

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

中图分类号:TP311.1

文献标识码:A

文章编号:1000—7180(2018)11—0028—05

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

1 引言

RF 模型是一种由数据驱动的非参数分类方法,也属于典型的机器学习方法.该方法自被提出以

来,就被广泛的应用于各个领域,例如数据分类^[1]、人体识别^[2]、图像分割^[3]等领域具有广泛的应用.但是利用传统的 RF 模型分类时,存在对非平衡样本集分类效果差、只能处理离散后的连续变量和投票

平均现象会造成算法停滞等问题,相关学者对于此类问题也进行了相应的优化研究^[4-5]. 本文在前人研究的基础上,提出改进训练样本集选择和分类结果最终决定方式来改进传统 RF 模型,在一定程度上解决了对非平衡样本集分类效果差和投票平局造成算法停滞现象的问题,提高了 RF 模型对非平衡数据集的分类精度.

2 RF 模型

2.1 RF 模型

RF 算法是由 Breiman L 等人于 2001 年基于随机子空间方法^[6]提出的一种机器学习算法^[7]. 该算法属于组合分类方法,它基于 Bootstrap 抽样原理从原始数据集中有放回的随机选择有差异的若干个训练样本子集;以无剪枝决策树为基分类器对每个 Bootstrap 样本子集进行建模,并将生成的多个决策树集成;利用所构建的决策树群对测试数据进行分类投票,最后由投票结果决定样本的最终分类或预测结果,RF 模型的建模分类流程如图 1 所示.

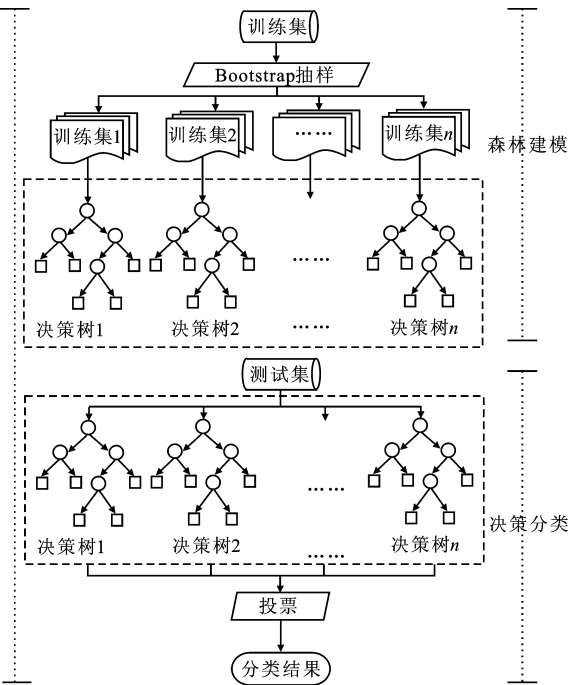


图 1 RF 分类模型

2.2 RF 模型分类的优势

RF 是一种智能机器学习算法,该方法集合了组合分类器、Bagging 随机抽样以及 CART 树构建时特征随机选择的优势,与其他的分类方法相比具有以下几类优点:(1)分类效果好,适应数据范围广;(2)分类过程中需要的人工干预少;(3)可以提供多种数据的刻画描述;(4)分类结果受样本噪声的影响

小;(5)分类运算速度快.

2.3 RF 模型分类的不足

随着应用范围增大与应用研究深入,RF 算法所存在的问题也逐渐暴露,RF 算法的不足之处主要体现在以下几个方面.(1)RF 对非平衡数据的处理效果差^[8];(2)RF 的分类时未考虑决策树分类性能的优劣性传统 RF 的分类结果投票时所有决策树的投票权重一致;(3)投票结果的平局现象没有得到有效的预防和处理. 传统 RF 模型对于以上几个问题均未给出相应的解决办法.

3 RF 模型优化

3.1 针对非平衡数据处理的样本选择方法的改进

若某个数据集中有一类数据在数量上所占比例与其他类别的数据数量相比明显的要少,则该类被称为少数类,反之,所占数量比例最大的那类被称为多数类,这种类型的数据集则被称为非平衡数据集. 在实际应用中,非平衡数据大量存在,若不进行处理,直接进该类数据分类预测由于少数类数据稀少,其特征无法得到充分学习,导致少数类的分类效果远低于多数类,从而导致分类器分类性能下降^[9]. 针对数据分布不平衡所导致的分类器分类效果不理想的问题,国内外目前的解决办法有改进算法和改造数据两大类. 改进算法是基于非平衡数据集的数据分布特征,提出新方法或优化已有方法来实现对非平衡数据集的高精度分类;改造数据主要是改变已有非平衡样本集中数据的分布^[10]. 本文主要从改进样本选择方法上提升样本数据质量和解决数据分布不平衡问题,以达到利用少量分类样本实现高精度分类地目的.

传统的 RF 建模时一般从原始数据集中随机抽取 20% 的样本构成训练样本集,剩余 80% 的样本作为测试样本集. 但是在实际应用中样本数据的获取会耗费大量的精力与时间,并且随机获取的样本数据没有考虑样本的特征典型性与多样性. 本文首先利用少量训练样本初始化机器学习模型,然后根据对模型的迭代状态和数据结构设定相应的规则选择最有价值的样本,并将样本添加到训练数据集中构造样本集.

本文样本选择方法改进基本思想为首先从原始数据集中随机抽取少量的样本(多数类和少数类样本数量相同)构成初始训练样本集,进行 RF 模型的训练与测试样本的分类;根据所构建各子树的投票结果计算各测试样本对象的分类结果,并用投票熵

来衡量各样本的分类不确定和多样性,选择投票熵较大的前 m 个样本数据;计算 m 个测试样本与所有训练样本对象之间基于特征参数的广义欧几里得距离 D ,并记录 D 最小时对应的训练样本的类别 $C_{D_{\min}}$,若 $C_{D_{\min}}$ 与 RF 投票结果一致,将该样本加入训练样本集,并将该样本从测试样本集中删除;如此迭代,不断增加训练样本集的样本数据,直到达到设定的训练次数,分类结束,输出分类结果.

投票熵 $H(s)$ 是由文献[11]提出用以度量测试样本数据分类不确定性的一种指标,某样本的 $H(s)$ 越大,其分类不确定性越大,根据这一规则所选择的样本所包含的信息量越丰富,样本越具有多样性. $H(s)$ 的定义如下:

$$H(s)=-\sum_i p_i(s)\times\log(p_i(s)) \tag{1}$$

式中, s 表示测试样本集中的一个样本对象; $P_i(s)$ 表示对于样本 s 各决策树对不同类的投票结果,本文中分别对应滑坡和非滑坡类.

分类不确定性越大的样本其分类结果的不准确性越大,为了提高样本分类的准确性,在此引入两个样本之间基于特征参数的欧几里得距离概念. 两样本之间基于特征参数的欧几里得距离越小,样本特征参数的相似性越大,两样本类别属性相同可能性越高(由于样本对象的各特征参数量纲不一致,参与计算基于特征参数的欧几里得距离的特征参数为归一化处理后的特征值). 计算投票熵较大的前 m 个测试样本距离训练样本基于特征参数的欧几里得距离,记录 D 取最小值时所对应的训练样本的类别 $C_{D_{\min}}$,若 $C_{D_{\min}}$ 与 RF 投票的分类结果一致,将该样本加入训练样本集,并从测试样本集中删除. 基于特征参数的广义欧几里得距离计算过程如下:

$$\text{Dist}_E=\sqrt{\sum_{i=1}^n(C_{\text{test}<i>-C_{\text{train}<i>})^2} \tag{2}$$

式中, C_i 为样本数据的第 i 个特征参数的归一化值; n 表示样本数据中的特征参数的个数.

3.2 投票平局现象的预防与处理

传统的 RF 分类结果利用多棵决策树投票法获得,以得票最多的类别属性作为样本最终的分类结果. 而在实际分类时会存在恰好各个类别在决策树投票时得票相等的情况,即投票结果平局现象. 此时若没有给定分类结果选择规则,最终的分类结果则无法确定,从而导致算法停滞. 传统的 RF 分类算法中,并没有针对投票平局现象提出一定的预防措施

或者解决方案,所以一旦投票平局现象出现,算法就会因为无法选择最终分类结果而陷入停滞状态. 本文设定当平局现象出现时用距离该测试样本的广义欧几里得距离最近的训练样本所属类别作为该测试样本的分类结果. 经本文优化改进后的 RF 分类具体技术流程如图 2.

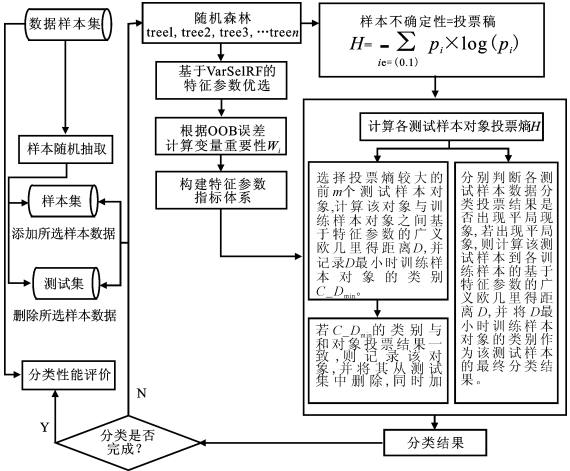


图 2 本文优化 RF 算法分类流程图

4 实验结果及分析

4.1 实验数据介绍

实验数据来源于 Martha T R 与 Kerle N 等人在文献[12]所提供的印度北阿坎德邦喜马拉雅山麓 Madhyamaheshwar 河流域的滑坡数据. 研究区面积约 28 km²,该地区是全球滑坡灾害多发地之一(研究区滑坡分布情况见图 3(1),黑色圈线表示滑坡边界). 数据源为来自印度资源卫星的 5.8 米分辨率(包括红、绿与近红外三个波段)的影像数据,以及从立体制图星-1 号 2.5 米分辨率数据生成的 10 米分辨率的 DEM. 首先利用面向对象分割软件 eCognition Deveopler8.9 对研究区的数据进行分割获得实验区的对象样本集. 经过多次实验,进行多尺度分割时将形状因子设置为 0.3,紧致度因子设置为 0.7,分割尺度设置为 10(分割结果见图 3(2)). 实验区共有分割为 28 127 个对象样本数据,其中滑坡对象 1 167 个,非滑坡对象 26 960 个,实验区样本集属于典型的非平衡数据集.

首先提取样本对象在遥感影像上的纹理、几何、亮度等遥感特征参数和坡度、坡向、顺坡性、坡形等地形特征参数共 150 个;然后利用 R 语言程序包 VarSelRF 进行特征参数优选. 共选取 Ratio_red, Ratio_nir, Min_pixel_, Max_diff, GLDV_Mean_,

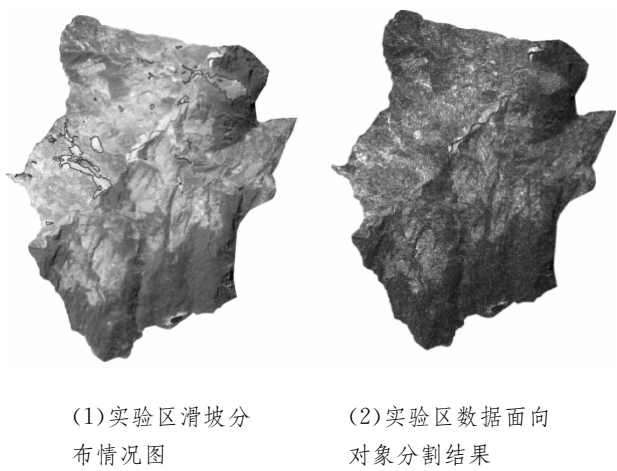


图3 滑坡分布情况

Mean_slope, Mean_red, Mean_pcl, Mean_ndvi, Max_pixell, Brightness, Shape_Area 等 27 个特征参数构成实验区最终的特征参数指标体系. 最后利用本文方法优化的 RF 模型进行研究区的滑坡分类.

4.2 实验结果分析

本文利用由 500 棵决策树构成的优化 RF 模型进行分类实验,首先随机抽取 10 个滑坡样本和 10 个非滑坡样本参与原始 RF 建模,经过 2 500 次迭代后,训练样本集包含滑坡样本为 477 个,非滑坡样本为 585 个,滑坡样本与非滑坡样本基本在同一个数量级,即样本的选择基本不受原始样本集为非平衡数据集的影响;最终参与 RF 建模的样本总数为 1 062 个,占样本数据集的 3.939 2%. 与传统 RF 一般从原始数据集中随机抽取 20% 的样本数据构成训练样本数据集,剩余 80% 的数据作为测试样本数据集相比,本文的样本选择方法既不受原始数据集的目标类与非目标类数据不平衡的影响,又可利用少量的样本数据参与训练减少了运行时间,并能获取较高的分类精度.

利用文献[13]所介绍的分类器性能评价指标来对比分析对本文优化 RF 模型的分类性能进行评价. 表 1 展示了在实验区利用本文改进的 RF 学习算法进行滑坡分类时所参与建模的滑坡和非滑坡样本数量以及分类精度变化趋势. 从表 1 可知迭代 2 500 次结束后, user’s accuracy 值大于 0.93, Accuracy 值大于 0.99, producer’s accuracy 相对较低约为 0.79, 分类器分类性能的综合评价指标 G. mean 值为 0.888 1. 因此,从各分类器评价指标综合可知本文优化 RF 模型在进行非平衡数据集的分类时可获得满意的分类结果.

表 1 实验区基于优化 RF 的滑坡分类性能评价结果

次数	样本数 (滑坡)	样本数 (非滑坡)	users. acc	prod. acc	Accuracy	G. mean
1	10	10	0.276	0.926	0.897	0.911
250	58	85	0.707	0.709	0.976	0.837
500	134	157	0.743	0.784	0.981	0.880
750	189	220	0.765	0.765	0.983	0.871
1 000	234	281	0.795	0.766	0.985	0.872
1250	301	368	0.838	0.765	0.987	0.872
1 500	388	473	0.901	0.762	0.990	0.872
1 750	443	532	0.921	0.780	0.992	0.882
2 000	462	566	0.923	0.788	0.992	0.887
2 250	467	576	0.930	0.788	0.993	0.887
2 500	477	585	0.936	0.789	0.993	0.888

为了对比分析本文的对非平衡数据集处理的优化效果,对实验区的数据集利用各参数与优化 RF 相同的传统 RF 模型随机从样本集中抽取 20% 的数据构成训练样本集(研究区一共有对象样本数据 28 127 个,抽取 20% 的样本数据集中有滑坡样本 198 个,非滑坡样本 5 427 个)进行 RF 建模,并对剩余 80% 的测试样本集进行分类. 传统 RF 模型与本文优化 RF 模型分类精度对比结果见表 2,由表 2 可知优化 RF 模型的各项分类器指标均高于传统 RF 模型,但是本文改进 RF 模型的缺点是该方法是通过多次迭代不断添加训练样本,因此对同一数据集分类时本文优化 RF 的平均耗时比传统 RF 的耗时要长.

表 2 传统 RF 模型与本文优化 RF 模型训练样本集选择与分类精度对比结果

	滑坡 样本	非滑坡 样本	users. acc	prod. acc	Accur acy	G. mean
传统 RF	198	5 427	0.824	0.579	0.977	0.759
优化 RF	477	585	0.936	0.790	0.993	0.888

本文采用文献[14]所提出的四个基于面积的滑坡分类精度评价指标 DP、QP、OE 和 CE 来计算滑坡分类精度,计算结果见表 3. 由表 3 可知实验区的滑坡识别指标 DP 与 QP 均超过 92%, 滑坡识别综合效果较好; OE 和 CE 相对均较小,但是 OE 略大于 CE,说明本文方法滑坡识别时非滑坡区域被识别为滑坡的区域较少,滑坡区域被识别为非滑坡的区域较多,如图 4 所示. 实验区 CE 值为 1.83%, 即滑坡误识别率比较小.

表 3 滑坡分类精度评价结果

TP/m ²	FN/m ²	FP/m ²	DP	QP	OE	CE
707 550.12	44 539.36	13 153.24	94.08	92.46	5.92	1.83

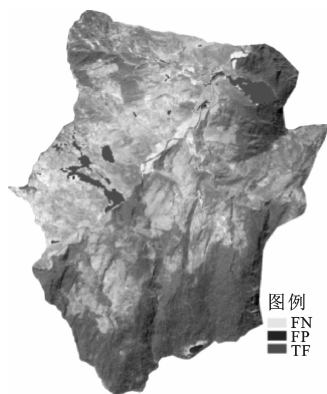


图4 实验区滑坡分类结果

5 结束语

本文对于传统 RF 模型中存在的对非平衡数据集的处理效果差以及需要的训练样本数据多的问题样本选择方式上进行改进,选择分类不确定性大的样本构建训练样本集,以增加样本的多样性;并且结合投票熵和基于测试样本和训练样本间基于特征参数的欧几里得距离确定最终的分类结果,既可以提高分类精度,又可以解决投票平均现象造成的算法停滞问题.实验证明本文的优化 RF 模型在对不平衡数据集分类时可以取得较好的效果.但是本文优化 RF 模型的缺点是该方法是通过多次迭代不断添加训练样本,因此相同样本集分类时本文优化 RF 的耗时比传统 RF 模型要长,这是将来要解决的问题.

参考文献:

- [1] Stumpf A, Kerle N. Combining random forests and object-oriented analysis for landslide mapping from very high resolution imagery[J]. *Procedia Environmental Sciences*, 2011, 3(1):123-129.
- [2] 马越, 贺光辉. 一种改进的基于随机森林的快速人眼定位方法[J]. *微电子学与计算机*, 2016, 33(7):1-4.
- [3] 刘亚东, 崔日鲜. 基于可见光光谱和随机森林算法的冬小麦冠层图像分割[J]. *光谱学与光谱分析*, 2015, 35(12):3480-3484.
- [4] 焦盛岚, 杨炳儒, 翟云, 等. 一种用于非平衡数据分类的集成学习模型[J]. *计算机工程与应用*, 2012, 48(29): 119-123.
- [5] 秦静, 钱雪忠, 王卫涛, 等. 一种处理不平衡大数据的并行随机森林算法[J]. *微电子学与计算机*, 2017, 34(4):22-27.
- [6] Ho T K. The random subspace method for constructing decision forests[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1998, 20(8): 832-844.
- [7] Breiman L. Random forests[J]. *Machine Learning*, 2001, 45(1):5-32.
- [8] Weiss G M. Mining with rarity: a unifying framework[J]. *Acm Sigkdd Explorations Newsletter*, 2004, 6(1): 7-19.
- [9] 尹留志. 关于非平衡数据特征问题的研究[D]. 合肥: 中国科学技术大学, 2014.
- [10] 陈荣鑫. R 软件的数据挖掘应用[J]. *重庆工商大学学报(自然科学版)*, 2011, 28(6):602-607.
- [11] 赵悦, 穆志纯. 基于委员会投票选择方法的主动学习的研究[J]. *太原理工大学学报*, 2006, 37(4):469-472.
- [12] Martha T R, Kerle N, Westen C J V, et al. Object-oriented analysis of multi-temporal panchromatic images for creation of historical landslide inventories[J]. *Isprs Journal of Photogrammetry & Remote Sensing*, 2012, 67(2):105-119.
- [13] 曹正凤. 随机森林算法优化研究[D]. 北京: 首都经济贸易大学, 2014.
- [14] Behling R, Roessner S, Kaufmann H, et al. Automated spatiotemporal landslide mapping over large areas using rapid eye time series data[J]. *Remote Sensing*, 2014, 6(9):8026-8055.

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.