计算机工程与应用 2010 46 (6): 139-143 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

基于判别分析的半监督聚类方法

陈小冬[1，2]，尹学松[2]，林焕祥[3]

1.浙江大学 计算机科学与技术学院，杭州 310027
2.浙江广播电视大学 信息与工程学院，杭州 310012
3.浙江科技学院 信息学院，杭州 310012

收稿日期 2008-9-8 修回日期 2008-12-15 网络版发布日期 2010-2-23 接受日期

摘要 与无监督聚类相比，半监督聚类是利用一部分先验信息来更好地挖掘和理解数据的内在结构，并紧密遵从用户的偏好。现有的典型半监督聚类算法仅仅适合于低维数据，文中提出一种新颖的基于判别分析的半监督聚类算法来解决高维数据聚类问题。新算法首先使用主成分分析来投影高维数据，进一步在投影空间中，使用基于球形K均值聚类算法对数据聚类；然后利用聚类结果，使用线性判别分析降维输入空间数据；最后在投影空间中对数据再次聚类。在一组真实数据集上的实验表明，所提出的算法不仅可以有效地处理高维数据，还提高了聚类性能。

关键词 半监督聚类 成对约束 主成分分析 线性判别分析

分类号 TP311

Semi-supervised clustering approach with discriminant analysis

CHEN Xiao-dong[1，2]，YIN Xue-song[2]，LIN Huan-xiang[3]

1.College of Computer Science，Zhejiang University，Hangzhou 310027，China
2.College of Information and Engineering，Zhejiang Radio & TV University，Hangzhou 310012，China
3.College of Information & Electronic Engineering，Zhejiang University of Science & Technology，Hangzhou 310012，China

Abstract

The semi-supervised clustering is to mine and help to understand better the structure of unlabeled data and to more closely conform to the user's preferences using those supervised data，in comparison with unsupervised clustering.Most existing semi-supervised clustering methods are designed for handling low-dimensional data.In this paper，a novel Semi-supervised Clustering Approach with Discriminant Analysis（SCADA） is presented for clustering the high-dimensional data.Specifically，the data are first mapped onto the low-dimensional space by principal component analysis such that constrained spherical K-means algorithm is used to cluster those transformed data.Secondly，linear discriminant analysis is used to reduce the number of the dimensionality of the data in terms of the clustering results.Finally，the data in the embedded space are clustered.Indeed，the experimental results on several real-world data sets show the SCADA method can effectively deal with the high-dimensional data and provides an appealing clustering performance.

Key words semi-supervised clustering pairwise constraint principal component analysis linear discriminant analysis

DOI: 10.3778/j.issn.1002-8331.2010.06.040

通讯作者 陈小冬