Research on Emotion Classification based on Complex Network and Ensemble Learning

CAO Qianqian1*, CHEN Xiangyang2, LAI Yuanzhe3, DENG Chenzhou4

1 School of Computer Science and Engineering, Wuhan Institute of Technology, Hubei 430205, China
21907010065@stu.wit.edu.cn

Abstract. Because the traditional feature extraction is based on the statistical information such as document frequency and word frequency, the selection of feature words is ignored, and the semantic correlation between words in the text is ignored. The feature selection method based on complex network takes into account the semantic association between words, but does not take into account the statistical information such as word frequency. The above methods are not satisfactory for the selection of feature words, which affects the effect of text classification. Therefore, this paper combines the two, proposes a new method for feature selection, and in order to solve the problem of low accuracy rate of single classification algorithm, USES integrated learning [1] to strengthen the classification algorithm. The results show that this method is feasible and achieves good classification effect.

1. Introduction

With the rapid development of information technology, more and more Internet applications have penetrated into every aspect of people's life. As the interaction between ordinary users and web applications becomes more and more frequent, the role of Internet users gradually evolves from the viewer to the creator of Internet content and information. With the rapid development of information technology, more and more Internet applications have penetrated into every aspect of people's life. As the interaction between ordinary users and web applications becomes more and more frequent, the role of Internet users gradually evolves from the viewer to the creator of Internet content and information. Because the traditional feature extraction is based on the statistical information[2] to select the feature words, but ignore the semantic correlation between the words in the text. The feature selection method based on complex network[3] takes into account the semantic association between words, but does not take into account the statistical information such as word frequency. Therefore, this paper combines the two, firstly using the traditional information gain method for feature selection, then extracting feature items from the original feature set based on the complex network comprehensive characteristics, and finally removing the duplicate items by taking the union of the two, that is, the final feature selection result set. Finally, the feasibility of the method was verified by experiments using integrated learning combined with naive Bayes classifier, and the comparison experiment showed that the combination of the two methods achieved the best classification effect.

2. Related works

2.1. Integrated learning with naive Bayes classifier
Naive Bayes Algorithm NB (Naive Bayes) is a very simple classification algorithm [10-11], which is based on Bayes algorithm. The basic idea is: for the given item to be classified, solve the probability of each category under the condition of the occurrence of this item, take the one with the highest probability, and consider which category the text to be classified belongs to. The AdaBoost[6] algorithm is applied to the naive Bayesian algorithm. During each iteration, the training samples are trained $x_i$. If misclassified, weight $w_i^{t+1}$ will increase, otherwise $w_i^{t+1}$ will reduce. During the iteration of AdaBoost, the weight assigned to each training sample is $w_i^t$. We then introduce it into the parameter $P(x_k \mid c_j)$. Then the naive Bayes formula that we had before would be:

$$P(x_k \mid c_j) = \frac{\prod_{i=1}^{n} w_{jk}e^{w_{ij}} + 1}{\sum_{j=1}^{p} \prod_{i=1}^{n} w_{jk}e^{w_{ij}} + |p|}$$

Therefore, with each iteration of AdaBoost, the sample weight will be updated each time, and the prior probability and posterior probability of naive Bayes will change, which will disturb the classification of naive Bayes classifier and increase the heterogeneity of the classifier.

| Table 1 | The first 6 items of Adaboost enhanced naive Bayes classification results table |
|---|---|
| id | Text (Weibo content) | category |
| 1 | Special, no two, wheat, zongzi, spring onion, two cakes, professional, happy birthday | 1 |
| 2 | Different, no two, wheat, zongzi, spring onion, two cakes, professional, happy birthday today, take, the first, hero, back to Jiangsu, increase wages | 1 |
| 3 | Fuck, ma, the room, the noise, the effect, it's bad | 0 |
| 4 | Like, state, instant, happy | 1 |
| 5 | Waterlogged, unwilling to drought, down, yeah | 1 |
| 6 | See, a lot, my friend, leave a message, said that yesterday evening, also, no, see enough, tonight, continue, to share, nicky wu, the fixation, luo, a new round, countdown | 1 |

| Table 2 | Statistics of the number of microblogs with URLS |
|---|---|
| Covariance item | result |
| The number of tweets with a URL | 1756 (total: 4780) |
| Average number of references | 2.72 |

As you can see from the table above, there is a lot of use for urls in microblogs. Therefore, before emotion analysis, it is still necessary to properly remove the junk microblog and remove the useless URL, so as to improve the efficiency of emotion analysis and avoid the analysis of these useless information.
3. Feature Selection

Feature selection should try its best to select emotional feature words that express strong emotional information in text to improve the accuracy of emotion classification. Firstly, by constructing the text in the text, to retain the semantic information and its structure, and then using the comprehensive property of node (that is, the center words), looking for the key node to as key words of the text, and remove those words less information, to reduce the number of nodes in complex networks, text to achieve the purpose of reduce complexity. Although the feature selection method based on complex network considers the interrelation between texts, word frequency is not considered. Therefore, this paper combines the two and proposes a new method for feature selection. Through the final experiment, it is found that the feature selection combining the two has the best classification effect.

3.1. Traditional methods

In emotion analysis, information gain\[5\] is transformed into characteristic items $w$ in $C_i$. The amount of information caused by the occurrence of a situation in a class is defined as follows:

$$IG(w) = E(C_i) - E(C_i | w)$$

$$= -\sum_{i=1}^{n} P(C_i) \log_2 P(C_i) + P(w) \sum_{i=1}^{n} P(C_i | w) \log_2 P(C_i | w) + P(C_i) \log_2 P(C_i | w)$$

$$= P(w) \sum_{i=1}^{n} P(C_i | w) \log_2 \frac{P(C_i | w)}{P(C_i)} + P(w) \sum_{i=1}^{n} P(C_i | w) \log_2 \frac{P(C_i | w)}{P(C_i)}$$

After the information gain of feature items is calculated, the first 500 words are extracted as global feature words in descending order according to the information gain value, and the text file of feature words is saved.

3.2. Feature extraction of complex network

Feature selection algorithm based on weighted complex network\[6\] by analyzing the comprehensive properties of weighted text nodes in complex networks, which considering the weighted degree, weighted aggregation coefficient of node and edge betweenness measure of key importance in the text, by constructing evaluation function response node comprehensive features, embody the node connection status, local intensity, as well as the impact on the global network, thus for text keyword selection, in order to achieve the purpose of feature selection.

The specific algorithm is as follows: Step1 preprocess document d. Step2 Establish a text-weighted complex network, take feature words as nodes, connect sentences with feature words whose span is less than or equal to 2, and merge the same feature words in different sentences. Step3 calculate $n_j$, the weighted degree, weighted aggregation coefficient and node interface of the node respectively, and conduct normalization processing respectively, and then construct the evaluation function $CF_i$ taking the function value as the comprehensive characteristic value of the node $n_j$.

$$CF_i = \beta_1 WD_i + \beta_2 WC_i + \beta_3 P_i$$

among, $\beta_i (1 \leq i \leq 3)$ is an adjustable parameter that represents the weight of the corresponding part. And has $\beta_1 + \beta_2 + \beta_3 = 1$.; the interface number of the node $n_j$. Step4 sort the function value of the node and select the feature words corresponding to the first $m$ nodes with large function value as the key words of the text.
In order to get better experimental results, repeated experiments were carried out, and $CF_i$, the mean values were $\beta_1 = 0.4$, $\beta_2 = 0.3$ and $\beta_3 = 0.3$.

4. Experiments

4.1. Contrast experiment

The experimental data set was from weibo network, and 15,446 microblog comments were crawled by crawler technology, among which the tagged ones were divided into training sets, which was a binary data set with a total of 13,712. The test set has 1,509.

| classes | IG | CN | IG-CN |
|---------|----|----|-------|
|         | Prec | Recall | F1(%) | Prec | Recall | F1(%) | Prec | Recall | F1(%) |
| 0       | 0.5103 | 0.8156 | 0.8980 | 0.6164 | 0.849 | 0.714 | 0.7226 | 0.894 | 0.9443 |
| 1       | 0.8765 | 0.6259 | 0.7699 | 0.8748 | 0.713 | 0.786 | 0.9256 | 0.792 | 0.8844 |
| average | 0.6934 | 0.7207 | 0.8339 | 0.7456 | 0.781 | 0.750 | 0.8241 | 0.843 | 0.9143 |

In the experiment, the training set was labeled into two categories, positive and negative. After text preprocessing, three groups of comparative experiments were conducted, namely, feature extraction using information gain, feature extraction based on complex network characteristics, and feature extraction method combining the two. Then, the naive Bayes classifier is trained by integrated learning, and the classification results are shown in the figure below.

As can be seen from the bar chart above, the classification tests performed by IG alone or based on the comprehensive characteristics of complex networks have similar effects, but the classification effect combined with IG is significantly better than that performed by IG alone.
5. Conclusion

This paper studies the classification of microblog emotions. First, Python is used to pre-process the data, and then three feature extraction methods are used to conduct training tests on the naive Bayes classifier of integrated learning. Through comparison experiments, it is found that the classification effect of the method proposed in this paper is the best, which verifies the feasibility of the method. However, this paper also has some shortcomings. Since the amount of data in the text corpus is not very large, the consumption of computing time is not very large. However, when the amount of data is too large, the cost of building a complex network will increase and the efficiency will decrease. Therefore, the efficiency will be further improved in the future research.

References

[1] Qin Li, Shaobo Li, Sen Zhang, Jie Hu, Jianjun Hu. A Review of Text Corpus-Based Tourism Big Data Mining[J]. Applied Sciences, 2019, 9(16).
[2] Dat Quoc Nguyen, Richard Billingsley, Lan Du, Mark Johnson. Improving Topic Models with Latent Feature Word Representations[J]. Transactions of the Association for Computational Linguistics, 2015, 3.
[3] Fu H P, Zou B J. A rapid training method of AdaBoost classifier [J]. Journal of yunnan university (natural science edition), 2020, 42(01):50-57.
[4] He Ming, Sun Jianjun, Cheng Ying. (in Chinese) A review of text classification based on naive bayes [J]. Information science, 2016, 34(07):147-154.
[5] HTROSH Ogr, HIROSHI Amano, MASATO Kond. Comaison of meris far featrselection in unb alanced text clsisitioi[J]D Expert Systemswith Applications, 2010.09:153-164.
[6] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, Donald Brown. Text Classification Algorithms: A Survey[J]. Information, 2019, 10(4).
[7] Park, Hong, Kim. A Methodology Combining Cosine Similarity with Classifier for Text Classification[J]. Applied Artificial Intelligence, 2020, 34(5).
[8] Donoho D L. De-noising by soft-thresholding [J]. IEEE Trans Inform Theory, 1995, 41 (3):613-627.
[9] Cancho Ramon Ferrer i, Solé, Richard V. The small world of human language[J]. Proceedings of the Royal Society B: Biological Sciences, 2001, 268(1482).
[10] Hu Mengya, Fan Zhongjun, Zhu Yue. [J]. Information and computer (theoretical edition), 2020, 32(12):71-73.