Application Research of text classification based on random forest algorithm

P.M. Jyothi Sailaja, P.Kavya, M.Navya Kranthi, N. Vishnukumari, Ms M. Amala
Swarnandhra college of Engineering
and Technology,Narsapur

Abstract—
In view of the poor classification effect of traditionalrandom forest algorithm due to the low quality of text feature extraction, a random forest method for text information is proposed. In view of the difficulty in controlling the quality of traditional random forest decision trees, a weighted votingmechanism is proposed to improve the quality of decision trees.This algorithm uses tr-k method based on text feature extraction to improve the quality and diversity of text features, and uses the latest Bert word vector generation model to represent the text. Experimental data in Python environment show that this method can achieve better results in text classification than IDF based random forest algorithm and original random forest algorithm.

Keywords:Random Forest;Text Classification;tr-k method

I. INTRODUCTION

With the rapid development of science and technology, since the 1990s, more and more data information has been generated, 80% of which is stored in text. Therefore, peoplecan't use the traditional manual filtering for huge amount oftext information. Text processing based on natural languageprocessing emerges as the times require. In recent years, there are more and more researches on text classification, mainly focusing on Naive Bayes, K-means clustering, SVMand other algorithms. Random forest algorithm is widely used in all walks of life due to its advantages of fast trainingspeed, easy parallel computing in the era of big data, stronganti-interference ability and excellent anti over fitting ability,and has achieved the effect of traditional methods.

For the study of text classification, many predecessors have done a lot of excellent work. For example, ZhouQingping proposed an improved KNN algorithm based on clustering; Yang, improved the feature selection function byconnecting the accurate coefficients of several feature selection functions to form a new feature selection function,and finally used SVM to classify; Zhang Xiang proposed animproved algorithm based on bagging's Chinese text classifier. Based on the increase of text information and the

development of text processing technology, the application of text classification is more and more. For example, public opinion monitoring, emotional analysis, commodityclassification, news classification, etc.

Many advantages of random forest algorithm (RFA)make experts and scholars carry out many improved application research on RFA. In 1995, tin Kam ho first proposed the concept of random forest. Later, Leo boeimanproposed that RFA is a classification and prediction model.M p Perrone, l n coope and others proposed that in the classification stage, RF class labels are synthesized from theclassification results of all decision trees, and are the most commonly used methods in voting and probability average.In terms of application, EI atta proposed a method to predictthe activity of cannabinoid receptor (CB2) agonist using RFin bioinformatics; in ecology, eruan et al. Studied air prediction using RFA; in genetics, retroria used RFA in generecognition. Moreover, RFA has achieved good results in biochip, information extraction and other fields.

II. ALGORITHM INTRODUCTION

The random forest algorithm (RFA) is composed ofmany decision trees. By voting the decision results of each decision tree, the category with the most votes is the result ofthe random forest algorithm. Because of the parallel computing, the random forest algorithm is easy to generalizeand not easy to over fit. It has been applied in biology, medicine, information retrieval and other fields. The main content of this chapter is to explain the basic construction process of RFA and the basic knowledge of decision tree, soas to pave the way for the subsequent optimization of RFA.

A. *Random forest algorithm*

First of all, we understand information entropy, which was put forward by Shannon in 1948 to solve the problem of quantitative measurement of information. The followingformula (1)

$1(X) = -P(u\ ) \log P(u\ )$sampling method, n samples are extracted to get a new

$\sum$ i $i=1$

(1)sample set.

In the third step, t(T<=t) attributes are randomly selected

Where I represents each message and R represents the number of messages.

In order to facilitate the later calculation, this paperproposes to delete the negative sign in front of the accumulation symbol, which is defined as purity. The lowerthe corresponding information entropy, the higher the purityof the corresponding data set.from the given t attributes. By using the optimal feature standard of a decision tree, the optimal classification node isselected so that all the sub samples are leaf nodes.

In the fourth step, repeat the third step of K, generate K decision trees, and get the final random forest.

In the fifth step, H(x) is the function model of theclassifier, the decision tree is represented by Hi, y is the target variable (classification label), and I (*) is the indicator

$$Ent(D) = \sum P_k \log_2 P_{k^1} \ (2)$$

function. The decision-making formula of random forest isas follows (6):

The above formula (2) is the purity calculation formula

defined in this paper. D represents a data set, K represents aclassification caused by a certain attribute, which is dividedinto y categories in total. We assume that due to attribute T,data set u can be divided into several classes, and then

$$H(x) = \arg\max \sum_{i=1} I(h_i(x) = Y)$$

C. Description of Bert word vector model

(6)

calculate the information gain of attribute t. As follows (3)

$$\frac{|D^U|}{|D|}$$

U

The Bert model was proposed by Google in November2018. It is called bidirectional encoder representations from

$$Gain(D,t) = Ent(D) - \sum$$

$$Ent(D^U\ )$$

transformers. Its model structure is shown in the figurebelow. Therefore, we can use formula (3) as our optimal featurestandard, and get ID3 algorithm. However, ID3 algorithm can not deal with continuous variable attributes, and in attribute preference, it tends to attribute classification attributes, which affects decision-making. Therefore, in 1993Quinlan proposed C4.5 algorithm, using information gain rate as the optimal feature standard, thus solving the attributepreference problem of ID3. The calculation formula of information gain rate is as follows (4), (5)

$$Gain\_rotio(D,t) = \frac{Gain(D,t)}{\ }$$

(4)

$$\frac{|D^U|}{|D|}$$

•••

$$IV(t) = -\sum^{U}$$

$$\frac{|D^U|}{|D|}$$

$\log_2)$

$IV(t)$

$U=1$

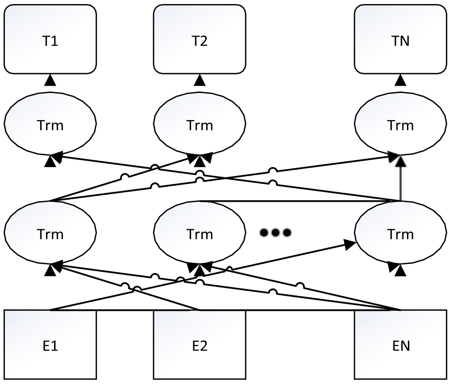

Figure 1. Bert model structure

In Figure 1, E1, E2,..., en represent Chinese short text

The main key point of RFA is two random sampling, oneis the use of bootstrap with put back sampling, so the new data set will contain 2 / 3 of the content of the old data set, thus generating out of bag data. Another random sampling, produced in the selection of features, each time we generatea decision tree, the features used are not exactly the same. The decision tree is generated by randomly extracting features less than the total number of features from the givenfeatures. Connecting the decision trees generated in front ofeach other becomes a random forest.

B. RFA algorithm steps

The algorithm steps of random forest algorithm in text classification are as follows:

The first step is text preprocessing. Firstly, the "noise" such as stop words and symbols in the text is removed. Using word

2 VEC word embedding model, the text information is vectorized and the training set is generated.

The second step is to assume that the training set containsn samples and T classification attributes. Using bootstrapinput. After bidirectional transformer encoder, we get the vector representation of text, that is, Bert and transformer's encoder network structure are exactly the same.

III. Improved RFA algorithm

Firstly, the traditional random forest algorithm doesn't consider the particularity of text data, and it can't improve the level of text classification because of the poor quality offeature extraction. Secondly, the random forest algorithm itself has the problem that similar decision tree will cover upthe real classification decision. This paper uses the most advanced word vector generation model Bert to improve it.

A. Tr-k method to improve the quality of text features

In the traditional random forest algorithm, the number and quality of feature selection are not prominent. But for books and other large capacity text classification, the more the number and quality of text features (classification decision tree attribute), the better the classification effect will be. Therefore, this paper proposes a tr-k method which

combines TF-IDF, textrank and K-means to improve theeffect of text classification.nodes pointing to other web pages.

$In(V_j)$

represents the

The full name of TF-IDF method is term frequency inverse document frequency, which is a common technology for information retrieval and data mining. TF value refers to number of nodes in the collection. Score the importance ofweb page as formula (10).

$S(V_j)$

the importance of the ith word in file J. $S(V_i) = (1-d) + d \times \sum$ (10)$_j \sum n_{k,j}$

(7)

$_{j \in In(V_i)}$ $|Out(V_j)|$

D is the damping coefficient. When no one visits the webpage in extreme cases, the formula will be meaningless.Therefore, the damping coefficient is added to avoid this

In formula (7), the numerator represents the frequency of the ith word in file J, and the denominator represents the totalfrequency of K words in file J.

$$\frac{|D|}{|\{j : t_i \in d_j\}| + 1}$$

situation.

Textrank is an extraction method for text keywords.Compared with web pages, the biggest difference is that it

log

generates directed weight graph. A variable window is usedto scan the sentences of the article. After removingthe stopwords, each word in the window is considered to be related

In formula (8), D represents the total number of files in

t_i corpus. d_j represents the ith word we want to test,

to each other, and the cosine similarity between adjacent words is calculated to calculate the weight between each word. Therefore, the calculation formula is as follows (11).

$|\{j : t_i \in d_j\}|$ represents the vocabulary set of file J

$WS(V_i) = (1-d) + d \times \sum$

W_{ji}

(11)

$W \times WS(V_j)$

containing, and represents the frequency of all filescontained. The reason why we need to add one operation in$_{j \in In(V_i)} \sum$

$_{jk}$

$_{V_k \in Out(V_j)}$

denominator is to avoid the occurrence of meaninglessdivision by zero.$WS(V_i)$ is the score of node V_i, W_{ji} is the weight of

node V to V calculated by cosine similarity.

$TFIDF_{ij} = TF_{ij} \times IDF_{ij}$(9) $_j$ $_i$

Through formula (8) and formula (9), high word frequency and low file frequency in a file set are analyzed togenerate TF-IDF with high weight. It can be seen from the conclusion that this algorithm tends to filter out common words and retain important words. The disadvantage is that the beginning and the end of the text have different importance to the semantics, and can not reflect the locationinformation of words.

Textrank algorithm comes from Google's PageRank algorithm, which is used to evaluate the importance of a webpage. It uses directed powerless graph to score. Set V_j as

the node of web page J, $In(V_j)$ as the collection of nodespointing to web page J, and $Out(V_j)$ as the collection of

This keyword extraction method has the advantages of low computational complexity and easy processing. K-mean algorithm is an unsupervised classification algorithm, whichis widely used because of its simple calculation and excellent clustering effect. But using k-means algorithm, the key is tofind the appropriate K value, that is, to initialize the center ofmass. If the selection is right, the efficiency and accuracy ofthe algorithm will be improved obviously. Therefore, this paper uses the feature set extracted by TF IDF and textrankas the k-value input of K-means algorithm, and conducts clustering analysis. To sum up, the flow chart of tr-k text feature extraction module proposed in this paper is shown inFigure 2.

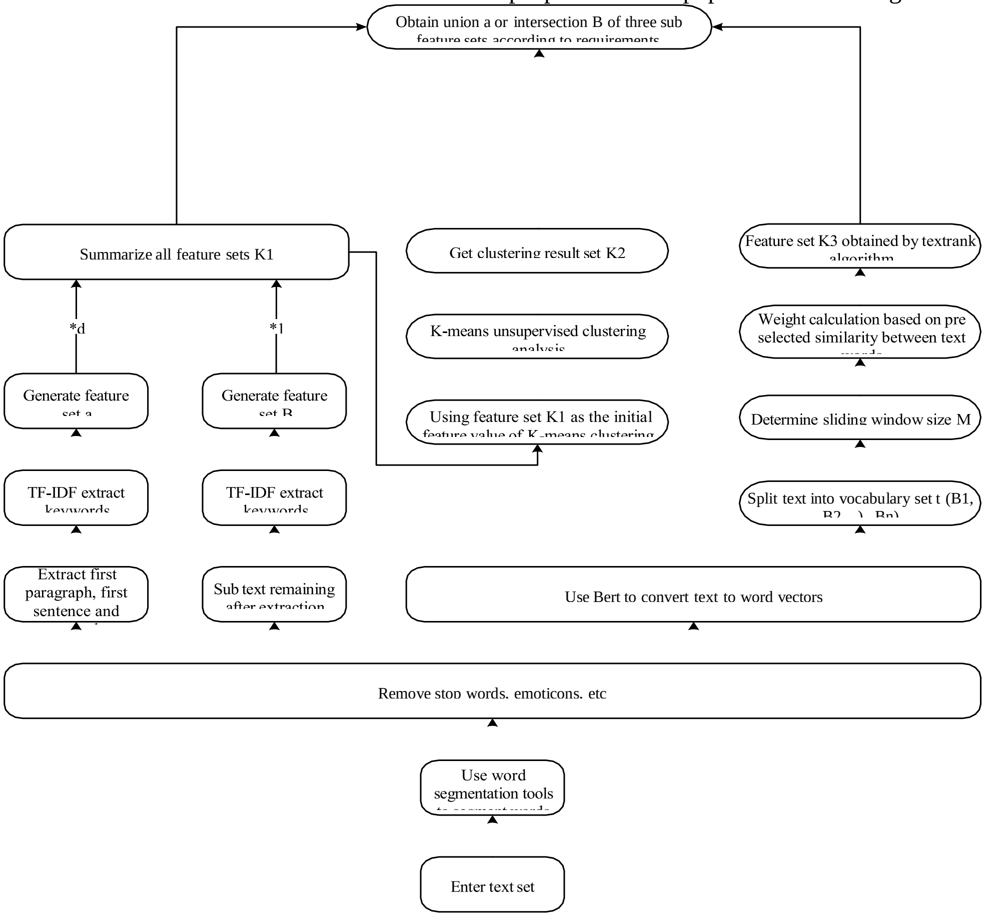
Obtain union a or intersection B of three sub feature sets according to requirements
Summarize all feature sets K1
*d
*1
Generate feature set a
Generate feature set B
TF-IDF extract keywords
TF-IDF extract keywords
Extract first paragraph, first sentence and
Sub text remaining after extraction
Get clustering result set K2
K-means unsupervised clustering analysis
Using feature set K1 as the initial feature value of K-means clustering
Feature set K3 obtained by textrank algorithm
Weight calculation based on pre selected similarity between text words
Determine sliding window size M
Split text into vocabulary set t (B1, B2, ..., Bn)
Use Bert to convert text to word vectors
Remove stop words, emoticons, etc
Use word segmentation tools to segment words
Enter text set

Figure 2. Structure of tr-k method for feature extraction

Tr-k method for the extraction of text features, first of all,preprocess the input text set, remove stop words. Then, feature set K1 is extracted by TF-IDF and feature set K2 is extracted by textrank. The intersection of feature set K1 andK2 is taken as the clustering center of K-means, and featureset K3 is obtained. By using three text feature extraction methods, we can get more diversified and high-quality text feature sets.

IV. EXPERIMENT SIMULATION AND ANALYSIS

A. Experimental environment

The experimental environment of algorithm performanceanalysis takes Windows 7 operating system as the support ofthe whole experiment, and uses Python as the compiling language of programming. Related configuration: Intel (R) core (TM) i5-

3470 CPU processor, 3.20ghz processor, 8gmemory, and pycharm Community Edition 2017.5.1development tool.

B. *Experimental data set*

This paper uses 20 newsgroups data sets, with 18000 articles, involving 20 topics, which are divided into trainingsets and test sets, and there is no cross between them.International standard data set, which is usually used for textclassification, information retrieval and text mining.

C. *Experimental design and analysis*

Through three effective model experiments, the traditional random forest algorithm and tf-rf algorithm whichonly uses TF-IDF to extract text features are compared withTrk Bert RF model designed in this paper. The experimentalcomparison is mainly divided into the following aspects: running time, classification accuracy and F1 value. In order

to ensure the stability and contrast of the experimental data,the comparative experiments are carried out when the trees of the decision tree are 50, 70, 100, 200, 300 and 400 respectively, and run 10 times under the same experimentalenvironment, taking the average value as the final experimental results.

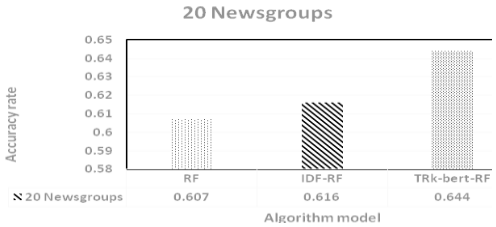

Figure 3. text classification accuracy under different models

V. CONCLUDING REMARKS

In this paper, the input text data set of random forest algorithm is processed to improve the classification effect. At the same time, we use Bert word vector model to improvethe quality of text representation, and then improve the classification accuracy of the final random forest. Experiments show that the model can effectively improve the classification accuracy and F1 value.

REFERENCES

Korde V, Mahender C N.Text classification and classifiers:A survey[J].International Journal of Artificial Intelligence&Applications, 2012, 3 (2) :85tkin L V , Konstantinov A V , Chukanov V S , et al. A weighted random survival forest[J]. Knowledge-Based Systems, 2019, 177(AUG.1):136-144.05Yang Y.Are-examination of text categorizationmethods[C]//International ACM SIGIR Conference on Research and Devel-opment in Information

Retrieval - ACM , 1999: 42-49 -

[1] Mantas C J , Castellano J G , Serafín Moral-García, et al. A comparison of random forest based algorithms: random credal random forest versus oblique random forest[J]. 2019.

[2] T.K.Ho. Random Decision Forest [J].In Proceedings of the 3rd International Conference on Document Analysis and Recognition.Montreal,Canada,1995,8:278-282.

[3] Breiman L.Random forests[J].Machine Learning, 2001, 45 (1) :5~32.

[4] L K Hansen, P Salamon.Neural network ensembles[J].Pat-tern Analysis and Machine Intelligence, 1990, 12 (10) :993~1001.

[5] M P Perrone, L N Cooper.When networks disagree:Ensem-ble method for neural net works[A].Artificial Neural Net-works for Speech and Vision[C].NewYork:Chapman&Hall, 1993.126~142.

[6] El-Atta A H A, Moussa M I, Hassanien A E. Predicting Biological Activity of 2,4,6-trisubstituted 1,3,5-triazines Using Random Forest[J]. 2014, 303:101-110.

[7] [10] Erwan Scornet, Gérard Biau, Jean Philippe Vert. Consistency of random forests[J]. Eprint Arxiv, 2015, 9(1):2015--2033.

[8] Petralia F, Wang P, Yang J, et al. Integrative random forest forgene regulatory network inference.[J]. Bioinformatics, 2015, 31(12):i197.

[9] Kimura S, Tokuhisa M, Okada-Hatakeyama M. Inference of genetic networks from time-series of gene expression levels using random forests[C]. Computational Intelligence in Bioinformatics and Computational Biology. IEEE, 2017:1-6.

[10] Janitza, Silke, Tutz, Gerhard, Boulesteix, Anne-Laure. Random forestfor ordinal responses: Prediction and variable selection[J]. Computational Statistics & Data Analysis, 96:57-73.

[11] Lee J , Yu I , Park J , et al. Memetic feature selection for multilabel text categorization using label frequency difference[J]. Information Sciences, 2019, 485:263-280.

[12] Tang X , Dai Y , Xiang Y . Feature selection based on featureinteractions with application to text categorization[J]. Expert Systemswith Applications, 2018.