一种基于人工神经网络的高校学生就业预测模型探

李 欣 周 丽
成都锦城学院 计算机与软件学院 四川成都 611731

摘 要：近年来高校应届毕业生人数持续走高，有效合理的缓解就业压力，为应届毕业生提供科学专业的就业指导显得至关重要。本文结合高校学生在校的期间的学情数据、素质教育数据、在线慕课学习行为数据、基础信息数据等方面构造维度特征筛选关键因素，进而利用人工神经网络算法发现隐藏在历史就业数据中的规则和知识，为科学指导学生就业提供了理论依据。

关键词：就业预测；特征筛选；人工神经网络

一、引言

随着我国高等教育的深度发展，2021年应届毕业生数量突破了900万，应届毕业生庞大基数和较高的薪资期望值极大的增加了应届生就业的难度。就业是民生之本，是国家发展的基石[1]。尤其是应用型大学，稳定的就业更是学校生存和发展的生命线。随着高校数字化、信息化教育的不断深入，越来越多的教育工作者尝试将大数据人工智能等技术引入到学生就业的指导领域。如何有效的挖掘毕业生海量就业数据，如何将就业数据和学生在校的个性化学习数据、长板素质教育数据结合，如何利用数据挖掘技术原理探索关键因素和预测应届高校毕业生就业能力，并个性化的定制学生的学习路路径和培养方案，是目前各个高校研究的热门课题。

本文以成都锦城学院计算机与软件学院的学生就业及培养数据模型为基础，从现有智慧校园信息化系统中，提炼出素质教育维度数据、学科成绩维度数据、在线学习维度数据、日常学习习惯维度数据、基础信息维度数据等四个方面的特征，使用随机森林算法筛选关键因素，利用人工神经网络算法构建预测模型，为应届毕业生就业提供科学的指导，为学校就业管理服务工作提供了有利的数据支撑和参考依据。

二、构建预测模型

1.模型框架

基于人工神经网络的高校学生就业预测模型整体框架可分为5个部分，即业务数据抽取、数据探索处理、建立机器学习识别模型、模型效果评价、模型部署上线五个部分。预测模型为离线模型，整合数据源可依据往届毕业生就业情况和在校数据，进行数据的探索和处理，数据源按照比例进行分割训练集和测试集，训练集构建识别模型，测试集用于效果评估，利用混淆矩阵、准确率、ROC曲线等方法，对就业模型进行最终评估，符合评估标准后，模型予以上线，进而作为就业指导依据。（见图1）

2.数据源获取

就业预测模型的数据来源主要来源于智慧校园信息化系统、超星慕课线上学习轨迹记录和高校学生管理机构如学生课记录的参与活动的综合素质评分等三个方面。其中学科成绩数据和基础信息数据主要来自于智慧校园信息化系统，学科成绩维度包括学生在校四年内的通识类公共课程、专业课、方向课、实验课期末综合成绩和平时成绩等，学生基础信息数据包括学生的入学成绩、统考方式、学生层次、自然属性、四六级英语成绩等，这部分数据比较客观，是学生四年真实学情数据；在线学习情况数据，主要来源与学校超星慕课的学习行为加工，可从学生的原始学习日志数据中提炼出学习偏好特征，如线上学习时间偏好、学习方式偏好、学习态度、学习能力等维度特征；最后是学生综合素质能力维度，

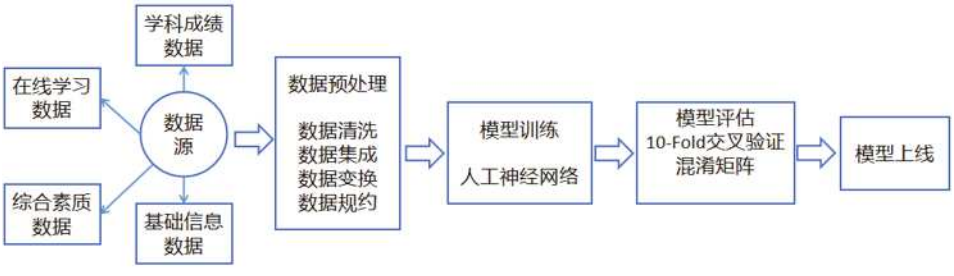

图1 高校学生就业预测模型整体框架

这部分数据主要来自与学生的直属管理机构如院系学生科老师或者辅导员的汇总记录，如综合能力、文体活动、科研能力、创新创业、沟通交流等维度，这部分数据是学生素质教育的成果体现，属于主观数据。

3.数据预处理

数据的预处理主要包括四个方面数据清洗、数据集成、数据变换、数据规约[2]。数据清洗方面，按照数据源的类型可以分别处理，离散类别型的数据主要通过该维度特征的众数来填充，如素质教育综合能力维度数据，属性取值为忧、良、中、一般、不及格等，部分样本维度缺失可用该属性众数填充；连续型数值数据，如线上学习时长，单位为小时，对于部分缺失的数据可采用连续性属性特征的中位数、平均值填充，也可以根据数据的规律，采用拉格朗日插值法或者样条插值法进行；当然有些样本属性特征缺失明显，对这种低质量的数据样本没有必要填充，在样本总量充分的前提下，可进行删除。数据集成方面，由于本模型的数据源主要来自信息化系统、线上超星软件系统、线下素质教育纸质版记录，数据来源较为分散，需要通过应届毕业班的学号ID主键，采用内连接的方式对各个结构化数据进行合并，数据集成过程重点关注消除实体冗余和属性冗余。数据变换方面，本模型主要是属性的构造和标准化。属性构造属于特征工程的领域，在素质教育数据方面，除了原有维度以外，可根据评分分布和相关性等，构造课外活动活跃度、课外活动偏好度、课外活动质量等综合特征；在线上学习数据方面，可深度挖掘学生的线上学习行为，按照授课学期构造学习能力、学习态度、学习方式偏好及统计维度特征。在学科成绩数据方面，可探索每学期成绩的综合分数，连续7个学期的变化趋势、斜率等动态信息。数据规约方面，在前面的数据预处理环节，模型尽可能的发散特征维度，但是该特征对预测目标变量的识别帮助是否明显，特别是学科成绩方面，大学4年学过的课程有几十门且课程之间也有一定的相关性，因此对这方面的维度数据可采用主成分分析PCA方法，提炼综合性的关键特征，即在保留原始信息90%以上的前提下，提取主轴特征，提高模型训练效率。

由于文本数据源的维度来源比较多，因此我们期望从原始的特征中筛选出能够识别影响学生就业的关键因素，进而利用有监督学习算法人工神经网络进行训练构建模型。在属性规约的过程中，我们主要采用随机森林进行特征筛选，将特征工程后的原始数据输入RandomForest模型，首先计算维度特征在构建森林中每棵树时的重要性，这里的重要性衡量主要包括分类的纯度、精度两个方面即基尼系数和袋外数据OOB错误率。其次考虑森林中决策树的规模计算重要性的集中趋势度衡量。最后将所有特征按照重要性降序排列。

4.模型训练

人工神经网络（ANN）起源于生物神经系统的研究[2]，由一组互相链接的节点和有向链构成，是一种具有自我学习能力的人工智能信息处理系统。结合本模型数据特征，采用BP单隐层前馈神经网络求解模型参数。

训练集$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，其中$x_i\in R^d$，$y_i\in R^m$，即输入数据有d个属性维度，这里的d维属性包括（学习成绩综合成分1，...，学习成绩综合成分i，学习趋势斜率，综合能力，...，沟通交流，课外活动活跃度，偏好度，课外活动质量，学习能力，学习态度，学习偏好，...，高考入学成绩，英语CET级别，...，统招方式等），v_{ih}（$i\in\{1,\cdots,d\},h\in\{1,\cdots,q\}$）为连接输入层维度特征和隐藏层神经元节点的权重，算法训练时，系统随机初始化权值向量。b_h为隐藏层神经元，h的取值范围为{1，...，q}，即隐藏层有q个神经元节点，隐藏层节点输入为$\alpha_h=\sum_{i=1}^{d}v_{ih}x_i$，经过Sigmoid激活函数得到中间层输出。$\omega_{hj}(h\in\{1,\cdots,q\},j\in\{1,\cdots,l\},)$为连接隐藏层和输出层的神经元的权重，同隐藏层节点输入构造类似，输出层节点的输入为$\beta_j=\sum_{h=1}^{q}\omega_{hj}b_h$经过激活函数得到神经网络的输出$\tilde{y}=\left(\tilde{y}_1,\tilde{y}_2,\cdots,\tilde{y}_l\right)$，即通过隐藏层转换后得到了l维数值向量，神经网络拓扑结构如下所示。

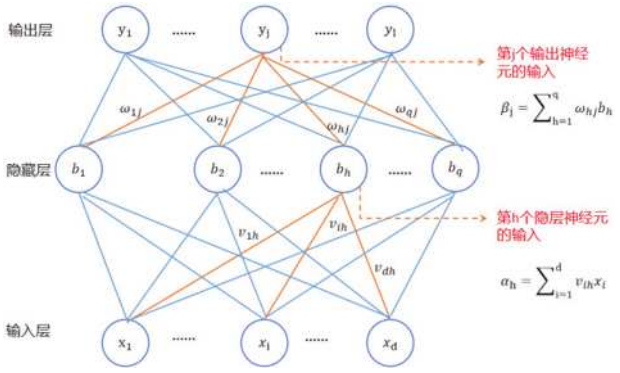

图2　预测毕业生就业神经网络拓扑结构

其中（$y-\tilde{y}$）为误差项，所有训练样本的误差平方和为$E(w)=1/2\sum_{j=1}^{l}\left(\tilde{y}_j-y_j\right)^2$。

构造出误差平方和后，我们的优化目标是使得误差

平方和最小，进而实现最小化损失函数，来求解神经网络权重系数参数。通常反向传播的BP算法是求解该问题的有效手段，该方法每次迭代分为两个步骤，向前阶段和向后阶段，即输入输出信息的向前传导和误差的反向传播。

5.模型评估

本文构建的预测模型是二分类目标变量识别，即每一个特征样本的类标签只有两个取值，即就业和未就业，二元化类别特征后，取值为0或者1。此处对于模型的评估可以从两个方面入手，一方面是模型识别算法适用性和泛化能力的评估，另一方面是模型有效识别能力的评估。针对算法适用性和泛化能力的评估这里可以采用K折交叉验证的方法，即将完整的数据集均分为互不相交的K个子集，模型预测时，每个子集都有参与评估的机会，循环K轮，每轮使用K-1个子集训练，余集评估测试。K折交叉验证避免了追求高准确率而在训练集上产生过拟合的情况，进而提升了模型的泛化能力。

针对模型有效识别能力的评估，对于二分类的有监督识别算法，我们先计算混淆矩阵，进而求解模型的精确率、准确率和召回率。同时也可绘制模型的ROC曲线，求解曲面积AUC值获得模型评估数字指标。

三、结语

本文以热点问题高校毕业生就业为目标变量构建预测，利用学生成绩维度、线上学习行为维度、素质教育维度、基础信息维度等数据构建属性特征，经过数据预处理探索，清洗数据质量、构造属性特征、通过PCA算法提取成绩数据主成分，通过随机森林算法筛选有效识别特征，最终通过单隐层BP人工神经网络训练模型参数，构造预测识别算法模型，在通过了10-Fold交叉验证的泛化能力评估和混淆矩阵等效果评估后，可上线使用。该模型对学生就业具有较好的预期指导作用提供了科学的理论依据，同时模型也存在优化空间，如目标变量的精细化分类，由二分类优化为多分类，如考研、就业、实习、未就业等，通过目标变量的多元化，提升就业预测模型的业务精度。

参考文献：

[1]李存岭等.切实抓好高校毕业生就业这项民生工程[J].山东教育（高教），2020（4）：9-11.

[2] Pang-Ning Tan.数据挖掘导论[M].人民邮电出版社，2013：150-151.