Research on New Word Recognition Based on Weibo Information and SVM

Yuanfang Xu

Inner Mongolia Normal University, Hohhot, China

xuyuanfang86@126.com

Abstract. In order to effectively recognize Weibo new words, this paper proposes a Weibo new word recognition method based on Weibo information and SVM. Firstly, positive and negative samples are extracted from the Weibo corpus and the training corpus annotated with part of speech. Next, the word features of these samples are vectorized and trained through support vector machines to obtain support vectors for new word classification. Finally, it is vectorized and inputted into the trained support vector machine classifier to obtain the recognized Weibo new words. By comparing the experimental results, the optimal fusion feature combination and SVM kernel function are obtained.

Keywords: Weibo new word recognition; SVM; Constraints; kernel function.

1. Introduction

With the rapid popularization and development of Weibo, various new words are constantly emerging on Weibo. The emergence of new words on Weibo represents the language trend of the Internet. How to better identify and utilize Weibo's new word resources has become an urgent problem that needs to be solved[1]. Some emerging vocabulary appearing on Weibo may be rearrangements of known vocabulary, which may be split into two seemingly unrelated individual words when recognized using existing segmentation techniques[2]. This article proposes a method that combines SVM and word features to identify new words in Weibo corpus. Through this method, the accuracy of Weibo corpus new word recognition can be improved on the basis of existing algorithms.

2. New Word Recognition Based on Weibo Information and SVM

2.1 Corpus preprocessing

We used web crawlers to capture 200000 hot topics on Sina Weibo in October 2020, covering various topics such as society, science and technology, education, and more. Compared to traditional Chinese textual materials, the content of this type of Weibo data is very chaotic and diverse [3]. Not only does it contain normal Chinese words and sentences, but there are also many irrelevant information, such as emoticons. The next step is to label these vocabulary and determine which ones are new Weibo vocabulary and which ones are not. Finally, based on the labeling results, they are divided into training sets and testing sets.

2.2 Selection and calculation of candidate word features

This method needs to combine the characteristics of the words themselves with Weibo corpus and test corpus to form feature vectors. The selected features of the words themselves in this article include mutual information (MI) [4], word formation probability (IWP) [5], morpheme productivity (MP) [6], frequency feature (F_F) [7], and contextual information (Context).

2.3 Problem transformation

The SVM classification function can be set to g (x), so:

$$g(x) = w \cdot x + b \tag{1}$$

If g (x) is a real function, then the output of the expression is a real number. In the previous statement, if g (x)>0, then the identification of the point is positive, usually 1. If g (x)<0, then mark the point as a negative class, usually -1. As a classification, there are only two types of outputs, one is positive and the other is negative. Therefore, for the output of g (x), a threshold can be added to judge. When g (x)>0, it is determined to belong to class C1, and when g (x)<0, it is determined to belong to class C2. That is to say, for binary classification g (x), a segmentation line is defined, which separates the two types to the maximum extent possible. In this way, a symbol function sgn() can be set as the judgment function:

$$f(x) = sgn\,(\,g(x)) \tag{2}$$

For the training set S of the binary linear problem, X represents the input domain, Y represents the output domain, and is the vector representation of the sample. In this paper, 7 attributes of the word feature are selected, so x here is a sample vector representation containing 7 attributes. When y_i=1, it represents that the input x_i sample is a non Weibo new word; When y_i=-1, it represents the input of x_i sample as a candidate for Weibo new words.

When classifying Weibo new words, the vectors of each sample can be expressed in the following form:

$$D_i = (x_i, y_i) \tag{3}$$

x_i is a high-dimensional text vector, and y_i represents classification labels. Therefore, when a sample point approaches the classification plane, its distance can be calculated:

$$\delta_i = y_i(wx_i + b) \tag{4}$$

Since the above results are always greater than 0, replacing the original w and b with w/| | w | | and b/| | w | | respectively, the interval can be written as:

$$\delta_i = \frac{1}{||w||}|g(x_i)| \tag{5}$$

The norm of a vector, commonly referred to as | | w | |, is a measure of its length. In fact, vector length usually refers to its 2-norm, rather than other more complex forms. The general representation of norm is p-norm:

$$||w||_p = \sqrt[p]{w_1^p + w_2^p + \_ + w_n^p} \tag{6}$$

Here, the definition of geometric interval needs to be introduced: for a given hyperplane, the distance between it and a specific set of points is calculated, which is the distance between the points closest to the hyperplane. Therefore, the problem is transformed into finding the minimum value of | | w | |:

$$min\,||w|| \tag{7}$$

This question can be translated into:

$$min\frac{1}{2}||w||^2 \tag{8}$$

2.4 constraint condition

In the process of classifying Weibo new words, this article expects non Weibo new word sample points to appear on the right side of H1, and Weibo new word sample points to appear on the left side of H2, as shown in Fig.1:

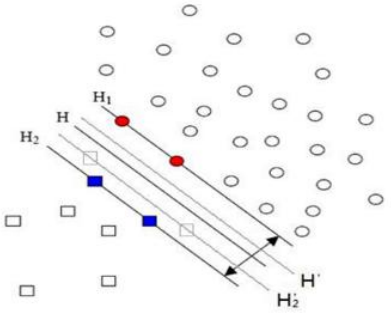

Fig. 1 Sample point distribution map

Considering the possibility that not every sample may appear within the H1 and H2 intervals, a constraint has been added to the model, which is to set the aforementioned spacing to be constant at 1. Specifically, set the distance between the closest Weibo new word and the non Weibo new word to 1, which means that if there are other gaps, their value should be at least equal to 1. According to the definition of interval, meeting these conditions is equivalent to always holding the following equation:

$$y_i[(wx_i) + b] - 1 \geq 0(i = 1,2,\dots l) \quad (9)$$

In this formula, all constraints are linear functions of w, and the independent variable is w. The problem is transformed into a quadratic programming problem. In fact, in the initial g (x), w is also the only variable, because x_i is the training sample vector and is known, while b can be obtained by simply bringing in any sample point after w is determined. The sample is set to w, which can be interpreted in mathematical terms as a combination of samples:

$$w = \partial_1 y_1 x_1 + \partial_2 y_2 x_2 + \cdots \partial_n y_n x_n \quad (10)$$

At this point, g (x) can be written as:

$$g(x) = < w, x > + b \quad (11)$$

In this formula, the Lagrange multiplier represents the maximum point corresponding to the parameter, and x_i represents the sample point. N represents the total number of samples, while the label y_i of the i-th sample is 1 or -1.

After calculating the Lagrange multiplier, it is concluded that only a very small number of values are not equal to 0. Therefore, the sample points x_i corresponding to these formulas that are not equal to 0 are exactly what is needed. These sample points will form a straight line, which is the expected support vector. The formula is simplified to obtain:

$$w = \sum_{i=1}^{n} (\partial_i y_i x_i) \quad (12)$$

At this point, g (x) is:

$$g(x) = \sum_{i=1}^{n} \partial_i y_i < x_i, x > + b \quad (13)$$

At this point, the problem has been solved. The x in the formula represents the sample that needs to be classified, which is the test sample. In this article, it is the scattered list of candidate Weibo new words that need to be classified.

3. Experimental results and analysis

Firstly, the role of different word features in new word recognition was evaluated. Select the radial basis kernel function of support vector machine as the core algorithm. In addition, three word features were used: MP, IWP, and word formation probability to construct a basic model F (B). Next, different combinations of mutual information, word formation probability, morpheme productivity, frequency features, and contextual information were integrated into the model, and experiments were conducted on the same test dataset. The experimental results are shown in Table 1:

Table 1. Comparison Table of Fusion Results of Different Features

Numble	Fusion feature types	Identified Weibo neologisms	Identify correct Weibo neologisms	accuracy R（%）
1	SVM combined with basic model F (Base)	32	6	18.76
2	F (base) fusion of contextual information	30	7	23.33
3	F (base) fusion mutual information	26	7	26.93
4	F (base) fusion of context and mutual information	27	8	29.63
5	F (base) fusion frequency features	23	7	30.43
6	F (base) fusion of contextual information and frequency features	16	6	37.5
7	F (base) fusion of mutual information	15	7	46.67

	and frequency features			
8	F (base) fusion of mutual information, contextual information, and frequency features	12	8	66.67

After experimental verification, the number of word features in a new word recognition system has a significant impact on accuracy and retrieval rate. When all possible word features were considered, including word formation probability, morpheme productivity, frequency features, contextual information, and mutual information, the optimal accuracy rate observed in the experiment was 66.67%. Therefore, future research will comprehensively use these word features for further experiments.

Next, while keeping other conditions unchanged, select all word features and conduct experiments using various kernel functions. The conclusions are shown in Table 2:

Table 2. Comparison Table of the Effects of Different SVM Kernel Functions on Experimental Results

Numble	Function type	Identified Weibo neologisms	Identify correct Weibo neologisms	accuracy R（%）
1	adial Basis Kernel Function (RBF)	12	8	66.67
2	Polynomial kernel function	13	7	53.85
3	Sigmoid kernel function	15	7	46.67

Through experiments, it was found that the accuracy of Weibo new word recognition system is optimal when the radial basis function RBF selected in this article is used.

4. Summary

This study proposes a Weibo new word recognition method based on Weibo data, which combines SVM and word features. After a series of comparative experiments, it has been proven that this method can effectively improve the accuracy of Weibo new word recognition. The next research focus is to further explore the practical application effects of this method in large-scale corpora.

Acknowledgements

Fund projects:
Research Project of Inner Mongolia Higher Education Institutions (NJZY21549)

References

[1] Fu Lina, Xiao He, Ji Donghong. New Emotional Word Recognition Based on OC-SVM [J], Computer Application Research, 2015,71946-1048.
[2] Li Chengcheng,Xu Yuanfang, Based on support vector and word features new word discovery research, proceedings of 2012 IEEE International Conference on Computer Science and Automation Engineering ,2012,166-168.
[3] Jian-Yun Nie, Unknown Word Detection and Segmentation of Chinese using Statistical and heuristic Knowledge. Communications of COLIPS,2008,5(I&2),47-57.
[4] Han Xiulong. Research on Weibo New Word Discovery Based on SVM and Feature Correlation [J], Computer Knowledge and Technology, 2018,14,66-69.
[5] Feng Yong, Li Hua. Based on Adaptive Chinese word segmentation and approximation of SVM text classification algorithm [J], computer science, volume thirty-seventh, 2010, first, 251-254, 293.

[6] Qian Qiuyin, Zhang Zhenglan. A method based on multiple SVM classification method of relevance feedback image retrieval [J], computer technology and development, 2009, volume nineteenth, issue eighth, 66-69.
[7] Huang Xiuli, Wang Yu.SVM in unbalanced data set [J], computer technology and development, 2009, volume nineteenth, issue sixth, 190-193.