计算机工程与应用 2007 43 (13): 159-162 ISSN: 1002-8331 CN: 11-2127/TP

数据库与信息处理

用于文本分类的改进KNN算法

王煜 张明 王正欧 白石

河海大学 南京师范大学数学与计算机学院 天津大学系统工程研究所 河北沧州市城建档案馆

收稿日期 2006-9-15 修回日期 网络版发布日期 2007-4-19 接受日期

摘要 采用灵敏度方法对距离公式中文本特征的权重进行修正；提出一种基于CURE算法和tabu算法的训练样本库的裁减方法，采用CURE聚类算法获得每个聚类的代表样本组成新的训练样本集合，然后用tabu算法对此样本集合进行进一步维护（添加或删除样本），添加样本时只考虑增加不同类交界处的样本，添加或删除样本以分类精度最高、与原始训练样本库距离最近为原则。

关键词 文本分类 KNN算法 灵敏度法 CURE聚类算法 tabu算法

分类号

An Improved KNN Algorithm Applied to Text Categorization

Yu Wang Ming Zhang ZhengOu Wang Shi Bai

Abstract

In this paper, based on the neural network theory, weights of features are adjusted firstly by using sensitivity method. A method is presented to prune training samples for KNN algorithm. First, representative samples set of training sets are acquired based on CRUE clustering algorithm. The representative samples set is taken as the initial set of tabu algorithm to further maintain. The method only considers the samples at different classes borders when samples are insert into new training set. The principles of delete or insert a sample are the higher categorization accuracy principle and the higher similarity with training set principle. The work of pruning and maintenance training samples set is decreased largely. Both satisfied speed and accuracy of classification can be acquired.

Key words text categorization KNN algorithm sensitivity method CRUE cluster algorithm tabu algorithm

DOI:

通讯作者 白石 baishi2005@126.com