计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(12):181−185 [doi: 10.15888/j.cnki.csa.006162]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

随机森林算法在小麦育种辅助评价中的应用①

邹永潘[1,2], 王儒敬[1], 李　伟[1]

[1](中国科学院 合肥物质科学研究院 合肥智能机械研究所, 合肥 230031)
[2](中国科学技术大学, 合肥 230026)

摘　要: 为了提高育种领域选种的准确率同时缩短品种培育年限, 利用改进的随机森林算法根据小麦育种历史数据构建评价模型. 在训练分类器之前, 利用改进的 SMOTE 算法来改善训练样本集中的非平衡现象; 在基分类器训练完成后, 测试单个分类器的性能并剔除性能较差的基分类器, 实现随机森林中基分类器的筛选. 实验结果表明, 文中提出的算法在小麦种质评价方面取得了不错的效果, 可以辅助育种工作者进行品种选育.

关键词: 小麦育种评价; 非平衡数据集; 随机森林; 改进的 SMOTE 方法

引用格式: 邹永潘,王儒敬,李伟.随机森林算法在小麦育种辅助评价中的应用.计算机系统应用,2017,26(12):181–185. http://www.c-s-a.org.cn/1003-3254/6162.html

Application of the Random Forest Algorithm in Wheat Breeding Evaluation

ZOU Yong-Pan[1,2], WANG Ru-Jing[1], LI Wei[1]

[1](Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China)
[2](University of Science and Technology of China, Hefei 230026, China)

Abstract: In order to improve the accuracy of seed selection and shorten the cultivation period of cultivars, the improved random forest algorithm is used to construct the evaluation model of the history data of wheat breeding. Before training the classifiers, the improved SMOTE algorithm is used to improve the non-balance of the training samples. After the training of the base classifiers, we test every classifier's performance and delete bad classifiers to realize the screening of the base classifier in random forest. The experimental results show that the proposed algorithm has achieved good results in wheat germplasm evaluation, which can help to breed varieties.

Key words: wheat breeding evaluation; imbalanced datasets; random forest; improved SMOTE

建国以来, 我国在小麦育种领域取得了卓越的成就, 选育出了数以千计的优良品种. 在先后经历了 20 世纪 50-60 年代以提高抗病稳产为主的育种阶段和 70-80 年代以矮化与高产为主的育种阶段之后, 从上世纪 90 年代开始, 我国小麦育种已进入了高产品种和优质品种并进的阶段[1]. 小麦育种是一个需要涉及多要素、受多方面因素综合影响的过程, 育种过程中各要素之间的相互关系以及各要素对育种结果的影响难以精确衡量, 因此科学有效的种质评价方法对于寻找优质品种显得至关重要.

传统的作物育种评价方法多是基于育种专家多年的育种经验对一个品种做出主观评价, 再通过来年种植下一茬作物来进行验证. 这种方法延长了品种的选育时间, 在多性状综合评价时由于人为因素干预过多, 往往导致评价的结果不甚理想. 部分育种工作者引入了层次分析法、模糊综合评价、灰色关联评价等方法来对品种进行综合评价, 这些方法在评价效果上各有优势, 有效提升了作物育种评价技术的数据化、信息化程度[2,3]. 但这些方法往往需要育种专家人为设置指标的权重来显性描述相关的专家经验, 进而来指导育

① 基金项目: 中国科学院战略性先导科技专项 (XDA08040110)
收稿时间: 2017-03-20; 修改时间: 2017-05-09; 采用时间: 2017-05-11

种评价的相关工作, 无法解释育种经验的合理性, 且模块化应用这些评价方法时难以实现. 刘忠强[4]将决策树算法应用到作物育种结果评价当中, 利用历史的育种数据记录, 建立对应的评价模型, 该模型综合考虑了各个育种性状和育种目标之间的关系, 同时体现了育种专家的历史选育经验, 可以辅助育种工作者进行育种评价. 但是, 基于决策树的评价方法需要进行大量的数据预处理工作, 且容易出现过拟合[5,6]. 随机森林算法(RF)[7]通过重采样技术构建多个弱分类器来对结果进行预测, 最终的评判结果取决于多个分类器的投票结果. RF 具有较强的容错能力且能很好的避免出现过拟合, 作为机器学习领域主流算法之一, 已经得到了十分广泛的应用[8-10].

小麦选种决策过程是从大量的已培育品种中选择出综合性能较好的品种, 可看做是一个非平衡数据集分类问题. 如果直接对原始数据进行建模, 难以得到理想的模型[11], 可以通过改造训练数据来提升训练数据的不平衡率, 主要实现方式包括随机过采样和随机欠采样. 随机过采样可能会导致最终的分类器过分的拟合训练数据, 而随机的欠采样则可能导致分类器在训练过程中失去一些多数类的信息, 从而使得分类结果对多数类不利. 针对过采样出现的问题, Chawla 等人于 2002 年提出了 SMOTE 算法[12], 该方法假设少数类样本的附近仍然是少数类, 为每个少数类样本确定其 K 个相邻的样本, 然后在该样本与其近邻样本连线上构造“人造样本”. 该方法解决了随机过采样中的过拟合问题, 但是在选取近邻样本时, 难以确定 K 的大小, 具有一定的盲目性, 此外改造后的数据集容易出现分布边缘化问题[13].

本文将一种改进的随机森林算法应用到小麦育种的种质评价阶段. 针对历史评价数据的不平衡现象, 在预处理阶段使用改进的 SMOTE 算法对训练数据进行改造, 使得训练数据中的正负类分布达到平衡; 在随机森林的决策阶段, 利用 OOB 数据计算每个基分类器的分类性能, 并剔除较差的分类器, 进一步提升分类器的综合性能. 实验结果表明, 该评价方法能够取得较准确的评价效果, 可以辅助育种工作者进行优质品种的选择.

1 相关算法介绍

1.1 随机森林分类算法

随机森林算法是由 Breiman 于 2001 年提出的一种机器学习算法[7], 实质上是由多个决策树构成的组合分类器, 其分类结果是由各个子分类器的结果共同决定, 通常是通过投票将决策票数最多的类别作为样本的最终所属类别. 随机森林的构建过程: 首先, 通过 Bagging(Bootstrap aggreating) 方法产生多个有差异的训练样本子集; 然后, 利用随机子空间划分 (Random subspace method) 方法选择部分属性采用 CART 算法无剪枝地构建多棵分类决策树.

自主抽样法是从含有 n 个样本的初始训练集中有放回的随机抽取 n 个样本形成新的训练样本子集的过程, 此处新的训练样本集大小和初始样本集相等. 因为初始训练样本集中的每个样本未被抽中的概率为 $(1-1/n)^n$, 当 n 趋向于无穷大时有:

$$\lim_{n\to\infty}(1-\frac{1}{n})^n \approx 0.368 \tag{1}$$

由上式, 初始训练样本中大约 36.8% 的样本不会出现在新训练样本集中. 初始样本集中未被抽取到的样本集合称为袋外数据 (Out of bag, 简称 OOB). 通过自助抽样法保证了子分类器之间训练样本的差异.

随机子空间的划分策略: 从拥有 M 个属性的数据集中随机抽取 m 个属性 ($m \ll M$) 作为候选属性. 在随机森林中, m 的建议取值为 $sqrt(M)$、$1/2sqrt(M)$ 或 $2sqrt(M)$[7].

对于数据集 D, 其纯度可以用基尼值来衡量:

$$Gini(D) = 1 - \sum_{k=1}^{|y|} p_k^2 \tag{2}$$

p_k 表示在数据集 D 中第 k 类样本占有的比例. $Gini(D)$ 反映了从数据集 D 中随机抽取两个样本类别不一致的概率, 值越小, 表明数据集的纯度越高.

在生成决策树的过程中, 根据属性的基尼指数进行结点的分类, 属性 a 的基尼指数定义为:

$$Gini\_index(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v) \tag{3}$$

在构建 CART 决策树时, 选择属性集合 A 中那个使得划分后基尼指数最小的属性作为最优划分属性, 即:

$$a_* = \arg\min_{a\in A} Gini\_index(D,a) \tag{4}$$

1.2 SMOTE 算法

SMOTE 算法 (Synthetic minority over-sampling technique) 其本质上是随机向上抽样算法的改进.

SMOTE 算法假设与少数类样本较近的样本也属于少数类, 通过在样本和其近邻样本连线上构造新的样本来提升训练数据的平衡率. 构造样本的过程根据公式 (5) 来完成:

$$P_{ij} = X_i + rand(0,1) \times (Y_{ij} - X_i) \quad (5)$$

其中, $X_i(i=1, 2, \ldots, n)$ 为少数类样本; $Y_{ij}(y=1, 2, \ldots, K)$ 表示与 X_i 的 K 个近邻样本中的第 j 个; P_{ij} 为 X_i 与第 j 个近邻样本合成的新样本; $rand(0, 1)$ 表示一个 0 到 1 的随机数. 假设数据集中少数类样本的个数为 N_+, 多数类样本的个数为 N_-, 采样率为 N.

SMOTE 算法的实现步骤如下:

Step 1. 计算并挑选出每个少数类样本的 K 近邻样本;

Step 2. 将每个少数类样本与其近邻样本随机地进行组合, 利用公式 (5) 产生新样本;

Step 3. 判断是否达到目标采样率, 若没有则转 Step2, 否则将所有产生的新的样本加入训练数据集中, 程序结束.

2 随机森林算法在小麦种质评价的应用

2.1 小麦种质评价流程

本文尝试将随机森林分类算法应用在小麦育种领域, 辅助育种工作者选择优质品种. 利用历史育种数据来训练分类模型, 并根据该模型实现对新培育材料的分类预测, 具体的步骤包括数据预处理、建立模型、新品种评价, 流程如图 1 所示.

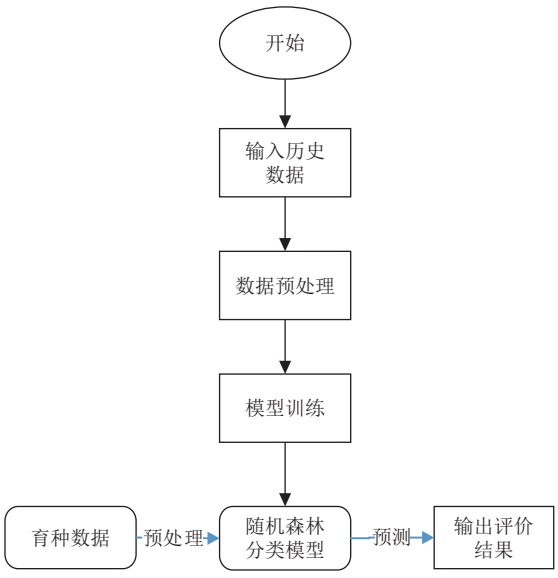

图 1 小麦种质评价流程

2.2 数据预处理

针对小麦育种记录数据, 本文进行的数据预处理包括规范化、异常值检测、缺失值填充、数据合成.

(1) 规范化

由于训练数据集可能是来自不同的育种机构, 对于同一个性状的记录可能会出现不同的描述形式, 因此需要首先对记录数据进行规范化. 主要包括计量单位的统一和表示形式的统一. 例如, 对于性状千粒重, 以克计量; 对于抗病性, 针对反应型以 1、2、3…表示等.

(2) 异常值检测

在实验过程中的异常检测主要是利用现有的育种记录经验来判断记录中是否存在不科学的记录结果, 由于育种数据来源于严谨的科研机构, 异常记录较少, 故直接删除含有异常值的记录.

(3) 缺失值填充

对于存在缺失的记录, 本文使用与给定元组属于同一类别的所有样本的均值进行填充.

(4) 数据合成

由于小麦育种数据集中的非平衡问题, 利用改进的 SMOTE 算法合成新的少数类样本, 改善训练样本集中得类别分布状况.

2.3 改进的 SMOTE 算法 (ISMOTE)

SMOTE 算法假设少数类样本的周围仍然是少数类, 并且在选择 k 近邻时存在一些盲目性. 事实上, 大多数情况下的样本分布并不满足上述假设, 这会导致经过 SMOTE 合成的样本集会出现样本重叠现象. 为了能够解决训练数据集中的非平衡问题, 同时使新合成的样本集能更加真实的反映初始数据集的分布, 本文提出了一种改进的 SMOTE 算法. ISMOTE 算法思路如下: 首先, 利用 k-均值聚类算法对少数类样本进行聚类, 得到 k 个聚类中心以及对应的簇; 然后, 利用每个样本和其对应的聚类中心合成新的样本. 具体实现流程如下:

Step 1. 对少数类样本利用聚类算法求得 k 个聚类中心 $X\_center_j(j=1, 2, \ldots, k)$, 将少数类样本集分成 k 簇样本 $Sub\_X_j(j=1, 2, \ldots, k)$;

Step 2. 任意抽取 Xi, 根据 Step 1 可得到对应的聚类中心, 利用如下公式合成新样本:

$$P_{ij} = X\_center_j + rand(0,1) \times (X_i - X\_center_j) \quad (6)$$

式 (6) 中, P_{ij} 表示少数类样本 X_i 与它对应的聚类

中心合成的新样本.

Step 3. 判断是否达到目标采样率, 若没有则转 Step 2, 否则将所有产生的新的样本加入训练数据集中, 程序结束.

经过 ISMOTE 算法处理之后, 整个预处理过程结束, 将使用新的样本集来训练分类模型.

2.4 改进的随机森林算法 (IRF)

在随机森林分类中, 最终的分类结果是由基分类器投票类别数最多的类, 没有考虑每个基分类器的分类性能. 随机的抽取样本和属性可能会导致某些基分类器的分类性能不理想甚至很差, 因此本文在利用 RF 进行分类决策之前先使用 OOB 数据对基分类器性能进行测试, 剔除性能相对较差的基分类器达到提升组合分类器性能的目的. IRF 的具体构造流程如图 2 所示.

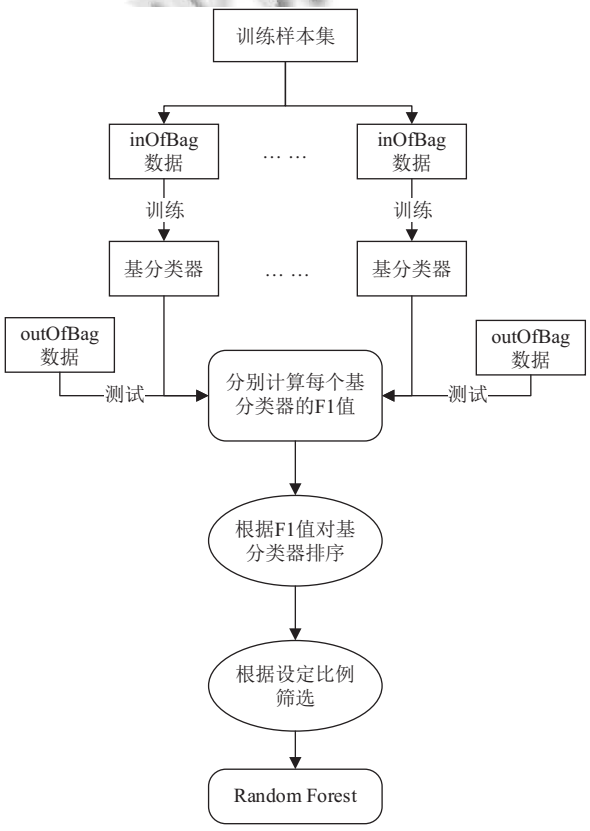

图 2 改进的随机森林构造流程图

3 实验

文中实验所用的原始数据来自于中国种业商务网的 1112 条小麦育种记录数据, 其中正类样本数为 115(假设好的品种为正类). 属性包括成熟期、株高、千粒重、亩穗数、硬度、容重、沉淀值等 24 个小麦育种过程中的常见性状, 这些性状在不同程度上反映了小麦品种的产量、抗病性和籽粒品质.

3.1 实验数据预处理

本文在预处理中的规范化主要包括计量单位的统一和量化方式的统一. 计量单位的统一针对的性状有: 株高 (cm)、千粒重 (g)、亩产 (Kg) 和容重 (g) 等. 量化方式的统一主要是针对枚举型数据, 根据性状的实际意义使用数值来进行量化表示. 例如, 锈病的反应型包括{免疫, 高抗, 中抗, 中感, 高感}, 可以使用{1, 2, 3, 4, 5}来进行量化表示. 实验中关于异常值检测和缺失值处理参照文中 2.2 节中的方法进行处理.

3.2 参数设置

在 2.2 节提出的 ISMOTE 算法中, 需要确定对少数类进行聚类的类别数 k, 根据经验方法, 对于 n 个数据集, 设置簇数 k 约为$\sqrt{n/2}$, 即实验中 k 取 8. 为了使最终的训练数据集尽可能的达到平衡, 在执行 ISMOTE 算法时, 设置采样率 N= (int)N_-/N_+. 随机森林算法中使用默认参数, 基分类器的筛选比例设置为 75%.

3.3 算法性能评估指标

由于小麦育种中更多的关注优质品种, 故在实验中只考量正类 (少数类) 有关的指标. 算法的性能评估是通过准确率 P(Precision)、召回率 R(Recall)、以及综合考虑指标 $F1$ 来衡量.

$$P=\frac{TP}{TP+FP},\ R=\frac{TP}{TP+FN},\ F1=\frac{2P*R}{P+R} \tag{7}$$

上式中, TP 表示正确分类的正例数目, FN 表示错分为负例的正例数目, FP 为错分为正例的负例数目.

3.4 实验及结果分析

利用 Java 语言在 eclipse 平台通过改进 weka 的库函数分别实现了 RF、SMOTE+RF、ISMOTE+RF 和 ISMOTE+IRF. 实验采用十折交叉验证的方式对样本集进行分析, 并基于准确率、召回率和 F1 值来对分类结果进行评估. 表 1 为利用四种方法进行实验的结果.

表 1 四种方法实验结果对比

	P	R	F1
RF	0.304	0.921	0.511
SMOTE+RF	0.764	0.755	0.759
ISMOTE+RF	0.809	0.786	0.797
ISMOTE+IRF	0.821	0.793	0.806

从表 1 可以看出, 由于小麦育种数据集存在非平衡问题, 直接使用 RF 算法进行处理得到的分类模型准确率很差, 也验证了随机森林算法在处理非平衡数据集分类问题上的局限性. 利用 SMOTE+RF 和 ISMOTE+RF 实验之后的结果在各项指标上均有不小的提升, 在一定程度上缓解了数据非平衡带来的影响. 但是后者相对前者的分类效果更好, 说明利用 ISMOTE 算法对少数类进行改造后的数据集比经 SMOTE 算法改造后的数据集更符合训练数据的原始分布情况. 在利用 IRF 算法考虑基分类器单独性能后, 算法的各项性能指标均得到了小幅的提高, 证明了在随机森林中考虑基分类器的性能、剔除不好的基分类器有助于提高随机森林的整体分类效果.

4 结语

本文尝试将随机森林分类算法应用于小麦种质评价中, 利用历史的选育评价数据训练分类器, 得到的组合分类器中可将每一个基分类器看做一个“专家”, 对新培育材料的最终评价结果由多个“专家”共同决定. 实验结果表明, 该评价方法能够取得较好的评价效果, 可以辅助育种工作者进行优质品种的选择. 然而本文的评价方法依然存在着不足, 主要体现在两个方面: 首先, 算法中的参数有待进一步优化, 从而提升算法的性能; 其次, 训练数据集的样本数量不够、属性集过小. 为了建立稳定的、具有代表性的分类评价模型, 需要进一步优化参数, 同时增加训练样本数据以及考虑包括基因型和表现型在内的更多的品种性状.

参考文献

1 李振声. 我国小麦育种的回顾与展望. 中国农业科技导报, 2010, 12(2): 1–4.

2 柏流芳, 吕黄珍, 朱大洲, 等. 农作物育种中的综合评判方法. 农业工程, 2013, 3(3): 112–119.

3 Smith AB, Lim P, Cullis BR. The design and analysis of multi-phase plant breeding experiments. The Journal of Agricultural Science, 2006, 144(5): 393–409. [doi: 10.1017/S0021859606006319]

4 刘忠强. 作物育种辅助决策关键技术研究与应用[博士学位论文]. 北京: 中国农业大学, 2016: 27–34.

5 Kubal C, Haase D, Meyer V, *et al*. Integrated urban flood risk assessment—adapting a multicriteria approach to a city. Natural Hazards and Earth System Sciences, 2009, 9(6): 1881–1895. [doi: 10.5194/nhess-9-1881-2009]

6 Liu XP, Li X, Liu L, *et al*. An innovative method to classify remote-sensing images using ant colony optimization. IEEE Trans. on Geoscience and Remote Sensing, 2008, 46(12): 4198–4208. [doi: 10.1109/TGRS.2008.2001754]

7 Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32. [doi: 10.1023/A:1010933404324]

8 赖成光, 陈晓宏, 赵仕威, 等. 基于随机森林的洪灾风险评价模型及其应用. 水利学报, 2015, 46(1): 58–66.

9 雷震. 随机森林及其在遥感影像处理中应用研究[博士学位论文]. 上海: 上海交通大学, 2012.

10 马玥, 姜琦刚, 孟治国, 等. 基于随机森林算法的农耕区土地利用分类研究. 农业机械学报, 2016, 47(1): 297–303. [doi: 10.6041/j.issn.1000-1298.2016.01.040]

11 职为梅, 郭华平, 范明, 等. 非平衡数据集分类方法探讨. 计算机科学, 2012, 39(6A): 304–308.

12 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321–357.

13 曹正凤. 随机森林算法优化研究[博士学位论文]. 北京: 首都经济贸易大学, 2014.