Text Classification Based on LDA and Semantic Analysis

Yongxia Jing, Heping Gou and Wei Sun

ABSTRACT

The quality of text features directly affects the text classification effect, in order to get the text features which have the high contribution to the text classification in class, this paper proposes a text classification method based on LDA model and category semantic similarity method. The method selects text document topic features by the LDA model and calculates the semantic similarity between these features and categories combined with the word vector model. According to the size of similarity, the weight of the text feature is obtained, and the text classification feature selection and text classification are realized. Finally, the feasibility, validity and correctness of the algorithm are verified by experiments.

1. INTRODUCTION

With the rapid development of big data and cloud computing technology, the data of Internet expands sharply, the big data era has arrived. How to get interesting and meaningful information in so much data is becoming a research hotspot. Text categorization is the effective means to realize data retrieval, which is widely applied in various fields, such as spam filtering, public opinion analysis, etc. Currently, common text classification methods mainly include decision tree [1],
K-nearest neighbor (KNN) [2], support vector machine (SVM)[3,4], neural network [5], rough set [6] and so on. Text classification needs training sample data to obtain the corresponding classification model, therefore, the high dimension of the sample data will directly influence the effect of text classification, feature selection is one of the effective methods to reduce the dimension of sample data [7], as a result, In this paper, a text classification method combining LDA model and semantic analysis is proposed to select the final classification features from two aspects: topic feature analysis of text and category semantic analysis of features, so as to reduce the computational overhead of this algorithm and improve the classification effect.

2. KEY TECHNOLOGY

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) algorithm was a method based on Bayes theory proposed by David Blei et al in 2003, which can find topics themes and characteristics of large-scale text documents, and has been widely applied in text mining, information retrieval and other fields [8]. LDA model is based on the assumption that a text document is composed of multiple topics, and each topic is a probability distribution of word set. A probabilistic graphical model composed of three levels of word-topic-document is shown in Figure 1.

K is the number of topics, M is the number of text documents, N is the word number of the text document, θ is the topic probability distribution, φ is the word probability distribution of the topic, α is the Dirichlet prior distribution parameter of the topic distribution θ, and β is the Dirichlet prior distribution parameter of the word distribution φ. For the word \( w_{i,j} \) of text document \( d_i \), the generation process includes three aspects:

Figure 1. Probabilistic graphical model for LDA.
In LDA model, the hyper-parameters α and β are set by practical experience, but parameter θ and φ need to be estimated, and the commonly used algorithms include Gibbs Sampling algorithm, variational Inference algorithm and so on. Gibbs Sampling method is simple to implement and can extract topics from text document set quickly and effectively. The LDA model of Gibbs Sampling is used for training and prediction process as follows.

Training process:
1) Set the value of α and β.
2) A topic number is randomly assigned to each word in each text document in the corpus.
3) Rescan the database, for each word, update its topic number according to Gibbs Sampling and update the number of the word in the corpus.
4) Repeat step 2) until Gibbs Sampling convergence.
5) Topic distribution θ is obtained according to the topic of words in each text document in the corpus, and topic-word distribution φ is obtained according to the distribution of subject words in the corpus.

For the new text document $d_i$, the prediction process is as follows:
1) For each word in $d_i$, randomly assign a topic number.
2) Rescan $d_i$ and resamples its topic number by Gibbs Sampling.
3) Repeat step 2) until Gibbs Sampling convergence.
4) The topic of each word in $d_i$ is counted, and the topic distribution of $d_i$ is the prediction result.

2.2 Feature weights based on TF-IDF

Feature weights were calculated using TF-IDF (term frequency-inverse document frequency) model [9]. According to this model, if a certain feature $t_i$ appears more frequently in document $d_j$ and less in other documents, then $t_i$ has better classification ability, then

$$TFIDF_{i,j} = TF_{i,j} \times IDF_{i,j}$$

(1)

Where, $TF_{i,j}$ represents the frequency of the feature $t_i$ in document $d_j$

$$TF_{i,j} = \frac{N_{i,j}}{N_j}$$

$N_{i,j}$ is the number of occurrences of the feature $t_i$ in document $d_j$, and $N_j$ is the number of all the features in document $d_j$.

$$IDF_{i,c} = \log \frac{N_c}{N_{tf}}$$
$N_C$ refers to the total number of text documents in sample set C, and $N_{f_i}$ refers to the number of text documents in sample set containing the $t_i$ feature.

3. TEXT CLASSIFICATION BASED ON LDA AND SEMANTIC ANALYSIS

In order to better select features, this paper proposes a feature selection algorithm combining topic model and semantic analysis. Topic model can select important topic features of text documents according to two matrixes: document-topic matrix and topic-word matrix. While method based on TF-IDF can select features with good classification ability, the value of TF-IDF will increase with the increase of word occurrence in the document and decrease with the increase of word occurrence in the corpus. For two different categories: document $d_j \in C_j$ and $d_k \in C_k$, for features $t_i$ and $t_j$, if they occur equally in documents $d_j$ and $d_k$, and not in other documents, then $t_i$ and $t_j$ have the same weight in both documents. That is, the classification ability are same, but in fact, because they are in two different classes, Their ability to distinguish categories should be different. The method based on TF-IDF feature weight algorithm and semantic similarity of category name is proposed to adjust feature weight.

Given a sample set C and its category name $c_n$, for the document $d_j (d_j \in C)$, for the feature word $t_i$, its semantic similarity degree $s(t_i, C)$ with $c_n$ is

$$s(t_i, C) = sim(w_i, c_n)$$  \hspace{1cm} (2)

Where, $sim(w_i, c_n)$ is calculated based on word vector. Word vector is able to reflect the meaning of the word, It can be adopted to get the semantic similarity between words and words. Therefore, word vector model needs to be obtained through word vector training. This paper adopts wikipedia corpus and uses word2vec model for word vector training [10]. The word vector representation $\overrightarrow{w_i}$ and $\overrightarrow{c_n}$ of $w_i$ and $c_n$ can be obtained by word2vec word vector model, and then the cosine similarity between the two vectors is calculated, that is

$$sim(w_i, c_n) = \cos(\overrightarrow{w_i}, \overrightarrow{c_n}) = \frac{\overrightarrow{w_i} \cdot \overrightarrow{c_n}}{||\overrightarrow{w_i}|| \cdot ||\overrightarrow{c_n}||}$$

Where $||\overrightarrow{w_i}||$ is the length of the vector, therefore, the weight of the feature $t_i$ is calculated by combining equations (1) and (2), and the formula is

$$tf\_idf\_class = TFIDF_{i,j} \ast s(t_i) = (TF_{i,j} \times IDF_{i,j}) \ast s(t_i, C)$$  \hspace{1cm} (3)
Given a corpus $C = \{C_1, C_2, \cdots, C_K\}$ contain k text classes, for any text class $C_k$ and all text document set $D_{C_k}$ in the class, the training process of text classification model based on LDA and category semantic analysis is described as follows:

Step 1: Use Word2Vec to conduct word vector training and obtain word vector model.

Step 2: Perform word segmentation on all text documents and stop word processing to convert text documents into word sequences.

Step 3: Vectorization of all text documents, such as text document $d_i \in D_{C_k}$, its word sequence is expressed as $W_{d_i} = (w_1, w_2, \cdots, w_m)$. The document $d_i$ word vector is obtained using the word vector model and the formula (1), then, for the text feature $w_i (i = 1, 2, \cdots, m)$, $Sim(w_i, D_{C_k})$ is calculated according to the word vector model and formula (2).

Step 4: Select the features of text document and obtain the feature set that can reflect the topic content of text document using LDA model. For example, for $d_i$, the new feature set $W'_d_{d_i} = \{w_1, w_2, \cdots, w_n\} (n \leq m)$ is obtained through LDA-based feature selection algorithm.

Step 5: Get the feature dictionary of sample set according to the feature set $W'_d_{d_i}$ of each text document, and then calculate the weight tf_idf_class of each feature by using formula (3) to achieve the vectorization representation of text document.

Step 6: Apply the same method to vectorize the new text document according to the acquired sample set feature dictionary. Then SVM is used for classification to obtain the final category of samples.

4. EXPERIMENT AND RESULT

In this paper, SVM text classification algorithm is adopted to verify the validity and correctness of the text classification algorithm proposed in this paper. Tc-corpus-answer of Chinese text set compiled by Li Ronglu of Fudan University is adopted as experimental data. A total of 653 texts were selected from the data set, consisting of 11 categories. In order to facilitate analysis, the classification algorithm using LDA theme model and TFIDF feature weighting method is expressed as lda-tfidf, and the text classification algorithm proposed in this paper is expressed as lda-tfidf-class. SVM classification algorithm is evaluated by traditional text classification performance evaluation indexes: precision, recall, and F1 (f1-measure). Wikipedia is used to train the word vector model. The experimental results are shown in TABLE I.
TABLE I. COMPARISON OF TEXT CLASSIFICATION RESULTS.

| Category   | precision lda-tfidf | recall lda-tfidf-class | f1-score lda-tfidf | precision lda-tfidf-class | recall lda-tfidf-class | f1-score lda-tfidf-class |
|------------|---------------------|------------------------|-------------------|---------------------------|------------------------|--------------------------|
| Transport  | 0.83                | 0.91                   | 0.87              | 0.87                      | 0.87                   | 0.87                     |
| Military   | 0.80                | 0.92                   | 0.86              | 0.86                      | 0.86                   | 0.86                     |
| Medical    | 0.83                | 0.83                   | 0.83              | 0.83                      | 0.83                   | 0.83                     |
| Philosophy | 0.79                | 0.92                   | 0.85              | 0.85                      | 0.85                   | 0.85                     |
| Education  | 0.77                | 1.00                   | 0.87              | 0.87                      | 0.87                   | 0.87                     |
| Literature | 1.00                | 0.50                   | 0.67              | 0.73                      |                        |                          |
| Law        | 0.60                | 0.55                   | 0.57              | 0.64                      | 0.57                   | 0.64                     |
| Electronics| 0.64                | 0.88                   | 0.74              | 0.78                      | 0.74                   | 0.78                     |
| Mine       | 1.00                | 0.58                   | 0.74              | 0.74                      | 0.74                   | 0.74                     |
| Energy     | 0.83                | 0.62                   | 0.71              | 0.80                      |                        |                          |
| Communication | 0.88            | 0.78                   | 0.82              | 0.82                      | 0.82                   | 0.82                     |
| Average    | 0.82                | 0.77                   | 0.78              | 0.81                      |                        |                          |

According to the experimental results, the precision, recall and f1-measure of the algorithm proposed in this paper are 2%, 3% and 3% higher than the lda-tfidf algorithm, indicating that the algorithm proposed in this paper can well consider the topic meaning and category semantics of text features and achieve good results.

5. CONCLUSIONS

In this paper, a text classification algorithm based on LDA and semantic analysis is proposed. The main purpose of this algorithm is to obtain feature words that can express text topics, but these feature words can only reflect the semantic information of text documents themselves, and cannot well reflect the category characteristics of sample sets. Therefore, word vectors based on word2vec model are adopted to obtain category names and text features, and the importance of text features to categories is measured by the similarity between word vectors. The greater the similarity, the greater the classification contribution and the higher the weight. Finally, the effectiveness of the algorithm is proved by experimental analysis.

ACKNOWLEDGMENTS

The Project was supported by Natural Science Foundation of Hainan Province, China (No. 617160, No. 618MS086), Educational and Teaching Reform Research
REFERENCES

1. Tong D., Qu Y. R., Prasanna, V. K. 2017. “Accelerating Decision Tree Based Traffic Classification on FPGA and Multicore Platforms,” *IEEE Transactions on Parallel and Distributed Systems*, 28(11):3046-3052.
2. Weinberger K. Q. 2009. “Distance Metric Learning for Large Margin Nearest Neighbor Classification,” *JMLR*, (10):207–244.
3. Saunders C, Stitson M. O, Weston J., et al. 2002. “Support vector machine,” *Computer Science*, 1(4):1-28.
4. Abd A. M., Abd S. M. 2017. “Modelling the strength of lightweight foamed concrete using support vector machine (SVM),” *Case Studies in Construction Materials*, 6:8-15.
5. Lam H. K., Ekong U., Liu H., et al. 2014. “A study of neural-network-based classifiers for material classification,” *Neurocomputing*, 144(1):367-377.
6. Pacheco F., Cerrada M., Sánchez R. V., et al. 2017. “Attribute clustering using rough set theory for feature selection in fault severity classification of rotating machinery,” *Expert Systems with Applications*, 71:69-86.
7. Zhou J. Z., Cui X. H. 2019. “Keyword extraction method based on word vector and TextRank,” *Application Research of Computers*, 36(4): 1051-1054.
8. Wang F. X., LI F. 2018. “Emotion Classification Method Based on Topic Weighted LDA Model,” *Journal of Southwest China Normal University (Natural Science Edition)*, 43(09):38-44.
9. Hou Y., Dai M.H. 2013. “News Recommendation Technology Combining Semantic Analysis with TF-IDF Method,” *Computer science*, 40(11A):267-300.
10. Tang, M., Zhu, L., Zou, X.C. 2016. “Document Vector Representation Based on Word2Vec,” *Computer Science*, 43(6):214-217.