一种 Spark 下分布式 DBN 并行加速策略

黄震 1，钱育蓉 1，于炯 1,2，英昌甜 1,3，赵京霞 1
(1 新疆大学 软件学院，新疆 乌鲁木齐 830008；2 新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830046；3 新疆大学 电气工程学科博士后科研流动站，新疆 乌鲁木齐 830000)

摘要：Spark 下分布式深度信念网络(Distributed Deep Belief Network，DDBN)存在数据倾斜、缺乏细粒度数据置换、无法自动缓存重用度高的数据等问题，导致了 DDBN 计算复杂高、运行时效性低的缺陷.为了提高 DDBN 的时效性，提出一种 Spark 下 DDBN 数据并行加速策略，其中包含基于标签集的范围分区(Label Set based on Range Partition，LSRP)算法和基于权重的缓存替换(Cache Replacement based on Weight，CRW)算法.通过 LSRP 算法解决数据倾斜问题，采用 CRW 算法解决 RDD(Resilient Distributed Datasets)重复利用以及缓存数据过多造成内存空间不足问题.结果表明：与传统 DBN 相比，DDBN 训练速度提高约 2.3 倍，通过 LSRP 和 CRW 大幅提高了 DDBN 分布式并行度.

关键词： 分布内存计算框架；缓存替换；范围分区；深度信念网络；数据倾斜

A Parallel Acceleration Strategy for Distributed DBN in Spark

HUANG Zhen1, QIAN Yu-rong1, YU Jiong1,2 ,Ying Chang-tian1,3, Zhao Jing-xia1
(1 School of Software, Xinjiang University, Urumqi 830008, China；
2 School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China；
3 Postdoctoral Research Station of Electrical Engineering, Xinjiang University, Urumqi 830000, China)

Abstract： DDBN(Distributed Deep Belief Network,DDBN) has many problems in Spark,such as data skew, lack of fine-grained data replacement, and unable to cache data with high re-usability automatically, resulting in high complexity and low timeliness of DDBN computing. In order to improve the timeliness of DDBN, a parallel acceleration strategy is proposed for DDBN in Spark, which includes LSRP(Label Set based on Range Partition,LSRP) algorithm and CRWS(Cache Replacement based on Weight Statistics,CRWS) algorithm. The problem of data skew is solved by LSRP algorithm, and CRW algorithm is used to solve the problem of RDD reuse and cached data caused by insufficient memory space. The results show that compared with the traditional DBN, the training speed of DDBN is increased by about 2.3 times, and the distributed parallelism of DDBN is greatly improved through LSRP and CRW.

Key words： distributed memory computing framework;cache replacement; range partition; deep belief network; data skew

作者简介：
黄震男，(1989-)，硕士研究生.研究方向为内存计算.
钱育蓉(通信作者)女（满族），(1980-)，博士，教授.研究方向为网络计算和遥感图像处理.
E-mail：qyr@xju.edu.cn.
于炯男，(1964-)，博士，教授，博士生导师.研究方向为网络安全,网格与分布式计算等.
英昌甜女，(1989-)，博士.研究方向为分布式并行计算、分布式并行系统、内存计算.
赵京霞女(满族)，(1995-)，硕士研究生.研究方向为深度学习在遥感图像处理方面的应用.