面向大数据的并行分类混合算法研究

陈学斌，王　师，董岩岩
（华北理工大学 理学院，河北 唐山 063009）

摘　要： 针对传统分类算法及技术在处理海量异构数据存在的系统性能拓展性低、计算量大、耗时长、分类效果不佳等问题，采用 Map-Reduce 与邻近分类算法融合设计适合大数据处理的并行分类混合算法，利用加权欧氏距离并行计算，达到提高海量数据分类效率、提高分类识别率和减小资源开销的目的，搭建 Hadoop 集群研究并在多个数据集上测试算法的可行性.实验结果表明，并行分类混合算法在海量数据分类中显现出较好的分类效果，是可行的海量数据分类模型.

关键词： 大数据；Map-Reduce；算法融合；并行分类

Research on Parallel Classification Hybrid Algorithm for Big Data

CHEN Xue-bin, WANG Shi, DONG Yan-yan

(College of Science, North China University of Science and Technology, Tangshan 063009, China)

Abstract： To solve the problem of the traditional classification algorithms and technologies in the huge amounts of heterogeneous data, such as low-expanding, large-calculating, time-consuming and poor classification. Parallel Classification Hybrid Algorithm is design by fusion of Map-Reduce and Nearest Neighbor Algorithm and parallel computation of weighted Euclidean distance, which improved the efficiency of mass data classification, improved the classification rate and reduced the cost of resource. The Hadoop platform is built for research and test the feasibility of the algorithm on multiple of data sets. The experimental results demonstrate that the Parallel Classification Hybrid Algorithm show a good classification effect in the massive data classification, is a feasible mass data classification model.

Key words： big data; Map-Reduce; fusion algorithm; parallel classification

作者简介：
陈学斌　男，（1970-），博士，教授.研究方向为大数据、云计算、物联网、网络安全等.E-mail：chxb@qq.com.
王　师　男，（1990-），硕士研究生.研究方向为云计算理论及应用.
董岩岩　男，（1990-），硕士研究生.研究方向为云计算理论及应用