计算机工程与应用 2010 46 (8): 123-126 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

基于seeds集和频繁项集挖掘的半监督聚类算法

赵 倩，尚学群，王 淼

西北工业大学 计算机学院，西安 710072

收稿日期 2008-9-18 修回日期 2008-12-4 网络版发布日期 2010-3-11 接受日期

摘要 半监督聚类在无监督学习中通过对少量监督信息的有效利用提高聚类性能。提出一种基于seeds集的半监督聚类算法，它采用Apiori算法对初始seeds集和扩大规模后seeds集的数据进行频繁项集挖掘，使得数据中存在的噪音数据和误标记数据得到净化、修正，以改善seeds集质量，提高聚类性能。该算法使用带权χ^2测试这一数学模型作为分类规则度量指标，以对无标记数据进行类标签值预测。实验结果显示，所提出的结合了频繁项集挖掘和带权χ^2测试的基于seeds集的半监督聚类算法不仅改善了seeds集质量，也提高了预测结果的精确度，优化了聚类性能。

关键词 半监督聚类 频繁项集挖掘 带权χ^2测试 seeds集

分类号 TP311

Semi_supervised clustering algorithm based on seeds set and frequent itemset mining

ZHAO Qian，SHANG Xue-qun，WANG Miao

School of Computer，Northwestern Polytechnical University，Xi’an 710072，China

Abstract

Semi_supervised clustering makes use of few supervised information in unsupervised clustering to boost the clustering performance.This paper proposes a semi_supervised clustering algorithm based on seeds set and frequent itemset mining，which mines frequent itemsets in the beginning seeds set and the enlarged seeds set for eliminating the noise data and correcting the mislabeled data to improve the quality of seeds set and enhance the performance of clustering.A weighted χ^2 measure，as a classification rule evaluation measure，is used to label unlabeled data and they are added into the initial seeds set to enlarge the scale.The experimental results show that the proposed approach effectively reduces the noise data，and not only makes the results more correct but also makes the performance of clustering more better.

Key words semi_supervised clustering frequent itemset mining weighted χ^2 measure seeds set

DOI: 10.3778/j.issn.1002-8331.2010.08.035

通讯作者 赵 倩 zhaoqian_qiezi@126.com