计算机科学
COMPUTER SCIENCE

基于分数线预测的多特征融合高考志愿推荐算法

王泽卿, 季圣鹏, 李鑫, 赵子轩, 王鹏旭, 韩霄松

引用本文

王泽卿, 季圣鹏, 李鑫, 赵子轩, 王鹏旭, 韩霄松基于分数线预测的多特征融合高考志愿推荐算法[J]. 计算机科学, 2022, 49(11A): 211100266-7.

WANG Ze-qing, JI Sheng-peng, LI Xin, ZHAO Zi-xuan, WANG Peng-xu, HAN Xiao-song. Novel College Entrance Filling Recommendation Algorithm Based on Score Line Prediction andMulti-feature Fusion [J]. Computer Science, 2022, 49(11A): 211100266-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100266

基于分数线预测的多特征融合高考志愿推荐算法

王泽卿[1] 季圣鹏[1] 李 鑫[2] 赵子轩[1] 王鹏旭[1] 韩霄松[1,3]
1 吉林大学软件学院 长春 130012
2 吉林大学原子与分子物理研究所 长春 130012
3 吉林大学计算机科学与技术学院 长春 130012
(wangzq5519@mails.jlu.edu.cn)

摘 要 近年来,随着我国高考人数逐年增多,考生对高考志愿填报服务的需求日益增加。面对海量的院校填报信息,考生往往很难在短期内做出比较符合自身意愿的合理选择,进而导致报考事故的发生。因此,针对高考志愿报名问题,在爬取历年高考录取数据的基础上,提出一种基于分数线预测的多特征融合推荐算法(Reco-PMF)。该算法首先利用历年高校最低投档位次,通过BP神经网络预测报考年份各高校最低投档位次以及最低投档分数线,然后根据考生分数进行院校初筛,进而构建3种与录取分数相关的特征,结合院校软科排名,通过遗传算法进行权值寻优,得到不同院校的录取概率,并在此基础上定义推荐度实现为考生进行不同录取风险层次的高校推荐,形成完整的推荐结果。实验结果表明,基于BP神经网络的高校录取分数预测算法在不同误差限下的表现均优于其他算法;相比百度和夸克的已有服务,所提算法在多层次测试分数下,平均录取率分别提升14.8%和24.1%,同时成功录取院校的平均位次分别提升了99名和87名。

关键词:高考志愿填报;分数线预测;参数优化;多权重;遗传算法
中图法分类号 TP391

Novel College Entrance Filling Recommendation Algorithm Based on Score Line Prediction and Multi-feature Fusion

WANG Ze-qing[1],JI Sheng-peng[1],LI Xin[2],ZHAO Zi-xuan[1],WANG Peng-xu[1] and HAN Xiao-song[1,3]
1 College of Software,Jilin University,Changchun 130012,China
2 The Institute of Atomic and Molecular Physics,Jilin University,Changchun 130012,China
3 College of Computer Science and Technology,Jilin University,Changchun 130012,China

Abstract In recent years,as the number of high school graduates growing,the demand of college entrance filling is increasing. But faced with massive amounts of college entrance data,students always cannot make reasonable decisions conform to their own will in a short time,resulting in filling accident. To address this issue,on the basis of crawling college entrance history data by web spider,a novel college entrance filling recommendation algorithm based on score line prediction and multi-feature fusion(Reco-PMF) is proposed. Firstly,BP neural network is applied to predict all the colleges admission lines of current year. Then,combining with colleges' rankings,an admission probability algorithm is constructed on the basis of three score related features. Genetic algorithm is employed to optimize the weights of above features. On this basis,recommendation-score is defined to measure admission risk. Finally,a college filling list with multi-admission risk is generated. Experiment results show that,the college admission line prediction algorithm based on BP neural network performs better than other algorithms under all error bounds. Compared with existing on-line services of Baidu and Kuake,Reco-PMF increases the acceptance rates by 14.8% and 24.1%,and improves the average ranking of recommended colleges by 99 and 87 in accepted colleges.

Keywords College entrance filling,Score line prediction,Parameter optimization,Multi-weight,Genetic algorithm

1 引言

志愿填报是高考招生中至关重要的一个环节。假如没有科学的指导和正确的填报方法,很容易出现考生盲目填报的情况,从而造成滑档,无法录取等事故。数据显示,70%的大学毕业生对自己当年填报的高考志愿感到后悔,有些考生在进入大学后会因专业不满意而选择复读[1],而有些考生会因为院校与预期不符而产生厌学情绪,甚至最后会因成绩不达标无法顺利毕业[2]。我国作为一个人口大国,自2019年全国高考人数重回千万量级后[3],近两年来持续增长,其中仅河南省

基金项目:国家自然科学基金(61972174);国家级大学生创新创业训练计划(202110183225)
This work was supported by the National Natural Science Foundation of China(61972174) and National Innovation and Entrepreneurship Training Program for College Students(202110183225).
通信作者:韩霄松(hanxiaosong@jlu.edu.cn)

2021年的考生便较2020年增长了10万人，达到了125万人，而全国高考总人数从2020年的1 071万人增长到了1 078万人，再创历史新高。数量众多的考生也带来了对志愿填报推荐服务的庞大需求[4]。我国在多年的高考以及高考招生的信息化建设中积累了非常丰富的高考信息数据资源[5]，但是在庞大的数据面前，考生填报志愿时却往往无所适从，以至于最后选择了一个不满意的院校。

当前，国内已有多种模型结合大数据和机器学习算法对考生志愿进行推荐，网络上可以找到各院校的相关信息以及录取的分数线等。但高考志愿推荐面临着诸多挑战，比如如何判定影响分数线的因素、如何获取真实可靠的高考历年数据、无法获取考生详细报考数据导致无法使用如协同过滤等传统推荐算法；尤其是近年来随着我国新一轮高考改革的开始，许多历史数据不再适用[6]。Ren通过支持向量回归机(Support Vector Regression，SVR)对分数线进行预测，结合模糊C均值聚类(Fuzzy C-Means，FCM)算法和遗传算法(Genetic Algorithm，GA)对院校进行个性化推荐，其系统中数据种类齐全，且对考生信息要求较少[7]。但该模型使用了高考分数而非更稳定和具有代表性的位次，导致部分数据预测误差较大。Yin采用了灰度预测对录取概率进行计算，对考生进行不同层次的院校志愿推荐，能够减少考生滑档的概率。生成的报考方案能够帮助考生对整体可报高校有较为宏观的了解，并且以录取概率的形式反馈给考生作为参考[8]，但是系统对数据的完整程度依赖性过强，无法应对部分院校数据缺失的问题。Wang等以成人高考网上报名系统中的数据为研究对象，通过基于信息增益率的协同过滤推荐算法设计志愿推荐系统[9]。Wu等运用基于知识与基于内容的推荐算法设计了一套包含专业推荐、学生推荐、院校及专业推荐共3个模块的自动问答系统[10]，该系统能够减少考生在填报志愿时面对海量信息投入的时间成本。Yu等通过C均值模糊聚类以及院校不同层面特征权重，给出院校录取概率计算公式，最后形成不同风险下的冲、稳、保3类推荐结果，能够让考生对录取结果有较为直观的预期，并具有很好的可解释性[11]。

当前的志愿推荐方法往往存在数据依赖较强、最低投档分数线预测误差较大、推荐结果不够直观的问题。针对上述问题，本文通过构建主题爬虫获得较为全面的院校信息，再通过BP神经网络对院校当年的最低投档位次进行预测，在此基础上，结合院校排名信息，利用遗传算法对本文提出的基于分数线预测的多权重高考志愿推荐算法(A Novel Recommendation Algorithm Based on Score Line Prediction and Multi-feature Fusion for College Entrance Filling，Reco-PMF)模型进行参数调优，在提高录取概率的前提下，最大化利用考生的分数，推荐排名位次较高的院校；最终根据考生的高考位次信息计算院校的录取成功率、推荐度，为考生填报志愿提供参考。由于难以取得较长时间范围的院校相关数据以及真实的考生报考数据，因此本算法尽量采用了较为可靠的预测以及推荐方法。

2 数据获取及处理

2.1 数据获取

本文从中国教育在线[12]、软科[13]、阳光高考[14]等教育咨询网站获取了多年的历史高考数据，主要包括2015－2021年各院校在各省最低投档位次(理科)、2015－2021年各省一分一档表(理科)、2021软科中国大学排名。本文以河南省理科高考数据为例介绍所提模型，并以此来模拟该算法在全国的应用情况。

2.2 数据处理

如表1所列，在获取到的最低录取分数线数据中，每个院校不同年份、不同类别的数据独立出现，这样的数据格式不便于后续算法的处理，需对其进行合并操作。在合并过程中，由于部分院校存在同一年份在同一批次中有不同类型招生的情况，为了实验数据的统一性以及与不同系统实验对比的统一性，我们仅保留其中的普通类用于实验，得到合并后的数据。合并后的数据存在部分院校数据缺失的情况。为了尽可能扩大推荐高校覆盖率，选取缺失数据院校的邻近年份数据对缺失年份进行横向填充，这是由于数据缺失院校在纵向上往往与相似院校数据相差较大。模型中使用到软科院校排名，故将软科2021排名中不包含的高校删除，最终得到825所高校的信息，部分数据如表2所列。

表1 合并前最低投档位次数据(部分)

Table 1 Lowest position data before merging(partial)

高校名称	年份	考生所在省份	文/理	录取批次	录取类型	最低投档位次
吉林大学	2021	河南	理科	本科一批	农林矿	23232
吉林大学	2021	河南	理科	本科一批	普通类	11699
吉林大学	2020	河南	理科	本科一批	农林矿	29485
吉林大学	2020	河南	理科	本科一批	普通类	11437
吉林大学	2019	河南	理科	本科一批	农林矿	25153
吉林大学	2019	河南	理科	本科一批	普通类	10464

表2 处理后最终数据(部分)

Table 2 Data after pre-processing(partial)

软科排名	高校名称	录取批次	2021年	2020年	2019年	2018年	2017年	2016年	2015年
24	吉林大学	本科一批	11699	11437	10464	11400	15197	10460	10915
53	郑州大学	本科一批	23896	24312	24522	23432	22210	23072	22155
97	河南大学	本科一批	36819	37744	40404	39974	41790	39011	37776

3 Reco-PMW模型

3.1 提档线预测

当前的高考志愿推荐系统往往基于高校历年的最低投档分数线来对未来的最低投档分数线进行预测。但是最低投档分数线往往随着每年试题难度等因素的变化而波动较大，因此本文对更为稳定且能够代表院校实际录取难度的最低投档位次进行预测，再将其转换为最低投档分数线。

构建BP神经网络模型如图1所示，根据历年已有院校最低投档位次来预测下一年的院校最低投档位次。我们将相邻两年的最低投档位次作为输入，而将第三年的最低投档位次作为输出。输入节点和输出节点都是历年的最低投档位次。故网络结构的输入层含有2个节点，输出层含有1个节点，隐含层节点数目设置为8，隐含层节点采用Sigmoid激活函数，输出层采用线性激活函数，损失函数设置为误差平方和损失函数。为防止过拟合，设置损失值下限，当损失值小于损失值下限时停止迭代。图2给出了各个院校训练100轮的平均损失曲线。

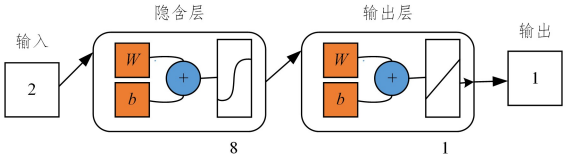


图 1　BP 神经网络结构

Fig. 1　BP neural network structure

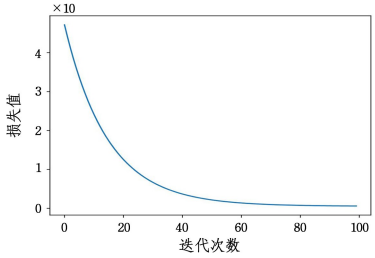


图 2　平均损失曲线

Fig. 2　Average loss curve

3.2　推荐策略

推荐策略如图 3 所示。考生输入分数后，根据其分数初步筛选出一定数量的院校，随后计算每一所院校录取该考生的概率；将录取概率大于 0.8 的院校划分为“保”；大于等于 0.6 且小于等于 0.8 的划分为“稳”，小于 0.6 的院校划分为“冲”。并在此基础上计算推荐度，根据院校的推荐度以及录取概率综合排名得到最终推荐结果。以河南省为例，每名考生在本科一批志愿中可以填报 6 所院校，因此对“冲”“稳”“保”每个推荐类别取其综合排名前两名的院校作为推荐结果。同时按照推荐度确定推荐顺序，生成考生志愿填报表。

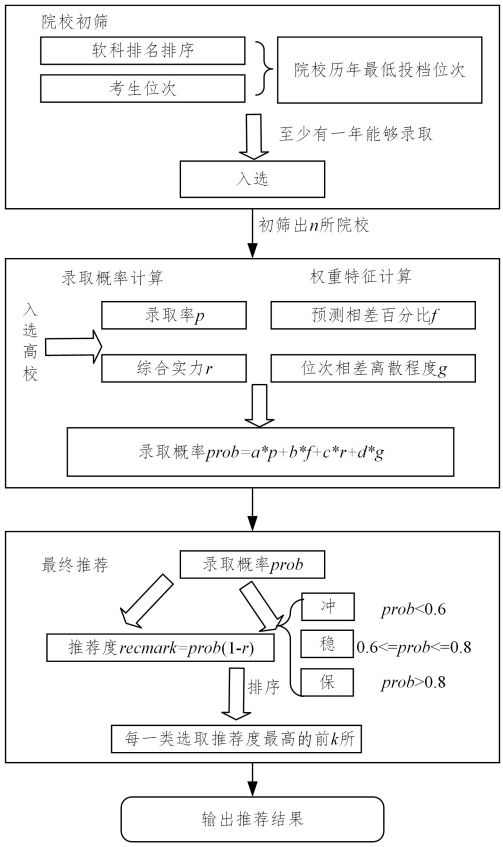


图 3　推荐策略

Fig. 3　Recommended strategy

3.2.1　院校初筛

根据考生的分数对院校进行初步筛选，筛选规则为：将院校根据 2021 年软科大学排名进行排序后，由高向低寻找高校；入选的高校在 2015－2020 年分数线所对应的位次中，至少有一年此考生能够被录取；由此筛选出 n 所高校。图 4 给出了 2020 年与 2019 年院校最低投档分数线总和分布图，由图中可以看出整体院校分布类似于正态分布，因此我们采用正态函数对其进行拟合，得到的正态函数均值 μ 约为 540，标准差 σ 约为 55。由此我们得到 n 的计算式如式(1)所示：

$$n=25+25*100*\frac{1}{\sqrt{2\pi}*55}e^{\left(-\frac{(s-540)^2}{2*55^2}\right)} \tag{1}$$

其中，s 为考生分数，由于该正态分布函数数量级与院校数量相差较大，因此通过系数调整其数量级。例如一名考生高考分数 s 为 618 分，则对其进行推荐时，经过计算得到 n 值为 31；因此需要为其筛选 31 所院校，并在此基础上，对这些院校进行录取概率、推荐度的计算，用于后续推荐。

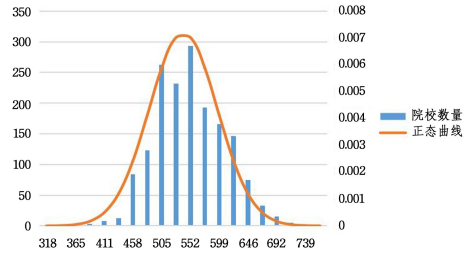


图 4　分数分布与正态拟合曲线

Fig. 4　Curve of score distribution and normal fitting

3.2.2　录取概率计算

研究人员发现，通过对院校的相关特征进行评分，并且根据这些评分指标，如近年来高校的最低投档位次、最低投档位次与考生分数对应位次之间的距离离散度等计算录取概率的方法具有较好的推荐效果[11]。受其启发，结合本文数据特点，我们确定以下 4 个评分指标：

(1)历年录取率 p。计算方法如式(2)所示：

$$p=\frac{n}{year} \tag{2}$$

其中，$year$ 为过往年份总数，本算法实验中过往年份为 2015 年至 2020 年，故 $year$ 值为 6；n 为考生分数对应位次能够满足高校最低投档位次的年份数目。以吉林大学为例，对于一位 2021 年分数对应位次为 11000 名的河南理科考生来说，此位次在 2017 年、2018 年、2020 年均可以考入吉林大学，则其 n 值为 4，计算其 p 值为 0.5。

(2)预测最低投档分数相差百分比 f。计算如式(3)所示：

$$f=\frac{s-y}{m-s} \tag{3}$$

其中，考生分数为 s，预测分数为 y，该省高考总分为 m，其中预测分数 y 由 BP 神经网络方法得到的预测最低投档位次转换得到，综合了该校近年来的录取情况。f 衡量了该考生的分数与预测得到的分数线之间的差距，能够判断考生分数相对于该院校的优劣势。以吉林大学为例，假设吉林大学 2021 年在河南得到的预测分数为 610 分，河南省高考满分为 750 分，对于一位 2021 年高考分数为 618 分的河南理科考生来说，其 s 为 618，y 为 610，m 为 750，

则其 f 值为 0.0606。

(3)软科综合实力排名 r。院校的软科综合实力排名计算式如式(4)所示。如吉林大学 2021 年软科综合实力排名为 24 名,则其 r 值为 24。

$$r=rank(school) \tag{4}$$

(4)位次差离散程度 g。计算式如式(5)所示。以吉林大学为例,假设吉林大学 2015 年至 2020 年最低投档位次为 10915,10460,15197,11400,10464,11437;考生分数对应位次为 10000 名,则其 g 值为 0.0947。

$$g=\frac{\sqrt{\sum_{i=0}^{n}\left(\frac{w-w_i}{w}\right)^2}}{year} \tag{5}$$

针对筛选出的所有院校确定 4 个参数后,对 f,r,g 参数进行归一化处理。由于 p 参数取值较为固定,彼此之间差距较小,且均在 0~1 范围内,故不对其进行归一化处理。得到概率计算式(6),其中 a,b,c,d 为该公式的未知参数,即每个评分指标的权重系数,其和为 1,$prob$ 为该高校录取此考生的概率。

$$prob=a*p+b*f+c*r+d*g \tag{6}$$

3.2.3 推荐度

得到高校的录取概率后,由于其考虑因素较为单一,不宜直接根据其高低进行推荐,因此我们定义某高校的推荐度 $rec\text{-}socre$ 如式(7)所示:

$$rec\text{-}score=prob*(1-r) \tag{7}$$

其中,$prob$ 为该院校的录取概率,r 为该校经过归一化后的软科综合实力排名。

3.2.4 推荐结果

得到筛选的 n 所高校各自的录取概率以及推荐度后,按推荐度对院校进行排序,分别找出“冲”“稳”“保”3 个类别中推荐度居前列的数所院校,作为最终推荐结果。其中具体选取的院校数量由考生所在省份志愿填报规则决定,如河南省本科一批可以填报 6 所院校,则在 3 类中选取推荐度最高的前两所院校作为推荐结果。

3.3 权重优化

3.3.1 遗传算法

遗传算法(GA)是一种通过模拟自然界中生物进化的人工智能技术,以自然选择和遗传理论为基础的人工智能算法,由 Michigan 大学的 Holland 教授于 1975 年提出[15],目前已经广泛应用于计算机科学、工程实践中[16]。

遗传算法可以对一个种群进行多方向的搜索,进而能够跳出局部最优解,能够有效避免陷入局部极值点[17]。利用遗传算法为本算法中概率计算公式中的权重进行参数寻优。

染色体数量代表目前解的个数。本算法中,对于每一个生成的染色体,都需要对其中的 4 个基因,即 4 个未知参数在解码后进行规范化处理,使其和为 1。每个基因编码为长度为 4 的二进制序列,解码时,将其转换到[1,16]范围内。如一个染色体的 4 个基因分别为[1,1,0,0],[1,0,1,0],[0,1,0,0],[0,0,1,0],先将各个基因按位解码为十进制数,即 12,10,8,2,随后对每个十进制数加 1,得到解码完成的 4 个基因:13,11,9,3。随后,将得到的十进制数进行规范化操作,使其和为 1,则上述解码完成后的十进制数转换为:0.36,0.31,0.25,0.08。解码流程如图 5 所示。

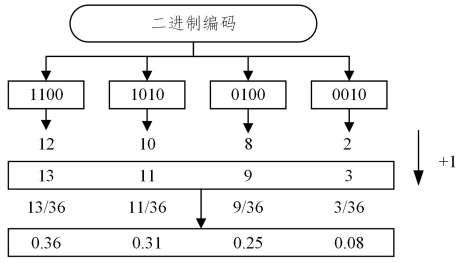


图 5 解码流程

Fig. 5 Decoding process

适应度函数是遗传算法的关键部分。对于最终的推荐算法来说,我们希望能够尽可能区分各院校的录取概率,以便后续推荐的进行。同时,对于考生来说,最终能够被录取是最重要的。综合考虑以上两点,期望在考生不出现滑档的同时,尽可能被较好的院校录取。因此,设计适应度函数如式(8)所示:

$$fitness=15*variance+0.8*hit@保 \tag{8}$$

其中,$variance$ 为在此种群下计算得到的各个院校录取概率的方差,此方差代表各个院校录取概率的区分度,方差越大,院校录取概率之间的区分度越高。$hit@保$ 为在“保”这一类别的实际录取率,系数为经验值,用于调整两者之间的数量级关系。

采用区间采样的方法计算每个种群的适应度。选取 475,500,525,550,575,600,625,650 共 8 个分数作为输入,分别得到其适应度 f_1-f_8,以其平均值作为该种群的最终适应度 f_{avg},如式(9)所示:

$$f_{avg}=\frac{\sum_{i=1}^{8}f_i}{8} \tag{9}$$

3.3.2 参数调优

以 2020 年的数据作为调优目标,2015-2019 年的数据作为已知历史数据。对于得到的含有 4 个未知参数的概率计算公式,规定其取值区间为[1,16],经过迭代后得到 1 组最优参数,遗传算法迭代过程如图 6 所示。

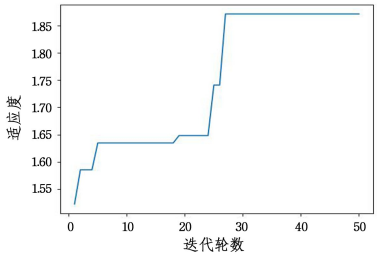


图 6 遗传算法迭代过程

Fig. 6 GA iterative process

通过迭代与计算,最终确定 4 个参数值如式(10)所示:

$$prob=0.73*p+0.13*f+0.09*r+0.05*g \tag{10}$$

优化得到的 4 个参数中,历年录取率 p 的系数为 0.73,权重最高,这是因为对于大部分院校来说,年份之间最低投档位次的变化不会太大,因此该参数对于考生能否被录取的影响也较大。而预测最低投档分数相差百分比 f 的参数和位次差离散程度 g 的参数较低,这是由于尽管这两者也能够反应出最低投档的位次变化趋势,但是不如历年最低投档位次数据直接,因此参数值较低。软科综合实力排名 r 的参数也较低,这是因为一所院校的最低投档位次并非与其综合实力成严格的正比关系,还有许多如地域、优势专业分布等其他

因素的影响，因此难以直接用线性系数的方式表现出其对最低投档位次的影响。

4 实验与结果

本文所有实验对象均为2021年高考河南理科考生，该省2021年理科一本线为518分，理科二本线为400分。由于难以取得真实报考数据，因此采用了模拟考生填报的方法进行实验。

4.1 最低投档位次预测算法对比

4.1.1 百分位模型

考虑到实际情况中最低投档分数线较低的院校每年最低投档位次变化较大，如内蒙古科技大学2020年最低投档位次为133199，2021年为296564，差值达163365，对这类院校最低投档位次的预测不具有现实意义，因此我们将2021年最低投档位次变化幅度超过近六年均值一定百分比的院校标记为异常，并在此基础上建立百分位模型。对未标记为异常的院校分别采用线性回归模型、梯度下降模型、BP神经网络模型进行对比。

4.1.2 分数线预测模型对比结果

通过2021年最低投档位次的预测结果与实际最低投档位次作差，建立百分位模型，将误差小于2021年实际最低投档位次一定百分比的院校作为合格院校。分别采用线性回归模型、梯度下降模型、神经网络模型，比较合格院校在误差限为5%，10%，15%，20%下无异常院校中的占比，结果如表3所列。其中误差限为预测最低投档位次与实际最低投档位次之差占实际最低投档位次的百分比。误差限反映了在一定接受程度下，最低投档位次预测的准确率。

表3 合格院校占比

Table 3 Proportion of qualified colleges

（单位：%）

模型\误差限	5	10	15	20
线性回归	45.1	56.7	69.2	71.7
梯度下降	45.9	55.8	67.6	71.7
BP神经网络	80.3	86.5	94.2	93.4

由表3可知，在测试的所有误差限下，BP神经网络预测效果均为最优。

4.2 推荐结果

4.2.1 实验标准

对于高考考生来说，最重要的是能够被录取并被录取到一所较好的院校。对于录取质量，我们采用录取比 p 作为评价指标，计算式如式(11)所示：

$$p=\frac{in}{rec} \tag{11}$$

其中，rec 为实际推荐的院校数目，in 为推荐的高校中实际能够被录取的院校数目。

对于是否被录取到一所较好院校的评价，本文采用推荐可录取高校的软科平均排名 q 对其进行评估，计算式如式(12)所示：

$$q=\frac{\sum_{i=1}^{m} rank_i}{m} \tag{12}$$

其中，$rank_i$ 为推荐院校中能够被被录取的第 i 所院校对应的2021年软科大学排名，m 为推荐院校中能够被录取的总数。不同于通过院校最低投档分数线进行评估，综合排名是一所院校综合实力的体现，院校最低投档分数线的高低很大程度上取决于其排名的高低；但是对于部分专业导向型考生，可能其考虑更多的是院校专业水平，而该校最低投档分数线可能会较低，那么对于这类考生，用最低投档分数线评价其选择很显然是不恰当的，因此我们通过推荐可录取高校的软科平均排名的方法来对其进行评价。

4.2.2 录取实验结果

最终推荐的院校数量与考生省份有关。本文以河南省为例，在每个类别中选取推荐度最高的前两所院校进行输出，最终推荐6所院校。以一名高考分数为575分的理科考生为例，其推荐结果如表4所列。

表4 系统输出推荐表(考生分数为575分)

Table 4 System outputs(examinee scored 575)

志愿顺序	推荐院校	批次	类别	是否录取
一	宁波诺丁汉大学	本科一批	冲	是
二	中国医科大学	本科一批	冲	否
三	河北医科大学	本科二批	稳	是
四	西交利物浦大学	本科一批	稳	是
五	哈尔滨医科大学	本科二批	保	是
六	北京中医药大学	本科二批	保	是

通过模拟考生数据的方法对各推荐类别进行录取率结果验证。分别在5个分数段[450－500]，[500－550]，[550－600]，[600－650]，[650－700]进行验证。每个分数段随机产生50名2021年河南理科考生的分数，验证其划定范围内全部高校各类别的平均录取率，实验结果如图7所示。

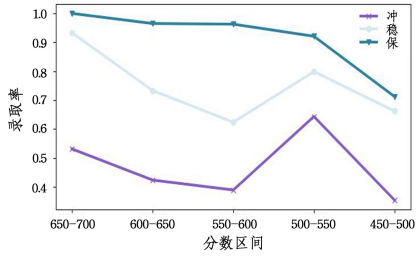


图7 各分数段各类别平均录取率

Fig. 7 Average admission rates of each score segment

从图7可以看出，低分段“冲”类别的录取准确率较高，而高分段的“稳”与“保”准确率较高，几乎达到了100%。这是由于高分段考生人数少，院校数量少，每年最低投档位次较为稳定，考生冲击到相对较高的院校难度较大，而低分段考生人数多，院校多，每年最低投档位次浮动也较大，因此冲击上高分院校的概率也较大，而低分段院校较大的分数波动也导致了“稳”与“保”类别录取率降低。由对比结果可见，较低分数段的推荐效果较差，而较高考分各类别的录取率均有提升。这是由于在较低分数段中院校数量较多，考生人数也较多，院校最低投档位次浮动较大，“信息过载”问题较突出，而在高分数段中，院校数量较少，考生人数较少，“信息过载”情况较少[18]。

4.3 与主流推荐系统对比

选取“百度”[19]与“夸克”[20]两家拥有广泛受众以及良好口碑的企业所提供的高考推荐服务进行对比，其推荐结果同样分为冲、稳、保3类。分别对比实际录取率与推荐高校平均排名。同样采用区间采样方法，分别对分数为450，475，500，525，550，575，600，625，650的2021年河南理科考生进行测试。同时，为了保证推荐的统一性，按照平台给出的每个类别的推荐顺序，在推荐的“普通类”公办高校中由高至低选择前

两所，即3个类别共6所高校，形成完整志愿表进行对比。需要注意的是，对于部分分数，对比系统或本系统并未给出两所及以上的符合要求的推荐，此时采用邻近类别作为补充，以保证选够6所院校。各系统部分推荐结果如表5所列，每所院校下方为其2021年河南理科实际最低投档分数线以及软科综合实力排名。

表5 各系统之间推荐结果对比(部分)

Table 5 Comparsion of recommended results between systems(partial)

平台	分数	第一志愿		第二志愿		第三志愿		第四志愿		第五志愿		第六志愿	
百度	500	太原师范学院		吉林农业大学		北方民族大学		内蒙古工业大学		张家口学院		辽宁科技学院	
		510	491	497	160	502	439	489	285	487	546	452	480
	525	南阳师范学院		内蒙古科技大学		沈阳工业大学		山西财经大学		北方民族大学		河南城建学院	
		534	351	537	323	515	216	517	288	533	439	519	454
	550	安徽工业大学		西北民族大学		辽宁工程技术大学		浙江师范大学		中国人民警察大学		黑龙江中医药大学	
		559	168	569	379	537	261	574	103	545	413	543	302
	575	大连海事大学		青岛理工大学		北京联合大学		山东科技大学		东北农业大学		青岛科技大学	
		600	104	521	254	566	259	577	133	585	116	572	161
夸克	500	西南医科大学		闽南师范大学		内蒙古财经大学		福建工程学院		仲恺农业工程学院		贵州财经大学	
		517	293	503	318	497	366	498	320	497	354	463	387
	525	上海体育学院		江汉大学		中国民用航空飞行学院		北方民族大学		西南医科大学		新疆医科大学	
		553	165	557	263	535	440	533	439	517	293	508	226
	550	沈阳药科大学		中南林业科技大学		扬州大学		中国人民警察大学		沈阳农业大学		中国民用航空飞行学院	
		559	194	558	210	580	77	545	413	529	141	535	440
	575	青岛大学		上海政法学院		北京第二外国语学院		汕头大学		安徽财经大学		扬州大学	
		581	100	581	285	580	188	575	131	565	224	580	77
本系统	500	南宁师范大学		内蒙古师范大学		内蒙古民族大学		黑龙江八一农垦大学		新疆师范大学		新疆农业大学	
		512	292	506	267	458	294	487	312	474	295	461	316
	525	河南师范大学		河北农业大学		福建农林大学		海南师范大学		吉林农业大学		沈阳师范大学	
		499	127	515	178	498	121	507	234	497	160	498	196
	550	河北医科大学		对外经济贸易大学		江苏师范大学		黑龙江大学		北京中医药大学		福建农林大学	
		554	105	642	47	527	122	516	138	546	103	498	121
	575	宁波诺丁汉大学		中国医科大学		河北医科大学		西交利物浦大学		哈尔滨医科大学		北京中医药大学	
		539	82	610	62	554	105	550	128	545	79	546	103

对比不同平台给出推荐结果的最终录取率及其高校平均位次，得到的结果如表6、表7所列。

表6 各系统在各分数的录取率

Table 6 Acceptance rate of each system at each score

分数	夸克	百度	本系统
450	0	0	0
475	0.333	0.667	0.667
500	0.667	0.667	0.667
525	0.333	0.500	1.000
550	0.500	0.500	0.667
575	0.333	0.500	0.833
600	0.667	0.833	0.500
625	0.667	0.667	0.833
650	0.333	0.333	0.833

表7 各系统在各分数的录取平均排名

Table 7 Average admission ranking of each system at each score

分数	夸克	百度	本系统
450	—	—	—
475	519.50	493.50	368.00
500	356.75	367.75	304.25
525	259.50	319.30	169.30
550	331.30	325.30	121.00
575	177.50	224.70	99.40
600	117.75	153.20	57.70
625	78.50	48.25	45.60
650	41.00	40.00	18.60

最终录取率对比中，3个系统在450分时的推荐结果均是未能成功录取到任何一所院校，这是因为我们将对比院校的范围划定到了公办院校的普通类中，而在该分数附近，公办院校数量较少，且分数线波动较大，很容易录取失败。而在其他8个分数中，本系统在5个分数中的录取率高于另外两个系统，且在一个分数达到了全部录取的效果。但是需要注意的是，较高的录取率并不一定代表了较好的推荐结果，如果推荐院校的录取率较高而推荐院校水平较低，很显然不符合考生的意愿，因此，还需要结合推荐院校的水平来综合考虑。

可录取院校软科平均排名对比中，由于在450分时，3个系统均未能成功推荐录取到任何一所院校，因此将其数据设置为“—”。而对其他分数，可以看到“夸克”平台推荐可录取院校软科平均排名大多优于“百度”平台，而本系统的推荐可录取院校软科平均排名均优于另外两个系统。

综上，所提算法在最终推荐中取得了较好的结果，验证了其有效性。

结束语 本文通过BP神经网络模型对最低投档位次进行预测，之后结合多特征计算的方法对各院校录取概率进行计算，并给出推荐度的概念，最终根据录取概率和推荐度得到最终推荐结果。所提算法能够从考生分数对应位次直接得到一份完整的志愿推荐表，能够很好地为考生提供志愿推荐服务。通过与当前的主流平台对比，验证了本文算法的有效性。

但本文算法还存在一些问题，比如基于最低投档位次预测时部分数据存在异常，应该结合院校当年其他维度信息，如计划招生人数等，对最低投档位次变化较大的院校进行处理。本系统推荐时仅考虑了考生的分数，没有结合考生的偏好，如地区偏好、专业偏好等辅助信息[21]，后期可以增加性格测试部分，根据测试得到的结果进行推荐，并增加院校的专业实力维度；也可以由考生自己指定录取概率计算公式中的系数，充分考虑考生意愿，为考生提供个性化推荐服务。在最终的志愿推荐中，不同省份的不同批次可以填报的院校数量不同，本文为了简化对比选定河南省并统一推荐6所院校，但实际

填报中，不同批次之间的填报规则并非完全相同。这些都将是后期的研究内容。

参 考 文 献

[1] YAN W. Research on fuzzy clustering mining technology and its application in college entrance examination voluntary filling service[D]. Changsha: Central South University College of Information Science and Engineering, 2009.

[2] FANG X F. Fill in application for college entrance examination scientifically and reasonably[J]. Shanxi Education, 2018(10): 7.

[3] XIONG B Q. What Is the Implication of Gaokao Examinees' Number Back to Ten Million? [J]. Shanghai Journal of Educational Evaluation, 2019, 8(4): 14-17.

[4] CHENG L L. Design and implementation of volunteer filling system based on multi-source heterogeneous data page rendering[D]. Shenyang: Shenyang Institute of Computing Technology, Chinese Academy of Sciences, 2021.

[5] MENG Z. Design and implementation of college entrance examination recommendation system based on Spark[D]. Jinan: Shandong Normal University. 2017.

[6] KANG L, HA W. The Effect of College Admission Mechanism Reforms on the Quality of Matching (2005—2011)[J]. Peking University Education Review, 2016, 14(1): 105-125, 191.

[7] REN J T. Application of recommendation algorithm in college entrance examination[D]. Kunming: Yunnan University of Finance and Economics, 2018.

[8] YIN H Y. Design and Implementation of university entrance examination volunteer recommendation system based on big data [D]. Chengdu: University of Electronic Science and Technology of China, 2018.

[9] WANG Y J. Research of Adult College Specialties Recommendation System Based on Data Mining and Collaborative Filtering [D]. Beijing: Beijing Forestry University, 2011.

[10] WU L. Design and Implementation of Intelligent Wish Filling System for College Entrance Examination[D]. Kunming: Yunnan University of Finance and Economics, 2018.

[11] YU K F, DUAN G H, SHI X. Recommendation algorithm of college entrance examination based on fuzzy clustering of multi-feature weights[J]. Journal of Central South University (Science and Technology), 2020, 51(12): 3418-3429.

[12] EOL[EB/OL]. [2021-10-10]. https://www.eol.cn/.

[13] ShangHaiRanking [EB/OL]. [2021-10-10]. https://www.shanghairanking.cn/.

[14] YangGuangGaoKao[EB/OL]. [2021-10-10]. https://gaokao.chsi.com.cn/.

[15] HOLLAND J H. Adaptation in Natural and Artificial Systems SIAM Review[EB/OL]. [2021-10-13]. https://dl.acm.org/doi/10.1137/1018105.

[16] ZHENG L P, HAO Z X. A Review on the Theory for the Genetic Algorithm[J]. Computer Engineering and Applications, 2003 (21): 50-53, 96.

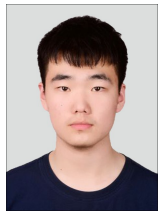
[17] BIAN X, MI L. Development on genetic algorithm theory and its applications [J]. Application Research of Computers, 2010, 27(7): 2425-2429, 2434.

[18] XU L J, LI S, YAN Z. College Entrance Examination Voluntary Recommendation System Based on Collaborative Filtering[J]. Computer System & Applications, 2015, 24(7): 185-189.

[19] Baidu[EB/OL]. [2021-10-10]. https://www.baidu.com/.

[20] Quark[EB/OL]. [2021-10-10]. https://www.myquark.cn/.

[21] LIU W D, LIU Y N. Variational autoencoder with side information in recommendation systems[J]. Journal of Tsinghua University (Science and Technology), 2018, 58(8): 698-702.



WANG Ze-qing, born in 2000, undergraduate, is a student member of China Computer Federation. His main research interests include machine learning and recommender system.

HAN Xiao-song, born in 1984, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include machine learning and optimization algorithm.