计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(7):212−216 [doi: 10.15888/j.cnki.csa.007482]

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

改进随机森林算法在人才培养质量评价中的应用①

毕瑶家, 刘国柱, 王华东, 孙　驰, 付兆殊

(青岛科技大学 信息科学技术学院, 青岛 266061)
通讯作者: 刘国柱, E-mail: lgz_0228@163.com

摘　要: 高校毕业生质量直接关系到高校的社会声誉与发展. 为了准确的评价高校的毕业生质量, 本文基于某高校计算机类毕业生的历史数据, 采用一种改进的随机森林算法构建人才培养质量评价模型. 在训练分类器之前, 利用RF Ranking 方法来度量特征重要性并选取 75% 的特征进行降维处理, 以此改善训练样本的非平衡现象; 通过对基分类器的训练, 测试各个分类器的性能, 依据性能的强弱对单个分类器作加权处理, 以此降低性能较差的分类器对结果的影响. 实践结果表明, 该算法提高了人才培养质量评价的准确率和精确度, 可以在高校人才培养方面起到指导作用.

关键词: 改进随机森林; 特征选取; 加权; 人才培养质量评价

引用格式: 毕瑶家,刘国柱,王华东,孙驰,付兆殊.改进随机森林算法在人才培养质量评价中的应用.计算机系统应用,2020,29(7):212–216. http://www.c-s-a.org.cn/1003-3254/7482.html

Application of Improved RF Algorithm in Quality Assessment of Personnel Training

BI Yao-Jia, LIU Guo-Zhu, WANG Hua-Dong, SUN Chi, FU Zhao-Shu

(College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: The quality of college graduates is directly related to the social reputation and development of colleges and universities. In order to accurately evaluate the quality of college graduates, based on the historical data of computer graduates in a university, this study uses an improved random forest algorithm to build a talent training quality assessment model. Before training classifiers, RF ranking method is used to measure the importance of features and select 75% of the features for dimension reduction, so as to improve the unbalanced phenomenon of training samples; through the training of base classifiers, the performance of each classifier is tested, and a single classifier is weighted according to the strength of performance, so as to reduce the impact of poor performance classifiers on the results. The practical results show that the algorithm improves the accuracy and precision of the quality assessment of personnel training, and can play a guiding role in personnel training in colleges and universities.

Key words: change random forest; feature extraction; weighting; personnel training

近年来, 国家对高校的关注点已经从数量的增长转移到学生培养质量的提升上来, 因此人才培养质量问题受到越来越多的关注. 学者关于人才培养质量已经有了一定的研究成果, 如彭建林根据就业质量的评价要求, 构建了大学生就业质量评价指标体系, 包括工作保障、薪资条件等 7 个一级指标, 并且细化为 10 个

① 基金项目: 山东省重点研发计划 (2017GGX10107)
Foundation item: Key Research and Development Program of Shandong Province (2017GGX10107)
收稿时间: 2019-10-24; 修改时间: 2019-11-20, 2019-12-17; 采用时间: 2020-01-07; csa 在线出版时间: 2020-07-03

二级指标, 并通过设计指标权重对大学生的就业质量进行了评价[1]; 宁东卫、范春梅等 (2016) 根据影响人才培养的相关因素、从个人与学校两个方面出发选取指标体系, 丰富了人才培养的指标体系, 提供了丰富的参考依据[2]; 宋俊秀、谢德刚提出了基于模糊综合评价法的大学生就业质量分析, 以安徽省为例探讨高校大学生就业质量, 构建合理的评价体系, 运用模糊综合评判法建构大学生就业质量评价模型, 对高校大学生就业质量进行总体、分学历、分学科层次多维评价[3]; 韩天才提出了基于层次法的大学生就业质及系统的设计与实现, 通过建立评价指标体系, 以层次分析方法为基础构建了评价模型并完成了毕业生就业信息管理系统[4].

国外学者将人才培养看作就业质量, 分为高质量就业和低质量就业. 对高质量就业的定义为: 在具有挑战性和满意的工作环境中通过体力劳动或者脑力劳动获得生存所需的酬劳. 同时国外专家认为收入虽然重要, 但是收入的高低不足以体现就业质量. 虽然美国、欧盟等国家对于“工作满意度”、“工作质量”, “生活满意度”等与就业质量相关的方向进行研究分析, 但是通过文献资料可以看出其研究对象主要是针对某一群体的劳动者, 而对大学生的研究相对较少[5–8].

1 RF (随机森林) 算法

随机森林算法[9]是数据挖掘技术中一种自然的非线性建模工具, 通过集成多棵决策树 (Decision Tree, DT) 使模型有较好的稳定性[10]. 随机森林算法的本质是一种组合分类器, 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[11].

算法 1. RF 算法思想

Input: 训练集 D, 待测样本;

Output: 待测样本的类别或回归值;

Step 1. 采用 Bootstrap 抽样从训练集 D 中抽取 k 个子训练集, 子训练集的大小和 D 一致;

Step 2. 每个节点分裂之前随机选择特征生成特征子集;

Step 3. 建立 k 棵决策树;

Step 4. 对于待测样本, k 棵决策树得出 k 个结果;

Step 5. 对 k 个结果进行一票制投票或取平均值得到结果.

采用 bootstrap 重采样方法时, 使用 bagging 方法从原始训练集 D (样本总数为 N) 中有放回地抽取样本, 形成一个样本集, 因此, 存在一些未被抽取到的样本. 训练集中每个样本未能被抽取到的概率为:

$$p = (1 - 1/N)^N$$

当 N 趋向于无穷大时, p 约为 0.368, 可以得出训练集 D 中约有 37% 的样本不会被抽到这部分样本为袋外样本 (Out-Of-Bag, OOB), OOB 既可用于误差估计, 也可用于特征重要性分析.

随机森林算法流程如图 1.

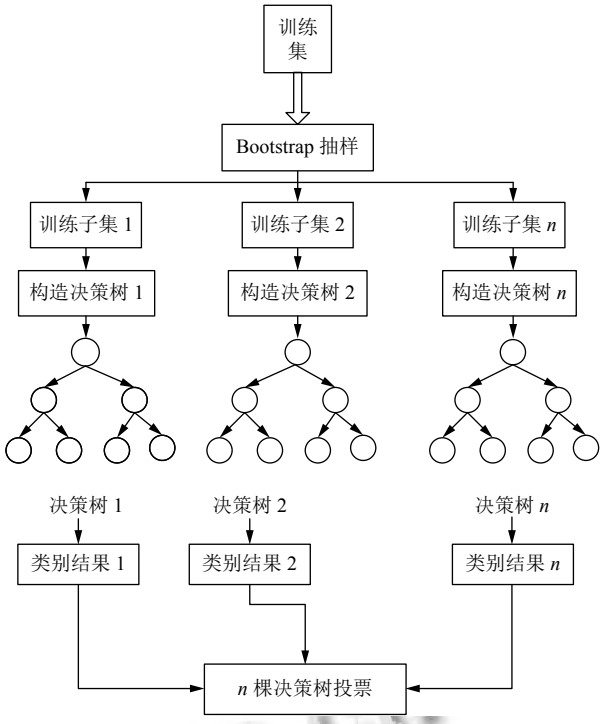

图 1 随机森林算法流程

随机森林算法用于人才培养评价具有很多优点, 首先它能够将弱分类器通过覆盖优化的手段进行综合, 使分类系统的整体能力得到了提升. 其次在生成决策树的过程中, 每一棵决策树都相互独立且同时生成, 提高了训练的效率. 另外在选择样本和构建决策树时随机选择的特征, 使该算法的抗噪能力大大提高.

当然随机森林算法也存在很多的不足之处. 随机森林算法在进行决策时采用平均投票的机制, 没有考虑到强弱分类器的差异, 其中弱分类器过多的参与决策过程会降低决策的准确率[12–15]. 另外由于采用了随机选择的方式选取样本特征, 因此在处理非平衡数据时无法消除样本数据所带来的影响[16–18].

2 RF 算法的改进

人才质量评价的过程是从学生在校表现的各项指

标中选择综合质量最好的人才, 可以看作是非平衡数据集的分类问题. 如果不考虑指标的平衡性直接对原始数据进行建模, 很难得到比较理想的模型, 可以需要通过训练数据来提升不平衡率, 主要实现方式为通过特征重要度度量来衡量数据指标的重要性, 以此为指标加权的标准, 本文选用 Ranking 方法对指标点的重要度进行衡量; 另一方面由于随机森林算法对分类器采用的是平均投票的机制, 这种投票机制使弱分类器对最终的评价结果产生影响, 本文采用 *F-measure* 算法对分类器进行加权, 降低弱分类器对于结果的影响.

特征样本的选取和决策树的投票过程是影响 RF 算法在人才培养评价中应用的主要方面, 本文就这两个方面提出了以下的改进方案.

1) 特征重要度度量算法 (WRF)

传统的学生评价指标的处理方法都是根据文献资料和专家意见对指标点进行选取和加权, 这种方法受到主观因素的影响较大, 没有考虑不同环境下指标影响力是不同的情况. 这种方法主要存在两个问题: 第一, 这种方法不仅效率低下而且也会由于认为因素影响最终的结果; 第二, 由于在标准随机森林算法中特征选择是完全随机的, 因此样本特征被选中的概率是一样的, 但实际上每一个特征的重要度是不同的, 也就是说在人才培养质量评价的过程中, 所涉及到的是非平衡数据.

为解决以上问题, 本文以每一个样本特征的重要度为依据, 选择重要度较高的选择样本特征, 降低弱分类器生成的可能性. 度量特征重要性的方法有很多, 本文选用随机森林排序算法 (Random Forest Ranking, RF Ranking) 计算特征重要度, 以下为该方法的主要步骤:

Step 1. 选取某一样本特征 X, 随机引入噪声数据并再次计算 OOB, 结果记为 $errOOB2$, 初始的 OOB 计算结果记为 $errOOB1$. 假设在随机森林中存在 N 棵决策树, 则特征 X 的重要度计算公式是:

$$I_X = \frac{\sum_j^N (errOOB2_j - errOOB1_j)}{N}$$

Step 2. 跟着 Step 2 得到的排好序的特征, 选择 75% 的特征, 在特征集合移除后面 20% 的特征.

Step 3. 重复上述两个步骤, 直到特征数目降为 M, 提前设定好的一个值. 最终得到 m 个最终特征集合.

2) *F-measure* 加权算法 (FRF)

传统随机森林方法在进行分类决策时, 采用的是平均多数投票法, 每一颗决策树输出自己的分类标签, 最终的结果为输出最多的类. 但是在分类过程中, 决策树的分类效果是不同的, 如果按照平均投票的方法, 每一个决策树都具有相同的投票权重, 就会导致效果好的分类器不能更好地发挥作用, 效果差的分类器对结果产生负面影响.

本文基于 *F-measure* 方法, 设计了一种新型的基决策树加权方法. *F-measure* 是 *Precision* 和 *Recall* 加权调和平均, 是 IR (信息检索) 领域的常用的一个评价标准, 常用于评价分类模型的好坏. 利用混淆矩阵计算分类器的召回率 *Recall* 和准确率 *ACC*:

$$Recall = \frac{TP}{TP + FN}$$

$$ACC = \frac{(TP + TN)}{TP + FP + FN + TN}$$

其中, *TP* 表示实际是高质量毕业生预测为高质量毕业生的人数, *TN* 代表的是实际是低质量毕业生预测为低质量毕业生的人数. *FP* 代表的是实际为低质量的毕业生预测为高质量的毕业生, *FN* 代表的是实际为高质量的毕业生预测为低质量的毕业生.

根据 *F-measure* 计算公式, 计算出组成随机森林分类器的每一颗决策树的 *F-measure* 值.

$$F - measure = \frac{2 \times recall \times precision}{recall + precision}$$

在上式中, *recall* 代表召回率, *precision* 代表准确率. 首先, 将验证集的数据输入到每一个决策树中, 然后每一个决策树对验证集中的每一个记录会有一个类别预测, 根据决策树预测的结果和真实的结果进行比对.

改进后的随机森林算法降低了平均投票机制所带来的影响, 降低了弱分类器对于结果的影响, 提高了算法的整体性能, 无论是在人才质量评价中还是在其他的应用中都可以应用.

改进后的随机森林算法流程图 2 所示.

3 应用研究

1) 数据来源与数据处理

本文的数据来源主要是青岛科技大学信息学院在国家工程专业认证过程中所收集的学生数据, 该数据由学院档案记录、问卷调查、综合测评成绩等多方面组成, 包含了 2008 年到 2017 年 2000 多名毕业生的详细数据, 每一条数据包含约 35 个字段, 共计 8 万条数

据. 根据人才培养质量评价的需求, 只需选择与评价内容密切相关的数据建立数据库即可. 最终只保留以下字段, 见表 1.

其中将 2008 到 2015 年共 8 年的样本数据作为原始训练集, 占总样本数的 80%, 2016 年和 2017 年两年的样本数据作为测试集.

2) 特征选取对于算法性能的影响

本文对 RF 算法做了两次改进, 为了验证两种改进都能对评价结果产生积极的影响, 本文对两种改进分别进行验证, 以证明两种改进各自的有效性. 为了验证特征选取对于算法性能的加强, 本文将不带有特征重要度加权的 RF 算法与带有特征重要度加权算法进行比较, 结果如表 2 所示.

从表 2 中可以看出, 在同一数据集中, 带有特征重要度加权的 RF 算法比原算法的准确率有了明显的提升, 在特征选取的过程中, 改进后的算法能够自动筛选出对评价结果有利的特征指标, 降低弱分类器的生成概率, 间接提高了评价模型的准确率.

3) *F-measure* 加权算法对算法性能的影响

为了验证 *F-measure* 加权算法对于算法性能的影响, 将普通投票机制的 RF 算法与带有 *F-measure* 加权投票机制的 RF 算法 (WRF) 在进行比较, 结果如表 3 所示.

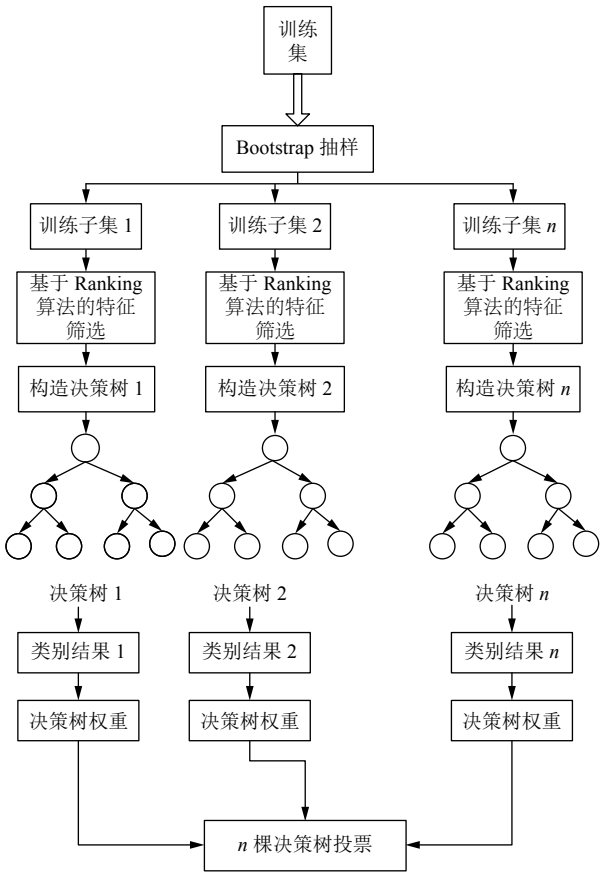

图 2 改进随机森林算法流程

表 1 处理后的数据所包含的字段信息

学号	姓名	性别	专业	专业技术能力
创新创业能力	知识学习能力	管理实践能力	综合发展能力	可持续发展能力
就业服务满意度	工作满意度	教学满意度	工作相关度	就业满意度

表 2 特征重要度方法对 RF 算法的性能影响

Algorithms	*ACC* (%)	*recall*	*F-measure*	ROC Area
RF	91.78	0.635	0.694	0.846
WRF	92.81	0.647	0.684	0.836

表 3 *F-measure* 加权算法对算法性能的影响

Algorithms	*ACC* (%)	*recall*	*F-measure*	ROC Area
RF	91.78	0.635	0.694	0.846
WRF	92.2	0.642	0.711	0.850

由表 3 可以看出, 通过加权投票机制改进的随机森林算法模型的准确率有了进一步的提升. 本文提出的基于 *F-measure* 加权投票机制的随机森林算法比传统的随机森林算法具有更高的性能.

4) FWRF 算法性能测试

为了证明 FWRF 算法在高校学生毕业质量评价方面的作用, 本文选取了几种经典的 RF 改进算法与 FWRF 算法进行横向比较, 实验数据为数据集中的所有类别. 经过最终的筛选, 本文选取了混合粒子群随机森林算法、混合遗传随机森林算法、混合鱼群随机森林算法以及原始随机森林算法为对比算法, 性能的评价指标主要是准确率. 利用 Python 语言在 PyCharm 开发平台上使用 scikit-learn 库实现了以上 4 种算法. 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 *F*1 值来对分类结果进行评估. 表 4 为 5 种方法的实验结果.

从表 4 中可以看出, 与几种经典的改机随机森林算法相比, 本文提出的改进随机森林算法在用于人才培养评价时, 在精确度和召回率上差异不大, 但是在准确率上有了一定程度的提高, 符合设计的要求.

表 4 5 种实验方法对比

Algorithms	*ACC* (%)	*recall*	*F-measure*	ROC Area
RF	91.78	0.635	0.694	0.846
混合离子	92.46	0.656	0.717	0.866
混合遗传	92.66	0.676	0.722	0.872
混合鱼群	92.55	0.676	0.722	0.872
FWRF	93.45	0.697	0.723	0.897

4 结束语

本文基于标准随机森林算法, 对其特征选择机制和决策树投票机制进行了改进, 使得抽样获得的决策树更加具有代表性. 结果显示, 改进后的模型在处理人才培养评价的问题时, 无论是相比于标准的随机模型还是经典的改进型随机森林算法, 其准确率有了一定程度的提高; 而且改进后的模型决策树的数量有所减少, 缩短了算法的运行时间, 在简化分析模型和提高模型准确度方面有一定的优势. 该算法能够解决高校毕业生的质量评价问题, 可以在高校的学生培养方面起到指导作用.

参考文献

1 李莉, 朱明珍, 洪云. 大学毕业生就业质量评价指标体系研究——以云南省高校为例. 昆明理工大学学报 (社会科学版), 2017, 17(1): 73–86.

2 宁东卫, 范春梅, 王莘. 大学生就业质量评价指标构建的探析. 云南农业大学学报 (社会科学), 2016, 10(4): 89–93.

3 宋俊秀, 谢德刚. 基于模糊综合评价法的大学生就业质量分析——以安徽省高校 2014-2017 年样本为例. 高校辅导员学刊, 2018, 10(3): 51–56.

4 韩天才. 基于层次分析法的大学生就业质量评价及系统的设计与实现[硕士学位论文]. 合肥: 安徽农业大学, 2018.

5 James G, Witten D, Hastie T, *et al*. An introduction to statistical learning with applications in R. New York: Springer, 2013. 3–26.

6 刘永垚, 第宝锋, 詹宇, 等. 基于随机森林模型的泥石流易发性评价——以汶川地震重灾区为例. 山地学报, 2018, 36(5): 765–773.

7 Brida JG, Lanzilotta B, Moreno L, *et al*. A non-linear approximation to the distribution of total expenditure distribution of cruise tourists in Uruguay. Tourism Management, 2018, 69: 62–68. [doi: 10.1016/j.tourman.2018.05.006]

8 Feng Y, Cui NB, Gong DZ, *et al*. Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling. Agricultural Water Management, 2017, 193: 163–173. [doi: 10.1016/j.agwat.2017.08.003]

9 周博翔, 李平, 李莲. 改进随机森林及其在人体姿态识别中的应用. 计算机工程与应用, 2015, 51(16): 86–92, 141. [doi: 10.3778/j.issn.1002-8331.1309-0162]

10 李欢, 熊梦莹, 聂斌, 等. 融合因子分析的随机森林研究. 计算机工程与应用, 2019, 55(23): 125–130. [doi: 10.3778/j.issn.1002-8331.1808-0266]

11 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. [doi: 10.15888/j.cnki.csa.006162]

12 刘勇, 兴艳云. 基于改进随机森林算法的文本分类研究与应用. 计算机系统应用, 2019, 28(5): 220–225.

13 王平, 单文英. 改进的随机森林算法在乳腺肿瘤诊断中的应用. 计算机应用与软件, 2016, 33(4): 252–257, 264. [doi: 10.3969/j.issn.1000-386x.2016.04.059]

14 李慧, 李正, 余堃. 一种基于综合不放回抽样的随机森林算法改进. 计算机工程与科学, 2015, 37(7): 1233–1238. [doi: 10.3969/j.issn.1007-130X.2015.07.002]

15 谢晓龙, 叶笑冬, 董亚明. 梯度提升随机森林模型及其在日前出清电价预测中的应用. 计算机应用与软件, 2018, 35(9): 327–333. [doi: 10.3969/j.issn.1000-386x.2018.09.058]

16 魏正韬, 杨有龙, 白婧. 基于非平衡数据的随机森林分类算法改进. 重庆大学学报, 2018, 41(4): 54–62.

17 孙悦, 袁健. 基于 Spark 的改进随机森林算法. 电子科技, 2019, 32(4): 60–63, 67.

18 王凯, 王菊香, 邢志娜, 等. 基于改进特征选择 RF 算法的红外光谱建模方法. 计算机应用研究, 2018, 35(10): 3000–3002. [doi: 10.3969/j.issn.1001-3695.2018.10.027]