Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 4, No. 7, 2024

Optimizing Text Classification Using Artificial Neural Networks and K-Fold Cross-Validation in High-Volume Data Processing

Callum Revere
Department of Computer Science, Redstone University, USA
callum.revere91@redstone.edu

Abstract:

With the rapid expansion of internet resources and advancements in big data and cloud computing, effective text classification has become essential for managing and extracting meaningful information from vast text datasets. Traditional, labor-intensive text categorization methods are increasingly inadequate for handling modern data demands. This study develops a robust automatic text classification method based on artificial neural networks (ANN) to enhance accuracy and efficiency. By employing web-crawling techniques, 1600 categorized articles across technology, finance, culture, and politics were gathered, and a vector space was created using the bag-of-words model. The ANN model was trained on these features, and K-fold cross-validation was applied to assess the model's performance. The resulting classification model demonstrated acceptable accuracy with a 15% error rate, suggesting that further refinement through K-fold cross-validation could enhance reliability. This approach has significant implications for large-scale, automated text processing across sectors, minimizing manual intervention and optimizing information retrieval.

Keywords:

Text Classification; Word Frequency Statistics; Vector Space; ANN

1. Introduction

Since the birth of the computer in the 1940s, with the continuous development of computer network technology, the information resources on the Internet are becoming more and more large and complex, and people tend to become powerless when dealing with it. Therefore, the technical field of text classification is becoming more and more important. At the same time, with the rise of big data, cloud computing and other technologies, a large amount of text data can be easily obtained and effectively managed. How to carry on the further mining and the information processing to these text data has become the main current work.

At present, the text categorization technology has been used in many fields since birth, such as web page classification, text sentiment analysis, information retrieval, text filtering, etc. Every technological applications are large improved the related social productivity, the traditional text classification is an important foundation for the work, but a large number of rely on artificial to complete, in order to reduce labor costs, improve efficiency, the current text automatic classification technology has become the focus of current research.

Based on ANN, this paper designs a faithful text classification method. By using the crawler technology, it crawls 1600 articles of science and technology, finance and economics, culture and politics on the People's Internet. The word bag model is used to create a vector feature space. Finally, the artificial neural network is designed, and the characteristic data are used for training. Finally, the method of K-fold cross test is utilized to test test the Buddha nature.

2. Related works

From the 1960s to the beginning of this century, text classification models based on shallow learning were still dominant, which meant statistism-based models such as Naive Bayes, K-Nearest Neighors and Support Vector Machines.Doing so is time-consuming and expensive, and using such methods often ignores contextual information and order structures in the underlying data.Therefore, since the 2010s, people have gradually improved the shallow learning model to the deep learning model, which avoids the manual design of rules and functions.

Deep learning based neural network models have achieved great success in many NLP tasks, including learning distributed words, sentence and document representation, parsing, statistical machine translation, sentiment classification, etc.Learning distributed sentence representation through neural network models requires little external domain knowledge, and can achieve satisfactory results in emotion classification, text classification and other related tasks.To sum up, given the text classification technology is widely used in the field of related applications, at the same time, along with the vigorous development of the neural network technology, to explore the characteristics of deeper relationship and obtain more accurate text classification, improve the accuracy of the classification results, therefore, the text classification based on artificial neural network research, has very important value and significance in the application.

3. Proposed method

3.1 Overall design

The process of using ANN for text classification in this article is divided into five steps:

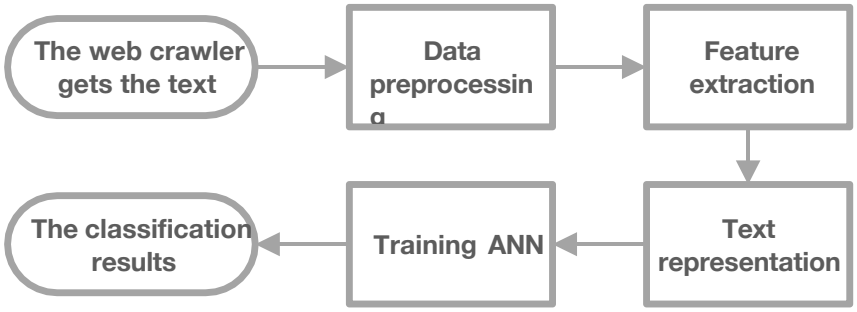

Figure 1. Text classification

Step1 The crawler technology was used to extract news about military, finance and economics, political and military from Xinhuanet, and about 400 pieces of news about each theme were crawled.

Step2 The crawled text is preprocessed to remove the unnecessary parts of the text, such as punctuation and prepositions, etc. Jieba word segmentation is used in each news article to extract the appearing nouns for constructing feature vectors.

Step3 Text is vectorized using the word bag model.

Step4 The results of word segmentation are counted and some of them are selected to construct new feature vectors.

Step5 Based on the constructed new eigenvector, ANN is trained and classified.

3.2 Pretreatment

Before preprocessing, the network structure of the text to be crawled should be analyzed first, and the text on the network should be obtained by writing scripts.The obtained text needs to be classified and stored, and then preprocessed for each type of text and extracted feature words for each text. The specific pre-processing steps are as follows:

(1) Wrote scripts to crawl and retrieve articles used for training, and classified and stored them

(2) Remove the punctuation marks in the text, and conduct word segmentation for each type of text

(3) Conduct statistics on the results of word participles

After crawling the text, in order to generate the feature vector representing the text, it must be segmented. In this paper, Jieba, a commonly used word segmentation database, is selected for word segmentation.

3.3 Feature extraction

Feature extraction is the key technology in text classification

Introduction to artificial neural network

Artificial Neural Network(ANN), also known as a neural network or quasi-neural network, its essence is a kind of artificial mathematical model or computational model for statistics and calculation that imitates the biological brain neural function. It is an adaptive learning system with certain learning function.

Using ANN can solve some problems of traditional text classification:

(1) When the classification method is too different from the actual classification results, ANN can use the learning algorithm to adjust the classification ability of the network itself; At the same time, the neural network system can realize the fuzzy reasoning naturally.

(2) Neural network itself has a strong robustness and fault tolerance, good for the association, comparison and promotion, for local operations will not affect the overall effect.

(3) Artificial neural network can find the connection between input and output through training data. It not only relies on prior knowledge and rules, but also has good adaptability.

(4) The speed of an artificial neural network is very fast. It can use parallel computing or GPU to calculate. For a large number of text classification features, it can run at a very high speed.

After that, the data are summarized in the form of tables, sorted into XLS format files, and input into a Python program for data preprocessing. After the data is processed, it is outputted in the form of "0" and "1".

Among them, "0" represents the occurrence of this kind of characteristic words in this paragraph, and "1" represents the absence of such words.In the model, the frequency of feature words is used to judge the topic of the text.

Artificial neural network model

The neural network model is divided into four layers, as shown in Figure 2 (the number of neurons can be adjusted, and it is temporarily adjusted to 6

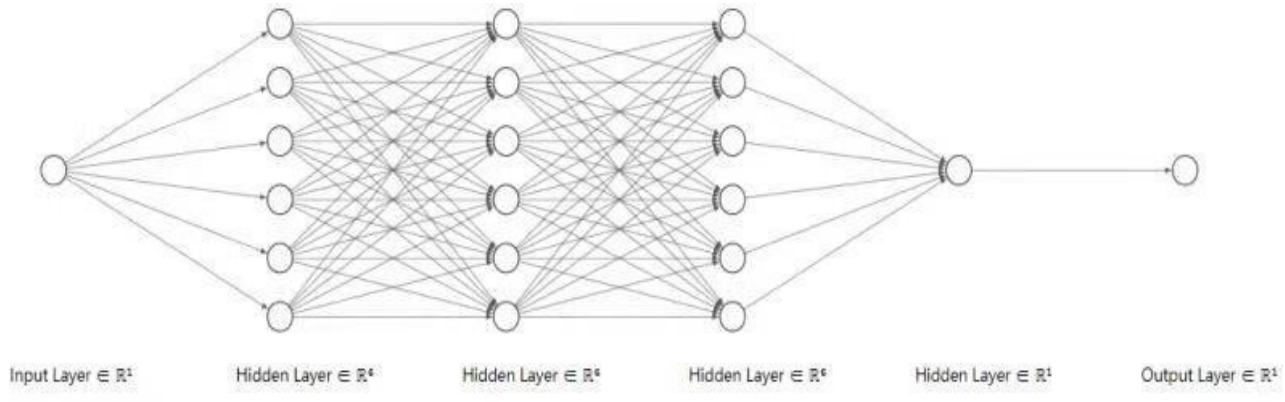

Figure 2. Internal structure of the neural network model

Among them, the first three layers are hidden layers, the fourth layer is output layer

Out of 366 sets of data, we selected 70% as training data, 15% as test data, and 15% as validation

data. We designed a MRE script as a program to compare the error results.

Script Introduction: Set the initial value of MRE to 0, then design a cycle, set the number of cycles to the amount of test data, calculate the result of each classification success, success result is 1, classification failure result is 0, after the end of the cycle, divide the failure result by the total number of calculation, get the corresponding error rate.

Model training results: the difference of the number of neurons and the number of hidden layers, the error results were slightly different. The MRE of the economy class was stable at 17%, the MRE of the entertainment class was stable at 16%, the MRE of the weather class was stable at 12%, and the MRE of the military class was stable at 13%.The results vary from time to time but generally fluctuate no more than two to three percent.

4. Conclusion

Through the above analysis, the model has achieved a satisfactory classification effect, with an error rate that remains around 15%. Although this error rate is not negligible, it is within an acceptable range for practical applications. To further enhance model accuracy and reduce the error rate of the Mean Relative Error (MRE) results, an improved approach can be implemented by adopting the k-fold cross-validation method. This technique allows the model to be trained and validated across multiple subsets of data, offering a more robust evaluation and enabling a reduction in variance. By refining the model with this method, it is anticipated that the accuracy can be further optimized, enhancing the reliability of the classification outcomes.

References

[1] Tao Kai,Tao Huang.A text classification model based on deep learning[J].Journal of Taiyuan Normal University (Natural Science Edition),2020,19(04):45-51.

[2] Wang Baipeng. Research and application of text classification methods based on deep learning [D]. Northern University for Nationalities, 2020.

[3] Wang, C., Dong, Y., Zhang, Z., Wang, R., Wang, S., & Chen, J. (2024). Automated Genre-Aware Article Scoring and Feedback Using Large Language Models. arXiv preprint arXiv:2410.14165.

[4] Yu You, Fu Yu, Wu Xiaoping. Overview of Chinese Text Classification Methods[J]. Journal of Network and Information Security, 2019, 5(5): 1-8.

[5] Dai Liuling, Huang Heyan, Chen Zhaoxiong. Comparative Research on Feature Extraction Methods in Chinese Text Classification[J]. Journal of Chinese Information Processing, 2004, 18(1): 27-33.

[6] Ge Meng, Ouyang Hongji, Liu Minna. An Intelligent Webpage Information Filtering Model Based on ANN[J]. Modern Computer (Professional Edition), 2009.

[7] Liu, S., Liu, G., Zhu, B., Luo, Y., Wu, L., & Wang, R. (2024). Balancing Innovation and Privacy: Data Security Strategies in Natural Language Processing Applications. arXiv preprint arXiv:2410.08553.

[8] Wang, B., Zheng, H., Liang, Y., Huang, G., & Du, J. (2024). Dual-Branch Dynamic Graph Convolutional Network for Robust Multi-Label Image Classification. International Journal of Innovative Research in Computer Science & Technology, 12(5), 94-99.