基于 AdaBoost-Bayes 算法的中文文本分类系统

徐　凯 1，陈平华 1，刘双印 2
（1 广东工业大学 计算机学院，广东 广州 510003； 2 广东海洋大学 信息学院，广东 湛江 524088）

摘　要： 针对中文文本分类准确率低，分类算法低效不稳定问题，提出基于自适应提升朴素贝叶斯算法.该算法采用 Naive Bayes 和 AdaBoost，并且通过优化组合结构，融合两种算法的优点.首先，使用 SMEL 序列组合成词算法对中文语料进行分词，提取文本特征词汇.然后，使用增强的贝叶斯分类器，通过较小的样本训练，提取出文本特征，生成训练分类矩阵.结合自适应提升算法对简单分类器进行加权，保证分类有平稳准确的效果.通过实验证明，该算法与其他算法相比，错误率更低，可以使分类准确率达到 98%以上，而且 F1 值也优于其他分类算法.

关键词： 中文分词；文本分类；AdaBoost；Bayes

A Chinese Text Classification System Based on Ada Boost-Bayes Algorithm

XU Kai1,CHEN Ping-hua1 ,LIU Shuang-yin2
(1 Faculty of Computer, Guangdong University of Technology, Guangzhou 510006， China；
2 College of Information, Guangdong Ocean University, Zhanjiang 524088, China)

Abstract：In view of the low accuracy of Chinese text classification algorithm, the classification algorithm is inefficient and the problem of low efficiency and low efficiency is proposed. Based on the adaptive algorithm, the proposed algorithm is proposed to improve the accuracy. The algorithm uses Bayes Naive and AdaBoost, and the advantages of the two algorithms are fused by the optimization of the structure. First, using the SMEL sequence of the word segmentation algorithm to segment the Chinese corpus and extract the feature words. Then, the enhanced Bias classifier is used to extract the feature of the text and generate the training classification matrix through the small sample training. Combined with the adaptive lifting algorithm, the simple classifier is weighted to ensure that the classification is stable and accurate. Experiments show that the error rate is lower than other algorithms, and the classification accuracy of the algorithm is more than 98%, and the F1 value is better than other classification algorithms.

Key words：Chinese word segmentation; text classification; AdaBoost；Bayes

作者简介：
徐　凯　　男，(1987-)，硕士研究生.研究方向为推荐系统、数据挖掘.
E-mail:504087493@qq.com.
陈平华　　男，(1967-)，教授.研究方向为云计算、Web 挖掘、推荐系统.
刘双印　　男，(1977-5),博士，教授.研究方向为智能计算、 智能信息处理、物联网.