研究、探讨

基于熵度量的空间邻域离群点查找

苏锦旗[1]，薛惠锋[1]，吴慧欣[2]

1.西北工业大学 自动化学院，西安　710072
2.华北水利水电学院 信息工程学院，郑州　450011

摘要　　离群点的查找算法主要有两类：第一类是面向统计数据，把各种数据都看成是多维空间，没有区分空间维与非空间维，这类算法可能产生错误的判断或找到的是无意义的离群点；第二类算法面向空间数据，区分空间维与非空间维，但该类算法查找效率太低或不能查找邻域离群点。引入熵权的概念，提出了一种新的基于熵权的空间邻域离群点度量算法。算法面向空间数据，区分空间维与非空间维，利用空间索引划分空间邻域，用非空间属性计算空间偏离因子，由此度量空间邻域的离群点。理论分析表明，该算法是合理的。实验结果表明，算法具有对用户依赖性小、检测精度和计算效率高的优点。

New approach of spatial neighborhood outliers detection based on entropy measurement

SU Jin-qi[1]，XUE Hui-feng[1]，WU Hui-xin[2]

1.Automation College，Northwestern Polytechnical University，Xi'an 710072，China
2.Dept. of Information Engineering，North China University of Water Conservancy & Electric Power，Zhengzhou 450011，China

Abstract

There are usually two classes of outlier detection algorithms.One is usually applied to statistical data and takes all attributes as multi-dimensional space，while not distinguish between geo-spatial dimensionality and non-spatial dimensionality in detecting process.Meaningless or incorrect outliers can be found if we use these approaches.The other outlier detection algorithms distinguish between geo-spatial dimensionality and non-spatial dimensionality，but they have poor efficiency or can't detect neighborhood outliers.To overcome these shortcomings，new approach of spatial neighborhood outliers detection based on entropy measurement is proposed.In this paper，the spatial attributes are used to determine spatial neighborhood，entropy theory is used to determine the weight of non-spatial attributes，and the non-spatial dimensions are used to compute the spatial neighborhood outlier factor，thus spatial neighborhood outliers can be captured. Theoretical analysis shows that the algorithm is reasonable.The experimental results show that the approach is practical.

通讯作者　苏锦旗 jinqisu@163.com