Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Information Processing and Management 43 (2007) 379–392

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

A hybrid generative/discriminative approach to text classification with additional information

Akinori Fujino *, Naonori Ueda, Kazumi Saito

NTT Communication Science Laboratories, NTT Corporation, 2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

Received 27 May 2006; accepted 25 July 2006
Available online 11 October 2006

Abstract

This paper presents a classifier for text data samples consisting of main text and additional components, such as Web pages and technical papers. We focus on multiclass and single-labeled text classification problems and design the classifier based on a hybrid composed of probabilistic generative and discriminative approaches. Our formulation considers individual component generative models and constructs the classifier by combining these trained models based on the maximum entropy principle. We use naive Bayes models as the component generative models for the main text and additional components such as titles, links, and authors, so that we can apply our formulation to document and Web page classification problems. Our experimental results for four test collections confirmed that our hybrid approach effectively combined main text and additional components and thus improved classification performance.

Keywords: Multiclass and single-labeled text classification; Multiple components; Maximum entropy principle; Naive Bayes model

1. Introduction

Text data samples such as Web pages and technical papers usually contain multiple components. For example, Web pages consist of main text and additional components such as titles, hyperlinks, anchor text, and images. Although the main text plays an important role when designing a classifier, additional components may contain substantial information for classification. Therefore, designing classifiers for dealing with multiple components is an important and challenging research issue in the field of machine learning. Recently, such classifiers have been developed for multiple components such as text and hyperlinks on Web pages (Chakrabarti, Dom, & Indyk, 1998; Cohn & Hofmann, 2001; Lu & Getoor, 2003; Sun, Lim, & Ng, 2002), text and citations in papers (Cohn & Hofmann, 2001; Lu & Getoor, 2003), and text and music (Brochu & Freitas, 2003). In this paper, we focus on probabilistic approaches to designing text classifiers that can deal with arbitrary additional components as studied in (Brochu & Freitas, 2003; Lu & Getoor, 2003).

* Corresponding author.
E-mail addresses: a.fujino@cslab.kecl.ntt.co.jp (A. Fujino), ueda@cslab.kecl.ntt.co.jp (N. Ueda), saito@cslab.kecl.ntt.co.jp (K. Saito).

0306-4573/$ - see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2006.07.013

Existing probabilistic approaches are generative, discriminative, and a hybrid of the two. Generative classifiers learn the joint probability model, $p(\boldsymbol{x}, y)$, of input $\boldsymbol{x}$ and class label y, compute $P(y|\boldsymbol{x})$ by using the Bayes rule, and then take the most probable label y. However, such direct modeling is hard for arbitrary components consisting of completely different types of media. In Brochu and Freitas (2003), under the assumption of the class conditional independence of all components, the class conditional probability density $p(\boldsymbol{x}^j|y)$ for each component is individually modeled, where $\boldsymbol{x}^j$ stands for the feature vector corresponding to the jth component. Hence, as described later, the joint probability density is expressed by the simple product of $p(\boldsymbol{x}^j|y)$.

Discriminative classifiers directly model class posterior probability $P(y|\boldsymbol{x})$ and learn mapping from $\boldsymbol{x}$ to y. Multinomial logistic regression (Hastie, Tibshirani, & Friedman, 2001) can be used for this purpose. However, such modeling without consideration of components may have an intrinsic limitation in terms of achieving good classification performance. In Lu and Getoor (2003), a class posterior probability $P(y|\boldsymbol{x}^j)$ for each component is individually modeled, and then the simple product of $P(y|\boldsymbol{x}^j)$ is used for predicting the class to which $\boldsymbol{x}$ belongs.

Hybrid classifiers learn a class conditional probability model for each component, $p(\boldsymbol{x}^j|y)$, and directly model class posterior probability $P(y|\boldsymbol{x})$ by using component generative models. Namely, each component model is estimated on the basis of a generative approach, while the classifier is constructed on the basis of a discriminative approach. Hybrid classifiers are constructed by combining the component generative models with weights determined discriminatively. This contrasts with pure generative and discriminative classifiers, which are based on the simple product of component models without weights. For *binary* classification problems, such a hybrid classifier has already been proposed and applied to documents consisting of two text components ("subject" and "body") (Raina, Shen, Ng, & McCallum, 2004). It has been shown experimentally that this hybrid classifier achieves higher accuracy than pure generative and discriminative classifiers.

We present a new hybrid classifier for *multiclass* and single-labeled text classification problems. More specifically, we design individual component generative models $p(\boldsymbol{x}^j|y)$ for main text and additional components. Then, by combining the trained component generative models based on the *maximum entropy* (ME) principle (Berger, Della Pietra, & Della Pietra, 1996), we design a class posterior probability distribution $P(y|\boldsymbol{x})$, where a combination weight is provided per component. We expect the way in which the components are combined to utilize additional information effectively and thus improve classification performance.

According to the ME principle, we can obtain another classifier formulation based on a combination of component generative models, where individual combination weights of components are provided per class. Since the different way in which components are combined would affect classification performance, we also explore the formulation.

To enable us to apply our hybrid classifier to documents and Web pages containing main text and additional components such as titles, authors, and hyperlinks, we employ naive Bayes (NB) models as their individual component generative models. We train the NB models of components with a leave-one-out cross-validation of the training samples to improve their generalization abilities.

The organization of the paper is as follows. In Section 2, we review the formulas for conventional generative, discriminative, and hybrid classifiers that deal with main text and additional components. In Section 3, we present the formulation for our hybrid classifier and the method for applying the hybrid classifier to document and Web page classification. In Section 4, our hybrid classifier is evaluated experimentally using four test collections. Our experimental results show the effect of dealing with additional information and of our hybrid approach on the performance of multiclass classification. Related work is reviewed in Section 5, and our conclusions are presented in Section 6.

2. Conventional approaches

In multiclass and single-labeled classification problems, a classifier categorizes a feature vector $\boldsymbol{x}$ into one of $K(>2)$ classes $y \in \{1, \ldots, k, \ldots, K\}$. Each feature vector consists of J separate components as $\boldsymbol{x} = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^j, \ldots, \boldsymbol{x}^J\}$. The classifier is trained on training sample set $D = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$. In the following, we derive basic formulas for the conventional approaches.