计算机科学与探索 2007 1 (3): 282-292 ISSN: 1673-9418 CN: 11-5602/TP

学术研究

基于树编辑距离的层次聚类算法

乔少杰[1,2], 唐常杰[1+], 陈 瑜[1], 彭 京[3], 温粉莲[1]

1.四川大学 计算机学院,成都 610065
2.新加坡国立大学 计算机学院,新加坡 117590
3.北京大学 信息科学技术学院,北京 100871

收稿日期 修回日期 网络版发布日期 2007-9-20 接受日期

摘要 为了识别犯罪嫌疑人伪造和篡改的虚假身份,利用树编辑距离计算个体属性相似性,证明了树编辑距离的相关数学性质,对属性应用层次编码方法,提出了一种新的基于树编辑距离的层次聚类算法HCTED（Hierarchical Clustering Algorithm Based on Tree Edit Distance）。新算法通过树编辑操作使用最少的代价计算属性相似性,克服了传统聚类算法标称型计算的缺陷,提高了聚类精度,通过设定阈值对给定样本聚类。实验证明了新方法在身份识别上的准确性和有效性,讨论了不同参数对实验结果的影响,对比传统聚类算法,HCTED算法性能明显提高。新算法已经应用到警用流动人口分析中,取得了良好效果。

关键词 树编辑距离 层次聚类 属性相似性 数据挖掘

分类号

A new hierarchical clustering algorithm based on tree edit distance

QIAO Shaojie[1,2],TANG Changjie[1+],CHEN Yu[1],PENG Jing[3],WEN Fenlian[1]

1.School of Computer,Sichuan University,Chengdu 610065,China
2.School of Computing,National University of Singapore,117590,Singapore
3.School of Electronics Engineering and Computer Science,Peking University,Beijing 100871,China

Abstract

In order to recognize the false status which has been forged and tempered by suspects,a new method is proposed to compute attribute similarities based on tree edit distance,and its mathematical properties are proved. The paper proposes a new clustering algorithm based on hierarchical encoding method named HCTED（Hierarchical Clustering Algorithm Based on Tree Edit Distance）. This method uses tree edit distance to compute attribute similarities with minimum cost,overcomes the shortage of traditional clustering algorithms and improves the precision of clustering according to the predefined threshold. Experiments demonstrate that the new method is accurate and efficient in identity recognition,discuss the effects of different experimental parameters,and show that HCTED is more accurate and faster than traditional clustering algorithms. The new algorithm has been used in data analysis of transient population for public security successfully.

Key words tree edit distance hierarchical clustering attribute similarity data mining

DOI:

通讯作者 乔少杰 E-mail:tangchangjie@cs.scu.edu.cn