数据库、信息处理

基于PCA及属性距离和的孤立点检测算法

张忠平, 宋少英, 宋晓辉

燕山大学　信息科学与工程学院，河北　秦皇岛　066004

摘要　　提出了一种基于主分量分析和属性距离和的孤立点检测算法。该方法首先通过主分量分析方法从众多属性中提取出满足累计贡献率的主分量，同时利用PCA变换矩阵把原始数据集转换到由主分量组成的新的特征空间上，之后对转换后的数据集用属性距离和的方法对孤立点进行检测。实验结果证明了基于主分量分析和属性距离和的孤立点检测算法的有效性。
关键词　　孤立点　主分量分析　累计贡献率　属性距离和
分类号

Algorithm for outlier detection based on principal component analysis and sum of attributes distance

ZHANG Zhong-ping,SONG Shao-ying,SONG Xiao-hui

College of Information Science and Engineering，Yanshan University，Qinhuangdao，Hebei 066004，China

Abstract

An outlier detection algorithm based on principal component analysis and the sum of attributes distance is proposed.The algorithm firstly extracts the principal components from many attributes satisfying accumulative contribution rate.Simultaneously，by the PCA matrix original dataset is transformed to a new feature space composed of principal component.Then outliers are detected using the approach of the sum of attributes distance in the transformed datasets.The results of the experiment show that the outlier detection algorithm based on principal component analysis and the sum of attributes distance is effective.

Key words　outlier　principal component analysis　accumulative contribution rate　the sum of attributes distance

通讯作者　张忠平 juegood@yahoo.com.cn