一种基于粒子群算法优化的加权随机森林模型

王 杰, 程学新, 彭金柱

(郑州大学 电气工程学院 河南 郑州 450001)

摘要: 随机森林是一种高效的分类算法,其模型中的投票选取机制会导致一些训练精度较低的决策树也拥有相同的投票能力,从而降低准确度,而且模型中的决策树棵数及其他参数通常难以选取. 为解决此问题,在投票时将每棵决策树乘以一个与其训练精度成正比的权重,并采用粒子群算法优化随机森林模型,通过迭代优化选取模型中包含的参数. 通过UCI数据库进行验证,结果显示提出的加权随机森林模型分类正确率高于一般的随机森林算法及传统的分类算法.

关键词: 随机森林; 决策树; C4.5算法; 粒子群

中图分类号: TP181 **文献标志码**: A **文章编号**: 1671-6841(2018)01-0072-05

DOI: 10.13705/j.issn.1671-6841.2017006

0 引言

随机森林(RF)算法是一种分类模型,其本质是将Bagging算法和random subspace算法结合起来[1-3],通过构造多棵决策树分类器对测试样本进行分类,然后对这些决策树采取投票选取机制确定最终的分类结果. 由于随机森林模型对噪声和异常值的容忍度较高,且随机森林直接通过数据进行分类,不需要分类样本的先验知识,因此可省略数据预处理的工作. 随机森林算法自提出之后,被广泛地运用于数据挖掘与分类问题中[4-6].

针对随机森林算法中存在的一些问题,学者们提出了不同的方案,对随机森林算法进行改进. 文献[7]从选取特征、训练样本等多个方面对随机森林进行改进,提升了训练准确率. 文献[8]提出了基于生存树的随机森林模型(RSF),证明了随机生存森林对于高维样本的分类能力大于普通随机森林. 文献[9]将分位数回归理论引入随机森林中,提出了分位数回归森林(QRF). 在处理生长较差的决策树方面,文献[10]将随机森林算法中每棵决策树按分类性能进行排序,并淘汰掉分类性能差的决策树. 但此方法的弊端在于容易淘汰掉仅对某一类别分类较好的决策树,从而影响该类别最终的分类效果. 文献[11]提出了加权投票的概念,但其权值的计算方式不够优秀,很容易出现特别大的权值,从而使随机森林的投票都集中于较少的几棵树. 随后,文献[12]将以上两种方法相结合,采用out-of-bag准确率作为决策树投票权重并保留70%的决策树. 此方法的弊端在于每棵决策树out-of-bag样本不同从而导致投票不能保证其公平性. 此外,至今为止较少有文献提及随机森林中各参数对模型的影响,决策树棵数或剪枝阈值的选取也没有理论上的支持,通常只能靠经验选取.

为解决以上问题,本文提出了一种基于粒子群算法优化的加权随机森林算法(PSOWRF),采用粒子群算法对随机森林模型进行优化,通过迭代优化的方式选取决策树棵数、剪枝阈值等参数. 同时,为解决投票权重问题,本文从训练样本中提取出一部分样本,作为预测试样本,用来计算每棵决策树的权值,从而保证其投票的公平性. 而本文对权值的计算方式加以简化,仅采用预测试样本的分类正确率作为每棵决策树的权值. 其原因在于过于复杂的权值计算方式会极大地增加训练时间,而采用较为简单的正确率加权方法,经过粒子群算法优化之后,同样可保证其分类精确度.

收稿日期: 2017-03-10
基金项目: 国家自然科学基金项目(61473266).
作者简介: 王杰(1959—),男,河南周口人,教授,主要从事模式识别与智能控制研究,E-mail: wj@zzu.edu.cn.

1　决策树算法

1.1　ID3 算法

随机森林算法是由多棵决策树分类器构成,故在研究随机森林算法之前,需要对决策树有一定的了解. 目前有很多种决策树生成算法,但最有影响力的是 Quinlan 的 ID3 算法. 该算法在决策树中引入信息论的概念,并定义了信息增益. ID3 算法的原理如下所示: 假设样本 S 可按照目标属性分成 c 个不同的类别,则其分类的熵的定义为

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i, \tag{1}$$

其中: p_i 为目标属性的第 i 个值所对应的样本在总样本 S 中所占的比例.

以上定义的分类熵可作为样本 S 的纯度判别标准,即 $Entropy(S) = 0$ 表示样本 S 属于同一种类别. 根据此定义,可进一步引伸出样本 S 中决策属性 A 的信息增益 $Gain(S,A)$ 的定义为

$$Gain(S,A) = Entropy(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} Entropy(S_v), \tag{2}$$

其中: $V(A)$ 是决策属性 A 的值域,$|S_v|$ 是属性 A 的值为 v 的样本数量,$|S|$ 是总样本数量.

ID3 算法就是每次都选取拥有最大信息增益的属性作为其分类属性进行分类,而分类属性临界值的选取也要最大化信息增益,直到所有结点的分类熵值为 0. 若样本的所有属性均参与分类后分类熵仍大于 0,则返回该样本中目标属性的众数所对应的分类作为该样本最终的分类结果.

1.2　C4.5 算法

ID3 算法虽可正确地进行分类,但仍存在许多问题. ID3 算法只能对离散的数据进行分类,无法处理连续数据. 且 ID3 算法没有剪枝的步骤,甚至可能导致每个叶结点只包含一个样本,产生过拟合现象. C4.5 算法是对 ID3 算法的一个改进,它采用信息增益率而非信息增益来选择决策属性. 信息增益率的定义为:

$$GainRatio(S,A) = \frac{Gain(S,A)}{\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}}. \tag{3}$$

此外,C4.5 算法还加入了前剪枝的步骤. 即在分类过程中,当某集合的样本数小于一个给定的阈值 ε 时,就直接将此集合看作一个叶结点,然后返回目标属性的众数作为分类结果. 阈值 ε 直接决定了决策树是否会出现过拟合现象或者出现分类不准确,但 ε 只能凭经验选取,并无理论上的支持.

2　随机森林模型及其优化

2.1　随机森林模型

随机森林模型的实质是一个有多棵互不相关决策树的分类器. 每棵决策树均采用 Bootstrap 方法进行采样,然后再从所有的 M 个决策属性中随机挑选出 m 个属性进行分类. 在整个训练过程中,一般 m 的取值不变. 训练完成后,当测试样本输入时,每棵决策树均对测试样本进行分类,并采取投票的方法决定该测试样本的最终分类结果. 假设对于一个测试样本 x,第 l 棵决策树的输出为 $f_{\text{tree},l}(x) = i, i = 1,2,\cdots,c$,即为其对应的类别,$l = 1,2,\cdots,L$,$L$ 为随机森林中的决策树棵数,则随机森林模型(RF) 的输出为

$$f_{\text{RF}}(x) = \arg\max_{i=1,2,\cdots,c} \{ I(f_{\text{tree},l}(x) = i) \}, \tag{4}$$

其中: $I(\cdot)$ 表示满足括号中表达式的样本个数.

2.2　随机森林模型加权

在传统的随机森林模型中,每棵决策树在投票时权重都相等,但又不能保证每棵决策树的分类精度一致. 因此,总会有一些训练精度不高的决策树投出错误的票数,从而影响了整个随机森林的分类能力. 为了降低训练精度不高的决策树对整个模型的影响,本文提出了一种加权随机森林模型. 其核心是将训练样本分为

两部分: 一部分作为传统随机森林模型的训练样本,对所有的随机数进行训练; 另一部分为预测试样本,在训练完成之后,对每棵决策树分别进行测试,并计算其分类正确率,

$$w_l = \frac{X_{\text{correct},l}}{X}, l = 1,2,\cdots,L, \tag{5}$$

其中: $X_{\text{correct},l}$ 为第 l 棵树分类正确的样本数,X 为预测试样本数.

将此正确率作为对应决策树的权重,每棵决策树在进行投票时,都要乘以此权值. 则加权随机森林模型(WRF) 的输出为:

$$f_{\text{WRF}}(x) = \arg\max_{i=1,2,\cdots,c}\{\sum_{l\in L, f_{\text{tree},l}(x)=i} w_l\}. \tag{6}$$

2.3 随机森林模型 PSO 加权优化

以上算法中,剪枝阈值 ε、决策树棵数 L、预测试样本数 X、随机属性个数 m 等参数对整个模型的输出具有一定的影响. 但所有参数均需要通过经验选取,并没有理论上的支持. 粒子群算法通过对鸟类捕食行为进行模拟,能够快速地选取最优解. 本文通过将粒子群算法引入模型,对加权随机森林算法中的参数进行迭代优化,最终达到了较好的分类效果.

粒子群优化加权随机森林算法的步骤如下:

Step1 确定算法的参数,随机设定出剪枝阈值 ε、决策树棵数 L、预测试样本数 X、随机属性个数 m 的初值;

Step2 采用 Bootstrap 算法采样,随机生产 L 个训练集,并在每个训练集中选出 X 个预测试样本;

Step3 利用每个训练集剩下的样本分别生成决策树,共 L 棵,在生成过程中,每次选择属性前,均从全部属性中选出 m 个属性作为当前结点的决策属性;

Step4 当结点内包含的样本数少于阈值 ε 时,将该结点作为叶结点,并返回其目标属性的众数作为该决策树的分类结果;

Step5 当所有决策树生成后,对每棵决策树进行预测试,并利用式(5) 计算其权值;

Step6 利用式(6) 计算出模型的分类结果;

Step7 将分类结果作为适应度值,采用粒子群算法对 Step1 中提到的参数进行迭代优化,确定最终模型的参数.

3 实验验证及分析

本文中用到的实验数据均来自于加利福尼亚大学的 UCI 数据库,并选取了其中 Abalone、Banks、Car Evaluation、Letter、Wine Quality 和 Yeast 共 6 个数据集. 为了验证模型参数对加权随机森林算法分类能力的影响,本文选取 Wine Quality 数据集作为验证数据集,分别对剪枝阈值 ε 和决策树棵数 L 进行验证. 实验 1 对剪枝阈值 ε 在 0 到 30 之间进行取值,并记录所得分类准确率. 实验 1 的结果如图 1 所示. 实验 2 对决策树棵数 L 分别取值为 10 到 100 之间的 10 的整数倍,其结果如图 2 所示.

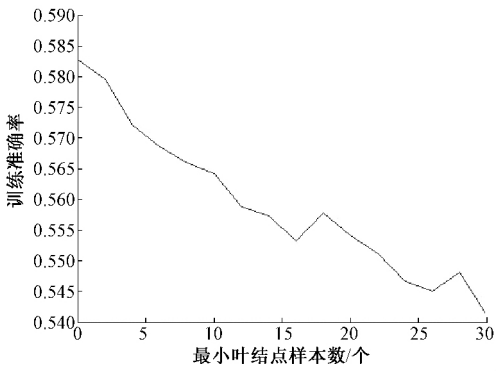

图 1 剪枝阈值对分类性能的影响

Fig. 1 Effect of pruning threshold on test accuracy

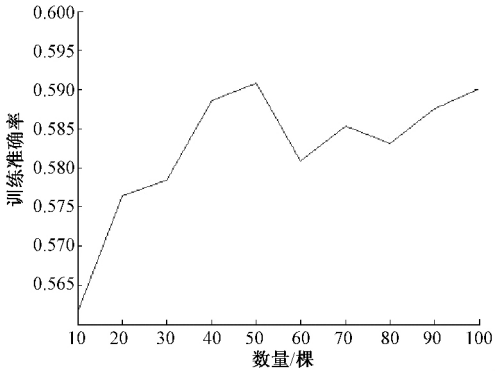

图 2 决策树棵数对性能的影响

Fig. 2 Effect of tree's number on test accuracy

通过图 1 可看出,随着剪枝阈值 ε 的不断增加,分类性能呈现一个下降的趋势. 因此对于数据集 Wine Quality 来说,剪枝阈值 ε 为 0 可取得最高的分类准确率. 从图 2 可发现,决策树棵数在 50 以后,分类的准确率在 0. 59 左右开始波动. 故对于数据集 Wine Quality,最佳的决策树棵数为 50. 因此可说明,Step1 中提及的参数对模型的分类性能具有一定的影响. 为保证选取到最优值,本文采用粒子群算法对模型进行优化,提出粒子群优化加权随机森林算法(PSOWRF) ,并在 6 组数据集上进行测试. 将其训练结果与文献[12]中提到的普通加权随机森林(WRF) 、分位数回归森林(QRF) 、随机生存森林(RSF) 、传统随机森林(RF) 、C4. 5 决策树分类器(DT) 、支持向量机(SVM) 和 BP 神经网络等传统分类器进行对比,结果如表 1 所示. 表 1 中记录了所有算法对 6 个数据集的平均分类正确率. 每个数据集名之后括号中的两个数字分类代表了该数据集的属性个数和类别个数.

表 1　不同算法分类性能比较

Tab. 1　The comparison of different algorithms

数据集	Abalone (8,3)	Banks (16,2)	Car Evaluation (6,4)	Letter (16,26)	Wine Quality (11,7)	Yeast (8,4)
PSOWRF	**0. 801 2**	0. 900 5	**0. 975 3**	**0. 756 7**	0. 603 7	0. 679 8
WRF	0. 784 1	0. 899 3	0. 965 7	0. 714 0	0. 590 5	0. 677 2
RSF	0. 762 7	**0. 904 2**	0. 967 3	0. 732 5	**0. 607 2**	0. 678 6
QRF	0. 770 6	0. 897 5	0. 971 0	0. 687 2	0. 589 4	0. 674 3
RF	0. 753 5	0. 897 7	0. 960 2	0. 666 4	0. 587 7	0. 676 0
DT	0. 718 4	0. 885 4	0. 931 3	0. 535 4	0. 497 7	0. 572 3
SVM	0. 776 8	0. 901 4	0. 972 8	0. 599 3	0. 556 8	**0. 682 2**
BP	0. 652 9	0. 864 3	0. 695 1	0. 067 0	0. 089 3	0. 451 6

注: 粗体代表了每个数据集的最优正确率,下划线代表了与最优正确率无统计学差异.

根据表 1 可以得到,PSOWRF 在 Abalone、Car Evaluation、Letter 3 个数据集上均取得了最优分类正确率,同时在 Banks 和 Wine Quality 两个数据集上的分类正确率与最优正确率无统计学差异. 对于 Abalone 数据集,PSOWRF 取得了最优结果,WRF、RSF、QRF 和 SVM 的表现情况相差不大,均比 RF、DT 和 BP 算法更加优秀. 对于高维的数据集 Banks 和 Wine Quality,RSF 充分展示了其对高维数据的分类能力,取得最优解,而 PSOWRF 也表现良好,与 RSF 无明显差异. 对于数据集 Car Evaluation,PSOWRF 和 QRF、SVM 同时取得较好的分类结果,优于其他算法. 对于多类别的 Letter 数据集,PSOWRF 远远领先于其他算法,取得最优结果. 最后,对于 Yeast 数据集,SVM 算法超过了所有的随机森林算法及改进. 综上所述,本文所提出的 PSOWRF 算法在除 Yeast 之外的 5 个数据集上均能够取得不错的表现,而且在 Abalone 和 Letter 数据集上要明显优于其他算法.

4　结论

本文对随机森林模型的投票机制做出了一定的改进,从训练样本中提取出一部分预测试样本,并将每棵决策树对预测试样本的分类正确率作为其投票权值. 为保证模型参数能够取得最优值,本文还利用粒子群算法对模型进行优化,提出了基于粒子群优化的加权随机森林算法. 该算法在 6 组实验数据集上均取得了良好的结果. 在今后的工作中,我们将会对随机森林模型的采样机制进行研究. 通过对比不同采样算法对分类性能的影响,决定出最优的采样算法,而不局限于仅用 Bootstrap 算法进行采样.

参考文献:

[1] BREIMAN L. Random forests[J]. Machine learning, 2001, 45(1):5-32.

[2] BREIMAN L. Bagging predictors[J]. Machine learning, 1996, 24(2):123-140.

[3] HO T. The random subspace method for constructing decision forests[J]. IEEE transactions on pattern analysis and machine intelligence, 1998, 20(8):832-844.

[4] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报, 2013, 50(4):1190-1197.

[5] 林成德, 彭国兰. 随机森林在企业信用评估指标体系确定中的应用[J]. 厦门大学学报(自然科学版), 2007, 46(2):199-203.

[6] 杨帆, 林琛, 周绮凤,等. 基于随机森林的潜在 k 近邻算法及其在基因表达数据分类中的应用[J]. 系统工程理论与实践, 2012, 32(4):815-825.

[7] ROBNIK-IKONJA M. Improving random forests[C]//15th European Conference on Machine Learning. Italy, 2004.

[8] ISHWARAN H, KOGALUR U B, BLACKSTONE E H, et al. Random survival forests[J]. Journal of thoracic oncology official publication of the international association for the study of lung cancer, 2008, 6(12):1974-1975.

[9] NICOLAI M. Quantile regression forests[J]. Journal of machine learning research, 2006, 7(2):983-999.

[10] CROUX C, JOOSSENS K, LEMMENS A. Trimmed bagging[J]. Computational statistics & data analysis, 2007, 52(1):362-368.

[11] AMARATUNGA D, CABRERA J, LEE Y S. Enriched random forests[J]. Bioinformatics, 2008, 24(18):2010-2014.

[12] XU B, GUO X, YE Y, et al. An improved random forest classifier for text categorization[J]. Journal of computers, 2012, 7(12):2913-2920.

A Weighted Random Forest Model Based on Particle Swarm Optimization

WANG Jie, CHENG Xuexin, PENG Jinzhu

(*School of Electrical Engineering, Zhengzhou University, Zhengzhou* 450001, *China*)

Abstract: The voting mechanism in the random forest (RF) model would reduce the correct rate. The number of decision trees and the other parameters in the random forest were difficult to select. To solve these problems, a weighted random forests model was proposed. In voting, each decision tree was multiplied a weight which was proportional to its training accuracy. The parameters contained were selected by the iterative optimization with PSO algorithm. The experimental results with the UCI database showed that the classification accuracy of the proposed model was higher than that of the original random forests and the traditional classification algorithm.

Key words: random forest; decision tree; C4.5 algorithm; particle swarm optimization

(责任编辑: 王浩毅)