Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 155–165
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Information gain and divergence-based feature selection for machine learning-based text categorization ☆

Changki Lee *, Gary Geunbae Lee *

Department of Computer Science and Engineering, Pohang University of Science and Technology, San 31 Hyoja dong, Nam Gu, Pohang 790-784, Korea (South)

Received 2 February 2004; accepted 11 August 2004

Abstract

Most previous works of feature selection emphasized only the reduction of high dimensionality of the feature space. But in cases where many features are highly redundant with each other, we must utilize other means, for example, more complex dependence models such as Bayesian network classifiers. In this paper, we introduce a new information gain and divergence-based feature selection method for statistical machine learning-based text categorization without relying on more complex dependence models. Our feature selection method strives to reduce redundancy between features while maintaining information gain in selecting appropriate features for text categorization. Empirical results are given on a number of dataset, showing that our feature selection method is more effective than Koller and Sahami's method [Koller, D., & Sahami, M. (1996). Toward optimal feature selection. In *Proceedings of ICML-96, 13th international conference on machine learning*], which is one of greedy feature selection methods, and conventional information gain which is commonly used in feature selection for text categorization. Moreover, our feature selection method sometimes produces more improvements of conventional machine learning algorithms over support vector machines which are known to give the best classification accuracy.

Keywords: Text categorization; Feature selection; Information gain and divergence-based feature selection

☆ This work was supported by the Brain Korea 21 Project (Ministry of Education) in 2003. The short and preliminary version of this paper was published in HLT/NAACL2004.
* Corresponding authors. Tel.: +82 54 279 5581; fax: +82 54 279 2299 (C. Lee), tel.: +82 54 279 2254; fax: +82 54 279 2299 (G.G. Lee).
E-mail addresses: leeck@postech.ac.kr (C. Lee), gblee@postech.ac.kr (G.G. Lee).

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2004.08.006

1. Introduction

Text categorization is the problem of automatically assigning predefined categories to free text documents. A growing number of statistical classification methods and machine learning techniques have been applied to text categorization in recent years (Yang & Liu, 1999).

A major characteristic, or difficulty, of text categorization problems is the high dimensionality of the feature space (Joachims, 1998). The native feature space consists of the unique terms (words or phrases) that occur in documents, which can be tens or hundreds of thousands of terms for even a moderate-sized text collection. This is prohibitively high for many machine learning algorithms. If we reduce the set of features considered by the algorithm, we can serve two purposes. We can considerably decrease the running time of the learning algorithm, and we can increase the accuracy of the resulting model. In this line, a number of researches have recently addressed the issue of feature subset selection (Lewis & Ringuette, 1994; Schutze, Hull, & Pedersen, 1995; Yang & Pedersen, 1997). Yang and Pederson found information gain (IG) and chi-square test (CHI) most effective in aggressive term removal without losing categorization accuracy in their experiments (Yang & Pedersen, 1997). They also discovered that IG and CHI scores of a term are strongly correlated.

Another major characteristic of text categorization problems is the high level of feature redundancy (Joachims, 2001). While there are generally many different features relevant to a classification task, often several such cues occur in one document, and these cues are partly redundant. Naive Bayes, which is a popular learning algorithm, is commonly justified using assumptions of conditional independence or linked dependence (Cooper, 1991). However, theses assumptions are generally accepted to be false for text. To remove these violations, more complex dependence models such as Bayesian network classifiers have been developed (Sahami, 1998), but they require complex models by trading efficiency.

Most previous works of feature selection emphasized only the reduction of high dimensionality of the feature space (Lewis & Ringuette, 1994; Schutze et al., 1995; Yang & Pedersen, 1997). The most popular feature selection method is IG. IG works well with texts and has often been used. IG looks at each feature in isolation and measures how important it is for the prediction of the correct class label. In cases where all features are not redundant with each other, IG is very appropriate. But in cases where many features are highly redundant with each other, we must utilize other means, for example, more complex dependence models.

In this paper, for the high dimensionality of the feature space and the high level of feature redundancy, we propose a new feature selection method which selects each feature according to a combined criterion of information gain and novelty of information. The latter measures the degree of dissimilarity between the feature being considered and the previously selected features. MMR provides precisely such functionality (Carbonell & Goldstein, 1998). So we propose MMR-based feature selection method which strives to reduce redundancy between features while maintaining information gain in selecting appropriate features for text categorization.

In machine learning field, some greedy methods that add or subtract a single feature at a time have been developed for feature selection (Koller & Sahami, 1996; Pietra, Pietra, & Lafferty, 1997). Pietra et al. proposed a method for incrementally constructing random field (Pietra et al., 1997). Their method builds increasingly complex fields to approximate the empirical distribution of a set of training examples by allowing potential functions, or features, which are supported by increasingly large sub-graphs. Each feature is assigned a weight, and the weights are trained to minimize the Kullback–Leibler divergence between the field and the empirical distribution of the training data. Features are incrementally added to the field using a top–down greedy algorithm, with the intent of capturing the salient properties of the empirical samples while allowing generalization to new configurations. However the method is not simple, and this is problematic both computationally and statistically in large-scale problems.