混合的大规模数据库中数值型数据聚类算法研究

何育朋
（广东财经大学 华商学院，广东 广州 511300）

摘　要：大规模数据库中的海量数据多具有混合属性，即数值型数据与其他类型的数据混合于一体、数据量庞杂、不易区分.传统算法往往忽视多种属性间的关联性，算法复杂、聚类速度慢、效果差.对此提出一种基于划分聚类的混合大规模数据库中数值型数据聚类算法.首先为降低传统算法的高复杂度，要从大规模的数据库中合理抽取多个小数据集，小数据集中要包含数据库中全部的自然簇；依据相似度定义构建小数据集的相似度矩阵，并分别进行数值型数据及其他类型数据的相似度计算；最后对抽样聚类的结果进行整合、均值更新和划分，实现混合的大规模数据库中数值型数据的聚类.仿真实验表明，提出的算法计算速度快、运算量相对较小、误差率低，能够得到更理想的聚类效果，适用于大规模的数据聚类处理.
关键词：大规模；数值型；数据聚类

Research on Numerical Data Clustering Algorithm in Hybrid Large Scale Database

HE Yu-peng
(Huashang College,Guangdong University of Finance and Economics,Guangzhou 511300, China)

Abstract：The mass data in the large scale database has mixed attributes, namely, the mixed data of symbolic data and numerical data, and the quantity of data is complex and difficult to distinguish. Traditional algorithms often ignore the correlation between the two attributes, the calculation is complex, the clustering speed is slow, the effect is poor. A numerical study of mixed database clustering in large-scale data clustering algorithm based on the traditional algorithm, firstly in order to reduce the high complexity, from reasonably extracting large-scale databases of multiple small data sets, all natural clusters contain database on small data sets; similarity matrix similarity is defined to construct small data set based on the then, the similarity data symbols and numerical data calculation; integration, the final result of the sample clustering updated mean and classification, clustering of numeric data mixed in large databases. Simulation results show that the proposed algorithm can get a better clustering result, and is suitable for large scale data clustering processing. The algorithm has a fast calculation speed, a relatively small amount of computation, and a low error rate.
Key words：large scale; numerical model; data clustering

作者简介：
何育朋　男，（1980-），硕士，讲师.研究方向为计算机应用技术.E-mail：530272504@qq.com.