Automatic Learning Features Using Bootstrapping for Text Categorization

Chen Wenliang, Zhu Jingbo, Wu Honglin, Yao Tianshun

Natural Language Processing Lab
Northeastern University, Shenyang, China
{chenwl,zhujingbo,wuhl,tsyao}@mail.neu.edu.cn

Abstract. When text categorization is applied to complex tasks, it is tedious and expensive to hand-label the large amounts of training data necessary for good performance. In this paper, we put forward an approach to text categorization that requires no labeled documents. The proposed approach automatically learns features using bootstrapping. The input consists of a small set of keywords per class and a large amount of easily obtained unlabeled documents. Using these automatically learned features, we develop a naïve Bayes classifier. The classifier provides 82.8% F1 while classifying a set of web documents into 10 categories, which performs better than naïve Bayes by supervised learning in small number of features cases.

1 Introduction

The goal of text categorization is to classify documents into a certain number of predefined categories. When provided with enough labeled training examples, a variety of techniques for supervised learning algorithms have demonstrated reasonable performance for text categorization [Yang, ml1997; Yang and Liu, 1999], such as Rocchio[Ittner et al, 1995; Lewis et al, 1996], SVM[Joachims, 1998], decision tree[Lewis, 1994], Maximum Entropy[Nigam et al., 1999], and naïve Bayes [McCallum and Nigam, 1998]. However, when applied to complex domains with many classes, these algorithms often require extremely large training sets to reach useful classification accuracy. Creating these sets of labeled data is tedious and expensive, because labeled documents should be labeled by hand. So in this paper, we consider learning algorithms that do not require such labeled documents [Nigam et al., 2000].

Obtaining labeled data is difficult; on the contrary, unlabeled data is readily available and plentiful. Unlabeled data can often be obtained by fully automatic methods. It can be easy to download unlabeled examples of news articles from Internet.

In this paper, we propose a new automatic text categorization method based on unsupervised learning. Without labeled documents, we automatically learn features for each category using a small set of keywords per class and a large amount of unlabeled documents. And then, using these features, we develop a naïve Bayes text classifier. Keywords are generated more quickly and easily than labeling even a small