

数据库、信号与信息处理

基于属性值信息熵的KNN改进算法

童先群, 周忠眉

漳州师范学院 计算机科学与工程系, 福建 漳州 363000

收稿日期 2009-10-12 修回日期 2009-11-30 网络版发布日期 2010-1-28 接受日期

摘要 为了克服传统KNN算法, 距离加权-KNN算法在距离定义及投票方式上的不足, 提出了一种基于属性值对类别重要性的改进算法Entropy-KNN。首先定义两个样本间的距离为相同属性值的平均信息熵, 此距离可通过重要属性值有效度量样本之间的相似程度, 其次算法Entropy-KNN根据上述定义的距离选取与待测试样本距离最小的K个近邻, 最后根据各类近邻样本点的平均距离及个数判断待测试样本的类别。在蘑菇数据集上的实验表明, Entropy-KNN算法的分类准确率高于传统KNN算法和距离加权KNN算法。

关键词 [分类](#) [KNN算法](#) [属性值](#) [信息熵](#)

分类号 [TP391](#)

Enhancement of K-nearest neighbor algorithm based on information entropy of attribute value

TONG Xian-qun, ZHOU Zhong-mei

Department of Computer Science & Engineering, Zhangzhou Normal University, Zhangzhou, Fujian 363000, China

Abstract

In order to improve traditional KNN and KNN with weighted distance, which is on the distance definition and test mode, an improved algorithm entropy-KNN based on the classification importance of an attribute value is proposed. At first, a distance of the two samples is defined as the average information entropy of the same attribute values. The distance can effectively measure the similarity degree of the two samples. Secondly, the Entropy-KNN selects the K nearest neighbors by the distance above. Finally, the class label of the test sample is decided by the average distance and the numbers on the respective class. The experimental results on mushroom data set show this approach has much better than traditional KNN and KNN with weighted distance.

Key words [classification](#) [K-nearest neighbor algorithm](#) [attribute value](#) [information entropy](#)

DOI: 10.3778/j.issn.1002-8331.2010.03.034

通讯作者 童先群 xq_tong@126.com

扩展功能

本文信息

- ▶ [Supporting info](#)
- ▶ [PDF\(528KB\)](#)
- ▶ [\[HTML全文\]\(0KB\)](#)
- ▶ [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“分类”的 相关文章](#)
- ▶ 本文作者相关文章
 - [童先群](#)
 - [周忠眉](#)