MapReduce 环境下处理多类别不平衡数据的改进随机森林算法

姚立，张曦煌
(江南大学 物联网工程学院，江苏 无锡 214122)

摘要：针对传统 MapReduce 环境下的随机森林算法在处理多类别不平衡数据问题时仍然以全局最优点作为划分点，忽略了少数类对分类准确率的影响，文本提出了一种 MapReduce 环境下处理多类别不平衡数据的改进随机森林算法（MR-RF-SHDSE）.该算法利用分层采样方法在各个类别中进行样本抽样，并以 HDDT 决策树作为基学习器以弱化数据偏置给分类准确率带来的影响，最后计算决策树的 GMean 值和不合度值，利用调和平均值作为衡量标准对决策树进行选择集成.通过实验证明，相比其他算法，MR-RF-SHDSE 能够有效提高了对多类别不平衡数据集的分类准确率.
关键词： MapReduce；随机森林；分层采样；HDDT 决策树；选择集成

An Improved Random Forest Algorithm for Multi-Class Imbalanced Data Problem Under Map Reduce

YAO Li，ZHANG Xi-huang
(School of IOT Engineering, Jiangnan University, Wuxi, 214122, China)

Abstract： Because the traditional random forest algorithm under the MapReduce still takes the global optimal point as the dividing point when dealing with the multi-class imbalance data problem, ignoring the influence of the minority class on the classification accuracy rate, this paper presents an improved random forest algorithm (MR-RF-SHDSE) for dealing with multi-class imbalance data under MapReduce. This algorithm uses the stratified sampling method to sample the samples in each category, and uses the HDDT decision tree as the learner to weaken the impact of data bias on the classification accuracy. Finally, the GMean value and the disagreement measure value of the decision tree are calculated, we use harmonic mean as a metric to select decision trees. Experiments show that compared with other algorithms, MR-RF-SHDSE can effectively improve the classification accuracy of multi-class imbalanced data sets.
Key words： mapreduce; random forest; stratified sampling; hellinger distance decision tree; selective ensemble

作者简介：
姚立男，(1991-)，硕士研究生.研究方向为数据挖掘、机器学习.E-mail：yl_jnu@126.com.
张曦煌男，(1962-)，博士，教授.研究方向为嵌入式系统、数据挖掘、网络协议、图形与图像处理.