

数据库、信息处理

面向中文文本分类的C4.5Bagging算法研究

张翔^{1,2}, 周明全^{1,3}, 耿国华¹, 侯凡¹

- 1.西北大学 可视化技术研究所, 西安 710127
- 2.西安建筑科技大学 信息与控制工程学院, 西安 710055
- 3.北京师范大学 信息科学与技术学院, 北京 100875

收稿日期 2008-5-20 修回日期 2008-8-1 网络版发布日期 2009-9-15 接受日期

摘要 对于中文文本分类问题, 提出一种新的Bagging方法。这一方法以决策树C4.5算法为弱分类器, 通过实例重取样获取多个训练集, 将其结果按照投票规则进行合成, 最终得到分类结果。实验证明, 这种算法的准确率、查全率、F1值比C4.5、kNN和朴素贝叶斯分类器都高, 具有更加优良的性能。

关键词 [Bagging算法](#) [C4.5算法](#) [中文文本分类](#)

分类号 [TP311](#)

C4.5Bagging algorithm for Chinese text categorization

ZHANG Xiang^{1,2}, ZHOU Ming-quan^{1,3}, GENG Guo-hua¹, HUO Fan¹

- 1.Visualization Technology Institute, Northwest University, Xi'an 710127, China
- 2.College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China
- 3.College of Information Science and Technology, Beijing Normal University, Beijing 100875, China

Abstract

Aiming at the problem of Chinese text classification, a new method of Bagging is developed. The decision tree C4.5 is selected as the weak classifier and multiple training sets are gained through re-sampling instance. Then, the outputs are combined by voting and the final classification results are obtained. The experimental results show that the classifier based on the C4.5Bagging gets higher precision, recall, F-measure and better performance than C4.5, kNN and Naive-Bayse.

Key words [Bagging](#) [C4.5](#) [Chinese text categorization](#)

DOI: 10.3778/j.issn.1002-8331.2009.26.039

通讯作者 张翔 zhangxiang1001@126.com

扩展功能

本文信息

- ▶ [Supporting info](#)
- ▶ [PDF\(618KB\)](#)
- ▶ [\[HTML全文\]\(0KB\)](#)
- ▶ [参考文献](#)

服务与反馈

- ▶ [把本文推荐给朋友](#)
- ▶ [加入我的书架](#)
- ▶ [加入引用管理器](#)
- ▶ [复制索引](#)
- ▶ [Email Alert](#)
- ▶ [文章反馈](#)
- ▶ [浏览反馈信息](#)

相关信息

- ▶ 本刊中 包含“[Bagging算法](#)”的 [相关文章](#)
- ▶ 本文作者相关文章

- [张翔](#)
- [周明全](#)
- [耿国华](#)
- [侯凡](#)