基于不平衡数据集的改进随机森林算法研究

刘耀杰,刘独玉
(西南民族大学 电气信息工程学院,四川 成都 610041)

摘 要:随机森林算法在多种应用场景与数据集中都实现了良好的模型分类效果,但该算法在应用于不平衡二分类数据集时,受限于样本数据量本身的好坏比倾斜与决策子树叶节点投票机制,对样本量占相对少数的小类属样本不能很好地对分类进行表决。对此,文中对原有随机森林算法的节点分类规则进行改进。在模型训练过程中,综合考虑度量节点样本分类占比与节点深度,增加有利于少量类样本分类信息,从而提高了少数样本类的分类准确率。通过在不同数据集上进行随机森林改进算法的效果测试,证明改进算法相对于传统算法在不平衡数据集上有更好的模型表现,大样本条件下少量类样本分类准确率有显著提升。

关键词:不平衡数据集;随机森林;决策树;节点分裂;分类准确率
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2019)06-0100-05
doi:10.3969/j.issn.1673-629X.2019.06.021

Research on Improved Random Forest Algorithm Based on Unbalanced Datasets

LIU Yao-jie, LIU Du-yu
(School of Electrical and Information Engineering, Southwest Minzu University, Chengdu 610041, China)

Abstract: Random forest algorithm has achieved a great classification effect in a variety of scenarios and datasets, but when applied in the unbalanced binary classification datasets, it is restricted to the imbalance of sample data itself and the leaf node voting mechanism, the sample which size of relatively few samples can't vote on classification very well. For this, we improve the node classification rules of original random forest algorithm. In model training, by considering sample classification proportion and the depth of the measurement nodes comprehensively, and increasing classified information in favor for the small amount of samples, the accuracy of the few sample classification can be raised. After testing on different datasets, it proves that the improved algorithm on unbalanced dataset has better performance than the traditional algorithm, and that the few sample classification accuracy has been increased significantly under the condition of large amount of dataset.

Key words: imbalance data; random forest; decision tree; node split; classification accuracy

0 引 言

随机森林算法(random forest, RF)是一种集成机器学习方法,利用随机重采样技术 Bootstrap 和节点随机分裂技术构建多棵决策树,通过投票得到最终分类结果[1]。RF 算法在含有缺失值和噪声的数据集上表现出良好的鲁棒性,并且可以利用并行计算方式加快学习速度,目前广泛应用于分类问题中。

分类是数据挖掘中最常见的任务,利用数据挖掘的方法充分发掘数据潜在信息应用于分类预测中,建立预测模型,可以对待解决问题进行有效预测[2]。在现实场景中,大量的分类问题数据集分布并不均衡,而且每个分类的重要程度也不尽相同。然而大量的实践经历和研究表明,随机森林算法在样本数量不均衡的情况下,由于算法追求全部样本集分类精度最大化,导致对少类样本分类和预测的准确率远低于对多类样本分类和预测的准确率,即算法偏向于多类[3-4]。国内

收稿日期:2018-07-21 **修回日期**:2018-11-14 **网络出版时间**:2019-03-06
基金项目:中央高校基本科研业务费专项资金项目(2017ZYXS09)
作者简介:刘耀杰(1992-),男,硕士研究生,研究方向为机器学习、数据挖掘、深度学习;刘独玉,博士,副教授,研究方向为数据挖掘、鲁棒稳定性控制。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190306.0952.068.html

外研究人员已经做了大量的工作和尝试,主要从两个方面来解决不平衡分类问题,分别为数据预处理的方法[5-6]和算法级改进的方法[7-8]。但这些方法都存在一些不足之处,例如数据预处理方法可能造成数据的完整性缺失或者数据冗余,算法级改进可能造成模型的局部过拟合或加大计算资源开销等。文中针对这一问题,对原有随机森林算法的节点分类规则进行改进:在模型训练过程中,综合考虑度量节点样本分类占比与节点深度,增加有利于少量类样本分类信息,从而提高少数样本类的分类准确率。

1　随机森林算法

1.1　算法简介

随机森林算法由 Leo Breiman(2001)[9]提出,通过自助法(Bootstrap)重采样技术,从原始数据集 N 中有放回重复随机抽取 b 个样本生成新的训练样本集合,通常抽样个数 b 等于数据集样本数,因为是有放回抽样,所以有些样本会被重复抽取,同时随机漏掉一部分样本,经过采样后的训练集样本大小通常为原始样本大小的三分之二;之后根据每个训练集分别建立决策树,将训练好的多个决策树分类器通过 Bagging[10]方法集成,形成"森林",通过投票的方法集成多个决策树的分类结果,输出最终结果。随机森林算法中的决策树在进行训练生长过程中,不进行优化剪枝,生长条件仅限制于最大深度和叶节点相关属性等生长控制条件,能有效防止决策树过拟合;另外在训练决策树过程中,对特征集也进行了随机抽样,通过无放回抽样,每次使用特征集中的一部分进行决策树构建,通过特征集与样本集的双重随机机制,构成随机森林算法的算法基本思想。

随机森林算法步骤如下:

输入:训练集 $S=\{(x_i,y_i),i=1,2,\cdots,n\},(X,Y)\in R^d\times R$,待测样本 $x_t\in R^d$;

For $i=1,2,\cdots,N_{\text{tree}}$;

1. 对原始训练集 S 进行 Bootstrap 抽样,生成训练集 S_i;

2. 使用 S_i 生成一棵不剪枝的树 h_i:

(1)从 d 个特征中随机选取 M_{try} 个特征;

(2)在每个节点上从各特征中依据 Gini 指标选取最优特征;

(3)节点分裂直到达到生长上限。

End

输出:树的集合 $\{h_i,i=1,2,\cdots,N_{\text{tree}}\}$;对待测样本 x_t,决策树 h_i 输出 $h_i(x_t)$。

1.2　算法优势

随机森林算法具有如下一些优点:

(1)随机森林算法可以处理高维数据,并且不用做特征选择。随机森林算法可以对特征的重要程度进行自排序和筛选,也可以对样本特征进行打分。

(2)模型泛化能力强,不容易过拟合。创建随机森林的时候,对 generalization error 使用的是无偏估计,通过多重随机,算法具有良好的泛化能力。

(3)训练速度快,可以并行计算。由于每个基分类器都是独立的,所以在内存允许的条件下可以进行并行计算,大大提高了算法效率。

(4)对样本缺失值不敏感,甚至在缺失值较多的情况下也可以获得较好的训练精度。

1.3　算法缺陷

在实际应用中,随机森林算法会有如下不足:

(1)对于有不同取值的属性的数据,取值划分较多的属性会对随机森林产生更大的影响,所以随机森林在这种数据上产出的属性权值是不可信的。

(2)随机森林算法在某些噪声较大的分类或回归问题上存在过拟合。

(3)对于不平衡数据集,虽然算法有一定的平衡效果,但由于分类结果倾向于最大分类正确率,少量类样本分类结果依然不理想。

2　不平衡分类问题研究现状

现实中针对不平衡数据集分类问题,国内外研究人员已经做了大量的工作和尝试,主要从数据层面[5-6]和算法层面[7-8]来解决不平衡分类问题。数据层面主要通过调整正负样本比例来平衡数据集。算法层面主要通过改进传统算法或设计新算法来适应不平衡数据集。

2.1　数据层面改进

数据层面的改进通常较简单,主要有三种方法:少量类样本过采样、大量类样本欠采样和混合采用。通过平衡数据集,有效改善传统分类器对少量类样本的辨识精度[3]。文献[6]提出了一种基于模糊样本修剪和非监督型的数据集欠采样解决方法,通过 K-means 模糊样本修剪技术处理样本集内部噪声数据和边界点,再利用非监督方法对大量类样本集进行欠采样,在缩小样本间数量差异的同时尽量降低数据集的信息损失。然而过采样会增大数据集规模,一方面使训练时间增长,另一方面容易使模型过拟合。欠采样则会造成整体数据集信息缺失。

2.2　算法层面改进

对于算法层面的研究主要是改进传统算法来更好地适应不平衡分类数据集,或者研究出新的算法让分类的规则更加适应不平衡分类数据集。传统算法一般追求整体分类精度最优,如果训练集是不平衡分类数据集,则分类器会提升整体准确率来进行建模,从而导致正样本的分类精度较低,负样本的分类精度较高。

文献[7]提出一种基于代价敏感机制的 GBDT 算法，针对样本分类间的不平衡性及重要程度引入代价敏感指标权重，在构建 GBDT 算法模型过程中加大了少量类样本在梯度提升过程中的权重，对于少量类样本分类精度的提升取得了一定效果。文献[8]在随机森林算法模型构建过程中对决策子树引入了权重指标，通过模型二次训练，在第一次训练给出决策子树权重的基础上提升二次训练模型的准确率，通过给出集成学习模型中不同子分类器的权重提升模型效果，但这种方法弱化了随机森林算法在充分随机化下的抗干扰能力，一定程度上增加了过拟合风险。

3 基于叶节点分裂规则改进的随机森林优化算法

3.1 算法思路

随机森林算法在训练过程中，为获得优良的泛化能力，通常会限制树的深度和参与训练的特征集大小，这意味着部分有用信息被随机性丢弃了。每次节点分裂都是利用当前可用特征所蕴含的信息，将节点划分为两个“纯度”更高的节点，也就是说会分裂出一个大量类样本占比相对父节点更高的子节点和另一个少量类样本占比相对父节点更高的子节点，即“诱导”森林中已经“生长”到条件限制深度的决策树偏向少量类节点继续“生长”。

随机森林中每个决策树的叶节点构成了分类预测打分的基础结构[11]，在对测试样本进行分类决策的过程中，每棵决策树都参与投票，在决策树进行分类过程中到达的叶节点中正负样本的比例构成了投票的基本打分[12]，然而由于数据集的不平衡性，叶节点中大量类样本占比通常占优势，这种状态下必然导致少量类样本误分率偏高。

原有算法针对不平衡数据集未做单独处理，文中针对不平衡数据集做了相应改进，提出了一种针对不平衡数据集的节点再分裂规则，对于少量类占比高于某一阈值的节点，进行再次分裂。由于随机森林算法的“诱导生长”过程依然存在一定的随机性[13]，并且在节点分裂过程中备选特征集依赖于随机森林算法中的特征随机化，所以改进算法在增加决策树最大深度的同时尽量减小了模型过拟合风险，通过这种方法可以进一步发掘蕴含少量类样本分类的有效信息，增加最终分类决策中少量类样本的投票权重。

算法为了充分利用少量类样本的信息，尝试在决策树的每个节点中都刻画出少量类样本所占比例，通过该比例说明对少量类样本的利用情况。为此，引入了一个新的定义，当前节点少量类样本数 D_L 在当前节点样本总数 D 中的占比为当前节点纯度 P（$P=\frac{D_L}{D}$）。P 值越大，说明当前节点少量类样本占比越大，当前节点信息熵值越高，代表该节点在进一步分裂时可能产生效果较好的划分。

设训练集样本总数为 N，训练集少量类样本数为 N_L，则二分类训练集偏斜程度 $B_{\text{Rate}}=\frac{N_L}{N}$。在算法实践的过程中发现，以 B_{Rate} 值作为进一步分裂节点参考的阈值存在这样一个问题：在某些数据集上，尤其是 B_{Rate} 值很小的数据集上，容易使大量节点满足分裂条件而导致决策子树规模明显变大，使模型复杂度快速提升，从而引发模型过拟合效应；经过实践发现，采用 B_{Rate} 值的开平方作为阈值较为理想，开平方函数在 0-1 范围内属于单调递增凸函数，相当于对自变量进行了“平滑”处理，例如当 $B_{\text{Rate}}=0.05$ 时，$\sqrt{B_{\text{Rate}}}=0.224$，经过开平方函数处理后 B_{Rate} 值扩大 4.48 倍，而当 $B_{\text{Rate}}=0.5$ 时，$\sqrt{B_{\text{Rate}}}=0.707$，$B_{\text{Rate}}$ 值仅扩大 1.414 倍，经过开平方函数处理后，削弱了极端值对整个算法稳定性的影响。

3.2 算法描述

算法流程如图 1 所示。

算法步骤如下：

(1)计算样本集少量类样本数占比；

(2)根据设定随机森林决策树数目进行样本与属性的双重抽样，同时记录每个抽样样本集备选属性集；

(3)构造决策树，依据设定决策树深度及最小叶节点规模利用抽样后的样本集进行模型训练，在决策树达到当前生长条件准备构造叶节点时，判断当前节点 P 值是否大于设定阈值，若满足条件，则加入备选属性集循环分裂节点直到节点深度达到特征值数量上限或当前节点 P 值小于设定阈值，构造叶节点；

(4)利用不同抽样样本集以步骤 3 构造多个决策树模型；

(5)使用测试集样本进行验证，利用已经构建的多棵决策树对每一个测试样本综合输出分类结果，其中每一个决策树模型单独输出一个分类概率，最终输出分类结果为多决策树模型输出结果的加权平均值。

4 实 验

4.1 实验描述

实验硬件平台为 Intel Core i7-4700MQ 型号 CPU 和 8 GB 内存的 PC；代码执行平台为 QT5，算法实现语言采用 C++；实验数据来源为标准数据集 UCI 上的不同数量级、不同属性个数、样本分布均衡程度不同的五个二分类数据集，数据集的数据分布情况见表 1。

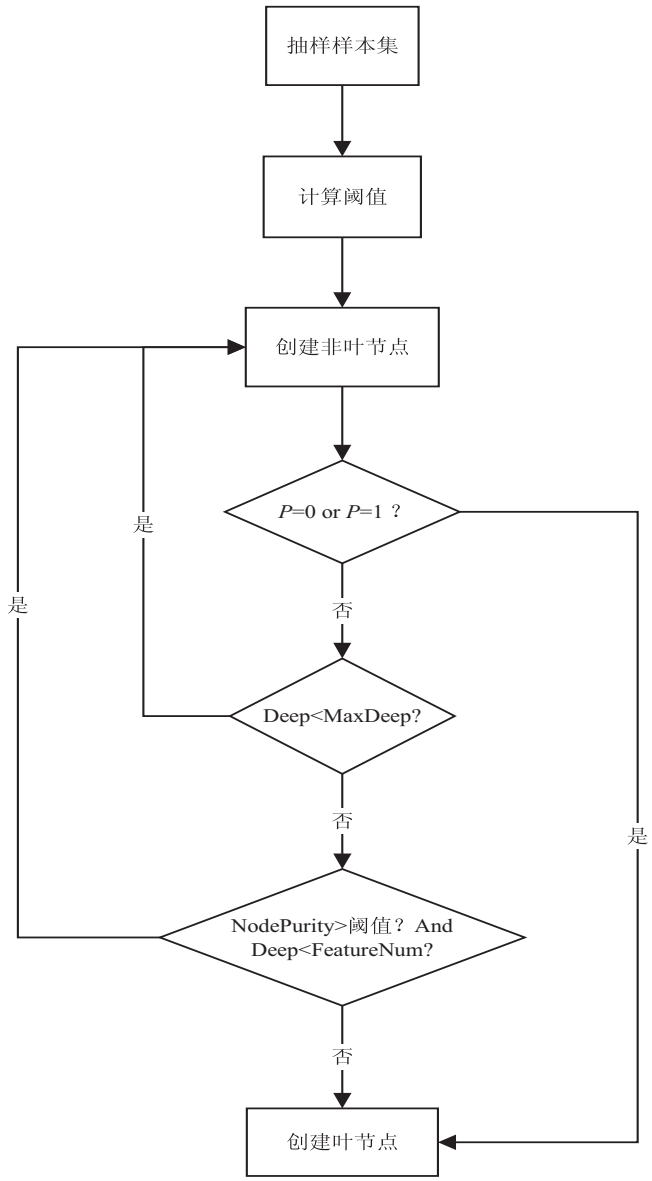

图 1　算法流程

表 1　样本集数据分布

编号	数据名称	样本个数	属性个数	正类样本占比
1	data_banknote_authentication	1 371	5	0.445
2	credit_card_clients	30 000	23	0.221
3	spambase	4 600	57	0.394
4	ionosphere	350	34	0.360
5	census-income	299 283	41	0.062

为了说明算法改进效果,实验在相同条件下分别验证标准随机森林算法与改进算法的分类结果。实验过程采用五折交叉验证,通过实验结果比对分析算法改进效果及性能。分类结果评价指标为误分率、召回率(TPR)、真负率(TNR)。定义样本数据经过模型分类后的四种结果为:TP:预测为正,实际为正;FP:预测为正,实际为负;TN:预测为负,实际为负;FN:预测为负,实际为正。定义 $\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$,表示分类器识别出的正类数占测试集正类数比重。定义 $\mathrm{TNR}=\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}$,表示分类器识别出的负类数占测试集负类数比重。

4.2　实验结果及分析

实验结果多指标对比见表 2。

表 2 实验结果多指标对比

编号	数据名称	算法	误分率	TPR	TNR
1	data_banknote_authentication	标准算法	0.092	0.866	0.94
		改进算法	0.097	0.888	0.914
2	credit_card_clients	标准算法	0.185	0.335	0.952
		改进算法	0.187	0.437	0.92
3	spambase	标准算法	0.049	0.917	0.973
		改进算法	0.070	0.957	0.913
4	ionosphere	标准算法	0.085	0.926	0.892
		改进算法	0.162	0.985	0.568
5	census-income	标准算法	0.050	0.267	0.995
		改进算法	0.048	0.356	0.991

通过实验结果对比发现:改进算法相对原标准算法 TPR 指标均有不同程度的提高,一般情况下训练样本越大,样本集的信息越丰富,也越能够反映真实情况下样本的分布情况;实验中样本集规模较大、特征丰富的样本集 TPR 指标提升效果最为明显,同时虽然 TNR 指标和整体误分率指标有轻微下降,但下降幅度相对 TPR 指标提升效果不明显;参与测试的六个样本集中有四个误分率指标变化控制在 5‰以内,TNR 指标变化控制在 5‰左右,其中样本数量最大的两个样本集(2 号样本集样本量为 30 000,5 号样本集为 300 000)的误分率指标变化都为 2‰,TNR 变化值分别为 3.2% 和 0.4%,而 TPR 提升值分别为 10.2% 和 8.9%。

结果表明:在样本类分布不均衡的 5 个实验样本中运用改进的随机森林算法,少量类样本分类准确率都有所提高,同时样本集整体误分率、TNR 指标波动相比 TPR 指标并没有产生明显的下降;改进的随机森林算法在大样本,高维度的不平衡数据集上效果明显,在取得良好的 TPR 指标提升效果的同时,数据集误分率与 TNR 指标下降波动低于 TPR 增加幅度一个数量级以上。

5 结束语

文中提出一种基于不平衡数据集的节点分裂规则改进随机森林分类器,用于解决样本不平衡对随机森林分类效果的不良影响。该算法在构建子树的过程中针对节点分裂机制引入样本类占比,针对训练样本集原有样本类占比值直接作为节点分裂参考指标产生的不稳定性,引入开平方函数,利用该函数一阶导数递减性,弱化样本类占比值直接作为阈值指标带来的潜在不稳定性,有针对性地引入一定比例的备选特征集应用于对于增加少量类样本分类权重有明显价值的少数节点,在挖掘少量类样本潜在信息的同时避免了决策子树深度增加带来的叶节点指数增长,从而降低了过拟合风险。在样本失衡比例不同的多个 UCI 数据集上的实验结果表明,相对于原有随机森林算法,该算法提高了少量类的分类准确率,样本量越大,特征空间维数越高,算法表现越好。但是算法在少量类低维样本数据集上的表现不稳定,效果相对较差,还有待进一步完善。

参考文献:

[1] 姚登举,杨 静,詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报:工学版,2014,44(1):137-141.

[2] 董跃华,刘 力. 基于均衡系数的决策树优化算法[J]. 计算机应用与软件,2016,33(7):266-272.

[3] 连克强. 基于 Boosting 的集成树算法研究与分析[D]. 北京:中国地质大学(北京),2018.

[4] 王日升,谢红薇,安建成. 基于分类精度和相关性的随机森林算法改进[J]. 科学技术与工程,2017,17(20):67-72.

[5] 徐少成. 基于随机森林的高维不平衡数据分类方法研究[D]. 太原:太原理工大学,2018.

[6] ANIS M. 基于邻近重采样和分类器排序的信用卡欺诈检测中不平衡数据研究[D]. 成都:电子科技大学,2018.

[7] 王天华. 基于改进的 GBDT 算法的乘客出行预测研究[D]. 大连:大连理工大学,2016.

[8] 叶 枫,丁 锋. 不平衡数据分类研究及其应用[J]. 计算机应用与软件,2018,35(1):132-136.

[9] BREIMAN L. Randomforest[J]. Machine Learning,2001,45(1):5-32.

[10] BREIMAN L. Baggingpredictors[J]. Machine Learning,1996,24(2):123-140.

[11] 唐耀先,余青松. 消除属性间依赖的 C4.5 决策树改进算法[J]. 计算机应用与软件,2018,35(3):262-265.

[12] 黄小猛. 异构代价敏感决策树与随机森林核心技术[D]. 桂林:广西师范大学,2013.

[13] 张 亮,宁 芊. CART 决策树的两种改进及应用[J]. 计算机工程与设计,2015,36(5):1209-1213.