

基于随机森林算法的高校学生就业岗位预测研究

李红梅

(广东白云学院 大数据与计算机学院软件工程系 广东 广州 510450)

【摘要】机器学习是人工智能的一个分支,被广泛应用于各种领域。本文针对高校毕业生就业岗位难以准确预测等一系列问题,提出了利用随机森林算法对毕业生就业岗位进行预测,并通过实验结果验证了算法模型的有效性,这为高校就业形势分析提供了可靠的参考依据,也为高校教育改革和发展提供了一定的参考。

【关键词】机器学习;随机森林;数据处理;就业岗位

引言

《中国教育报》显示,高校毕业生就业形势复杂严峻,需多方联动,拓宽大学生就业渠道,高校毕业生的就业问题受到各界的广泛关注。而高校大学生就业情况数据是动态且复杂的,如果用传统方法简单对就业数据进行统计分析,很可能无法探究到问题的本质。机器学习是能通过经验的不断累积自动进行改进的人工智能算法,非常适合用于对复杂就业数据的分析研究,得出的结论更加合理、有参考性^[1-2]。

本文通过我校近几年计算机类专业的就业数据,运用机器学习中的随机森林算法对学生就业岗位进行预测,通过数据处理、建立模型、实验分析可较精准的进行学生就业岗位预测。这为高校教育领域以及高校就业形势提供重要的参考依据,也为高校专业设置调整人才培养模式和教学的改革,提供重要依据。

一、随机森林算法在高校学生就业岗位预测中的研究

(一) 随机森林算法

随机森林^[3]是一种典型的监督式学习算法,通过集成学习的思想将多棵树集成的一种算法,它的基本单元是决策树,模型结构如图1所示。

该算法的核心是将大量的弱学习器的预测结果融合,即组合多个弱监督模型得到一个更好更全面的强监督模型,即便某一个弱分类器得到了错误的预测,其它弱分类器也可以将错误纠正回来,提高预测的准确率。

该算法的优点:

1. 采用了集成算法,本身精度比大多数单个算法要好,即准确性高;
2. 两个随机性引入,使得随机森林不容易陷入过拟合(样本随机,特征随机);
3. 两个随机性引入,使得随机森林具有一定的抗噪声能力,对比其它算法具有一定的优势;
4. 能够处理很高的维度(feature很多)的数据,并且不同做特征选择,对数据集的适应能力强:即能处理离散型数据,也能处理连续型数据;
5. 在训练过程中,能够检测到feature间的互相影响,且可以得出feature的重要性,具有一定的参考意义;因此本文选取随机森林算法对高校学生就业岗位进行预测。

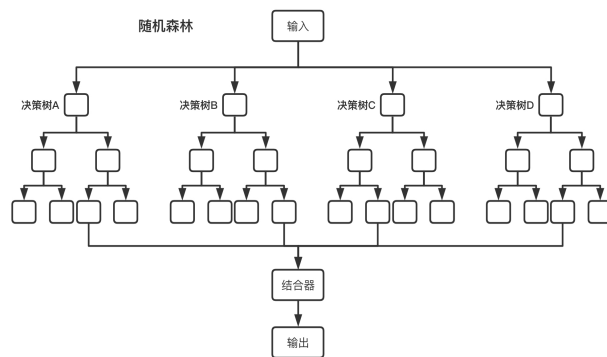


图1 随机森林算法模型结构图

(二) 数据来源

本研究以我校计算机相关专业毕业生就业数据和成绩数据为研究对象,并进行初步人工抽取处理,形成包含如下特征的数据:学号、性别、专业、政

基金项目:广东白云学院项目(2022BYKY15)。

作者简介:李红梅(1978-),女,湖南湘潭人,副教授,硕士研究生,研究方向:演化计算和机器学习。

表 1 计算机相关专业毕业生数据

专业	学号	性别	政治面貌	平均成绩	毕业时间	就业岗位
计算机科学与技术	17231503131	男	党员	85	2021	软件开发
计算机科学与技术	17331503261	男	群众	66	2022	招投标与成本控制
软件工程	17231543071	男	群众	73	2021	软件测试工程师
数据科学与大数据技术	17231547021	男	群众	80	2021	软件工程师
软件工程	17231543200	女	群众	78	2022	测试工程师
计算机科学与技术	17231503060	男	群众	78	2022	测试工程师
软件工程	17231543206	男	党员	75	2022	软件测试工程师
信息管理与信息系统	17231538048	男	群众	73	2022	游戏体验师

治面貌、平均成绩、毕业时间、就业岗位 7 个属性列，其中就业岗位是本研究算法模型的标签列，数据样本共 2735 条（见表 1）。

（三）数据预处理

数据预处理是对原始数据进行一系列工程处理，将其提炼为特征，作为输入供算法和模型使用。它的方法主要包括去除唯一属性、处理缺失值、属性编码、数据标准化正则化、特征选择、主成分分析等。

1. 文本数据数值化

文本数据需要特殊处理，然后才能开始将其用于预测建模。本研究中的性别、专业、政治面貌这些特征数据和就业岗位标签数据都要做相应的数值化处理。

性别：将性别特征数据的“男”设为 0，“女”设为 1。

专业：我校计算机相关专业主要有：软件工程、计算机科学与技术、大数据技术、信息管理与信息系统，分别处理成：0、1、2、3 数值。

政治面貌：将政治面貌特征数据的“党员”设为 0，“群众”设为 1。

就业岗位：

（1）软件工程专业核心岗位：Java 软件开发、前端开发、移动开发、软件测试，设为数值 0。

（2）计算机科学与技术专业核心岗位：网络管理、系统运维、网络工程，设为数值 1。

（3）大数据技术专业核心岗位：数据处理、数据分析、数据挖掘，设为数值 2。

（4）信息管理与信息系统专业核心岗位：信息收集、组织和管理或信息系统的分析、设计、实施和维护，设为数值 3。

（5）所有专业非核心岗位设为数值 4。

（6）未就业设为数值 5。

2. 数据标准化

为使不同维度的特征在数值上更具比较性，并提高分类器的准确性和收敛速度，本研究利用 sklearn 库中的 preprocessing 模块，采用标准差标准化方法对特征数据进行标准化处理^[4]，实现代码如下：

```
x_datas = preprocessing.scale(x_datas)
```

3. 数据集拆分

为了做模型的评估，需要将数据集拆分成训练集和测试集。本研究利用 sklearn model_selection 模块中的 train_test_split 方法将数据样本按照 8：2 的比例拆分为训练集数据和测试集数据，实现代码如下：x_train, x_test, y_train, y_test=train_test_split(x_datas, y_datas, test_size=0.2)

4. 构建岗位预测模型

利用 scikit-learn 库的 RandomForestClassifier 分类器构建岗位预测模型。代码及核心参数如下：RandomForestClassifier(n_estimators=100) n_estimators 这个参数代表随机森林中决策树的数量，本研究利用交叉验证法来找到此参数的最优值。实现时应用了 sklearn.model_selection 模块中的 cross_val_score 方法，代码如下：

```
a=[]
for i in range(1, 200):
    RFC = RandomForestClassifier(n_estimators=i, min_samples_split=6, min_samples_leaf=1)
    RFCscores = cross_val_score(RFC, x_train, y_train, cv=6)
    a.append(RFCscores.mean())
```

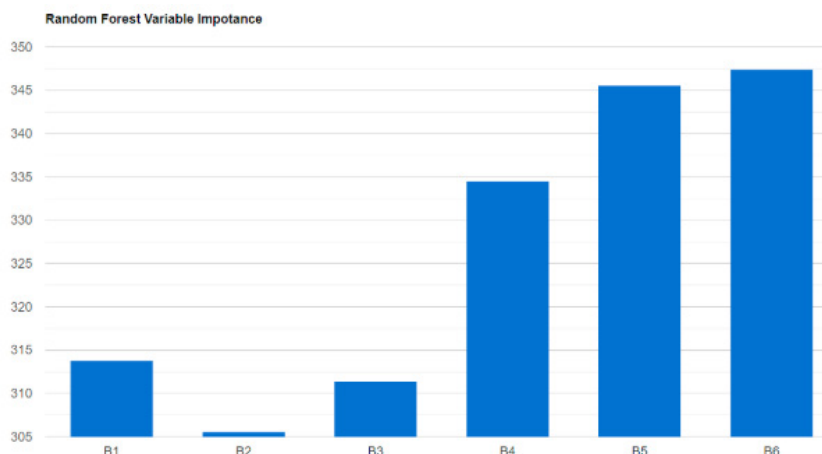


图2 特征重要性可视化

```
print(a.index(max(a))+1)
```

模型训练时将“学号”“性别”“政治面貌”“专业”“平均成绩”“毕业时间”6个样本特征作为输入值，将“就业岗位”作为最终的分类结果。模型训练完后，利用模型对象的 feature_importances_ 属性，实现数据特征重要性可视化。

横坐标表示样本数据的特征名称，其中：B1 表示性别，B2 表示学号，B3 表示政治面貌，B4 表示毕业时间，B5 表示平均成绩，B6 表示专业；纵坐标表示特征的得分值，分值越高，属性越重要。从图中可知，特征得分大小排序结果为：专业>平均成绩>毕业时间>性别>政治面貌>学号。

5. 模型评价

为验证利用随机森林算法建立的高校学生就业岗位预测模型的实用性，以测试集数据为基础，运用召回率、准确率和 F 值和精确度 4 个评价指标对决策树算法的性能进行评价。召回率越高，算法的灵敏度越好，准确率越高，算法准确度越高；F 值越大，算法的整体性能越好；精确度越高，算法的精确度越好^[5]。

算法	召回率	准确率	F值	精确度
随机森林	93.29%	84.12%	0.9103	0.8705

6. 在校生实际使用评价

从我校计算机相关专业学生随机选择 65 名在校学生对本系统进行评价打分，根据实际体验情况作出“没有帮助”“有帮助但帮助不大”“很有帮助”3 种情况的选择，结果如图 3 所示。

二、结论

针对高校毕业生就业岗位难以准确预测等一系列问题，本研究运用随机森林算法实现了对高校毕

业生就业岗位的预测，同时挖掘也出影响学生就业岗位的关键因素专业，其次是平均成绩和毕业时间。其中毕业时间属于客观因素，学生不能改变，但是学习成绩是可以努力提高。对于核心因素专业，学校可以宏观调控，限制或扩大专业招生人数。本研究也存在不足之处：没有考虑企业的招聘偏好，这是下一步的研究工作。

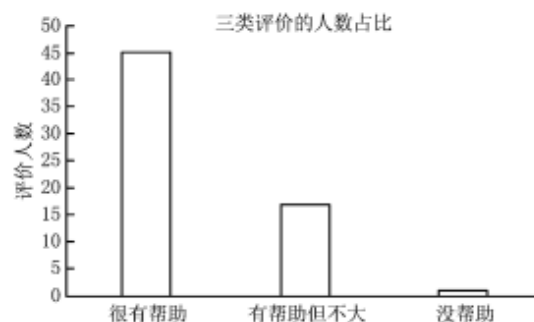


图3 在校生实际使用评价统计结果

参考文献：

- [1]李静,徐路路.基于机器学习算法的研究热点趋势预测模型对比与分析—BP神经网络、支持向量机与LSTM模型[J].现代情报,2019,39(4):23-33.
- [2]牟风云,杨猛,林孝松,等.基于机器学习算法模型的巫山县洪水灾害研究[J].中山大学学报(自然科学版),2020,59(1):105-113.
- [3]吕红燕,冯倩.随机森林算法研究综述[J].河北省科学院学报,2019,36(3):37-41.
- [4]赵伟杰,陈海军,张毅.浅谈大数据下的机器学习算法[J].科技创新导报,2019,16(24):250-251.
- [5]浮盼盼,司琪,王鑫赛.机器学习算法的超参数优化：理论与实践[J].电脑编程技巧与维护,2020(12):116-117+146.