Journal of Novel Applied Sciences
Available online at www.jnasci.org
©2013 JNAS Journal-2013-2-11/619-621
ISSN 2322-5149 ©2013 JNAS

Efficient Automatic Text Categorization Using A Hybrid Method

Mohammad Behrouzian Nejad[1*], Iman Attarzadeh[2] and Mehdi Hosseinzadeh[3]

1- Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Kerman, Iran
2- Department of Computer Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran
3- Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

Corresponding author: Mohammad Behrouzian Nejad

ABSTRACT: Automatic text categorization (ATC) is the process of classifying text documents into predefined categories based on their content. ATC are dividing into two main steps namely feature selection and learning step. In this paper we proposed a hybrid method for ATC. In the feature selection step we use from a filtering method to reduce the complexity and use from J48 in the learning step. The proposed method is a homogeneous classifier and uses from same classifiers with different sampling with replacement from the training set. For better evaluation of proposed method we compare it with single J48 and Naive Bayes classifiers. The results show that the proposed method has better performance than single classifiers.

Keywords: Automatic Text Categorization, J48, Naive Bayes, Hybrid Method.

INTRODUCTION

ATC is an important issue in the text mining. The task is to automatically classify text documents into predefined classes based on their content (Sebastiani, 2002; Kamruzzaman and Haider, 2004). The fast expansion of the Internet globally also has increased the need to ATC (Al-Mubaid and Umair, 2006). The classification is usually done on the basis of significant words or features extracted from the text document. Since the classes are pre-defined it is a supervised machine learning task (Dalal and Zaveri, 2011). Feature selection and learning are two important steps in the ATC. Several methods have been proposed for the text categorization. Among the proposed methods can be cited text categorization based on unorganized data with extracted information (Manne and Fatima, 2011), text classification based on features (Nirmala and Pushpa, 2012), Naive Bayes method (Kim *et al.*, 2006; Meena and Chandran, 2009), text categorization using association rules (Rahman *et al.*, 2003). In this paper we proposed a hybrid method to improve efficiency of text categorization. The results show that the proposed method has better performance than single J48 and Naive Bayes classifiers.

ATC Process

The main steps of the process of text categorization can be classified into three main stages involves preprocessing, feature selection and learning steps (Korde and Mahender, 2012). In the preprocessing stage, usually on the input data operations are separating words, remove stop words, stemming and term weighting. *Tfidf* is a most popular method for term weighting (Lan and Tan, 2007). Feature selection step refers to select important feature from all features. Irrelevant features can reduce performance. Feature selection methods for ATC are divided into two classes: filtering and wrapper methods. Filtering methods regardless of learning algorithm and using statistical methods to do feature selection and have time complexity lower than the wrapper methods. Wrapper methods uses from learning algorithm as the evaluation function. These methods have higher time complexity and accuracy than filter methods (Jensen, 2005). In this paper to reduce complexity, we use from the filtering method. Many filtering methods have developed for ATC like Gini index, Gain Ratio, Information Gain (IG), and Mutual Information (MI) (Sebastiani, 2002; Dave, 2011).

In this paper we use from information gain technique in the feature selection step. Information gain value measures “the number of bits of information obtained for category prediction by knowing presence of absence of a term in a document”. Information gain values were calculated as relation (1) which $P(t,c)$ shows number of text documents in category *c* which has term *t* (Dave, 2011):

$$\mathrm{IG}(t_k,c_i)=\sum_{c\in\{c_i,\bar{c}_i\}}\sum_{t\in\{t_k,\bar{t}_k\}}p(t,c)\log_2\frac{P(t,c)}{P(t)P(c)} \qquad (1)$$

In the learning step, classifiers from preprocessed text, act to learning (Dalal and Zaveri, 2011). One of the most popular classification techniques is J48 that used in this paper.

Proposed Method

In this paper proposed a hybrid method to increase efficiency of classification. In our method we used from three J48 classifier with linear kernel and finally, combine classifiers outputs using majority vote. The proposed method is homogeneous and uses same classifiers with different sampling with replacement from the training set. For preprocessing we first transform all characters of text to lower case and convert text to the separate words. Then eliminate the stop words like and, the, for and etc and use porter stemmer for stemming. Finally we use n-gram to make the text as a series of consecutive words with length n. We used from n-gram with n=2. For term weighting we used from *tfidf* method and information gain technique used for feature selection. For implementing the proposed method we use from RapidMiner tools, version 5.2. This software is an open source data mining tools and written by Java language. We use from R (8) subset of reuter-21578 dataset for input text documents.

Evaluation of Proposed Method

In this paper for evaluate, we compare proposed method with single J48 and Naive Bayes classifiers. Evaluation criteria are precision, recall and F1. Calculating of these criteria shows in the relation (2) to (4) respectively. If we have two positive and negative classes, FP shows number of incorrect classified documents under positive class; TN shows number of correctly classified documents under negative class, TP is number of correct classified documents under positive class, and FN is number of documents which incorrectly classified under negative class.

$$Precision_i=\frac{TP_i}{TP_i+FP_i} \qquad (2)$$

$$Recall_i=\frac{TP_i}{TP_i+FN_i} \qquad (3)$$

$$F1_i=\frac{2*Precision_i*Recall_i}{Precision_i+Recall_i} \qquad (4)$$

Were i index represent that these parameters should be calculate for each category i. For better evaluation, we use from averaging for these criteria. Results are per percent. Table (1) shows the results of proposed method and J48 and Naive Bayes classifiers.

Table 1. results of proposed method and J48 and Naive Bayes classifiers

	Proposed Method	J48	Naive Bayes
Average Precision	83.56	67.88	64.30
Average Recall	79.96	66.56	68.90
Average F1	81.76	67.20	66.52

The results show that proposed method have better performance than J48 and Naive Bayes classifiers with %83.56 for average precision, %79.96 for average recall and %81.76 for average F1. Also for single classifiers, J48 have better performance than Naive Bayes classifier with %67.88 for average precision and %67.20 for average F1 but Naive Bayes has better performance than J48 with %68.90 for average recall. According to this note that average F1 calculated by average precision and average recall, we see that better performance related to the proposed method, J48 and finally Naive Bayes respectively.

CONCULSION

Assign the text documents to pre-define categories called Automatic text categorization. This is important which this work to be with high performance. In this paper to improve efficiency of text categorization we proposed a hybrid method which uses from a filtering method to reduce the complexity of feature selection and uses from J48 in the learning step. The proposed method uses same classifiers with different sampling with replacement from the training set. We compared proposed method with J48 and Naive Bayes single classifiers. The results show that the proposed method has better performance than single classifiers in average precision, average recall and average F1 criteria.

REFERENCES

Al-Mubaid H, Umair SA. 2006. A New Text Categorization Technique Using Distributional Clustering and Learning Logic, IEEE Transactions On Knowledge and Data Engineering, 18(9): 1-10.

Dalal MK, Zaveri MA. 2011. Automatic Text Classification: A Technical Review, International Journal of Computer Applications, 28(2): 37-40.

Dave K. 2011. Study of feature selection algorithms for text-categorization, University of Nevada, Las Vegas, UNLV Theses/Dissertations/Professional Papers/ Capstones, Paper 1380.

Jensen R. 2005. Combining rough and fuzzy sets for feature selection, PhD Thesis, University of Edinburgh, UK.

Kamruzzaman SM, Haider F. 2004. A HYBRID LEARNING ALGORITHM FOR TEXT CLASSIFICATION, Paper Presented at the 3rd International Conference on Electrical & Computer Engineering, Dhaka, Bangladesh.

Kim S, Han K, Rim H, Myaeng SH. 2006. Some effective techniques for Naive Bayes text classification, IEEE Transactions on Knowledge and Data Engineering, 18(11): 1457-1466.

Korde V, Mahender CN. 2012. TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY, International Journal of Artificial Intelligence & Applications, 3(2): 85-99.

Lan M, Tan CL. 2007. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization, Journal of IEEE Pami, 10(10): 1-36.

Manne S, Fatima SS. 2011. A Novel Approach for Text Categorization of Unorganized data based with Information Extraction, International Journal on Computer Science and Engineering, 3(7): 2846-2854.

Meena MJ, Chandran KR. 2009. Naive Bayes text classification with positive features selected by statistical method, Paper Presented at the IEEE international conference on Advanced Computing, USA.

Nirmala K, Pushpa M. 2012. Feature based Text Classification using Application Term Set, International Journal of Computer Applications, 52(10): 1-3.

Rahman CM, Sohel FA, Naushad P, Kamruzzaman SM. 2003. Text Classification using the Concept of Association Rule of Data Mining, Paper Presented at the International Conference on Information Technology, Kathmandu, Nepal.

Sebastiani F. 2002. Machine Learning in Automated Text Categorization, ACM Computing Surveys, 34(1): 1–47.