一种处理不平衡大数据的并行随机森林算法

秦　静，钱雪忠，王卫涛，谢国伟，宋　威

(江南大学 物联网技术应用教育部工程研究中心，江苏 无锡 214122)

摘　要：基于 MapReduce 机制的并行随机森林算法 MR_RF 在处理不平衡大数据时，由于数据本身正类密度低且算法又以全局最优标准选择分割点，而导致正类有误分为负类的趋势，降低了分类效率.本文提出了一种改进的并行随机森林算法(SBWMR_RF)，该算法利用分层自助抽样方法增大对少数类的抽样数量，同时考虑正负类不同的误分代价，动态计算每个分区的代价敏感矩阵，将其引入到构建基分类器的关键步骤，弱化数据偏置的影响.实验证明 SBWMR_RF 算法提高了对不平衡大数据的分类准确率，没有出现过拟合现象，在极不平衡环境下优势明显.

关键词：不平衡大数据；MapReduce；随机森林；代价敏感；分层自助抽样

A Algorithm for Unbalanced Big Sata Using Paralleled Random Forest

QIN Jing, QIAN Xue-zhong, WANG Wei-tao， XIE Guo-wei, SONG Wei

(Engineering Research Center of Internet of Things Technology Applications,
Ministry of Education, Jiangnan University, Wuxi 214122, China)

Abstract：Paralleled random forest algorithm based on MapReduce(MR_RF) which constructing trees on partitions is a classic ensemble algorithm for big data classification.However,when encountering imbalanced big data,it's performance will decrease with the tendency of positive samples misclassified because of the low density of positive samples themselves and the algorithm's global optimal criteria for choosing split points.In this paper,An improved paralleled random forest called SBWMR_RF is proposed.It adopts stratified bootstrap to increase the minority during sampling.At the same time,cost-sensitive thought is applied through the key steps of tree construction,modify the distribution of the minority.The experiments prove that SBWMR_RF can effectively classify unbalanced big data especially in extremely unbalanced data scenario without overfitting but high speedup.

Key words：unbalanced big data;MapReduce;random forest;cost-sensitive;stratified bootstrap

作者简介：

秦　静　女，（1992-），硕士研究生.研究方向为数据挖掘.

E-mail：emily139617@126.com.

钱雪忠　男，（1967-），副教授.研究方向为数据库、数据挖掘.

王卫涛　男，（1989-），硕士研究生.研究方向为数据挖掘.

谢国伟　男，（1991-），硕士研究生.研究方向为计算机应用技术、数据挖掘.

宋　威　男，（1981-），博士，副教授.研究方向为数据挖掘、人工智能和模式识别、信息检索.