一种处理非平衡数据集的优化随机森林分类方法

马海荣 1，程新文 2
(1 湖北省农业科学院 农业经济技术研究所，湖北 武汉 430064；2 中国地质大学(武汉) 信息工程学院，湖北 武汉 430074)

摘要： 利用传统随机森林(random forest，RF)模型进行分类时存在分类精度受不平衡样本集的影响，以及投票平局现象会导致算法停滞等问题.本文对 RF 模型进行了优化改进，首先随机抽取等量的少数类与多数类样本构建训练样本集进行 RF 建模，然后根据投票熵与基于样本特征参数的广义欧几里得距离逐步添加具有最大投票熵的样本到训练样本集，解决传统 RF 模型随机抽取样本时训练样本集中包含不同类别样本数不平衡问题.对于分类过程中可能出现投票结果的平局现象，利用测试样本与邻近训练样本的广义欧几里得距离决定其分类结果，以消除投票平局现象造成的停滞问题.实验结果表明，本文优化 RF 模型对于非平衡数据集的分类可以取得较好的分类结果.

关键词： 随机森林；最大投票熵；广义欧几里得距离；不平衡数据集

A Method for Unbalanced Big Data Classification Based on Optimization Random Forest

MA Hai-rong1，CHENG Xin-wen2
(1 Hubei Academy of Agricultural Science, Wuhan 430064, China;
2 Faculty of Information Engineering, China University of Geosciences, Wuhan 430074,China)

Abstract： When utilized traditional random forest (RF) model for classification, there were following problems exist: for example, the classification accuracy was affected by the unbalanced sample set, equality votes of each class would lead to algorithms stalling. We improved the traditional RF model, first of all, we randomly selected the same number of samples from minority class and majority class to build training sample set for RF modeling. Then, according to the voting entropy and the generalized Euclidean distance based on the sample characteristic parameters gradually add the sample with maximum voting entropy to the training sample set. This could solve the problem that in traditional RF model training samples randomly selected contained too few minority class samples. In the classification process when the voting draw occurs, we utilized the generalized Euclidean distance between the test samples and the adjacent training samples to determine the classification result, this would eliminate the stagnation caused by the equality votes of each class. The experimental results show that the optimized RF model in this paper could achieve better classification results for unbalanced data sets.

Key words： random forest; max entropy voting model; generalized euclidean distance ; unbalanced data set

作者简介：
马海荣女，（1986-），博士，助理研究员.研究方向为遥感影像信息提取、数字图像处理和农业遥感应用.E-mail：mahairong1008@126.com
程新文男，（1956-），教授，博士生导师.研究方向为 3S 技术集成及应用.无线通信网络中快