计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(6):203−208 [doi: 10.15888/j.cnki.csa.006928]

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

基于 HMIGW 特征选择和 XGBoost 的毕业生就业预测方法①

李　琦[1,2], 孙　咏[2], 焦艳菲[3], 高　岑[2], 王美吉[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[3](沈阳高精数控智能技术股份有限公司, 沈阳 110168)
通讯作者: 李　琦, E-mail: hellodake1117@163.com

摘　要: 为了使高校的就业指导工作更具针对性, 可以有针对性地培养学生, 本文收集了毕业生的相关信息及其各自的就业情况, 构建了基于 HMIGW 特征选择和 XGBoost 的分类预测建模算法, 并将其应用于毕业生就业预测. 本文首先考虑到学生信息数据具有离散型和连续型混合的特点, 提出一种适应于就业预测的基于互信息和权重的混合 (Hybrid feature selection based on Mutual Information and Gain Weight, 以下简称 HMIGW) 特征选择算法, 该方法先对学生数据的特征做相关性估值, 然后采用前向特征添加后向递归删除策略进行特征选择, 最后基于选择后的最优特征子集数据用 XGBoost 预测模型进行训练与结果预测. 通过对比不同算法的结果, 本文采用的预测方法在准确率和时间等评价指标上有较好的表现, 对于毕业生培养就业指导具有积极作用.

关键词: 毕业生就业预测; 分类算法; 特征选择

引用格式: 李琦,孙咏,焦艳菲,高岑,王美吉.基于 HMIGW 特征选择和 XGBoost 的毕业生就业预测方法.计算机系统应用,2019,28(6):203–208. http://www.c-s-a.org.cn/1003-3254/6928.html

Graduates Employment Forecasting Method Based on HMIGW Feature Selection and XGBoost

LI Qi[1,2], SUN Yong[2], JIAO Yan-Fei[3], GAO Cen[2], WANG Mei-Ji[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computer Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[3](Shenyang Golding NC Technology Co. Ltd., Shenyang 110168, China)

Abstract: In order to provide more effective employment guidance work in colleges and universities, and train students in a more targeted manner, this study collects the relevant information of graduates and their employment situations, constructs a classification prediction modeling algorithm based on HMIGW feature selection and XGBoost, and applies it in graduates' employment forecasting. In consideration of the mixed discrete-continuous characteristics of the student information data, the study proposes an HMIGW feature selection algorithm suitable for employment prediction. This method firstly correlates the characteristics of student data, then adopts forward-increasing backward recursive deletion strategy to conduct feature selection. Finally, the XGBoost prediction model is used for training and result prediction based on the selected optimal feature subset data. By comparing the results of different algorithms, the prediction method adopted in this study has a better performance in evaluation indexes such as accuracy and time, and has a positive effect on employment guidance of graduates.

Key words: graduate employment forecast; classification algorithm; feature selection

① 收稿时间: 2018-12-07; 修改时间: 2018-12-25; 采用时间: 2019-01-08; csa 在线出版时间: 2019-05-25

随着高校的不断扩招, 毕业生就业形势愈加严重, 面对日益严重的就业压力, 很多高校都开展了就业指导活动, 但是大部分就业辅导工作都普遍存在着缺乏针对性和流于形式等问题, 并不能让真正需要帮助的学生受益. 很多高校都有着完善的学生学籍管理系统, 毕业生信息管理系统等各种学生系统, 存储了大量的学生数据, 但这些系统目前还只是用于存档查询, 如果可以充分利用这些系统包含的信息, 找到影响就业的主要因素, 就可以为就业指导提供更有针对性的指导工作.

关于毕业生就业问题, 为了有效提升就业质量, 国内外教育专家都从不同角度做过各种类型的分析统计. 例如, 文献[1]利用基于信息增益比的决策树方法来对就业情况进行分析, 抽取规则知识来分类预测. 在文献[2]中, 作者通过分析总结, 发现中医药院校毕业生的就业渠道狭窄, 就业偏向较为严重, 该文献中作者首先采用了 C4.5 决策树算法构造决策树, 然后通过随机森林建立多棵树来提高预测准确率, 但是其每棵树之间仍然是独立的, 导致模型偏差较大.

在预测问题方面, 很多学者都做了各种尝试, 预测用过的算法有以下几种, 包括文献[3]的神经网络模型, 文献[4]的自然邻居分类模型等. 文献[5,6]分析了影响毕业生的因素. 文献[7]介绍了大型数据集上建模的一些可供参考的思路方法. 上述文献大多采用单个全局优化模型, 单分类器模型性能有限, 存在泛化能力弱和容错性较差等问题. 本文提出采用集成学习模型, 即通过多个分类器的结果得到更为准确的预测结果. 集成学习模型一般预测准确性和模型稳定性都较高. 集成学习最常见的是文献[8]的随机森林模型以及梯度提升模型. 随机森林是采用 Bagging取样, 模型的方差较小, 但是偏差较大, 所以需要基分类器具有相对较高的准确性. Boosting 策略则可以降低模型偏差. 原理上讲, 即使采用了准确度相对较低的基学习器, 也可以通过逐步提升的方式, 提高整个模型的准确性. 而且 Boosting 策略可以看作串行化学习过程, 如果数据量比较大且属性比较复杂, 会出现数据不能全部加载, 模型计算的复杂度高, 预测结果精度不够等一系列问题, 所以考虑采用基于 Boosting 策略的 XGBoost 算法. XGBoost 支持在特征粒度上进行多线程并行, 有效提高模型运算效率, 实现上在目标函数里面引入了正则项, 进一步提高模型泛化能力, 文献[10]详细说明了 XGBoost 的优点.

针对以上问题, 本文提出一种基于特征选择和集成学习算法的预测建模方法, 其中, 特征选择步骤可以有效消除冗余特征, 得到最优特征子集, 进而提高模型准确率; 使用 XGBoost 能有效提高分类模型的容错性和泛化能力, 从而降低模型的误判率. 然后将此模型应用于毕业生就业预测, 对就业指导工作具有积极作用.

1 基于毕业生就业预测关键步骤算法介绍

1.1 特征选择算法 HMIGW

在文献[9]的基础上, 针对毕业生数据的特点, 以互信息和权重为基础, 综合过滤式和包裹式特征选择算法的优点, 本文提出一种基于互信息及权重的混合特征选择算法, 即 HMIGW 特征选择算法, 该方法包括过滤 (Filter) 和包裹 (Wrapper) 两个阶段, 分别如下:

(1) 针对冗余的无关特征进行过滤, 对于每个特征, 依次计算其信息度量, 求出相关性估值 I_x.

对于数据序列 $X=(x_1, x_2, \cdots, x_i, \cdots, x_m)$, 求熵公式如下:

$$H(X) = -\sum_{x_i \in X} p(x_i)\log(p(x_i)) \tag{1}$$

其中, $p(x_i)$ 表示 x_i 在 X 中的概率密度. 两个变量联合熵大小表示两个变量在一起的不确定性度量. 条件信息熵表示已知其中一个变量情况下求另一变量 $Y=(y_1, y_2, \cdots, y_i, \cdots, y_m)$ 情况下信息熵大小, 两者公式如下:

$$H(X,Y) = -\sum_{x_i \in X}\sum_{y_j \in Y} p\left(x_i, y_j\right)\log\left(p\left(x_i, y_j\right)\right) \tag{2}$$

$$H(Y|X) = -\sum_{x_i \in X}\sum_{y_j \in Y} p(x_i, y_j)\log(p(x_i, y_j)) \tag{3}$$

其中, 式 (2) 中 $p(x_i, y_i)$ 表示两个变量组合概率密度函数, 式 (3) 中 $p(x_i, y_i)$ 表示在已知 y_j 情况下 $p(x_i)$ 的条件概率密度. 由式 (1) 和式 (3) 可得式 (4) 表示如下:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{4}$$

熵大小表示变量之间的稳定性, 而互信息大小能够表示变量之间的相互依赖程度, 互信息定义如式 (5):

$$I(X;Y) = \sum_{x_i \in X}\sum_{y_j \in Y} p(x_i, y_j)\log\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{5}$$

其中, $I(X; Y)$ 表示 X, Y 两者之间的共享信息度量. $I(X; C)$ 越大说明 X, Y 相关性越强. 式 (5) 通过式 (1)、(2) 转换, 互信息可表示为熵的形式如式 (6):

$$I(X;Y)=H(X)+H(Y)-H(X,Y) \quad (6)$$

记 $I(X;Y)$ 为特征 X 的相关性估值 I_x.

(2) 采用前向特征添加后向递归删除策略进行特征选择

根据 (1) 中求得的相关性估值 I_x 对特征空间的特征降序排序, 使用前向添加策略对特征空间进行遍历, 每次都增加一个特征, 逐次增加相应的特征集合为 X_1, X_2, …, X_m(m 为特征子集所包含的特征数目), 然后通过 XGBoost 算法对当前特征集合进行分类精度计算, 求出分类精度为 a_i. 如果 $a_i<a_{i-1}$, 那么从特征集合 X 中去除特征 x_i, 依次循环往复直至结束.

上述策略的优点是: 通过先求出相关性估值 I_x, 然后基于 I_x 值, 再使用分类精度来二次对每个特征对预测结果的贡献权重进行评估, 这样就能有效降低特征波动性而且不牺牲预测精度, 根据评估结果删除权重较小的特征. 每次有特征被删除以后, 都会对特征集合进行重新遍历生成新的特征集合, 重复上述步骤, 就可以得到冗余最小、性能最优的特征集合.

与单纯向前、向后查找相比, 上述方法首先对特征进行降序排序, 并在其基础上以当前特征子集分类预测精度为评价指标, 递归去除掉冗余和分类精度较低的特征, 有效降低了特征波动性. 与单纯使用 Filter 或者 Wrapper 方法相比, HMIGW 算法在过程上先进行过滤然后根据过滤结果递归包裹, 保证了所选出的特征子集冗余最少性能最优.

算法 1. HMIGW 算法

输入: 数据集 D, 特征集合 $X=\{x_i|i=1,\cdots,v\}$, $r_{max}=0$, $X_{best}=\Phi$.
输出: 最优特征子集 X_{best}.

1) 分别计算特征 x_i 关于类别特征的 I_i, 若 $I_i=0$, 则删除特征 x_i, $X=X-\{x_i\}$;
2) 将上一步计算得到的 I_i 值记为综合评估值, 根据综合评估值 I_i 对特征进行降序排序;
3) 使用 XGBoost 算法综合评估, 对步骤 2) 降序排序后的特征子集采用前向特征添加策略, 即子集搜索策略遍历特征空间, 然后计算出算法在该特征子集 X_i 上的精确度 a_i, 其中 i 表示特征子集中元素的个数:
boolean flag=false;
for (a_i (i=1, …, v)) do
if (a_i <a_{i-1})
flag = true;
从集合中删除特征 x_i, 并记录删除特征 x_i 以后算法精度为 a_{tmp};
if(a_{max}<a_{tmp}) then
$a_{max}=a_{tmp}$, $x_{best}=X$;
end if;
break;
end if;
end for
util *flag*=false 达到终止条件

1.2 XGBoost 算法

集成学习的主要策略有 Bagging 与 Boosting, 两种策略都有各自的优缺点. 随机森林模型采用 Bagging 取样策略, 模型方差较小, 但是偏差较大, 因此要求基学习器具有相对较高的准确性. Boosting 策略则可以降低模型偏差. 原理上讲, 即使采用了准确度不高的基学习器, 也可以通过逐步提升来提高整个模型的准确性. XGBoost 算法是通过一组分类器的串行迭代计算实现更高精度的分类效果, 其基学习器是分类回归树 (CART), 在预测时将多个基学习器的预测结果综合考虑得出最终的结果. 假设有 K 个树, 树集成模型为:

$$\hat{y}_i=\sum_{k=1}^{K}f_k(x_i), \quad f_k\in F \quad (7)$$

其中, f_k 是函数空间 F 里的一个函数, F 是包含所有回归树的函数空间.

集成模型的参数有每棵树各自的结构以及叶的分数, 可抽象得使用函数 f_k 作为参数, $\Theta=\{f_1,f_2,\cdots,f_k\}$. 目标函数包含损失函数和正则化项, 损失函数评估每个真实类别 y_i 和诊断类别 $\hat{y}_i$ 的差异, 正则化也就是对于模型进行惩罚, 如果模型越复杂, 它的惩罚就越大.

$$Obj=L(\theta)+\Omega(\theta)=\sum_{i=1}^{n}l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \quad f_k\in F \quad (8)$$

式中, 损失函数为 $l(y_i,\hat{y}_i)=(y_i-\hat{y}_i)^2$, 正则化项 Ω 包含 L1 和 L2 正则, 单棵树的复杂性如下:

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\|\omega\|^2 \quad (9)$$

其中, T 是叶子节点数量, $\|\omega\|^2$ 表示 L_2 正则, ω 表示需要正则的参数, γ 和 λ 是系数.

正则化项可以使得模型结构和函数更为简单. 更具体地说, 第一项用来惩罚树的复杂结构 (叶子越少则 Ω 越小), 然而, 第二项惩罚项则惩罚单棵树, 使得单棵树不过重, 以防失去平衡的树去支配模型. 因此, 第二项会使得学习树的权重更平滑, 泛化能力更强.

对训练数据进行学习的时候, 每次都会在原有模型不变的基础上, 加上一个新的函数 f, 并观察当前目标函数, 如果新加入的函数会使目标函数的值比加入前更小, 那么就将该函数加入到模型中.

XGBoost 采用 Boosting 策略均匀取样, 同时 XGBoost

在迭代优化的时候使用了目标函数的泰勒展开的二阶近似因此基于 Boosting 的 XGBoost 精度更高[10]. XGBoost 在进行节点的分裂时, 支持各个特征利用 CPU 开多线程进行并行计算, 因此计算速度也更快.

2 实验分析

构建的预测模型在毕业生就业预测中实现. 就业预测建模可以分为 4 个步骤来实现, 如下图 1 所示, 首先是数据采集, 对数据进行预处理, 主要包括缺失值、异常值的处理、数据清洗规约等过程. 通过预处理成功构建数据集, 将数据集通过上述建模算法进行建模, 本文先采用了 HMIGW 特征选择算法选择出最优特征子集, 然后基于 XGBoost 算法进行预测, 最后进行结果分析, 验证本文所采用方法的有效性.

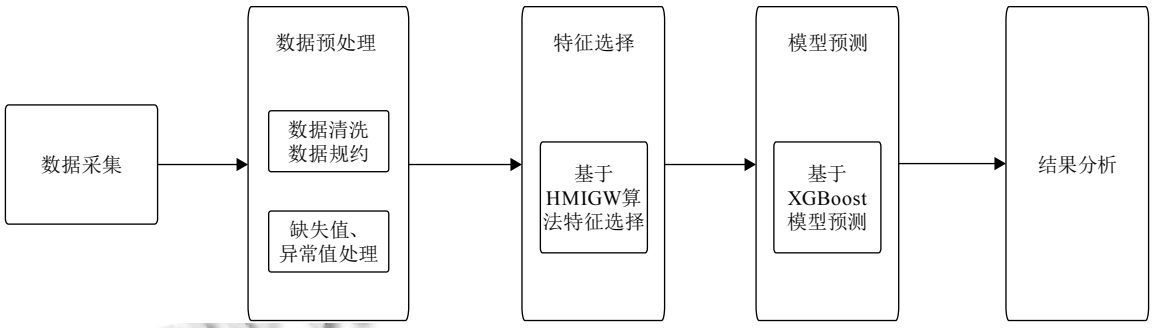

图 1 基于毕业生就业预测模型构建

2.1 数据采集

实验数据来源于实验室项目“沈阳航空航天大学实践教学管理系统”就业管理平台, 数据集包含了 2016 年到 2018 年三年的毕业生信息.

2.2 数据预处理

毕业生就业数据包含来自学籍管理系统以及毕业生就业管理系统等多个数据库来源的数据, 需要对影响就业的学生基本属性及就业单位属性进行筛选和归纳, 将学生的高考类别、入学成绩、学生地域、应往届、年龄、性别、综合素质评价、是否顺利毕业、企业单位地址、单位性质等信息进行数据清理转换和整合, 对缺失值进行删除插补等处理, 针对异常值进行填充或者删除.

2.3 特征选择

毕业生数据的信息量大而复杂, 特征维度较高, 而且各属性之间趋向离散, 且既有离散型特征也有连续型特征, 且冗余特征较多, 而在高维数据分析过程中, 当样本存在冗余特征时会大大增加问题分析复杂难度, 因此数据分析前从中剔除冗余特征尤为重要. 特征选择是指在保证特征集合分类性能的前提下, 通过一定方法从原始特征集合中选出具有代表性的特征子集, 从而将特征空间变为最小冗余、决定最优性能的过程. 根据依赖于模型与否, 特征选择方法有两种, 分别为过滤式 (Filter) 和包裹式 (Wrapper). 过滤式是根据数据本身的特性对当前特征子集进行评估, 独立于模型, 该方法一般运行效率较高, 但分类性能较差; 包裹式将模型分类精度同步于特征选择过程, 精度相对较高但是运行效率低下. 因此传统的特征选择方法不适用于这种情况.

根据本文提出的 HMIGW 算法, 分别计算每个特征和类别的相关性估值 I 如表 1.

表 1 各特征和类别相关性估值

特征属性	I	特征属性	I
1.获奖学金次数	1234.7420	10.计算机水平	354.6354
2.实习次数	1046.2240	11.大学英语等级	320.6521
3.综合成绩	856.2340	12.户口	207.2634
4.专业成绩	635.1523	13.托福/雅思/GMAT	175.2364
5.参加社团个数	596.2156	14.性别	9.6521
6.专业热门程度	537.5426	15.学校类别	8.0001
7.担任学生干部次数	527.1256	16.是否担任学生干部	7.0300
8.担任学生干部级别	482.1563	17.是否参加社团	7.0430
9.政治面貌	367.1253	18.民族	2.6530

根据 HMIGW 算法, 在选择最优特征子集的时候, 首先基于表 1 各特征的相关性估值可以看出每个特征和类别的相关程度, 再通过计算特征之间的相关性估值去除了和特征 7 冗余的特征 16 以及和特征 5 冗余的特征 17, 然后采用前向特征添加后向递归删除策略即子集搜索策略去除特征 15 和特征 18, 最后选出最优特征子集如下: 获奖学金次数、实习次数、综合成

绩、专业成绩、参加社团个数、专业热门程度、担任学生干部次数、担任学生干部级别、政治面貌、计算机水平、大学英语等级、户口、托福/雅思/GMAT、性别.

2.4 模型预测

对于就业情况需要进行量化, 本文基于该应用建立了一个二级分类模型, 首先根据是否就业分为就业、未就业两类, 然后对类别为就业的数据集进行二级分类, 该类别根据就业单位进行划分, 包括国企、外企和私企三类. 对类别为未就业的数据集进行二级分类, 类别包括升学以及未找到工作两类.

验证由两部分组成: 验证 HMIGW 算法的有效性和验证本文模型的有效性.

(1) 验证 HMIGW 算法的有效性

本实验中选用 CFS (Correlation-based Feature Selection) 和 WFS (Wrapper Feature Selection) 算法与本文提出的 HMIGW 算法进行性能对比. 使用 XGBoost 集成学习算法对上述三种特征选择算法选出的特征子集进行分类预测, 随后采用交叉验证方法计算分类模型的分类精度. 对实验结果进行分析比较, 验证本文特征选择方法的有效性.

(2) 验证本文模型的有效性

在数据集上首先使用 HMIGW 特征选择算法选出最优特征子集, 然后分别使用 XGBoost 算法和随机森林算法进行分类预测, 同样使用交叉验证计算模型的精确度. 对比实验结果, 验证本文采用 XGBoost 算法建模的有效性.

2.5 评价指标

通过计算评价指标来评价预测模型的质量是实验部分的重要环节, 本文选取的评价指标有准确率 (*Precision*)、召回率 (*Recall*)、F_1 值. 评价指标定义如下:

$$precision = \frac{T_p}{T_p + N_p} \tag{10}$$

$$recall = \frac{T_p}{p} \tag{11}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{12}$$

其中, p 为阳性样本总数, T_p 为正确预测的阳性样本数量, N_p 为错误预测的阳性样本数量, F_1 分数是精度和召回的调和平均值, 达到最优值为 1, 最差值为 0.

此外, 如果样本数据量比较大, 计算速度也应该作为一个评价指标. 本实验使用 python 的 time.clock() 来记录模型计算时间.

2.6 实验结果分析

(1) 实验中, CFS 和 WFS 算法分别采用最佳优先搜索 (Best First, BF) 和贪心单步搜索 (Greedy Stepwise, GS) 算法进行特征子集选择; 而 HMIGW 则采用前向特征添加后向递归删除策略进行特征选择.

表 2 列出了在数据集上上述 3 种算法来特征选择, 然后将得到的特征子集使用 XGBoost 算法来训练, 并且进行 5 次交叉验证后结果比较. 其中, F_{num} 表示特征的个数.

表 2 基于不同特征选择算法构建毕业生预测模型对比

特征选择算法	F_{num}	*Precision*	*Recall*	F_1 值
CFS	9	0.9468±0.0332	0.945	0.9459
WFS	35	0.9654±0.0203	0.963	0.9642
HMIGW	14	0.9732±0.0200	0.970	0.9716

注: 表中±后的数据分别表示 5 次测试结果的方差

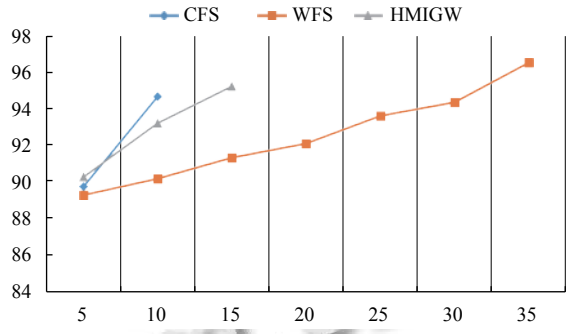

图 2 不同特征选择算法分类精度对比

从表 2 中可以看出, 本文提出的 FMIGW 特征选择方法分类精度为 97.324%, 召回率为 97.0%, 均优于其他特征选择方法. CFS 在降维方面表现较优, 但分类精度较低; WFS 在特征降维方面性能低于其他方法, 然而与本文提出的 HMIGW 算法相比综合性能较差且时间代价较大.

在冗余特征判断准确性方面, 根据表 1, HMIGW 算法筛选出了个别的相关特征, 例如属性“是否担任学生干部”和属性“是否参加社团”, 他们分别与“担任学生干部次数”和“参加社团个数”两个属性相关, 而 CFS 和 WFS 方法均没有去掉这两个相关属性, 但是 HMIGW 可以成功地选择出这些数据属性, 对冗余特征判断的准确性好于另外两种方法.

在对特征信息完整度影响方面, 根据表 1 的相关性估值排名以及可以看出, 15 以后的特征相关性估值

I 远远低于 175, 在 HMIGW 算法的前向特征添加后向递归删除过程中得出添加对应特征得出的精确度 a_i 小于等于不添加其的精确度, 且冗余信息不影响有效特征信息完整度, 所以使用 HMIGW 算法可以准确判断特征集合中的冗余信息, 且得出的特征子集能够最大程度保留有效特征信息完整度.

通过对比表明, 本文 HMIGW 特征选择方法能够选出特征维度相对较低, 分类性能最优的特征子集.

(2) 用测试集对 XGBoost 算法建立的预测模型进行性能评价. 为了比较更加直观, 添加了随机森林算法作为比较. 随机森林是通过建立多棵决策树, 每棵树单独对样本进行分类, 最终分类结果由每棵树各自的分类结果通过投票决定.

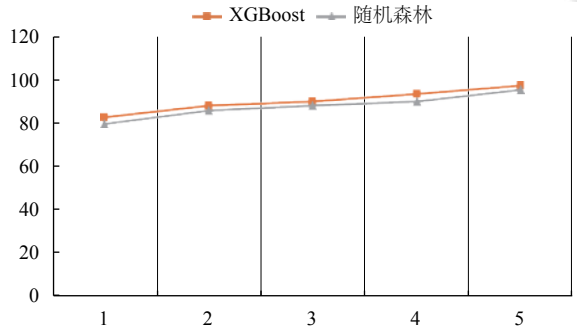

图 3 基于 HMIGW 的 XGBoost 和随机森林性能对比

结合表 3 中 2 个模型下的各项性能指标, 可以看出, XGBoost 模型在预测准确率, 召回率, F1 值方面均优于随机森林模型. 在模型训练时间方面, XGBoost 也优于随机森林.

表 3 XGBoost 和随机森林模型预测结果对比

模型名称	*Precision*	*Recall*	F_1 值	训练时间 (s)
XGBoost	0.9734±0.0252	0.973	0.9148	0.030
随机森林	0.9552±0.0303	0.892	0.8778	0.038

注: 表中±前面和后面的数据分别表示 5 次测试结果的方差

(3) 综合以上, 本文采用的毕业生就业预测方法通过先求出特征相关性估值 I, 在其基础上再使用分类精度来二次对每个特征对预测结果的贡献权重进行评估得出最优特征子集, 这样做有效降低征的波动性, 且不会降低预测精度. 然后采用 XGBoost 算法对得出的最优特征子集数据集进行分类预测, 通过串行迭代计算实现更高精度的分类效果, 在预测方面达到了 97.34% 的准确率, 在进行节点的分裂时利用 CPU 开多线程进行并行计算, 有效提升计算速度, 训练时间控制在了 0.03s, 比文中讨论的别的预测方法具有较大的性能提升.

3 总结

本文以毕业生就业预测为应用研究背景, 针对学生数据情况, 提出了一种适用于该应用的特征选择算法 HMIGW, 并采用 XGBoost 算法进行分类预测, 利用其并行运算速度快、精度高、灵活性强、鲁棒性好的特点, 构建了毕业生就业预测模型. 实验研究结果表明, 该模型能够根据毕业生的相关信息, 预测毕业生就业情况以及就业类型. 且该模型在准确率、召回率、F_1 值和模型训练时间指标上表示较优, 可以为毕业生就业提供强有力的指导, 具有非常积极的意义.

参考文献

1 孙晓璇, 杨家娥, 李雅峰. 基于决策树 ID3 算法的高职生就业预测分析. 电脑编程技巧与维护, 2015, (2): 15–16, 35. [doi: 10.3969/j.issn.1006-4052.2015.02.005]

2 唐燕, 王苹. 基于 C4.5 和随机森林算法的中医药院校毕业生就业预测应用研究. 中国医药导报, 2017, 14(24): 166–169.

3 吴振磊, 刘孝赵. 一种基于 BP 神经网络的就业分析预测模型. 轻工科技, 2016, 32(9): 70–71, 104.

4 朱庆生, 高璇. 应用自然邻居分类算法的大学生就业预测模型. 计算机系统应用, 2017, 26(8): 190–194. [doi: 10.15888/j.cnki.csa.005906]

5 Burnasheva S, Zhuravleva I, Kustov T, *et al*. Creation of the effective system for students' and graduates' employment promotion at the university: ETU "LETI" experience. Proceedings of 2016 IEEE V Forum Strategic Partnership of Universities and Enterprises of Hi-Tech Branches. St. Petersburg, Russia. 2016. 72–73.

6 Baskakova DY, Belash OY, Shestopalov MY. Graduates' employment: Expectations and reality. Proceedings of 2017 IEEE VI Forum Strategic Partnership of Universities and Enterprises of Hi-Tech Branches. St. Petersburg, Russia. 2017. 128–131.

7 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]

8 Chen TQ, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA. 2016. 785–794.

9 陈宇韶, 唐振军, 罗扬, 等. 皮尔森优化结合 XGBoost 算法的股价预测研究. 信息技术, 2018, (9): 84–89.

10 毛莺池, 曹海, 平萍, 等. 基于最大联合条件互信息的特征选择. 计算机应用, 2019, 39(3):734-741.