

结合相似度的朴素贝叶斯自训练算法的就业预测模型

宗慧敏 邢亚璇 刘微 黄江枫

河北化工医药职业技术学院

DOI:10.12238/jief.v6i10.10554

[摘要] 实现高校毕业生更加准确的就业预测对于促进高校毕业生更高质量就业具有重要的战略性意义。本文提出了结合相似度的朴素贝叶斯自训练分类方法应用于河北省高校毕业生的就业预测。通过对比其他就业预测算法,本文提出的方法稳定性好,并且具有更高的预测准确率,能够更好地解决当前河北省高校毕业生就业预测中存在的一些问题,对于河北省高校毕业生的就业指导与管理工作具有指导意义。

[关键词] 自训练; 朴素贝叶斯; 就业预测; 欧氏距离; 余弦相似度

中图分类号: F241.4 **文献标识码:** A

Employment Prediction Model Based on Naive Bayes Self-Training Algorithm Combined with Similarity

Huimin Zong Yaxuan Xing Wei Liu Jiangfeng Huang

Hebei Chemical & Pharmaceutical College

[Abstract] Achieving more accurate employment prediction for college graduates is of great strategic significance for promoting their higher-quality employment. This paper proposes the application of the Naive Bayes self-training classification method combined with similarity to the employment prediction of college graduates in Hebei Province. Compared with other employment prediction algorithms, the method proposed in this paper has good stability and higher prediction accuracy, and can better solve some problems existing in the current employment prediction of college graduates in Hebei Province. And it also has guiding significance for the employment guidance and management work of college graduates in Hebei Province.

[Key words] Self-training; Naive Bayes; Employment Prediction; Euclidean Distance; Cosine Similarity

引言

随着我国高等教育的不断发展与普及,高校招生规模逐年扩大,学生人数大幅度增加,高校毕业生的就业形势更加严峻。党的二十大报告中特别指出“实施就业优先战略”,而高校毕业生作为就业的主体,促进高校毕业生更加充分更高质量就业具有重要的战略性意义。因此,越来越多的学者开始研究高校毕业生就业的影响因素与预测,并取得了一定的成果。金延军^[1]使用Apriori算法探索了受教育程度与就业的关系。王亚如^[2]应用决策树算法找出了研究生就业的影响因素,对于就业指导工作具有实质性帮助。吴亚娟^[3]运用聚类分析方法对大学生就业满意度统计与预测进行了研究。刘洪超^[4]利用灰色理论预测高校毕业生的就业,但是由于其中简化了就业问题,因此预测结果在实际就业情况下缺乏可靠性。黄艳^[5]提出了基于神经网络的高校毕业生就业预测方法,但是神经网络需要问题的先验知识,因此建模较为复杂。此外,也有不少学者侧重从某一社会范围对毕业生就业率进行预测分析的研究,例如张稳^[6]以新

疆维吾尔自治区质量技术监督培训中心毕业生为对象,运用多元线性回归方法预测其就业率,但是不具有普适性。因此,如何对河北省高校毕业生进行准确高效的就业预测具有重要的研究意义。

1 算法思想

半监督自训练方法作为机器学习中的一种学习方式,其主要目的在于通过少量有标记样本集的训练得到一个基础分类器,进而预测大量的无标记样本数据,从中选取可信度比较高的数据添加到训练集中,然后通过扩充后的有标记样本集重新学习得到更好的分类器,重复上述过程直到满足终止条件。目前,常用的基础分类器有K邻近算法、支持向量机与朴素贝叶斯算法。其中,朴素贝叶斯算法是基于贝叶斯定理及特征条件独立假设的分类方法,相较于其他学习方法,能够直观地看到样本的分类概率,能够方便地选择出较高可信度的样本,具有求解速度快、抗噪声强、可以处理大样本问题等优势。因此本文选用朴素贝叶斯算法作为自训练算法的基础分类器。

然而，朴素贝叶斯自训练算法容易受到初始训练样本的数量与分布情况的影响，分布不佳的初始训练集有可能发生错误标记，容易选出置信度较低的无标记样本去进行分类器的训练，从而降低分类器的精度。董立岩等人^[7]提出通过计算最大后验概率与平均后验概率的差值来选择置信度较大的K个样本；Piroonsup等人^[8]提出通过聚类分析选取置信度高的样本；Gan等人^[9]提出使用半监督模糊C均值方法来改进自训练过程。

2 结合相似度的朴素贝叶斯自训练算法

为了提高自训练方法的分类精度，本文将相似度计算方法引入朴素贝叶斯算法自训练算法中。欧氏距离和余弦相似度是常用的两种度量相似度的方法。其中，欧式距离在考虑样本各维度计量单位影响的基础之上，通过计算n维空间上两点间的真实距离来衡量样本的相似度。样本点l到样本点u的欧氏距离计算公式为：

$$D(l, u) = \sqrt{\sum_{i=1}^n (l_i - u_i)^2} \quad (1)$$

余弦相似度是通过计算两个样本特征属性的余弦值来衡量样本的相似度，余弦值越高，样本之间的相似度越大。余弦相似度计算公式为：

$$\cos(q) = \frac{\sum_{i=1}^n (l_i \times u_i)}{\sqrt{\sum_{i=1}^n (l_i)^2} \times \sqrt{\sum_{i=1}^n (u_i)^2}} = \frac{l \times u}{\|l\| \times \|u\|} \quad (2)$$

本文综合采用了欧式距离和余弦相似度两种度量相似度的方法。通过计算无标记样本与有标记样本的相似度，赋予该无标记样本相似度最高的有标记样本的属性，并通过对比朴素贝叶斯分类器的分类结果，如果类别号结果一致并且满足相似度阈值，则将该样本认为是可信度高的样本并加入至训练集中，重复上述过程，就可以得到最终训练好的朴素贝叶斯分类算法。

结合相似度的朴素贝叶斯自训练算法的具体流程：

第1步：输入有标记样本集L、无标记样本集U、有标记样本类标号 T_{label} ，进行朴素贝叶斯分类得到分类标签矩阵L：

第2步：根据(1)式计算欧式距离，得到欧氏距离类标号矩阵 D_{Label} 、欧式距离最小值矩阵 D_{min} 和平均最小欧式距离 d_{means} ；

第3步：根据(2)式计算余弦相似度，得到余弦相似度类标号矩阵 C_{Label} 、余弦相似度最大值矩阵 C_{max} 和平均最大余弦相

似度 C_{means} ；

第4步：对比L、 D_{Label} 、 C_{Label} 三个矩阵中无标记样本 u_i

的类标号，若判别一致，且满足 $D_{min}(i) > d_{means}$ ， $C_{max}(i) > C_{means}$ ，

得到满足条件的 u_i 的矩阵U'；

第5步：有标记样本集L中加入U'，无标记样本集U中减去U'，并将类标号加入到 T_{label} ；

第6步：重复以上步骤，直至U'为空，算法结束。

3 模型建立与结果

3.1 数据处理

本文选取了河北省S高校2022届、2023届与2024届近三年毕业生就业数据作为研究对象，根据影响当前毕业生就业的个人因素与社会因素，着重收集了学生基本信息、学生成绩信息、学生就业信息三个方面的数据，共计12526条。以2023届毕业生为例，三个原始数据表的部分数据分别如表1、表2、表3所示。

表1 2023届毕业生基本信息

学号	姓名	政治面貌	干部情况	生源所在地
2023080224	刘××	共青团员	无	河北省石家庄市
2023030115	李××	共青团员	班长	河北省邢台市
2023050218	张×	共青团员	无	河北省张家口市
2023100205	王××	预备党员	无	河北省邯郸市
2023020310	周××	共青团员	无	河北省石家庄市
2023130309	赵××	共青团员	无	河北省保定市

表2 2023届毕业生就业信息

学号	姓名	学院	单位性质	就业状况	就业单位	单位所在地	单位行业
2023080224	刘××	化工学院	国有企业	已就业	沧州旭阳化工有限公司	河北省沧州市	制造业
2023030115	李××	经济管理学院	其他企业	已就业	石家庄恩成财务咨询有限公司	河北省石家庄市	金融业
2023050218	张×	计算机学院	其他企业	已就业	河北天兵电子科技有限公司	河北省石家庄市	信息技术服务业
2023100205	王××	制药工程学院	国有企业	已就业	石家庄四药有限公司	河北省石家庄市	制造业
2023020310	周××	化工学院	国有企业	已就业	万华化学集团股份有限公司	山东省烟台市	制造业
2023130309	赵××	质量检测与管理学院	待就业	待就业			

表3 2023届毕业生学生成绩信息

课程代码	课程名称	课程性质	学号	姓名	成绩	学院
2310001	高等数学	公共基础必修课	2023080224	刘××	80	化工学院
2210009	财务管理	专业必修课程	2023030115	李××	69	经济管理学院
2110105	数据结构	专业必修课程	2023050218	张×	93	计算机学院学院
1920221	大学英语	公共基础必修课	2023100205	王××	95	制药工程学院
1810028	化工单元操作	专业必修课程	2023020310	周××	86	化工学院
1910096	药物分析检验技术	专业必修课程	2023130309	赵××	78	质量检测与管理学院

在学生成绩信息表中,学生所修的每门课程都对应一条单独的数据,需要进行数据整合才能实现与另外两个表的统一。因此,按照课程性质将课程分为公共基础必修课、公共基础选修课、专业必修课、专业选修课、专业实践课和社会实践课六个课程类型,通过计算每名学生在每个课程类型下所有成绩的平均值,得到每名同学在每个课程类型下的成绩数据,形成整合后的学生成绩信息表。通过学号属性将三个表格中的信息进行集成,并删除掉姓名、就业单位所在地等冗余属性,得到含有12维属性的学生综合信息集成表,如表4所示。

表4 学生综合信息集成表属性及其字段简称

学生基本属性	字段简称	学生基本属性	字段简称
学院	XY	公共基础必修课成绩	GGJCBXKCJ
就业状况	JYZK	公共基础选修课成绩	GGJCXXKCJ
单位行业	DWHY	专业必修课成绩	ZYBXKCJ
政治面貌	ZZMM	专业选修课成绩	ZYXXKCJ
干部情况	GBQK	专业实践课成绩	ZYSJKCJ
生源所在地	SYSZD	社会实践课成绩	SHSJKCJ

对学生综合信息集成表进行数据转换与规范化处理,具体处理方式如下:学校用1-11分别表示各个学院;政治面貌用1表示党员(含正式党员、预备党员),0表示非党员(含群众、团员、入党积极分子);干部情况用1表示是,0表示否;生源信息1表示河北地区,0表示非河北地区;公共基础必修课、公共基础选修课、专业必修课程、专业选修课、专业实践课程和社会实践课程六个课程类型用1表示80分≤成绩≤100分,0表示60分≤成绩<80分,-1表示60分以下;就业情况1表示已就业,2表示待就业,3表示升学,4表示自主创业;单位类别用1表示部队机关,2表示事业单位,3表示企业。

对样本进行清洗操作,剔除掉由于数据错误、数据不完整、数据重叠等因素产生的异常数据,最终得到2022届与2023届毕业生的完整数据共7865条。

3.2模型建立与分析

本文利用毕业生数据首先进行朴素贝叶斯建模分析,然后通过计算毕业生与准毕业生之间的相似度,进而选择可信度较高的毕业生样本加入到训练集中共同训练贝叶斯基分类器,迭代上述过程,直到完成分类器的训练,最后对所剩样本进行分类,实现毕业生的就业预测。

其中,2022届3898条毕业生数据作为有标记样本集 L ,2023届3967条毕业生数据随机按照8:2的比例分别作为无标记样本集 U 和测试样本集 T 。

为了验证算法的有效性与可行性,实验选择传统朴素贝叶斯算法(NB)、K邻近算法(KNN)、朴素贝叶斯自训练算法(NBST)、结合相似度的朴素贝叶斯自训练算法(NBSTCS)四个对比算法进行仿真,每种算法进行5次实验并取5次实验的分类准确率平均值和总体平方差来对比四种算法之间的差异。其中,分类准确率(Accuracy)的计算公式为:

$$Accuracy = \frac{T_{true}}{T_{true} + T_{false}} \times 100\%$$
 (3)

其中, T_{true} 为判断类别正确的样本数量, T_{false} 为判断类别错误的样本数量。

每种算法进行5次实验的分类准确率及其平均值与总体平方差的结果如表5所示。

表5 四种算法的分类准确率与总体平方差

	NB	KNN	NBST	NBSTCS
实验1准确率(%)	89.407314	88.77679697	89.28121059	91.42496847
实验2准确率(%)	84.74148802	90.16393443	89.5334174	90.54224464
实验3准确率(%)	87.26355612	88.27238335	91.42496847	92.1815889
实验4准确率(%)	88.52459016	89.78562421	90.16393443	91.29886507
实验5准确率(%)	86.63303909	87.64186633	90.41614124	90.79445145
准确率平均值(%)	87.31399748	88.92812106	90.16393443	91.24842371
总体平方差	3.232890567	1.09247212	0.707642057	0.402322338

从表中可以看出,本研究提出的结合相似度的朴素贝叶斯自训练算法(NBSTCS)的准确率平均值约为90.8%,并且总体平方差约为0.69%,说明5次实验中算法比较稳定。因此,结合相似度的朴素贝叶斯自训练算法(NBSTCS)能够较好地完成毕业生就业预测工作。

3.3 预测结果

本文使用结合相似度的朴素贝叶斯自训练算法(NBSTCS)对河北省S高校2024届毕业生进行就业预测应用。将数据集按照学院进行划分,将2023届与2024届的毕业生样本作为有标记样本,导入2024届毕业生的数据,预测结果如表6所示。

表6 NBSTCS算法应用于河北省S高校2024届毕业生预测的结果

学院	就业率(%)	学院	就业率(%)
化工学院	94.33	装备制造学院	89.33
环境工程学院	92.17	智能工程学院	86.33
制药工程学院	93.67	计算机学院	91.67
质量检测与管理学院	91.33	电子与信息工程学院	88.67
药学院	90.75	经济管理学院	87.33
建工学院	91.17	整体就业率	90.61363636

结合河北省S高校近三年的就业率,本文使用结合相似度的朴素贝叶斯自训练算法(NBSTCS)进行的就业率预测符合预期效果,如表7所示。

表7 河北省S高校2021-2023届毕业生就业率

年份	2021年	2022年	2023年	2024年(预测值)
就业率	89.94%	90.87%	90.33%	90.61%

4 结论

本文提出的结合相似度的朴素贝叶斯自训练分类方法稳定性好,具有更高的分类准确率,能够较好地应用于河北省高校毕业生就业预测工作。根据预测结果可以对就业比较困难的毕业生采取心理访谈、技能提升等方法实施个性化就业指导,提升大学生综合就业能力,为河北省高校毕业生的就业工作提供指导意义。

[基金项目]

本文系2024年河北省社会科学发展研究课题项目“河北省

高校毕业生就业预测模型研究及应用(课题编号:202402101)”的阶段性研究成果。

[参考文献]

[1]金延军.数据挖掘技术在高校学生就业指导中的应用[J].黑龙江工程学院学报(自然科学版),2009,23(01):60-63.

[2]王亚如.基于决策树算法的大学生就业预测模型及应用研究[D].华中师范大学,2018.

[3]吴亚娟.基于因子-聚类分析的大学生就业满意度统计及预测[J].南京信息工程大学学报(自然科学版),2010,2(06):510-513.

[4]刘洪超,滕鑫鑫,白浩.基于大数据的高校智能就业平台建设与应用[J].现代教育技术,2020,30(02):111-117.

[5]黄艳,田辉玉,王建农.高校毕业生就业竞争力的评价与发展趋势预测[J].统计与决策,2014,No.400(04):114-116.

[6]张稳,恰汗·合孜尔.毕业生就业率分析与预测系统[J].计算机与数字工程,2008,No.224(06):95-98.

[7]董立岩,隋鹏,孙鹏,等.基于半监督学习的朴素贝叶斯分类新算法[J].吉林大学学报(工学版),2016,46(03):884-889.

[8]PIROONSUP N,SINTHUPINYO S.Analysis of training data using clustering to improve semi-supervised self-training[J].Knowledge-Based Systems,2018,143(1):65-80.

[9]GAN H T,SANG N, HUANG R,et al.Using clustering analysis to improve semi-supervised classification[J].Neurocomputing, 2013,25(3):290-298.

作者简介:

宗慧敏(1979--),女,河北怀安人,河北化工医药职业技术学院公共教学部副主任,硕士学位,副教授,研究方向为数学与应用数学,发表《高职人才培养体系中数学教学理念与方法的改革探索》等20余篇;著作4部,主持省级课题4项。邢亚璇,河北栾城人,河北化工医药职业技术学院公共教学部教师,硕士学位,讲师,研究方向为统计学。刘微,河北化工医药职业技术学院公共教学部教师,硕士学位,副教授,研究方向为拓扑学。黄江枫,河北化工医药职业技术学院公共教学部教师,硕士学位,助教,研究方向为概率统计学。