Research of Text Categorization on WEKA

Li Dan, Liu Lihua, Zhang Zhaoxin

Hebei Software Institute, Baoding , Hebei, 071000, China
Lidan8583@126.com

***Abstract*—The choice of algorithm is a key text categorization problem. In order to evaluation synthetically, analyzed three popular text categorization algorithm that are naive Bayes (NB), decision tree(DT) and support vector machines(SVM). Carried on simulation experiment used the open source data mining tool of Weka. Experimental results show some significant conclusions: The performance of three classification methods are better, including Support vector machine classification of the best performance, highest precision and recall, naive Bayes second, the minimum Decision tree. Also found that classification performance associated not only the choice of the classification algorithm but also the differences between corpus categories.**

Keywords-Text categorization; Naive bayes; Decision tree; Support vector machines; Weka

I. INTRODUCTION

Since the technology of computer and network appeared, it had been developed very rapidly. Network has becoming one of the most mainly-used information sources. Because most of the information in the network is text data type, automatic text categorization which is the important basic of effective organization and management text data has become an important study field. Automatic text categorization for short text categorization (TC) is an important intelligence information processing technology. This technology has high value in information filtering, information retrieval, text databases, digital libraries, and other aspects.

Automatic text categorization is to sort documents to one or more categories automatically. F. Sebastiani thinks the text classification tasks can be understood as a function: $\Phi: D\times C\to\{T,F\}$, where $D=\{d_1,d_2,\cdots d_{|D|}\}$ is a domain of documents and $C=\{c_1,c_2,\cdots c_{|c|}\}$ is a set of pre-defined categories. A value of T assigned to $<d_j,c_i>$ indicates a decision to file d_j under C_i, while a value of F indicates a decision not to file d_j under C_i. More formally, the task is to approximate the unknown target function, that describes how documents ought to be classified, the function called the classifier. Text categorization two key issues: a representation of the text and the other is the design of the classifier.

There is a lot of learning algorithms which has been applied to text classification including naive Bayes (NB), decision tree(DT), k-nearest neighbor (k-NN), support vector machines (SVM), neural networks (NNet), and linear least squares fit (LLSF). In this paper we discuss three popular algorithms for text categorization: naive Bayes method, support vector machine classifier and decision tree. Comparison of the classification performance of these three algorithms carried on simulation experiment.

II. TEXT CATEGORIZATION ALGORITHM

A. Naïve Bayes classifier

Naive Bayes classifier is based on the Bayesian theory，which is accepted as simple and effective probability classification method and has become one of the important contents in the text categorization. The basic idea is the use of feature items and categories of conditional probability to estimate a given document category probability. The naive Bayes classifier is the simplest of these models, in that it assumes that all attributes of the examples are independent of each other given the context of the class.

Assuming that d_i is an arbitrary document, which belongs to a certain category of the document class $C=\{c_1,c_2,\cdots c_{|c|}\}$. According to the Bayesian Classification:

$$P(\mathrm{c_j}\mid \mathrm{d_i})=\frac{P(\mathrm{c_j})P(\mathrm{d_i}\mid \mathrm{c_j})}{P(d_i)}$$

Naive Bayes classification key is to calculate $P(C_j)$ and $P(d_i\mid c_j)$. The method of calculation of $P(d_i\mid c_j)$ can be divided into maximum Likelihood model (MLM), multivariate Bernoulli model (MBM), Multinomial model (MM), Poisson model (PM) and so on. The commonly used is Multinomial model.

In the multinomial model, a document is an ordered sequence of word events, drawn from the same vocabulary V. We assume that the lengths of documents are independent of class. We make as assumption: that the probability of each word event in a document is independent of the word's context and position in the document. Thus, each document d_i is drawn from a multinomial distribution of words with as many independent trials as the length of d_i. Define N_{it} to be the count of the number of times word wt occurs in document d_i. Then, the probability of a document given its class is the multinomial distribution:

$$P(d_i\mid c_j)=P(\mid d_i\mid)\mid d_i\mid!\prod_{t=1}^{|V|}\frac{P(w_t\mid c_j)^{N_{it}}}{N_{it}!}$$

B. Decision tree classifier

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. Text classification, the root node of the decision tree is

978-0-7695-4923-1/12 $26.00 © 2012 IEEE
DOI 10.1109/ISDEA.2012.266

the properties of the sample; the branch of the tree structure is the values of the property. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable. At this point, we can grow a sequence of trees by recursively choosing splits which is going to maximize the homogeneity of the classes of the node. Tree growing will be going on until the stopping criteria are satisfied. After that point, a decision tree is going to be selected by pruning back the tree based on certain criteria. The distribution of classes in each terminal node determines the predictions. There is a lot of decision tree algorithms which has been applied to text classification including ID3, C4.5, CART and SLIQ,etc.

C. Support Vector Machines

The support vector machine is a classifier, originally proposed by Vapnik.That finds a maximal margin separating hyper plane between two classes of data. There are non-linear extensions to the SVM, but Yang found the linear kernel to outperform non-linear kernels in text classification. In our own informal experiments, we also found that linear performs at least as well as non-linear kernels.

Support vector machines are supervised learning methods used for classification, as well as regression. The advantage of Support Vector Machines is that they can make use of certain kernels in order to transform the problem, such that we can apply linear classification techniques to non-linear data. Applying the kernel equations arranges the data instances in such a way within the multi-dimensional space, that there is a hyper-plane that separates data instances of one kind from those of another.

The kernel equations may be any function that transforms the linearly non-separable data in one domain into another domain where the instances become linearly separable. Kernel equations may be linear, quadratic, Gaussian, or anything else that achieves this particular purpose.

Once we manage to divide the data into two distinct categories, our aim is to get the best hyper-plane to separate the two types of instances. This hyper-plane is important because it decides the target variable value for future predictions. We should decide upon a hyper-plane that maximizes the margin between the support vectors on either side of the plane. Support vectors are those instances that are either on the separating planes on each side, or a little on the wrong side.

III. TEXT CATEGORIZATION PROCESS

Text classification process is usually divided into two major step. The first step is to have a training set which composed by category known document, using the training establish classification model; the second step is to use the model to classify unknown category of document. Figure 1 shows the structural framework of the text classification process. First, preprocess the text, the text with vector space model to represent, and feature selection; then establish and train the classifier; and finally, use the classifier to classify the new text.

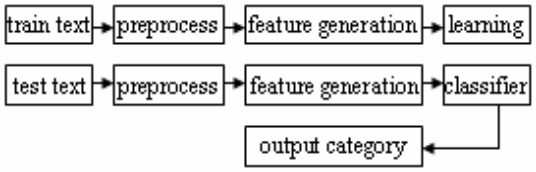

Figure 1. The structure framework of the text classification process

A. Text preprocessing

Text preprocessing is to represent the text into a form that can be processed by the classification algorithm. Preprocessing including Chinese word segmentation, textual representation and feature selection. Through the document processing, documents will be expressed as vector space model, that is, machine learning algorithm for feature vector. Due to the text characteristic item is a high dimensional feature, it is easy to appear a pattern recognition of the "dimension disaster" phenomenon, so the need for feature selection. Feature selection is often performed as a preprocessing step for the purpose of both reducing the feature space and improving the classification performance. Current feature selection method is more, commonly used methods are: document frequency (DF), information gain (IG), x2-statistics (CHI), mutual information (MI), etc.

B. Feature Weighting

Feature selection is often performed as a preprocessing step for the purpose of both reducing the feature space and improving the classification performance. After a feature selection, in order to measure the degree of importance and the ability to distinguish between the strength of the feature item in the document, we need to feature weighting. The popular feature weighting methods are: Boolean weights, absolute word frequency, TF-IDF, entropy weight, etc.

C. Classifier design

After complete the above steps, document form the feature vector. We can through training and learning to construct classifier, and finally, the use of classifiers to classify the new text.

IV. EXPERIMENTAL RESULT

This experiment is carried on the open source data mining tool of Weka. It was developed at the University of Waikato in New Zealand, and the name stands for Waikato Environment for Knowledge Analysis. The system is written in Java. It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

A. Dataset

Our experiments were performed on the datasets of the

Fudan natural language processing laboratories Corpus. This collection is the most widely used benchmark dataset for the text categorization research. We selected 6 categories of art, computer, agricultural, economy, politics, and sports. Each category selects 300 documents, a total of 1800 documents to experiment. Use cross-validation with 10. The entire data set were randomly divided into 10 parts, to take nine combined as the training set, the remaining 1 as a test set. The total run 10 times, as the final classification results averaged.

B. Experiment Step

1) Document segmentation with computing technology, Chinese lexical analysis system (ICTCLAS).

2) Stop words, low-frequency words removed from the document, the rough dimension reduction.

3) By writing a program to document into document vector, feature weighting using the TF-IDF.

4) Information gain evaluation function for feature selection; characteristics of items reduced to 1/3.

5) The text features vector into the Arff format which Weka can identify and sparse data.

6) Arff file loaded into Weka, use Experimenter interface experiment comparing three text classification algorithms.

C. Evaluation Measure and Results

The performances are evaluated using popular recall and precision .

Table 1. Experimental results comparison

category	NB		DT		SVM	
	P	R	P	R	P	R
art	0.971	0.990	0.954	0.967	0.960	0.983
computer	0.984	0.997	0.911	0.953	0.990	0.997
agricultural	0.931	0.983	0.946	0.940	0.970	0.963
economy	0.937	0.843	0.724	0.707	0.902	0.890
politics	0.868	0.940	0.765	0.750	0.912	0.930
sports	0.986	0.913	0.960	0.950	0.962	0.940
average	0.946	0.944	0.876	0.878	0.950	0.950

Table 2. The result of DT classification of the confusion matrix

	DT					
category	art	computer	agricultural	economy	politics	sports
art	290	0	0	0	7	3
computer	0	286	0	7	7	0
agricultural	0	2	282	16	0	0
economy	3	17	14	212	47	7
politics	7	9	2	55	225	2
sports	4	0	0	3	8	285

In table 1, we can observe that SVM classification accuracy and recall rate is highest, naïve Bayes take second place, decision tree minimum, but can also be achieved an average of about 87%. The results of this experiment than [7] [8] classification accuracy is high; the analysis reason is due to the feature selection method caused by different. Experiments economic and political two kinds of classification effect is not very good, especially the decision tree classification accuracy when only 74% or so. Table 2 shows using decision tree classification economic class document have 47 document (15.7%) was wrong to political class, But the political class document 55 document (18.3%) was wrong to economic class. Use naive Bayes and SVM classification, these two types of classification effect is not ideal, analyzes main reason is because two types of document of high repetition rate, category differences is not big, which affects the classification performance. We can see that the accuracy of the classification not only is the influence of the classification method, at the same time and the corpus document character also has the very big relations.

V. CONCLUSION

In this paper, we analyze three kinds of commonly used text categorization algorithm, and the comparative experiment, three kind of classification method can reach the better classification effect. Text categorization algorithm selection is the key to text classification problem, but classification accuracy is not only related with the classification algorithm, at the same time also will be affected by the influence of the difference between corpus category. The difference between category larger has higher classification performance, and the difference between categories smaller recognition performance remains to be further improved.

REFERENCES

[1] Zong chengqing, Statistical natural language processing. Beijing: the book concern of Qinghua University, 2008.(in Chinese)

[2] F.Sebastiani. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 2002, 34(1).

[3] S. Eyheramendy, D. Lewis, D. Madigan, On the Naïve Bayes Model for Text Categorization. Artificial Intelligence & Statistics, 2003.

[4] V Vapnik, Statistical Learning Theory. New York:Wiley-Interscience Publication. John Wiley and Sons, Inc, 1998.

[5] Gong zhile, Zhang dexian, Hu mingming. An Improved SVM algorithm for Chinese Test Classification. Computer Simulation, 2009, 26(7):164-167. (in Chinese)

[6] Dai liuling, Huang heyan, Chen zhaoxiong. A Comparative Study on Feature Selection in Chinese Text Categorization.Journal of Chinese Information Processing. (in Chinese)

[7] Lu wei, Peng ya, Performance Comparison and analysis of Several General Text Classification Algorithms. Journal of Hunan University (Natural Sciences), 2007, 34(6): 67-69. (in Chinese)

[8] Zhou wenxia, Modern Test Categorization Technology Analyses. Journal of Chinese People's Armed Police Force Academy, 2007,23(12): 93-95.(in Chinese)