基于 Spark 的并行 K-means 算法研究

许明杰，蔚承建，沈航
（南京工业大学 计算机科学与技术学院， 江苏 南京 211816）

摘要： 针对 K-means 算法在海量数据的处理过程中，由迭代计算次数加大导致的内存不足的问题，提出 Spark 并行化的 K-means 算法.将粒子群优化（PSO）与 K-means 结合，利用 PSO 来提高 K-means 的全局搜索能力，得到初始聚类中心.利用 Spark 的迭代计算能力，将 K-means 算法与 Spark 并行框架结合，提升该算法模型对数据的处理速度，缩短算法的整体运行时间.通过疾病检测数据进行实验，结果表明 Spark 并行化的 PSOK-means 算法在保证准确率的同时大大提高了算法的效率，本算法对于海量数据的聚类研究有着很好的应用场景.

关键词： Spark;K-means;PSO；迭代计算

Research on K-means Algorithm of Spark Parallelization

XU Ming-jie，WEI Cheng-jian, SHEN Hang
(College of Computer Science and Technology Nanjing Technical University, Nanjing 211816, China)

Abstract： In view of the problem of insufficient memory caused by the increase of iterative computation in the process of mass data processing in K-means algorithm, this paper proposes K-means algorithm of Spark parallization.the algorithm uses particle swarm optimization (PSO) to improve the global search ability of K-means to get the initial clustering center. Through the iterative computing power of Spark, the K-means algorithm is combined with the Spark parallel framework to improve the processing speed of the model and reduce the overall running time of the algorithm. The experiment was carried out by disease detection data, the experimental results show that the Spark parallelized PSOK-means algorithm greatly improves the efficiency and accuracy of the algorithm,It will be good application scenarios for the clustering of massive data.

Key words： spark; k-means;PSO;iterative computing

作者简介：
许明杰男，（1992-），硕士研究生.研究方向为云计算、机器学习、人工智能等.E-mail:1582677459@qq.com.
蔚承建男，（1957-），博士，教授，硕士生导师.研究方向为分布式计算、人工智能等.
沈航男，（1984-），博士，讲师.研究方向为云计算、移动互联网、无线多媒体通信协议等.