文本分类中基于 CHI 改进的特征选择方法

宋呈祥，陈秀宏，牛强
(江南大学 数字媒体学院，江苏 无锡 214122)

摘要： ：针对传统卡方统计量(CHI)方法在全局范围内做特征选择时忽略词的频度、词的分布等问题，提出了一种改进的文本特征选择方法.该方法通过定义特征词频度分布相关性系数，选择局部出现的强相关性特征，同时考虑特征词类间分布差异性来提升不均衡数据集的分类指标.结果表明，改进的方法不仅在分类效果上有明显的提高，而且性能更加稳定.

关键词： 文本分类；卡方统计量；特征选择；不均衡数据集

Improved Feature Selection Method Based on CHI for Text Categorization

SONG Cheng-xiang, CHEN Xiu-hong, NIU Qiang
(School of Digital Media, Jiangnan University,Wuxi 214122, China)

Abstract： Because the traditional Chi-square methodchooses the feature in the global scope and ignores theinformationof word frequency and distribution, this paper proposes animproved feature selection method. The method selects anumber of strong featureswith thedefined feature distributioncoefficient, andtakes into account featuredistribution thatimproves theperformances of Chi-square method in theunbalanced dataset.The results of the experimentshow that the improved algorithmnot only shows a significant improvement in classification efficiency, but also has more stable performance.

Key words： text categorization; Chi-square; feature selection; unbalanced dataset

作者简介：

宋呈祥男，(1989-)，硕士研究生.研究方向为机器学习与自然语言处理.E-maial:scx929@163.com.

陈秀宏男，(1964-)，博士（后），教授.研究方向为模式识别，图像处理及人工智能等.

牛强男，(1992-)，硕士研究生.研究方向为机器学习与图像处理.