Multi-label classification of technical articles based on deep neural network for CCC2019

Qiuhan Zhao[1], Wenchuan Yang[2]

[1,2]School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China

E-mail: allen_zqh@bupt.edu.cn

Abstract

This paper uses word-embedding and deep neural networks to build a multi-label classification model based on technical articles. In this paper, we use deep learning algorithms to train word vectors based on numerous technical articles, and then with the abstracts and corresponding CNKI labels of these articles as input of network is trained for compare and research the prediction results of different networks, finally determining the threshold by statistical distribution for label screening. Through parameter tuning, model fusion and data augmentation, the accuracy of multi-tag prediction network reaches 92.05%. Multi-label classification based on deep neural network has advantages in simple preprocessing, high accuracy and computational efficiency.

Focus

At present, many search websites for technical document classify each scientific and technical article according to the Chinese Library Book Classification with a large amount of labor costs. This is not only leads to a waste of labor costs, but also the classification accuracy is firmly related to the proficiency and knowledge level of the checkers. Realizing automatic multi-labels based on the content of technical articles can effectively solve this problem.

Methodology

We crawled included 168 sub-categories about 500 articles each to ensure that the number of articles in each category was evenly distributed, totally 84,000 articles from China National Knowledge Infrastructure (CNKI). Moreover, the total number of CNKI labels is 241 according to statistics.

The abstract part of a technical article is the essence of the full text. This paper uses the abstract as the input feed into the neural network and uses document corresponding multi-label combination as the real value so that making labels the network predicted are as close as possible to the real labels through training.

Because the number of sub-category in each major category is different, in order to avoid the impact that some categories have more articles and higher label frequency on result accuracy, this paper uses two data enhancement methods for the categories with less proportions. It is shown as the following.

Methodology

Table 1: Data Preprocessing and Argumentation

Title	“一带一路”背景下高校茶艺课程双语教学模式初探		
Operation	Pretreatment	Data argumentation	
		Random replacement	Mess up the order
Result	茶艺 一门 应用 技能型 学科 茶文化 重要 组成部分 一带 一路 背景 高校 茶艺 课程 开展 双语	茶艺 空 空 技能型 学科 空 重要 组成部分 一带 空 背景 高校 空 空 开展 双语	技能型 重要 茶艺 背景 一带 开展 茶艺 一路 应用 双语 高校 一门 茶文化 课程 组成部分 学科

The above is just a simple example due to the length of the poster. Long word length will result in too much padding characters which reduce the computational efficiency. On the contrary, it will miss important meanings. After experimental testing, we chose a word length of 64.

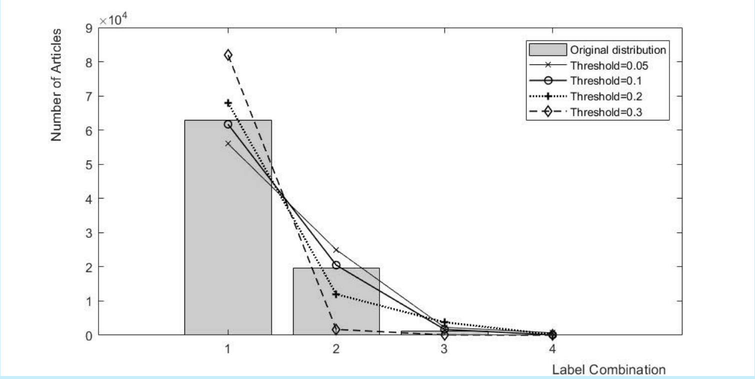
Result

This paper uses different networks for testing and tuning, and fuses different networks of optimal parameters. We tuned the model by setting parameters such as the number of convolution kernels and hidden layers, and obtained the following main results.

Table 2: Results of CNN-LSTM

Model	**ROC-AUC**	**Coverage**
CNN	91.38%	3.54
LSTM	91.47%	3.52
CNN-LSTM	92.05%	3.15

After using the compounded model to predict all articles in the dataset, We need to set a threshold to determine how likely the label needs to be retained. we select the probability thresholds respectively 0.05, 0.1, 0.2, and 0.3 for screening, and count the number of articles included in each label combination. Finally, the threshold is selected as 0.1, it is closest to the original distribution of the label combinations.



Conclusion

The accuracy rate finally reaches 92.05% by CNN-RNN model. At last, the selection of the threshold is determined by examining the statistical distribution method. However, this paper only uses the secondary label of CNKI. If we add a primary label and consider the parent relationship of secondary label, we may have further improvement in accuracy. In addition, the threshold can be refined and the optimal solution can be found in the 0.1-0.2 interval to achieve better filtering.

References

[1] Kim Y. Convolutional neural networks for sentence classification. 2014.

[2] Kang Liu et al. Siwei Lai, Liheng Xu. Recurrent convolutional neural networks for text classification. 2015.