2015 IEEE International Conference on Computational Intelligence & Communication Technology

Research of Text Categorization Model based on Random Forests

Dashen Xue 1st Affiliation
Transportation and Management Department Dalian
Maritime University
Dalian P. R. China
e-mail: xds59@dlmu.edu.cn

Fengxin Li 2nd Affiliation
Transportation and Management Department Dalian
Maritime University
Dalian P. R. China
e-mail: lifengxin@dlmu.edu.cn

Abstract—**Due to the good performance in computation speed and efficiency, Random Forest (RF) algorithm as a famous integrated learning algorithm has been widely applied in many fields. In addition, because of the rapid development of Internet, text categorization has become the key technology to process and organize large scale documents. It is appealing and important to employ RF algorithm to deal with text documents categorization problem. This paper introduces the details of RF algorithm and assess the text documents categorization model by using RF algorithm.**

Keywords-component; Text Categorization, Random Forests, Decision Tree

I. INTRODUCTION

Internet has been developing rapidly and has become an enormous information center since its emergency. It is important to apply information processing method to dig useful information within such center. Because most of the network information is stored in the form of text in different servers' hard disks or in databases, and most of them is semi-structured text data; text categorization has become a hotspot in the research of the modern information processing. Text categorization [1] is defined as determining a category for each document in a collection of documents according to the predefined topic category. Through text categorization, text can be classified, and thereby the efficiency of information search by end users can be greatly improve. Until now, there are a number of text categorization models have been proposed such as k neighbor algorithm, naive Bayesian algorithm, support vector machine (SVM) algorithm, neural network and decision tree algorithm. Among them, Random Forests (RF) [2] algorithm is a famous integrated learning algorithm by taking the decision tree as basic categorizer. The algorithm does not require a priori knowledge, and has high categorization accuracy without over-fitting problem. The specific details of RF algorithm will be introduced in following sections.

II. RANDOM FORESTS

A. Principle

It is well known that basic categorizer has some limitations and hence cannot be efficiently applied. To improve these limitations, the Integrated Learning (IL) has been developed and used. By combining multiple basic categorizers and analyzing the result of each basic categorizer, IL is able to determine the category of target sample with improved accuracy and can effectively improve the generalization ability of basic categorizer. However, not any combination of more than one basic categorizers can be called integrated learning. Two requirements need to be satisfied for IL: 1) each basic categorizer has to be valid;2) employed basic categorizers are different from each other

RF [2] is an integrated learning model proposed by the American scientist Leo Breiman. It is based on the K decision tree $\{h(X,\theta_k), k=1,2,...,K\}$ as a basic categorizer. The categorization result of the RF is decided by each decision tree through a simple voting way. The $\{h(X,\theta_k), k=1,2,...,K\}$ is a random sequence and determined by the following tow random ideas:

a) Bagging ideas: Assuming that the original sample set has a capacity of M, Bagging ideas is to randomly select samples from the original sample set as the single decision tree training set $\{T_m, m=1,2,....M\}$. *It should be noted that.*

b) Feature subspace ideas: In the process of constructing a decision tree, when a node is split, an attribute subset has to be first extracted from the all the nodes attributes. Then, an optimal attribute is selected from the subset to split the node.

In the process of constructing a decision tree, both training set and attribute subset are selected randomly. Hence, $\{\theta_k, k=1,2,...,K\}$ is a random sequence and makes RF have very strong generalization ability. Fig. 1 illustrates the flow chart of decision tree construction process.

978-1-4799-6023-1/15 $31.00 © 2015 IEEE
DOI 10.1109/CICT.2015.101

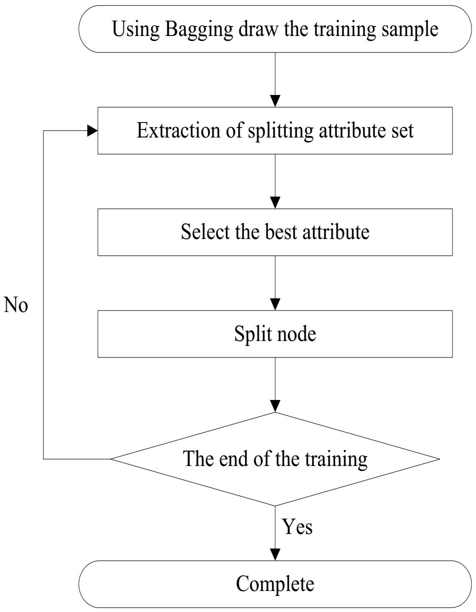
Using Bagging draw the training sample
Extraction of splitting attribute set
Select the best attribute
Split node
The end of the training
No
Yes
Complete

Fig. 1 The construction process of decision tree

B. *Theoretical basis*

- Overfitting

Suppose that there are k classifiers $\{h(x,\theta_1),h(x,\theta_2),\dots,h(x,\theta_K)\}$, and the training set random vector are Y, X distribution. Edge function is defined as [4]:

$$m_g(X,Y)=av_k I(h_k(x)=y)-\max_{j\neq y} av_k I(h_k(x)=j) \quad (1)$$

where $I(*)$ is indicator function, $av_k(*)$ is average function. Edge function correctly classified under portrayed Y and X number of votes exceeds the maximum extent of the votes of other categories. The confidence and the value of classifier is positive correlation. So the generalization error is:

$$PE^*=P_{X,Y}(m_g(X,Y)<0) \quad （2）$$

The subscript X, Y shows the definition of probability space.

In the random forests, $h_k(x)=h(x,\theta_k)$. When the number of decision tree is large, the law of Large Numbers will be obeyed. Hence, we have following Theorem:

Theorem 1[5]. With the increase of the number of decision tree, for all the sequence θ_i, PE^*, almost everywhere converge to:

$$P_{X,Y}(P_\theta(h(X,\theta)=y)=y-\max_{j\neq Y} p_\theta(h(X,\theta)=j)<0) \quad (3)$$

where θ is the random vector of the decision tree, $h(X,\theta)$ is the output based on X and θ.The reference [5] proved that with the increase of the size of the decision tree in the RF, the generalization error of the RF develops towards a upper bound, which shows that RF has a very strong generalization ability.

- Generalization error

The edge function of random forests is defined as:

$$mr(x,y)=P_\theta(h(x,\theta)=y)-\max_{j\neq Y} P_\theta(h(x,\theta)=j) \quad （4）$$

The strength of the categorizer $\{h(x,\theta)\}$ is:

$$s=E_{x,Y}mr(x,y) \quad （5）$$

Assuming that $s\geq 0$, based on chebyshev inequality, we obtain that

$$PE^*\leq \mathrm{var}(mr)/s^2 \quad （6）$$

Inequality (6) requires that $\mathrm{var}(mr)$ has the following form:

$$\begin{cases}\mathrm{var}(mr)=\rho(E_\theta sd(\theta))^2\\ \mathrm{var}(mr)\leq \rho E_\theta \,\mathrm{var}(\theta)\end{cases} \quad （7）$$

At the same time

$$\begin{cases}E_\theta \,\mathrm{var}(\theta)\leq E_\theta(E_{x,y}mg(\theta,x,y))^2-s^2\\ E_\theta \,\mathrm{var}(\theta)\leq 1-s^2\end{cases} \quad (8)$$

By the (6), (7) and (8) can get the following conclusion:

Theorem 2. The generalization error upper bound for random forests:

$$PE^*\leq \rho(1-s^2)/s^2 \quad (9)$$

where ρ is the average correlation coefficient, s is the strength of the tree.

According to Theorem 2, the generalization error upper bound can be calculated based on the strength of the decision tree in the forest and the dependence between the decision trees.

III. TEXT CATEGORIZATION MODEL BASED ON RANDOM FORESTS

Steps of the algorithm are shown as follows:

- The establishment of vector space model (VSM)

Using the word segmentation, the weight calculation method and the feature selection algorithm, the training set is converted to feature vector.

- Structure random forests classifier

Random forests is consisting of multiple decision trees. Hence, the forest constructing is mainly the process of constructing decision tree. By the principle of random forests, the steps to construct the decision tree are:

a) Using Bagging to extract the training sample set of decision tree from the original sample set.

b) Using CART algorithm to build a decision tree in the way of completely growth. It should be noted out that in the process of constructing a decision tree, when a node is split, an attribute subset should be first extracted from all the nodes attributes. Then an optimal attribute is selected from the subset to split the node. In this article, the size of the attribute subset is $\lfloor \log_2(N)+1\rfloor$ *(N is the size of all attributes).*

- Classify

Convert the testing sample set to feature vector, then input the feature vectors into the model. The output of the model is the categorization result of testing sample set.

IV. Experimental design and analysis

A. Experimental design

The Sogou laboratory text categorization corpus is used in this paper as benchmark corpus including five categories: IT, economics, health, tourism and sports. The size of training corpus is 2636, and the size of testing corpus is 6224.

The Bigram [6] method is used as the word segmentation method. This is because the language of Sogou laboratory text classification corpus is simplified Chinese, and about 70% is a two-character words in Chinese according to relevant statistics. So by using Bigram method, the complexity of the algorithm can be reduced and good categorization effect can be achieved..

TFIDF weights deformation formula is used as the weight calculation method, which is calculated as follows:

$$TFIDF_{ik} = TF_{ik} \times IDF_k = TF_{ik} \times \ln(\frac{N+1}{n_k}) \qquad (10)$$

The information gain algorithm is used as for feature selection.

The precision rate (Precision) and recall rate (Recall) are often used to evaluate the performance of text categorization system as:

$$\mathrm{Re}call = \frac{A}{A+C} \times 100\% \qquad (11)$$

$$\mathrm{Pr}ecision = \frac{A}{A+B} \times 100\% \qquad (12)$$

where A, B, C specific meanings are shown in Table 1.

They are from two different aspects to measure the quality of classification. In order to better reflect the quality of categorization, a new evaluation index - F1 value is created as:

$$F_1 = \frac{2 \times \mathrm{Pr}ecision \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \times 100\% \qquad (13)$$

This paper also uses micro average F1 and macro average F1 to evaluate the overall performance of the system. The micro average F1 equal considers each document, and is mainly affected by common class; Macro average F1 considers every category equally, and is mainly affected by the influence of rare class.

Table 2 shows that feature dimension has an impact on the performance of the model. From the dimension of value from 400 to 900, it can be observed that with the increase of the dimension, the performance of the model will be enhanced. Between the dimension of value 900 and 1000, when the feature dimension reaches a critical value; with the increase of the dimension, model performance will be slightly reduced. When size arrived in 500, the macro average and the average values are stable. When the size is 900, best results are achieved. To further investigate the impact of the size of decision tree on the performance of random forests, feature size is set as 900.The size of decision tree is ranging from 100 to 400, and the results are shown in Table 3.

Table 3 shows that the number of decision trees has a certain impact on the performance of the model. From the number of decision trees from 100 to 250, it can be known that the increasing number of decision tree model will improve the performance. From the number of decision trees from 250 to 400, it can observed that when arriving at a critical value, even though the number of decision tree is increased, the model performance remains unchanged. When the size arrived at 250, the macro average and the average values are stable. According to Table 2 and Table 3 we can know that random forests have a good performance in text categorization, because Bigram can only count up to 70% Chinese vocabulary.

V. Conclusion

Random forests is a famous integrated learning algorithm with good performance in categorization. Study of the performance of random forests for text categorization has certain significance. This paper first introduced random forests, including the principle and theoretical basis. In this section we can see that when the number of tree in random forest is big enough, generalization error has an upper bound, which relies on the strength of each of decision tree as well as the degree of dependence between the decision trees. Then establish a text categorization model based on random forest, and designed the experiment to evaluate its performance. The results of the experiment show that both the dimension of feature and the number of decision tree have certain influence on the performance of the categorization model. At the same time, it can be concluded that RF algorithm has a good performance in text categorization.

Acknowledgement

The authors would like to thank Huawei- Zhang for his guidance and help to accomplish this paper.

References

[1] Jiang Chengyi, Li Xia, Zheng Qi. Data mining theory and practice. Beijing: Electronic Industry Press, 2011.8.

[2] Dong Shishi, Huang Zhexue. A Brief Theoretical Overview of Random Forests [J]. Integrated Technologies, 2013, 2(1): 1-7.

[3] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140.

[4] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140.

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

[6] Yu Jinkai, Wang Yingxue, Chen Huaichu. An Improved Text Feature Extraction Algorithm Based on N-Gram [J]. Library and Information service, 2004, 48(8): 48-50.

Table 1 Facts and algorithm

Algorithm	Facts	
	Belongs to the class	Do not belong to the class
Belongs to the class	A	B
Do not belong to the class	C	D

Table 2 Results of different feature size

Size	400	500	600	700	800	900	1000
Macro average	0.8761	0.8962	0.8958	0.8992	0.8970	0.8998	0.8995
Micro average	0.8740	0.8941	0.8922	0.8943	0.8920	0.8960	0.8960

Table 3 Results of different tree size

Size	100	150	200	250	300	350	400
Macro average	0.8998	0.9000	0.9014	0.9053	0.9051	0.9036	0.9033
Micro average	0.8960	0.8959	0.8985	0.9015	0.9018	0.9004	0.9002