log p(c_j) + p(t) \sum_{j=1}^{m} p(c_j|t) \log p(c_j|t) + p(\bar{t}) \sum_{j=1}^{m} p(c_j|\bar{t}) \log p(c_j|\bar{t}) \quad (1)$$

Where c_j means the category attribute, $p(c_j)$ is the probability of the ith class value, $p(t)$ is the probability that feature t occurs, and $p(\bar{t})$ is the probability that feature t does not occur, $p(c_j|t)$ is the conditional probability that the class belongs to c_j when the feature t is included, while $p(c_j|\bar{t})$ is the conditional probability that the class belongs to c_j when the feature t is not included. m represents the number of categories.

2.2 Convolutional Neural Network

Convolutional Neural Network (CNN) is a feedforward neural network composed of various combinations of convolutional layer, pooling layer (also known as the subsampling layer) and full connection layer. The spatial local correlation can be utilized by implementing a local connection mode between neurons in adjacent layers. At present, CNN has been widely used in image understanding, computer vision, language recognition, natural language processing and other fields [8]. Generally, CNN consists of one or more pairs of convolution layers and pooling layers and ends up with a completely connected neural network. The typical convolutional neural network structure [9] is shown in Fig. 1.

As we can see from Fig. 1, each neuron in the convolution layer is locally connected to the input layer. Convolution operation is the weighted sum of the convolution kernel parameters and the corresponding local input, then plus the offset value. The value will input to the activation function, and the output of the activation function is the value of the node in the next layer. This process is equivalent to the convolution process, hence the name of CNN [10]. The size and number of convolution kernels

need to be customized. Each element on the convolution kernel is corresponded to a weight coefficient *w* and a bias vector.

The pooling layer carries out a subsampling operation on the output feature graph of the convolution layer. It can compress the number of data and parameters, reduce overfitting and improve the fault tolerance and training speed of the model. Common pooling methods include Max Pooling and Average Pooling [11]. Max Pooling is taking the maximum value point in the local acceptance domain. Average Pooling is calculating the mean value of all values in the local acceptance field.

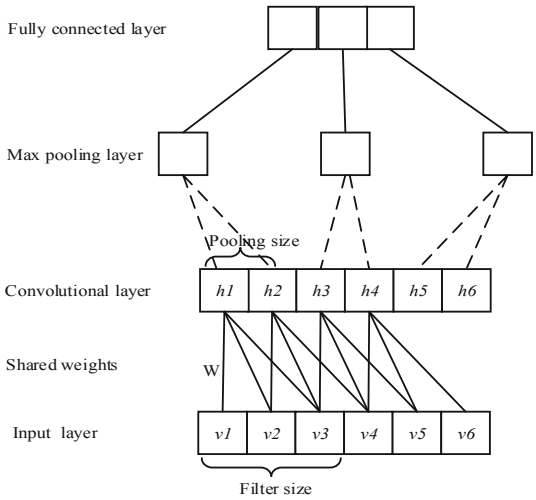

Fig. 1. Structure diagram of convolutional neural network.

The full connection layer can integrate the local information in the convolution layer or pooling layer. Then connecting all the local features and sending them to the classifier for classification [8].

CNN adopts local connection and weight sharing technology, which can not only extract feature information better, but also reduce network parameters and facilitate model training.

3 Improved Information Gain Algorithm

3.1 Disadvantage of IG

The IG value calculated by traditional algorithm only considers the document frequency of feature words and ignores the importance of word frequency information to classification. Therefore, it is easy to filter out the words with low IG value but high occurrence frequency and strong text recognition ability. To explain this, we assume that there are two feature words, w_1, w_2, and two document categories, c_1 and c_2. The document and frequency of feature words are shown in Table 1.

The feature words w_1 and w_2 in the table appear in the three documents D_1, D_2 and D_5. Calculated according to formula (1), $p(w_1) = p(w_2) = 3/8$, $p(\overline{w_1}) = p(\overline{w_2}) = 5/8$. The conditional entropy is $p(c_1|w_1) = p(c_1|w_2) = 2/3$, $p(c_1|\overline{w_1}) = p(c_1|\overline{w_2}) = 2/5$, $p(c_2|w_1) = p(c_2|w_2) = 1/3$, $p(c_2|\overline{w_1}) = p(c_2|\overline{w_2}) = 3/5$. The IG values calculated by formula (1) are the same, but the occurrence frequency of w_1 in the same document is significantly higher than that of w_2, which is more representative and IG values should be larger. The calculated results are inconsistent with logic. This may lead to the loss of important feature words in the text and the selection of words with lower frequency as feature words, resulting in worse classification effect.

Table 1. The distribution of feature words w_1 and w_2.

Feature words	c_1				c_2			
	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
w_1	12	10	0	0	5	0	0	0
w_2	3	2	0	0	1	0	0	0

3.2 Improved IG by Three Parameters

Intra-class Word Frequency. The higher the occurrence frequency of feature items in a certain category y, the stronger the classification ability and the greater the weight should be. Set the feature set F = {w_1, w_2, w_3, …., w_m}. There are $d_{ik}(1 \le k \le N_i)$ texts in class $c_i(1 \le i \le n)$ in the training set, N_i is the total number of texts in class c_i, and the occurrence frequency of feature $w_j(1 \le j \le m)$ in text d_{ik} in class c_i is $tf_{ik}(w_j)$. Formula (2) uses data "min-max" standardization to linearly change the original data, so that the result value is mapped to between [0-1], and then there are weight parameters:

$$\alpha_{ij} = \sum_{i=1}^{N_i} \frac{tf_{ik}(w_j) - \min\limits_{1 \le l \le m}(tf_{ik}(w_l))}{\max\limits_{1 \le l \le m}(tf_{ik}(w_l)) - \min\limits_{1 \le l \le m}(tf_{ik}(w_l))} \tag{2}$$

Considering the difference of the number of texts in different categories, formula (2) is normalized by "z-score" method.

$$\alpha = \frac{\alpha_{ij}}{\sqrt{\sum\limits_{j=1}^{m} \alpha_{ij}^2}} \tag{3}$$

Formula (3) reflects that the larger the word frequency is within the category, the larger the weight a corresponding to the feature item is, the stronger the classification ability is.

Inter-class Separation Degree and Intra-class Dispersion Degree. If most of the feature words appear in a certain category and less in other categories, it means that the feature words can be well identified in this category. The separation degree of feature words is calculated as shown in formula (4):

$$CO(w) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i(w) - \overline{X(w)}\right)^2}}{\overline{X(w)}} \tag{4}$$

Where n represents the number of classes, $\overline{X_i(w)}$ represents the average number of features w appearing in all classes, and $X_i(w)$ represents the number of features w appearing in class c_i. From the above equation, it can be seen that the higher the value of the degree of separation between classes, the better the classification effect.

If a feature word appears in most articles of a certain category, it will be more representative of the class. The dispersion of characteristic words is calculated as shown in formula (5):

$$CI(w, c_i) = \frac{\sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(X_{ij}(w) - \overline{X_i(w)}\right)^2}}{\overline{X_i(w)}} \tag{5}$$

Where, m represents the number of documents in class c_i, $\overline{X_i(w)}$ represents the average of the occurrence of feature word w in each document in class c_i, and $X_{ij}(w)$ represents the occurrence of feature word w in the jth document in class c_i. As can be seen from the above equation, the smaller the value of dispersion within the class, the better the classification effect.

To sum up, it can be seen from the above two equations that the greater the degree of separation between classes and the smaller the degree of dispersion within classes, the better the classification effect is. Therefore, the degree of distinction of class c_i by the feature word w can be defined as shown in formula (6):

$$\beta(w, c_i) = \frac{CO(w)}{CI(w, c_i)} \tag{6}$$

3.3 Improved IG Algorithm

Through the analysis of the first two sections, we introduce word frequency parameter α and distribution factor β. Not only the word frequency of feature words, but also the influence of feature word distribution on classification are considered. A new formula for calculating the information gain value is obtained, as shown in formula (7). The higher the word frequency, the higher α. The larger the degree of separation between

classes and the smaller the degree of dispersion within classes, the larger β will be. For such words, the larger the IG value calculated by formula (7) is.

$$IG_{\text{new}}(t) = -\sum_{j=1}^{m} p(c_j) \times \log p(c_j) + (p(t) \sum_{j=1}^{m} p(c_j|t) \log p(c_j|t) + p(\bar{t}) \sum_{j=1}^{m} p(c_j|\bar{t}) \log p(c_j|\bar{t})) \times \alpha \times \beta \tag{7}$$

4 I-CNN Model Training

This section will explain the model training procedures proposed in this paper step by step. The text classification process is mainly divided into four steps: text preprocessing, feature extraction, text representation and classifier classification.

In the first step, "jieba" word segmentation tool is used in this experiment to segment the text. After word segmentation, use the stop word list to remove noise and filter out numbers, symbols, or other nonsense words from the text.

In the second step, feature extraction will use the improved IG algorithm proposed in this paper. It will set a threshold to remove the feature words with too high or too low information gain value.

The third step is that, after word segmentation and feature extraction, the text is a set of words or phrases. In order to represent the text and combine it into a two-dimensional matrix which is similar to image, word vectorization is required. "Word2vec" is an open source Google word vector generator. It maps words into low-dimensional spaces where semantically similar words are similar by modeling the context and semantic relations of words. Word2vec [12] has two models: CBOW and Skip-gram. In this paper, Skip-gram model is used to train word vectors, and the probability of context words is predicted by the central word. This model trains each word into a distributed word vector, such as: $[w_{i1}, w_{i2}, w_{i3}, \ldots, w_{ik}]$, i represents the ith word in the text, and k refers to the dimension of word vector. Thus, a text can be represented as a two-dimensional matrix A_{lk}, such as: $\begin{pmatrix} w_{11} & \cdots & w_{1k} \\ \vdots & \ddots & \vdots \\ w_{l1} & \cdots & w_{lk} \end{pmatrix}$. l stands for l words in a text. As shown in Fig. 2, assuming a short essay this after word segmentation and feature selection, extract the eight feature words {Sun, Yang, Asian, Games, winner, Chinese, gold, medal}. Each word will be trained as a 5 - dimensional word vector. The final text can be expressed as 8×5 two-dimensional matrix. After word vector processing, each short text can be expressed as a two-dimensional matrix similar to image, which can be used as the input layer of convolutional neural network.

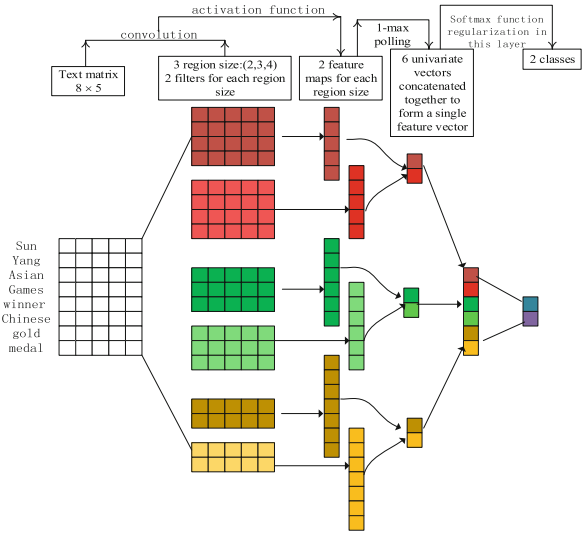

Fig. 2. Schematic diagram of model training.

The fourth step is to use convolutional neural network as classifier. The convolution kernel in the convolution layer will automatically extract more valuable features again. As shown in Fig. 2, convolution kernel with size (2, 3, 4) is set to convolve two-dimensional text matrix with two filters each. After convolution, 2 feature maps for each size. Among, ReLu [13] activation function is introduced for nonlinear processing to obtain convolution results, as shown in formula (8):

$$p = \mathrm{Re}LU(A_{lk} \cdot w + b) \tag{8}$$

Where, w is the parameter weight and b is the bias. During the training, the values of w and b are constantly adjusted until the model converges. In the experiment, the feature graph after convolution is still very large and easy to over-fitting. Therefore, a pooling function needs to be introduced. In this experiment, the maximum pooling method is adopted to sample the Feature Map obtained by the convolution layer, as shown in formula (9):

$$p' = \max\{p\} \tag{9}$$

p' is the Feature vector after pooling.

In Fig. 2, the pooled feature vectors need to enter the full connection layer to connect all the learned deep features. After several layers of full connection layers, the last layer is the softmax output layer, and you will get the probability that the text belongs to a class.

The above is the specific classification training steps in this paper. Feature selection method adopts the improved IG algorithm proposed in this paper to eliminate irrelevant and redundant features. So as to reduce the number of features and training or running

time, then improve model accuracy. The classifier selects the CNN model. Since the convolution layer and input are locally connected and the weights are Shared [14], the number of parameters that CNN needs to learn is greatly reduced. Furthermore, the parameters are set manually and randomly. The CNN model can train the parameters by itself, which greatly improves the training speed. Therefore, the classification accuracy and training time of the I-CNN model are proposed in this paper.

5 Experiments and Results

5.1 Experiment Environment

In this experiment, a personal laptop is used to verify and test the I-CNN model, which combines the improved information gain feature selection algorithm and the text classification of convolutional neural network. Details of other experimental equipment are shown in Table 2:

Table 2. Details of experimental equipment

The operating system	Windows 10 (64bit)
Central processing unit	Xeon E5-2620 v4
Memory	64G
The framework	TensorFlow
Integrated development environment	PyCharm

5.2 Experiment Preparing and Parameter Settings

According to the procedure proposed in the previous Sect. 3, some super parameters should be set before the experiment. According to practical experience, the word vector dimension is set to 300. The number and size of convolution kernel, the number of convolution layer and pooling layer all affect the final classification results. Therefore, in this experiment, in order to obtain more local feature information of text through different convolution kernels, the hyperparameters were set as: 128 convolution kernels of 3, 4 and 5 respectively, and the learning rate of 0.01 [15]. The convolution layer and the pooling layer are placed alternately in two groups and classified through a layer of full connection layer.

The optimization objective of the experiment is to minimize the loss function. We will adopt the “cross entropy loss function” commonly used in the classification model of CNN. With the optimization target, it is necessary to update the weight parameters iteratively and train the accuracy of the model. Therefore, the optimization algorithm adopts the familiar “Stochastic Gradient Descent” (SGD) to conduct parameter training in the CNN classification process.

Meanwhile, in order to prevent over-fitting during training, the regularization method of "dropout" [16] was introduced into the full connection layer. The dropout's keep-prob ratio was set to 0.5.

This paper will make three experimental. In order to compare the effect of the improved IG algorithm on the selection of important feature items and the text classification effect of the I-CNN model. In experiment 1, SVM and KNN classical text classification model are used to verify the effects of different feature selection methods. In experiment 2, the traditional information gain algorithm is compared with the improved information gain algorithm, and CNN is used as the classification model for training. In experiment 3, the improved IG+SVM text classification model and classical CNN classification model are compared with the I-CNN model proposed in this paper.

5.3 Experiment and Result Analysis

The data set of this experiment adopts THUCNews Chinese short text data set provided by Tsinghua University. THUCNews is generated by filtering the historical data of Sina news RSS subscription channel from 2005 to 2011, including 740,000 news documents (2.19 GB), all of in utf-8 plain text format. This paper randomly selected five text categories: finance, education, current politics, entertainment and sports. 1000 documents for each category. The text set of each category is divided into training set, testing set and distributed randomly in the form of training set: test set rate equals 2:1. The training set is mainly used to train a classification model which need to be verified the performance by the test set.

The common evaluation indexes of text classification model performance include precision rate P, recall rate R and F1 value. P is the proportion of the number of documents properly classified to the test data set documents. R is the proportion of the number of documents properly classified to the actual number of properly classified documents. F1 value is the harmonic average of precision rate and recall rate. It can comprehensively evaluate the prediction accuracy and recall sample situation of the classification model on each type of text set. This paper will adopt these three indexes to evaluate the experimental results. The values of F1 are calculated as shown in formula (10):

$$F1 = \frac{2 \times P \times R}{P + R} \tag{10}$$

Experiment 1. Traditional feature selection methods such as IG, CHI and MI perform well in English text, but DF algorithm performs better in Chinese text. Therefore, Figs. 3 and 4 of this experiment respectively show the F1 mean curves of three feature selection methods on SVM and KNN classifier. The upper limit of feature number is set as 30000.

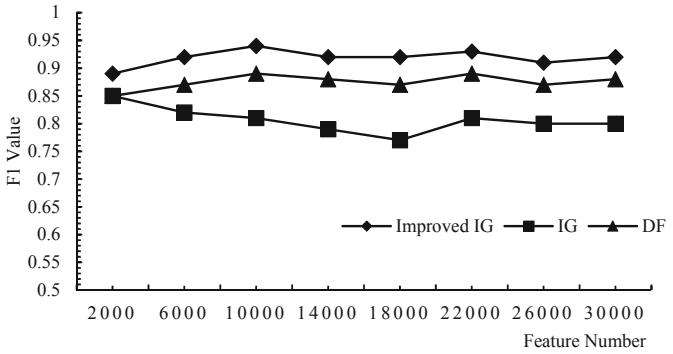

Fig. 3. Comparison of feature selection methods on SVM.

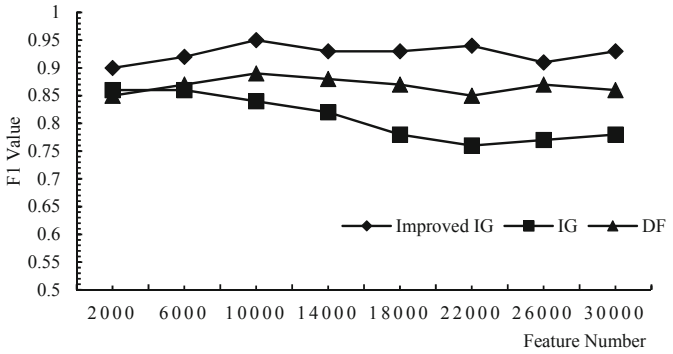

Fig. 4. Comparative performance of feature selection methods on KNN.

It can be seen from the figure that the improved IG algorithm not only performs better but also stably. F1 value in both classifiers is higher than other feature selection algorithms. However, the traditional IG algorithm is theoretically proved to be the best feature selection method, but its performance is the worst. The biggest reason is the neglect of word frequency information. Words that appear in many documents may not be representative words, but may be noise words.

Looking at the two figures, it can be found that the performance of the classifier reaches a maximum when the dimension of feature space is valued in the range of 5000 to 10000. It indicates that in the process of feature extraction, the dimension of feature space can be compressed to the original 10%–20%. And the remaining unnecessary words can be removed. This not only greatly reduces the dimension of feature space and memory occupation, but also improves the efficiency of classifier training without causing errors in classification results.

Experiment 2. In experiment 2, the traditional IG algorithm is compared with the improved IG algorithm. CNN is used as the classification model for training. It verifies whether the improved IG algorithm can effectively select representative words and

improve the classification effect. The experimental test set selects 200 articles of finance, education, current politics, entertainment and sports respectively from the "THUC-News" data set. Use recall rate R and precision rate P as evaluation indexes. Correct classification number refers to the number of classified texts that really belong to this class. Actual classification number refers to the number of 1000 texts classified into this class after I-CNN model classification. According to the criteria of experiment 1, in the process of feature extraction, about 80–90% of unimportant feature words can be discarded. Some representative words are left as important features of the classification.

Table 3. Test results of traditional IG algorithms in test sets (unit: paper)

Category	Test data set	Traditional IG algorithm			
		Correct classification number	Actual classification number	P (%)	R (%)
Finance	200	178	199	89.00	89.45
Education	200	164	181	82.00	90.61
Current politics	200	171	246	77.50	69.51
Entertainment	200	147	162	85.50	90.74
Sports	200	191	212	95.5	90.09
Total	1000	851	1000	85.9	86.08

Table 4. Test results of improved IG algorithms in test sets

Category	Test data set	Improved IG algorithm			
		Correct classification number	Actual classification number	P (%)	R (%)
Finance	200	176	207	88.00	85.02
Education	200	187	211	93.50	88.63
Current politics	200	159	185	79.50	85.95
Entertainment	200	172	186	86.00	92.47
Sports	200	189	211	94.50	89.57
Total	1000	883	1000	88.30	88.33

As can be seen from Tables 3 and 4, in the current political data set, the actual classification number of the traditional IG algorithm is 246, while the correct classification number is only 171, resulting in the accuracy rate of only 77.5%. However, the actual and correct classification number of the improved IG algorithm are 159 and 185 respectively, and the accuracy rate is 79.5%. Also, the correct classification number, R and P of the improved IG algorithm in financial and sports categories have been decreased. But the average recall rate and average precision rate are higher than traditional algorithms. According to the calculation, the average recall rate of the improved IG algorithm increased by 2.25%. The average precision rate increased by

2.4%. The number of correctly classified samples increased by 32. Therefore, it can be seen that the improved IG algorithm has not only stable performance but also better feature selection effect.

Experiment 3. In experiment 3, an improved IG+SVM text classification model and a classic CNN classification model are compared with the I-CNN model. The F1 value of experimental results is shown in Fig. 5.

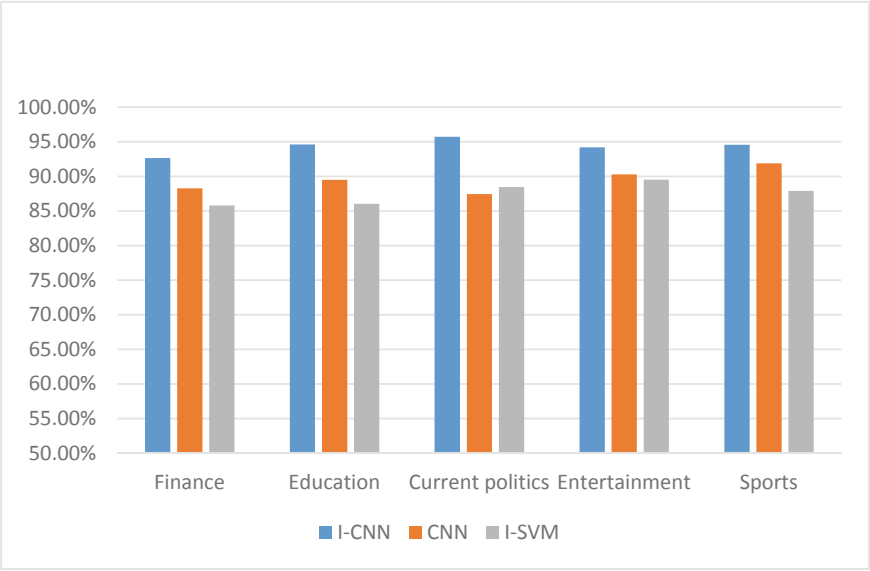

Fig. 5. Classification model F1 value comparison diagram.

It can be seen from the comparison experiment between CNN and I-CNN models, the CNN model lacks the feature extraction stage of improved IG algorithm. Although CNN can automatically extract features, the F1 value of CNN model is lower than that of I-CNN model. The F1 value of the I-CNN model in the current politics category reaches the highest value of 95.72%. While the CNN model in sports category only reaches 91.87%. And the classification effect is far less than that of the I-CNN model. Therefore, it can be seen that feature extraction is an indispensable part of the text classification process. Without feature extraction, there are a large number of noisy words and stop words in the text, which interfere with the training effect of the classifier.

According to the comparison experiment between I-CNN and I-SVM, when both have IG feature selection processing, the F1 value of CNN model is better than SVM model. The F1 value of the I-CNN model in the current politics category reaches the highest point of 95.72%. The F1 value of the I-SVM model in the entertainment category reaches only 89.52%. Moreover, the features of CNN local connection and weight sharing greatly reduce the number of parameters, make the model simpler and

train faster. However, the traditional text classification model requires manual feature extraction. And weight is not shared, resulting in very large parameters. In contrast, the CNN model is fast in training. With the least parameters, the best classification effect can be trained, which is the reason why CNN is adopted as the classification model in this paper.

Therefore, this paper adopts CNN as classifier and combines it with the improved IG algorithm. According to the experimental results, the classification effect of the model is better than that of the traditional classification model.

6 Conclusion and Future Work

As a key technology of text classification, feature selection has a direct impact on the classification performance. In view of the shortcomings of traditional IG algorithm, this paper introduces the word frequency, inter-class separation degree and intra-class dispersion degree parameters. Then proposes a new improved IG algorithm. This algorithm solves the problem that the traditional IG algorithm neglects the word frequency distribution to weaken the classification. And added the parameter about the distribution of feature words. The classification accuracy has been improved. In this paper, we use CNN as the model. Because CNN can automatically identify local features, which makes up for IG algorithm's consideration of feature term relation. Combined with the features of weights sharing and local connection of convolutional neural network, the model training and classification are faster. The experimental results show that the I-CNN model is better to the traditional classification model.

Although the I-CNN model proposed in this paper has a better effect on text classification than the previous existing models, the classification effect is not accurate and fast enough. It is expected that the text classification model can well identify the categories of huge amount of text data on the network. And classify them accurately and quickly. So as to make the network information more organized and classified. Deep learning model is the general trend in the field of text classification. But the end-to-end thought of deep learning model makes the deeper things inside the model still need to be explored. Therefore, text classification is still a worthy research direction.

References

1. Agnihotri, D., Verma, K., Tripathi, P.: Pattern and cluster mining on text data. In: 2014 Fourth International Conference on Communication Systems and Network Technologies, pp. 428–432. IEEE (2014)
2. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML, vol. 97, pp. 35, 412–420 (1997)
3. Xu, J., Jiang, H.: An improved information gain feature selection algorithm for SVM text classifier. In: 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 273–276. IEEE (2015)
4. Chen, Y., Chen, Y.: Attribute reduction algorithm based on information entropy and ant colony optimization. J. Chin. Comput. Syst. **36**(3), 586–590 (2015)

5. Ming, H.: A text classification method based on maximum entropy model based on improved information gain feature selection. J. Southwest Normal Univ. (Nat. Sci. Ed.) **44**(03), 119–124 (2019)
6. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
7. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
8. Chang, L., Deng, X.M., Zhou, M.Q., et al.: Convolution neural network in image understanding. Acta Automatica Sinica **42**(9), 1300–1312 (2016)
9. Sainath, T.N., Mohamed, A., Kingsbury, B., et al.: Deep convolutional neural networks for LVCSR. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8614–8618. IEEE (2013)
10. Lecun, Y., Bottou, L., Bengio, Y., et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
11. Zeiler, M.D., Fergus, R.: Stochastic pooling for regularization of deep convolutional neural networks. arXiv preprint arXiv:1301.3557 (2013)
12. Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
13. O'Shea, K., Nash, R.: An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458 (2015)
14. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Hinton, G.E., Srivastava, N., Krizhevsky, A., et al.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)