基于改进的局部异常因子检测的优化聚类算法

张丹丹 [1]，游子毅 [1]，郑 建 [2]，陈世国 [1]
（1 贵州师范大学 物理与电子科学学院，贵州 贵阳 550001；2 贵州省农村信用社联合社，贵州 贵阳 550081）

摘 要：聚类分析在无监督学习领域中一直备受国内外学者关注.针对 K-means 聚类算法对初始聚类中心点敏感、簇内数据相关性差以及收敛到局部最优的缺点，提出了一种基于离群因子的优化聚类算法.该算法采用信息熵加权欧式距离作为相似性度量依据，以更明显地区分数据对象间的差异，然后利用 k 距离参数自调整的局部异常因子检测算法计算出各数据点的离群因子并筛选出初始聚类中心的候选集，最后根据其离群因子加权距离法优化聚类中心.通过在 UCI 数据集上的实验测试结果表明，优化算法的准确率比 K-means++算法、OFMMK-means 算法、FCM 算法更高，运行速度比 FCM 算法更快.该算法能够更好地应用于入侵行为检测、信用风险评估以及多故障诊断等领域.

关键词：聚类；Kmeans；加权欧式距离；LOF 算法；优化

Optimal clustering algorithm based on modified local outlier factor detection

ZHANG Dan-dan [1], YOU Zi-yi [1], ZHENG Jian [2], CHEN Shi-guo [1]
（1 School of Physics and Electronic Sciences,Guizhou Normal University, Guiyang 550001, China; 2 Guizhou Rural Credit Cooperative Association, Guiyang 550081, China）

Abstract：Cluster analysis has been concerned by scholars at home and abroad in the field of unsupervised learning. Aiming at the disadvantages of K-means clustering algorithm for initial clustering center point sensitivity, poor data correlation in clusters and convergence to local optimization, an optimized clustering algorithm based on outlier factor is proposed in this paper. The algorithm firstly takes the information entropy weighted European distance as the basis of similarity measurement, in order to distinguish the difference between the data objects more obviously, then calculates the outlier factor of each data point by using the k distance parameter self-adjusting of the Local Outlier Factor algorithm and selects the candidate set of the initial clustering center, and finally optimizes the clustering center according to the outlier factor weighted distance method. The experimental results on UCI DataSet show that the accuracy of the optimization algorithm is higher than that of k-means++ algorithm, OFMMK-means algorithm and FCM algorithm, and its running speed is faster than the FCM algorithm. The algorithm can be better used in intrusion behavior detection, credit risk assessment and multi-fault diagnosis.

Key words：clustering; Kmeans; weighted european distance; LOF algorithm; optimization

作者简介：
张丹丹 女，(1994-)，硕士研究生.研究方向为智能信息处理与传输技术.
游子毅(通讯作者) 男，(1982-)，博士，教授. 研究方向为智能控制算法、网络技术.
E-mail：357534271@qq.com.
郑 建 男，(1979-)，博士，副教授.研究方向为机器学习与数据挖掘.
陈世国 男，(1967-)，博士，教授.研究方向为信号处理.