Hierarchical Classification Algorithm Based on FastText

Yaxin Ran, Hongqi Han, Yunliang Zhang, Xiaorui Zhai and Mengjuan Weng

ABSTRACT

At present, methods for automatic assigning labels for literature using the Chinese Library Classification system are mostly based on machine learning methods, and many use the build knowledge base to improve classification effect. These methods are applicable to small-scale dataset, and as the categorical numbers increase, the classification effect will decrease and the time cost will increase dramatically. The paper proposes a hierarchical classification algorithm based on fastText text classification tool to solve classification indexing problem on large-scale literature dataset. The algorithm trains the two-layer classifier from the top-down method and designs an optimization algorithm to reduce the error transmission between the two layers. Experimental results showed that the optimized hierarchical classification technique improved the classification effect compared to the traditional classification technique.

1. INTRODUCTION

At present, most of the domestic library and information fields use the Chinese Library Classification (CLC) system to classify the information resources such as books, journal articles, and web pages [1]. The CLC system is the most widely used large-scale comprehensive classification in China. It uses a hierarchical structure to
classify literature resources, and its tertiary categories reach thousands. With the increasing enrichment of network data resources, how to accurately classify and index the large-scale categories of literature based on the CLC system has become one of the urgent problems to be solved in the field of library and information.

As a common algorithm for text classification, generally, the traditional machine learning classification algorithm is applied to the CLC system. Li and Ma [2] established a hierarchical classification model earlier in the CLC system and used machine learning algorithms to classify scientific paper data. Zhang et al. [3] used the hierarchical structure to refine the classification and finalize the classification based on the CLC system. Most of these methods use machine learning classification technology. As the number of data sets and the number of categories increase, the machine learning classification algorithm will face time overhead and classification accuracy.

In recent years, the classification algorithm based on neural network has shown excellent effect in the computer vision field, and more and more researchers have applied it to the Natural language processing (NLP) field. Guo [4] introduced the convolutional neural network [5] into the automatic classification of the literature for the first time and trained the classification model based on the CLC system. The final experiment achieved good results in the four-level classification. However, the model has a huge time overhead. The Dai et al. [6] first applied the fastText algorithm based on neural network structure to Chinese text classification. The experiment achieved a macro accuracy of 0.87 on 830,000 text data of 14 major categories, and the average training time was 1 minute. Compared to other deep learning classification algorithms, FastText solves the problem of the large time overhead.

This paper attempts to introduce the fastText algorithm into the automatic classification of literature based on the classification scheme of the CLC system and trains the hierarchical classifier and designs the optimization algorithm to solve the problem of transmission of error of hierarchical classification.

2. RELATED WORKS

For the case of a large number of dataset categories, the hierarchical classification methods are used for text classification mostly. This is based on the premise that the more the number of documents, the worse the classification effect is. For the CLC system, its hierarchical structure just fits the idea of hierarchical classification.

The text classification of the machine learning algorithm applied to the CLC earlier was from the Hou [1] research team. The team proposed combining the indexing experience with the machine learning method to classify the text and achieve better results. Li Sen and Ma Jun [2] applied the hierarchical classification idea to the text classification under the CLC system. The results show that the
hierarchical classification model oriented to the scientific and technological field can effectively improve the classification accuracy. Su et al [7] discussed the key issues of dataset skew, multi-layer classification and algorithm scalability for machine learning text classification technology, and proposed rationalization suggestions.

Most of these methods apply traditional machine learning classification algorithms such as support vector machine [8], k-nearest neighbor [9], etc. to the classification scheme of the CLC system. However, with the increase of the size and category of data sets, there will be three problems when using machine learning algorithm classification: 1) when the data set category is more, the classification effect is worse; 2) when the data set size is large, the training time Slower; 3) the data distribution is uneven under each category, increasing the difficulty of classification.

In recent years, the key technology of text categorization - text representation has made a big breakthrough. It has gone through the process from one-hot representation [10] to distributed representation [11]. The latter's advantage over the former is to solve the feature dimension disaster problem and consider the connection between semantics. Distributed representations have derived a variety of models, such as word2vec [12] (2013), GloVe [13] (2014), and so on.

The text classification algorithms based on these text representation models have also made great progress, and the neural network-based classification algorithms have become mainstream. Guo introduced the convolutional neural network into the automatic classification of the literature and achieved good results in the training of about 1.7 million data sets in the National Newspaper Index. However, this neural network model consumes a lot of time.

In 2016, Facebook open source a fastText [14] tool focused on text categorization. It provides a simple and efficient text classification method and the experiment shows that it is often on par with deep learning classifiers in terms of accuracy, and many orders of magnitude faster for training comparable to that of a deep learning classifier. The Dai et al. applied the fastText algorithm based on neural network structure to Chinese text classification for the first time. Experiments show that the fastText algorithm excels in efficiency and classification.

Based on this, this paper will use the fastText algorithm to study the hierarchical classification algorithm based on CLC system.

3. HIERARCHICAL CLASSIFICATION METHOD BASED ON FASTTEXT

The design method of the hierarchical classification consists of three parts: data pre-process, training classifier and test classifier. The goal of data pre-process is used to clean up the collected original literature data and transform data to required format for training. Then the stage of training classifier will train several classifiers in two layers based on fastText algorithm, the first-layer classifier is called level-1 classifier and the second-layer classifier is called level-2 classifier that includes 21
classifiers. The stage of test performs a test on the two-layer classifier with the test dataset, and finally obtains the classification result (Figure 1.).

Figure 1. An overview of hierarchical classification system.

3.1 Data Pre-process

The original literature should be cleaned and transformed to required format for training. Firstly, extracting key fields from the original literature, such as title, summary, and the CLC number. The CLC number is the label of the literature. After word segmentation and stop words removing, the data will be transformed to required format for fastText training, and the format of each text record is: "__label__, text", then the text records will be saved in a file of '.txt' format for the training requirements.

3.2 Training and testing Classifier

After the data pre-process, all the literature data will be divided into training dataset and test dataset according to a certain proportion (the value of proportion will be determined in the experiment). And the training data will be used as the input parameter of the text classification interface, provided by fastText, to train the classifier.

One level-1 classifier and 21 level-2 classifiers will be obtained by training. Then performing the test with the test dataset for hierarchical classification system. Firstly, the test dataset enters the level-1 classifier, obtains the level-1 CLC number, and then enters the corresponding level-2 classifier, and finally obtains the level-2 CLC number as the final classification result. The above is the normal method. However, the way may lead to error transmission, if the level-1 CLC number is
wrong, then final result must also be wrong. Thus, the following design method are proposed.

If the number of level-1 category in the training dataset is $N$, then after training, $N + 1$ classifiers are obtained. The level-1 classifier corresponds to the first layer classifier, denoted as $L_1$, and the level-2 classifier corresponds to the second layer classifier, denoted as $L_{2-i} (i = 1, 2, 3, ..., n)$.

Suppose document $D$ is classified by $L_1$ to obtain several categories $C_i (i = 1, 2, 3, ..., n)$, then $P_1(C_i)$ indicates the probability of prediction as $C_i$. When the document $D$ is classified into several categories $C_i$ by $L_1$, the $m_1$ categories whose probability is ranked in the top-$m_1$ are selected, and the document $D$ is respectively sent to the corresponding $m_1$ level-2 classifiers, and then each $L_{2-i}$ selects the $m_2$ categories with the highest probability. This finally results in $m_1 * m_2$ classification results, and then we select the category with the highest probability, which is the final classification result. Therefore, the mapping formula for defining $C_{k=\text{max}}$ and prediction probability is shown in (1):

$$C_{k=\text{max}} = f (\max_{1 \leq j \leq m_1 \ast m_2} P_j(C_k))$$

(1)

Where $j$ represents one of $m_1 \ast m_2$ results, and $P_j(C_k)$ represents the probability that document $D$ is assigned to $C_k$ under result $j$. The method is also known as $M$-double maximum class (MDMC) algorithm.

4. EXPERIMENT

4.1 Experiment Data

The experimental data is derived from the actual text classification project, which is process of training nearly 510,000 Chinese literature and getting a classifier, and then automatically indexing the tertiary CLC number for new literature. After pre-processing dataset, the data of 21 first categories and 932 tertiary categories is obtained. Then the dataset is divided into train dataset and test dataset according to the rate of 5:2.

4.2 Experiment Results

In order to verify that the fastText-based hierarchical classification algorithm is better than the machine learning-based hierarchical classification algorithm, the experiment uses SVM and KNN classification algorithms as benchmark tests. After training, a total of 22 classifiers are obtained. For Confirming the MDMC algorithm is better than the normal method, we add a contrast test of fastText model. After the hierarchical classification test, Figure 2 shows the test results on 21 categories of test
dataset. TABLE I lists the macro accuracy of hierarchical classification test and the time overhead of training and testing.

![Graph showing test results of hierarchical classification.](image)

**Figure 2.** Test results of hierarchical classification. We implement the test with three classification algorithms on 21 categories of test dataset, the X-axis indicates the 21 categories and the Y-axis indicates the range of accuracy.

**TABLE I. THE CONTRAST OF MACRO ACCURACY AND TRAIN TIME. THE FASTTEXT WITH MDMC METHOD ACHIEVES THE BEST MACRO ACCURACY OF 0.644.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Macro accuracy</th>
<th>Train time/s</th>
<th>Test time/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.562</td>
<td>25934</td>
<td>1768</td>
</tr>
<tr>
<td>KNN</td>
<td>0.543</td>
<td>19073</td>
<td>1243</td>
</tr>
<tr>
<td>FastText</td>
<td>0.635</td>
<td>1923</td>
<td>244</td>
</tr>
<tr>
<td>FastText(MDMC)</td>
<td><strong>0.644</strong></td>
<td><strong>1923</strong></td>
<td><strong>582</strong></td>
</tr>
</tbody>
</table>

From the results in Figure 2, there is a little difference between the accuracy of SVM model and KNN model, but the accuracy of fasttext model in almost all categories is higher than SVM and KNN model.

As can be seen from the **TABLE 1**, the macro accuracy of fastText is 0.635, which is 0.073 higher than SVM, and SVM is 0.02 higher than KNN. That reveals that fastText has better performance. In addition, the fastText model with MDMC algorithm increases 0.01, it verifies the MDMC algorithm is effective. Then the train
time of fastText model is almost 10 times faster than KNN and much more 10 times faster than SVM. Meanwhile, it also shows the test time of fastText is faster. It demonstrates that the fastText model is more efficient than SVM and KNN.

5. CONCLUSIONS

Aiming at the huge number of categories under the classification system of the CLC system, this paper proposed a hierarchical classification algorithm based on the fastText. In addition, the paper designs an optimization algorithm called MDMC algorithm to solve the problem of error transfer between hierarchical levels. Experiment shows that the performance of hierarchical classifiers based on fastText algorithm is improved compared some machine learning algorithm, and the accuracy of hierarchical classification is improved by using optimization algorithm.

ACKNOWLEDGEMENTS

This work is mainly supported by the Program of the China Knowledge Centre for Engineering Science and Technology (CKCEST-2019-2-2) and the Sentiment Analysis Research Based on Barrage Comments and User Reviews (QN2019-11).

REFERENCES


899