文章编号：2095-2163(2021)10-0067-06 中图分类号：TP183 文献标志码：A

# 基于随机森林算法的职位薪资预测

彭义春，张 捷，覃左仕
（玉林师范学院 计算机科学与工程学院，广西 玉林 537000）

**摘 要：** 采用爬虫技术获取当前主流招聘网站的招聘信息，数据经清洗和标准化后，采用 Python 语言分别应用线性回归、支持向量机（SVM）、决策树、随机森林等算法完成训练，根据平均绝对误差、均方误差和 $R$ 平方误差对预测结果评价效果得出随机森林算法效果最好，最终选用随机森林算法来构建职位薪资的预测模型。随机森林算法对职位薪酬预测结果能给招聘者发布招聘信息和求职者查询适合自己的岗位提供一个合理、科学的参考依据，可极大提高招聘求职的成功率。

**关键词：** 职位薪资预测；随机森林；参数调优；网格搜索法；熵值法

## Prediction of position salary based on random forest algorithm

PENG Yichun, ZHANG Jie, QIN Zuoshi
(School of Computer Science and Engineering, Yulin Normal University, Yulin Guangxi 537000, China)

**【Abstract】** The crawler technology is used to obtain the recruitment information of the current mainstream recruitment websites. After the data is cleaned and standardized, the Python language is used to apply linear regression, support vector machine (SVM), decision tree, random forest and other algorithms to complete the training, according to the average absolute error, mean square error and $R$ square error, the random forest algorithm is the best, and the random forest algorithm is used to construct the prediction model of job salary. The algorithm of random forest can provide a reasonable and scientific reference for job seekers to publish recruitment information and search for suitable positions, which can greatly improve the success rate of recruitment.

**【Key words】** position salary forecast; random forest; parameter tuning; grid search method; entropy method

## 0 引 言

随着互联网的迅速发展和普及，网络招聘基本取代传统招聘形式，已成为招聘者和求职者的首选方式。网络招聘具有信息量大、不受时空限制、招聘成本低、便捷高效等优点；但也因信息量的激增和信息难以核实，带来了信息爆炸、信息过剩、信息失真、薪水不透明等问题[1]。因此，如何从琳琅满目的数据中提取有价值的信息成为关键。若能通过某些方式了解到本行业中类似岗位的薪资范围，就能对岗位的薪资是否合理有个准确的判断，以及对未知薪资的岗位有一个预判。薪资市场是一个较为复杂的非线性动力学系统，同时薪资数据中包含大量的关于职位本身的描述数据。随着人工智能算法的发展，灰色理论、朴素贝叶斯、doc2vec/word2vec[2]、回归模型[3]、移动平均模型[4]、神经网络、协同过滤[5]、深度学习[6]、K 最近邻[7]、决策树、支持向量机等机器学习算法已被广泛应用于预测领域。随机森林算法具有处理高维样本[8]、预测精度高、学习速度快、调节参数少及不产生过拟合[9]等优点，已被广泛应用于回归和分类问题。近年来，也有学者将随机森林算法用于薪资预测领域，如，文献[10]中采用随机森林模型预测和分析了云南省物流人才岗位薪资；文献[11]提出了一种基于随机森林模型对求职者和企业互惠就业推荐算法；文献[12]采用随机森林模型对农信金融企业员工工资进行预测等等。本文在对比分析 SVM、决策树、随机森林等机器学习算法的基础上，提出了基于随机森林的薪资预测模型。经验证结果表明，此模型在薪资预测中产生的误差较小、效果较好，预测结果既能更好地帮助求职者选择更适合自己的职位和判断职位薪资的合理性，也能帮助招聘者制定合理的职位薪资，招聘到合适的人才。

**基金项目：** 玉林师范学院 2021 年度高层次人才启动项目（G2021ZK05）；2018 年广西自然科学基金项目（2018JJA170050）；2020 年玉林师范学院大学生创新创业训练计划区级项目（202010606126）。

**作者简介：** 彭义春（1974-），男，博士，副教授，主要研究方向：特征选择算法、人工智能、GIS 与 RS；张 捷（1974-），男，博士，教授，主要研究方向：脑网络分析、单细胞数据分析、智能计算。

**通讯作者：** 彭义春 Email：515012487@qq.com

**收稿日期：** 2021-08-01

## 1 基于随机森林的薪资预测模型构建

随机森林指的是利用多棵树对样本进行训练并预测的一种分类器,属于一种集成算法,在分类、回归和聚类等方面应用效果较好。本课题的薪资预测是回归预测,需要用分类回归树(CART)作为基本单元进行构建森林,依次循环训练每一棵 CART,每棵 CART 的训练样本都是从原始训练集中进行可放回抽样(Bootstrap)得到。CART 较容易过拟合,但因随机森林经过 Bootstrap 和 Aggregate(聚集)这 2 个过程(又被称为袋装 Bagging)解决了过拟合问题,同时也因随机性而增强了模型的泛化(Variance)能力。根据国内主流招聘网站招聘公告的职位描述,薪资预测关联度较高的特征包括职位语言、职位类别、所属城市、学历、工作经验、公司规模和所属网站等。对多个特征值需要收集所有特征的最佳切分点进行对比,选出最好的特征划分点,采用平均或投票的方式对所有决策树做集成操作。

假设共有 $M$ 个样本,$n$ 个特征的数据集,最多须构建 $t$ 棵决策树,每棵决策树的特征个数为 $K$,则随机森林算法实现过程如下:

**步骤 1** 在训练数据集所在的输入样本中,对每个样本的每个特征进行遍历,递归地将每个区域划分为 2 个子区域。利用公式(1)计算 $n$ 个特征及其相应切分点下的残差平方和,找到一对$(j,s)$,且满足:分别最小化左子树和右子树的残差平方和,并在此基础上再次最小化二者之和[13]。式(1)的数学表示如下:

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{1}$$

其中,$R_1$、$R_2$ 代表被划分的 2 个子集(回归树为二叉树只有 2 个子集),$c_1$、$c_2$ 分别表示 $R_1$ 和 $R_2$ 样本的均值,$j$ 代表工作城市、职位名称、职位类型等样本特征,$s$ 表示划分点,$y_i$ 表示样本目标变量的真实值。

**步骤 2** 用选定的$(j,s)$来划分区域,并决定相应的输出值,求样本均值公式[14]为:

$$\hat{c}_m=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i,\ x\in R_m\quad m=1,2 \tag{2}$$

其中,$R_1(j,s)=\{x\mid x^{(j)}\leqslant s\}$,$R_2(j,s)=\{x\mid x^{(j)}\leqslant s\}$。

**步骤 3** 继续对 2 个子区域调用步骤 1、2,直到不能继续划分为止。

**步骤 4** 将输入样本划分为 $M$ 个区域,即:$R_1$,$R_2$,…,$R_M$ 生成决策树[15]。其公式如下:

$$f(x)=\sum_{m=1}^{M}\hat{c}_mI(x\in R_m) \tag{3}$$

其中,$c$ 代表对应区域的平均值;$I$ 代表是否符合条件,符合为 1,否则为 0。

**步骤 5** 采用有放回抽样,从原数据集中经过 $M$ 次抽样,获得有 $M$ 个样本的数据集(可能有重复样本)。从 $n$ 个特征里,采用无放回抽样原则,去除 $K$ 个特征作为输入特征。对新数据集重复上述过程 $t$ 次,构建 $t$ 棵决策树[16]。

**步骤 6** 对生成的 $t$ 棵决策树采用求平均的方法,最终得到一个随机森林模型。

随机森林构建流程如图 1 所示。

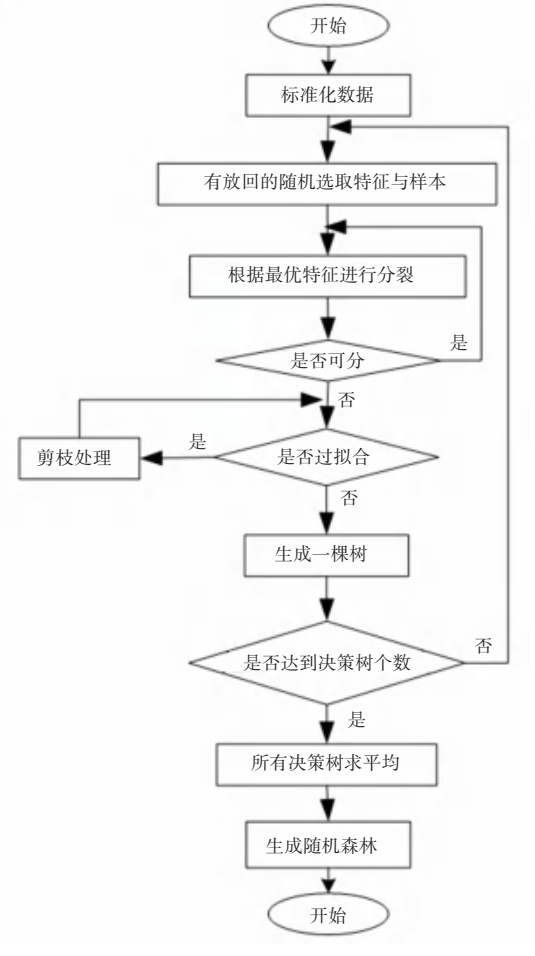



**图 1 随机森林构建流程图**

**Fig. 1 Flow chart of random forest construction**

## 2 样本数据采集、预处理与存储

采用 Python 爬虫技术爬取了猎聘网、拉勾网、Boss 直聘和前程无忧四大主流招聘网站 2021 年 1

月~5 月的 IT 行业招聘信息。为了提高薪资预测过程的高效性和结果的精准性，采用 Pandas、Numpy 模块，对爬取到的数据信息进行离群值检测、缺失值处理、异常值处理、字段分割、标签编码、重复值剔除等数据预处理。为了将文本类型的数据转换成数字型数据，首先遍历去重后的每一特征值的字符并给其赋一序列号，然后再次遍历此文件，把序列号映射回原来没有去重的列，相同文本则对应同一个序列号。最后，将处理过的数据以 csv 文件格式保存并存储到 MySQL 数据库中。

## 3　对比实验

### 3.1　不同格式数据训练结果对比

通过数据清洗后，将数据集划分为标签和特征值，并按照比例划分训练集和测试集。先运用 sklearn 模块构建决策树、线性回归、SVM、随机森林四个经典算法的预测模型，并观察模型在训练集的准确度（*accuracy*）表现。再分别对数据进行标准化处理和数据平滑处理，获取各算法模型的准确度表现，见表 1。

**表 1　4 种模型对不同格式数据训练得分对比**

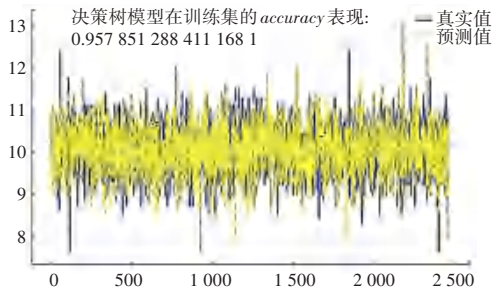
**Tab. 1　Comparison of training scores of four models on different format data**

| 模型 | 非标准化数据 | 标准化数据 | 平滑处理数据 |
|---|---|---|---|
| 线性回归 | 0.080 | 0.080 | 0.216 |
| SVM | −0.036 | −0.028 | 0.890 |
| 决策树 | 0.907 | 0.907 | 0.958 |
| 随机森林 | 0.850 | 0.849 | 0.898 |

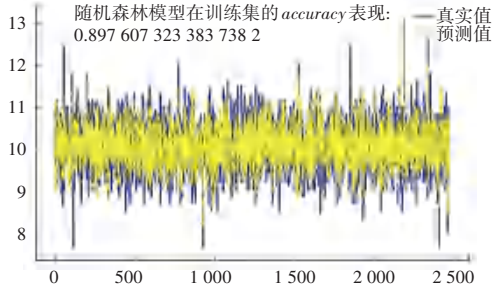
由表 1 中数据可见，数据标准化处理对于线性回归、决策树和随机森林模型的结果基本没什么影响，只有 SVM 略微提高，但依旧是负数；但对预测值数据平滑处理后，4 个算法模型得分都有明显上升，尤其是 SVM 算法，从负数提升到 0.89，而线性回归算法的得分仅有 0.216，说明线性回归算法并不适用于本课题的场景。

### 3.2　不同模型预测准确率对比

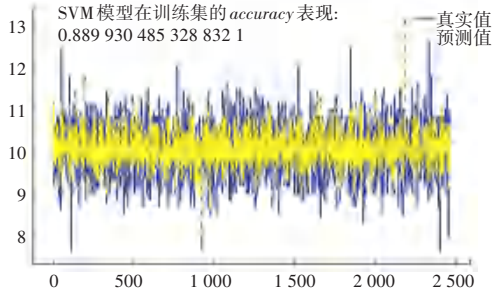
对比真实值与预测值的误差，可用来评估模型预测结果的准确率。图 2(a)~(c)分别为决策树、随机森林和 SVM 回归结果可视化图。对图 2 分析可知，决策树的准确率最高(0.957 85)，除了部分预测值和真实值相差较大外，绝大多数与实际偏差不大，其次是随机森林(准确率为 0.897 60)，最后为 SVM(准确率为 0.889 93)。




(a) 决策树回归结果




(b) 随机森林回归结果




(c) SVM 回归结果

**图 2　3 种算法的回归结果比较**

**Fig. 2　Comparison of regression results for the three algorithms**

### 3.3　不同模型 $R^2$ 值对比

预测结果的准确性并不能完全判定一个算法模型效果的好坏，还需要观察模型的拟合优度和泛化能力。$R^2$ 值是最常用的回归模型拟合程度的指标，其值的计算方法如下。

设 $y$ 为待拟合数据，$y$ 的均值为 $\bar{y}$，拟合函数计算结果为 $\hat{y}$，则：

(1)总体平方和 $SST$：

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4}$$

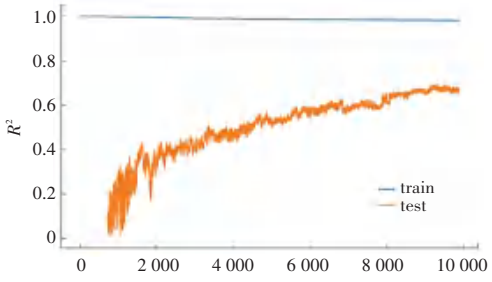
(2)残差平方和 $SSE$：

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$

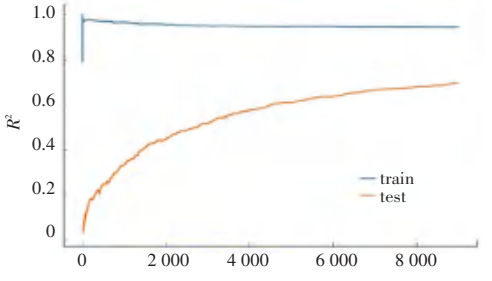
(3)拟合度公式 $R^2$：

$$R^2 = 1 - \frac{SSE}{SST} \tag{6}$$

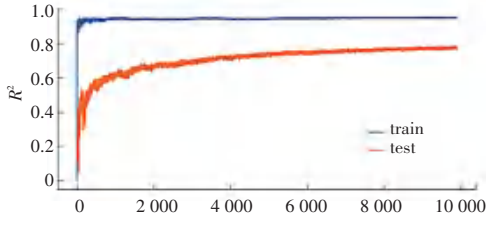
在 Python 中,通过记录训练集、测试集在训练和测试过程的 $R^2$ 值,并绘制如图 3 所示的曲线图。图 3(a)~(c)分别为决策树、SVM 和随机森林的训练集与测试集 $R^2$ 值图。




(a) 决策树训练集与测试集 $R^2$ 值图




(b) SVM 训练集与测试集 $R^2$ 值图




(c) 随机森林训练集与测试集 $R^2$ 值图

**图 3 3 种算法的 $R^2$ 结果比较**

**Fig. 3 Comparison of *R* square for the three algorithms**

经对比分析可见,决策树在训练集表现得很好,但在测试集表现较差,过拟合问题最严重;SVM 的拟合效果较好,但仍存在过拟合问题;随机森林的拟合效果最好,但 $R^2$ 值不高,必须通过参数调优来提高拟合度。

综合上述预测准确率和拟合度对比分析结果得出:决策树预测模型虽然预测的准确率最高,但拟合效果不佳,模型在预测新的数据集时准确率会大大下降,得到的预测结果不准确;SVM 预测模型拟合效果一般,准确率最低,且模型响应时间长,运用到实际应用中,用户体验会大打折扣;随机森林的准确率较高,特别是拟合效果最佳。因此,本研究最终采用随机森林算法构建职位薪资预测模型。

## 4 基于随机森林算法的职位预测实验

### 4.1 随机森林算法参数调优

随机森林是一种机器学习算法,算法参数的设置不仅影响模型的预测准确率,而且影响模型的训练效果的好坏。因此,模型构建之前对参数调优很有必要。其参数择优包括框架的参数择优和决策树的参数择优。本文采用改进的网格搜索(GridSearch)法[17]来完成参数调优。具体步骤如下:

(1)确定决策树个数 *n_estimators* 和划分时考虑的最大特征数 *max_features* 范围。先设定步长(即权重缩减系数 $\nu$,取值范围为(0,1]),在 *n_estimators* 和 *max_features* 坐标系上建立二维网格。网格节点就是相应的 *n_estimators* 和 *max_features* 的参数对。

(2)对网格节点上的每一组参数构建随机森林,并利用 OOB 数据估计残差平方均值。

(3)选择误差最小参数 *n_estimators* 和 *max_features*。若误差或者步长满足要求,则输出最优参数和残差平方均值,否则缩小步长。重复上述步骤,继续搜索。

在 Python 中,通过 *GridSearchCV*( ) 方法并使用十折交叉验证法求得模型的最佳参数组合,先增大 *n_estimators* 以提高模型拟合能力。这里当 *n_estimators* =110,拟合能力再无明显提升时,则再按照步长为 1 增大 *max_features* 来提高每个子模型的拟合能力,进一步提高模型的拟合能力。当 *max_features* = 5 时对应的拟合优度最大,残差平方均值最小。见表 2。

**表 2 随机森林参数表**

**Tab. 2 Parameter table of random forest**

| 参数 | 描述 | 最佳取值 |
|---|---|---|
| *n_estimators* | 决策树个数 | 110 |
| *oob_score* | 是否采用袋外样本来评估模型的好坏 | True |
| *bootstrap* | 是否对样本集进行有放回抽样来构建树 | True |
| *max_depth* | 决策树的最大深度 | 29 |
| *min_samples_split* | 内部节点再划分所需最小样本数 | 2 |
| *min_samples_leaf* | 叶子节点最少样本数 | 1 |
| *max_features* | RF 划分时考虑的最大特征数 | 5 |

表 1 中,前三行为框架参数,后四行为决策树参数。

将得到的最佳参数组合代入算法模型中，求出模型的平均绝对误差、均方误差、$R^2$ 值和袋外样本得分。职位薪资预测模型参数调优前后的模型预测效果对比，见表 3 。

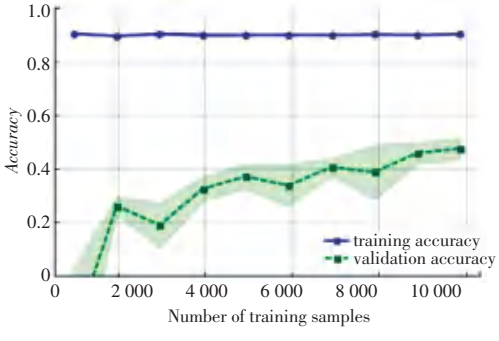
**表 3　参数调优前后评估指标对比**

**Tab. 3　Comparison of evaluation indicators before and after parameter tuning**

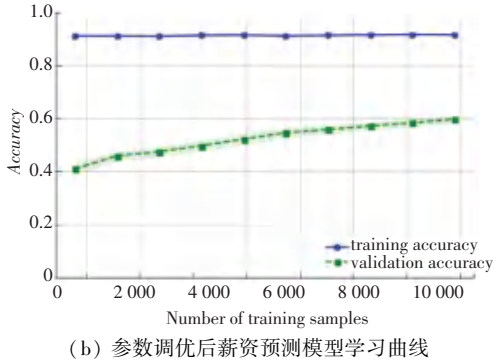
| 评估指标 | 参数调优前（默认参数） | 参数调优后 |
|---|---|---|
| 平均绝对误差（*MAE*） | 0.263 | 0.111 |
| 均方误差（*MSE*） | 0.138 | 0.027 |
| *R* 平方值（$R^2$） | 0.596 | 0.921 |
| 袋外样本得分（*oob_score*） | −2.343 | 0.651 |

从表 3 可明显得出，参数调优后模型的平均绝对误差、均方误差均有所下降，模型准确率更高。$R^2$ 值从 0.596 上升到 0.921，模型拟合效果较好；袋外样本得分由原先的−2.343 提升到 0.651，模型泛化能力显著增强。对比结果说明，参数调优后的职位薪资预测模型是十分有效的模型。

通过绘制参数调优前后模型的学习曲线，观察模型具体的拟合程度，如图 4 所示。参数调优前，职位薪资预测模型误差较大，过拟合程度较为严重；参数调优后，职位薪资预测模型学习曲线收敛，误差减小，过拟合程度明显下降。




（a）默认参数薪资预测模型学习曲线




（b）参数调优后薪资预测模型学习曲线

**图 4　参数调优前后模型学习曲线**

**Fig. 4　Learning curve of model before and after parameter tuning**

## 4.2　基于熵值法的特征重要性评估

职位薪资预测模型是以职位名称（*jobName*）、职位类别（*jobType*）、工作城市（*jobCity*）、学历（*jobEdu*）和工作经验（*jobExper*）为特征构建。为了构建预测模型，首先须确定各特征在模型中的权重。鉴于特征之间的相关性以及对薪资非线性影响的特点，采用熵值法确定各个特征在模型中的权重[18]。步骤如下：

（1）设数据有 $n$ 行记录，$m$ 个特征列，则数据可用一个 $n \times m$ 的矩阵 $\boldsymbol{A}$ 表示：

$$\boldsymbol{A} = [X_1, \cdots, X_m];$$

（2）数据归一化处理：

$$x_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{7}$$

（3）计算第 $j$ 项指标下第 $i$ 条记录所占比重：

$$P_{ij} = \frac{x_{ij}}{\sum_{1}^{n} x_{ij}} \quad j = 1, \cdots, m \tag{8}$$

（4）计算第 $j$ 项指标的熵值：

$$e_j = -k * \sum_{1}^{n} P_{ij} * \log(P_{ij}), k = \frac{1}{\ln(n)} \tag{9}$$

（5）计算第 $j$ 项指标的差异系数：

$$g_j = 1 - e_j \tag{10}$$

（6）计算第 $j$ 项指标的权重：

$$\omega_j = \frac{g_j}{\sum_{1}^{m} g_j} \tag{11}$$

在 Python 中，使用 Pandas 和 Numpy 库求出各特征的权重，见表 4。

**表 4　各特征在薪资中的重要性和权重**

**Tab. 4　The importance and weight of each feature in salary**

| 特征变量 | 重要性 | 权重 | 特征变量 | 重要性 | 权重 |
|---|---|---|---|---|---|
| *jobCity* | 0.917 055 | 0.422 464 | *jobExper* | 0.973 198 | 0.136 506 |
| *jobName* | 0.966 245 | 0.171 924 | *jobType* | 0.978 225 | 0.110 905 |
| *jobEdu* | 0.968 939 | 0.158 202 | | | |

由此可见，这 5 个特征的重要性都达 90%以上。将这 5 个影响因素指标作为随机森林模型的最优输入特征变量，最后做预测的特征中权重性排在最高的为工作城市，其次岗位类别，学历、工作经验和职业类型的重要程度基本相同。因此，工作城市和岗位是职位薪资高低的重要影响因素。

# 5　结束语

在对比分析线性回归、SVM、随机森林、决策树

等几种经典的机器学习算法后,因随机森林具有精度高、稳定性好、学习速度快等优势,故而选择随机森林构建职位薪资预测模型,并以 IT 行业为例,对职位薪资进行了模拟预测分析。结果表明,随机森林模型能合理有效地预测薪资,可作为中短期职位薪资预测的新途径。

然而,爬取的招聘信息时间间隔较短,数量有限,变量的类型上也有出入;另外薪资高低也受到外界客观因素的影响等。因此,随机森林模型对这些因素数据的预测效果不太理想,有待进一步深入研究。

## 参考文献


[1] 王洪艳,谢东洋. 浅议网络招聘存在的问题及对策[J]. 知识经济,2016(8):94-95.

[2] 潘博,张青川,于重重,等. Doc2vec 在薪水预测中的应用研究[J]. 计算机应用研究, 2018,35(1):155-157.

[3] 李媛. 多元线性回归在平均工资预测中的应用研究[J]. 信息通信,2018(1):31-33.

[4] 王宁宁. 基于 ARIMA 模型职工平均工资的分析与预测[J]. 智能计算机与应用, 2020,10(5):240-243,247.

[5] 张浩宇. 基于文本相似度与协同过滤的岗位薪资预测[D]. 长沙:中南财经政法大学,2018.

[6] 谷承维. 基于深度学习的垂直行业职位薪水分析与预测[D]. 北京:北京邮电大学,2018.

[7] ZHANG Junyu, CHENG Jinyong. Study of employment salary forecast using KNN algorithm[C]// Proceedings of the 2019 International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019). Wuhan: Atlantis Press, 2019,5: 175-179.

[8] 曹泽涛,方子东,姚瑾,等. 基于随机森林的黄土地貌分类研究[J]. 地球信息科学学报,2020,22(3): 452-463.

[9] 吴奉亮,霍源,高佳南. 基于随机森林回归的煤矿瓦斯涌出量预测方法[J]. 工矿自动化,2021,47(8):102-107.

[10] 宋倩楠. 云南省物流人才岗位薪资影响因素分析[D]. 昆明:云南大学, 2019.

[11] 高境辰,丁乐,王琦. 基于随机森林模型的互惠就业推荐算法[J]. 电子元器件与信息技术, 2021,5(3):133-135.

[12] 余顺坤,宋宇晴. GRA-RF 组合算法在农信金融企业工资要素优选及测算中的应用[J/OL]. 中国管理科学:1-11[2021-10-25].https://doi.org/10.16381/j.cnki.issn1003-207x.2021.0108.

[13] 甄亿位,郝敏,陆宝宏,等. 基于随机森林的中长期降水量预测模型研究[J]. 水电能源科学, 2015,33(6):6-10.

[14] 黄晗,孙堃,刘达. 基于随机森林的电力系统小时负荷预测研究[J]. 智慧电力,2018,46(5):8-14.

[15] 张蓓蓓,胡敏. 基于网格搜索改进随机森林的顾客满意度预测[J]. 北京信息科技大学学报(自然科学版),2021,36(4):50-53+58.

[16] 刘兴,王艳,纪志成. 基于随机森林的风电功率短期预测方法[J]. 系统仿真学报,2021,33(11):2606-2614.

[17] 温博文,董文瀚,解武杰,等. 基于改进网格搜索算法的随机森林参数优化[J]. 计算机工程与应用, 2018,54(10):154-157.

[18] 刘玉琪,胡庆武,程钢,等. 基于灰色关联分析与熵值权重的避难所适宜性评价[J]. 武汉理工大学学报(信息与管理工程版),2017,39(6):669-673,678.


(上接第 66 页)


[11] WAND M, KOUTNIK J, SCHMIDHUBER J. Lipreading with long short-term memory[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE press, 2016:6115-6119.

[12] CAMPBELL J P, REYNOLDS D A. Corpora for the evaluation of speaker recognition systems [C]//1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258). Phoenix, AZ, USA: IEEE press, 1999:829-832.