Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 393–405
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Efficient implementation of associative classifiers for document classification

Yongwook Yoon *, Gary Geunbae Lee

Department of Computer Science and Engineering, Pohang University of Science and Technology, San 31, Hyoja-Dong, Pohang 790-784, Republic of Korea

Received 25 May 2006; accepted 25 July 2006
Available online 12 October 2006

Abstract

In practical text classification tasks, the ability to interpret the classification result is as important as the ability to classify exactly. Associative classifiers have many favorable characteristics such as rapid training, good classification accuracy, and excellent interpretation. However, associative classifiers also have some obstacles to overcome when they are applied in the area of text classification. The target text collection generally has a very high dimension, thus the training process might take a very long time. We propose a feature selection based on the mutual information between the word and class variables to reduce the space dimension of the associative classifiers. In addition, the training process of the associative classifier produces a huge amount of classification rules, which makes the prediction with a new document ineffective. We resolve this by introducing a new efficient method for storing and pruning classification rules. This method can also be used when predicting a test document. Experimental results using the 20-newsgroups dataset show many benefits of the associative classification in both training and predicting when applied to a real world problem.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Text classification; Associative classifier; Feature selection; Rule pruning; Subset expansion

1. Introduction

An associative classifier is a classifier using classification rules that are produced through a frequent pattern mining process from a training data collection. This process is the same one used in traditional data mining for large log data of transactional database. Utilizing associative classifiers in the area of classification task (Agrawal & Srikant, 1994; Bekkerman, El-Yaniv, Tishby, & Winter, 2001; Yin & Han, 2003) has a relatively short history compared to other classifiers such as Naïve Bayes, *k*-NN, or SVM. It seems more difficult to find a study in which an associative classifier is applied in the text classification task.

* Corresponding author. Tel.: +82 54 279 5581; fax: +82 54 279 2299.
E-mail addresses: ywyoon@postech.ac.kr (Y. Yoon), gblee@postech.ac.kr (G.G. Lee).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.07.012

When performing a text classification task in a real world situation, the ability to provide abundant interpretation on the classification result is often as important as the ability to classify new documents exactly. Classification by a concrete form of rules ("*Features* → *Class*") has many benefits including this easy interpretability. The associative classifier is one of the rule-based classifiers. In contrast, some classifiers such as SVM or Neural Network cannot provide this easy interpretation for the classification result, though they may achieve excellent classification accuracy.

We can acquire several additional advantages from using rule-based classifier. One is that since the rules can be expressed in a very intuitive form, humans can easily understand them and can even edit them directly after the rules are produced by some inductive learning process. A human expert could delete the weak rules from the original rule set and add new rules that they carefully handcrafted. This can improve the classification accuracy remarkably with a little bit of added effort. Another is that the rules can be updated incrementally by other machine learning processes later.

Another benefit of the associative classifier is that it can exploit the combined information of multiple features as well as a single feature, while SVM or k-NN classifiers consider only the effects of each single feature. This means that in document classification tasks it is possible to use phrase occurrence information as well as word occurrence information.

To apply an associative classifier to the text classification problem in the real world, however, we need to remove several obstacles encountered during the training and testing phase. One of those is a high dimensional feature space. Dataset in the area of text classification, in many cases, has a very large number of features that are distinct lexical words. For example, the 20-newsgroups test collection has more than one hundred thousand lexical word features. Most documents of the 20-newsgroups have more than one hundred words; they are sparsely distributed in their word feature space. In associative classification, however, we consider all subsets of those words. Therefore, the effective number of features grows exponentially, and we cannot take into account all of them due to computational intractability.

To overcome this problem we adopt a feature selection-based dimensionality reduction technique at the same time maintaining necessary performance in classification. Many well-known methods of dimensionality reduction exist (Sebastiani, 2002). We used the mutual information measure of the information theory. From the training dataset we calculated the mutual information between the word and the class variables. And we selected words that have high mutual information, and used only those in classifying and neglected the others.

Another obstacle in associative text classification is the large number of classification rules that are produced in the training phase. Since using all of them becomes both inefficient computationally and ineffective in classifying, we should select a part of those rules that have high quality. This process has been called *Pruning* in associative classification. Liu, Hsu, and Ma (1998) proposed a pruning by database coverage, which is a kind of validation process using the training set for the purpose of choosing the best classification rules among others. Li, Pei, and Han (2001) refined the concept of the database coverage. In addition, they proposed two other pruning methods. One is to prune low-ranked rules in terms of the confidence and support of the rules. The other is to prune the rules in which the correlation between the pattern and the class variables is weak. In this paper, we adopted the pruning methods of Li et al.'s and improved them to work for text classification.

Related issue of the rule pruning is the prediction of a new document using classification rules. With a large number of rules, the prediction result of a test document often shows a split decision between different classes. A method is needed to select one correct class among many in an efficient and effective way. It is not a simple problem because if we extract relatively small portion of the rules to avoid many contradicting rules for a document, we might lose latent candidate classes that may be the correct answer. To handle this problem, Li et al. (2001) used the weighted chi-square method. We try to resolve this problem by simple efficient voting on the different answer classes.

In Section 2 we introduce the general aspects of the associative classification. In Section 3 we explain the overall architecture of our text classification system using association rules and address the issues such as dimensionality reduction, and rule pruning and prediction from multiple rules. Experimental results and analyses of text classification using a large dataset are presented in Section 4, and we conclude our works in Section 5.

Download English Version:

https://daneshyari.com/en/article/515206

Download Persian Version:

https://daneshyari.com/article/515206

Daneshyari.com