文章编号:1000-2367(2022)02-0073-06 **DOI**:10.16366/j.cnki.1000-2367.2022.02.009

融合智能审核的高考志愿推荐模型

刘行兵[a,b],王英英[a],孙钦英[a],柴斌[a],李冉[a]

(河南师范大学 a.计算机与信息工程学院;b."教育人工智能与个性化学习"河南省重点实验室,河南 新乡 453007)

摘 要:为解决考生志愿填报抉择困难、志愿合理性无从审核的问题,提出一种融合智能审核的高考志愿推荐模型.依据高考志愿填报策略和梯度划分思想,对考生高考志愿进行智能分析与合理评估,筛查不合理的志愿表单,指出问题所在,警示考生.并结合考生初始志愿表单,提取考生个性化标签,根据考生选择偏好修改、完善考生志愿,实现志愿智能审核和推荐.实例表明模型能有效降低志愿填报风险.

关键词:审核;梯度划分;Jaccard 相似度;高考;志愿填报

中图分类号:TP319 **文献标志码**:A

高考志愿填报是高考招生工作的重要环节[1],但由于高考数据冗杂繁多,考生个人定位不明确,志愿填报规则复杂[2-3]等问题,考生和家长难以在短时间内精准进行信息筛选并做出合理的志愿填报方案选择.一旦选错志愿,做错决策,将影响考生未来的就业之路[4],严重时还会造成高考落榜,带来巨大的损失.因此,利用相关技术手段帮助考生快速审核高考志愿填报方案的合理性是解决此问题的关键.

目前,高考志愿填报的相关研究主要包括:高校录取成绩预测[5-7],高考志愿推荐[8],高考志愿决策系统[9]等.沈小娟等[10]基于高考志愿录取机制,依据考生志愿选择偏好和考生位次,建立志愿填报概率模型,计算出考生被高等学校录取的概率,提高考生志愿填报的有效性.周井芝[11]针对报考中出现的问题,提出一种基于数据分析的志愿决策模型,根据考生的分数、选择偏好等因素为考生志愿填报提供决策支持.余奎锋等[12]基于 C 均值模糊聚类的多特征权重模糊均值聚类算法构建了高考志愿推荐原型系统,更好地利用考生分数,满足考生个性化志愿需求,降低了志愿填报风险.

虽然以上工作已经取得很好的效果,但现存模型大多数只实现了基础的志愿推荐功能,并没有解决如何审核考生所选高校的合理性问题.综合考虑高校特征及考生偏好对高考志愿填报的影响,实现高考志愿智能审核和个性化推荐是一项重大挑战.为了解决以上问题,提出了一种融合智能审核的高考志愿推荐模型,利用梯度划分原则和高校录取成绩预测模型对考生初始志愿进行审核,并在考生初始志愿中提取考生偏好特征,结合考生偏好来修改、完善考生志愿.最后,通过实例验证了模型的有效性.本研究为解决高考志愿填报问题提供了一个新思路.

1 融合智能审核的高考志愿推荐模型

融合智能审核的高考志愿推荐模型利用高考志愿填报策略和智能化审核方法对考生志愿进行评估,提出合理的建议和意见.算法主要运用了梯度划分思想和高校录取成绩预测技术对考生志愿进行审核,提出修正意见,并结合志愿修正意见和从考生初始志愿表单中提取的考生偏好为其再次推荐个性化志愿表单.模型技术路线如图 1 所示.

收稿日期:2021-02-24;**修回日期**:2021-07-13.

基金项目:国家自然科学基金(U1804164);河南自然科学基金(182300410306);河南省教育厅自然科学项目(17A520039).

作者简介:刘行兵(1973-),男,河南信阳人,河南师范大学副教授,博士,研究方向为机器学习、人工智能、推荐算法等,E-mail:liuxingbing@htu.edu.cn.

通信作者:王英英,E-mail:2268414907@qq.com.

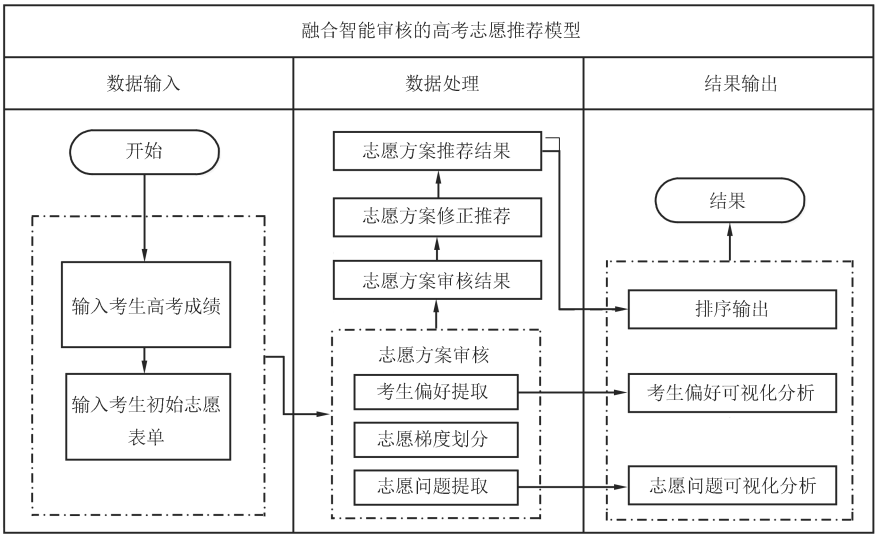

图1 模型技术路线图

Fig.1 Model technology roadmap

1.1 梯度划分

高考录取按照“分数优先,遵循志愿”的规则,考生可以填报多个高校(不同省份规定填报的高校数目不同),模型将考生填报的高校进行梯度划分,按照冲、稳、保规则进行填报.设某省份某批次最多可填报的志愿数为 R,将其分为冲、稳、保 3 个梯度,每个梯度的志愿数为$<R/3>$($<R/3>$是对 $R/3$ 取整的结果),按照同位分预测法预测每个高校的最低录取线和高校平均录取线.同位分预测法是根据院校往年的录取位次在所预测年份的一分一段表中找到对应的分数,将其作为所预测年份院校录取线预测值的方法.如(1)式所示:

$$X_F = G(F, g, h, e), \tag{1}$$

其中,$G(F,g,h,e)$ 为将位次映射为分数的函数,F 为院校录取位次,X_F 为院校录取分数,g 为考生所在省份,h 为文理科,e 为高考年份.在 g,h,e 相对固定的情况下,该映射可以通过查询对应的一分一段表获得.

根据高校最低录取线和高校平均录取线预测结果计算考生 $U_n(n=1,2,3,\cdots,n)$ 对每个高校的录取概率.概率 p 计算如(2)式所示:

$$p=\begin{cases}50\%, M=S_m,\\ \dfrac{M-S_m}{S_a-S_m}, S_m<M<S_a,\\ 99\%, M=S_a,\end{cases} \tag{2}$$

其中 S_m 表示高校最低录取线的预测值,S_a 表示高校平均录取线的预测值,M 表示考生成绩. 根据录取概率将高校划分在 3 个梯度中,第 1 个梯度为冲刺类高校,录取概率在$(0,50\%]$;第 2 个梯度为稳定类的高校,录取概率在$(50\%,80\%]$;第 3 个梯度为保底类高校,录取概率$(80\%,99\%]$.

1.2 考生偏好提取

考生志愿表单中隐含考生的选择偏好,将从志愿表单中获取的考生偏好作为考生的个性化标签.根据考生填写的初始志愿表单,在高考数据库中挖掘出高校地区、高校类型、专业类别等信息,根据此信息为考生出具志愿评价结果.若某省份某批次可以选报 m 个高校,每个高校可以填报 n 个专业.则每个考生填报的高校集合为 $C=(c_1,c_2,c_3,\cdots,c_m)$,专业为 $m\times n$ 矩阵 $\boldsymbol{A}$,如(3)式所示:

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & & \vdots\\ a_{m1} & a_{m2} & \cdots & a_{mn}\end{bmatrix}. \tag{3}$$

根据 C 找到考生志愿中每个高校的高校特征 $\mu=(\mu_1,\mu_2,\cdots,\mu_i)$,包含:高校类型特征 $t=(t_1,t_2,\cdots,t_m)$,$t\in\mu$,地区特征集合 $l=(l_1,l_2,\cdots,l_m)$,$l\in\mu$,以及高校所在城市级别特征集合 $o=(o_1,o_2,\cdots,o_m)$,$o\in\mu$ 等.其中 i 指高校特征总数.根据 $\boldsymbol{A}$ 找到考生志愿的专业特征矩阵:专业类型矩阵 $\boldsymbol{B}$,如(4)式所示:

$$\boldsymbol{B}=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix}. \tag{4}$$

1.3　志愿问题分析及推荐

结合梯度划分思想审核考生志愿,将志愿表单中不符合梯度的高校列出,分析原因,并为考生志愿提出合理的修正建议,使用融合录取概率和考生偏好信息的高考志愿推荐模型进行志愿推荐,指导考生修改和完善志愿方案.同时,为了解决有相似偏好的考生得到相同的推荐结果,导致众多考生扎堆填报同一高校的问题,本文将随机采样和 top-k 法相结合,生成最终的高校推荐结果.

(1)针对考生分数,计算考生被高校录取的概率,产生考生对应的高校候选集 Q.

(2)从考生志愿表单中提取考生偏好.构建考生对应的志愿个性化标签 $Y=(y_1,y_2,y_3,\cdots,y_m)$,其中,$m\leq R$,$y=\mu$ 为考生初始志愿表单中的高校特征集合.

(3)利用 Jaccard 系数[13]结合志愿特征计算相似度.Jaccard 系数定义为:给定两个集合 D,E,Jaccard 系数为 D 与 E 交集的大小和 D 与 E 并集的大小的比值,如(5)式所示:

$$J(D,E)=\frac{|D\cap E|}{|D\cup E|}=\frac{|D\cap E|}{|D|+|E|-|D\cap E|}, \tag{5}$$

当集合 D,E 都为空时,$J(D,E)$定义为 1.

分别计算考生志愿与候选集中每个高校的相似度.当考生对应的志愿个性化标签 $Y=(y_1,y_2,\cdots,y_m)$,候选集 Q 中高校 j 的高校特征为 $S=(s_1,s_2,\cdots,s_i)$时,考生志愿与候选集中高校的相似度为:

$$\mathrm{sim}(Y,S)=\frac{\sum_{m=1}^{R}J(y_m,S)}{R}=\frac{\sum_{m=1}^{R}\dfrac{|y_m\cap S|}{|y_m\cup S|}}{R}=\frac{\sum_{m=1}^{R}\dfrac{|y_m\cap S|}{|y_m|+|S|-|y_m\cap S|}}{R}. \tag{6}$$

(4)产生推荐集.将相似度计算后的高校集按照相似度降序排列,得到推荐列表 W,最终为用户推荐的高校数量为 I,I 的取值根据考生可填报的高校数 R 来决定,一般将 I 定为 R 的两倍.为了缩小考生选择范围,扩大选择空间,同时避免考生扎堆填报相同高校的问题,将其方法与梯度划分原则相结合,依据录取概率将推荐列表 W 也分为冲、稳、保 3 个梯度,根据已经计算好的相似度和 top-k 法(取 $k=10$),在每个梯度中,选取前 10 个高校作为初始推荐表,再用随机抽样的方法,在每个梯度中为考生随机抽取 Z 个高校,得到最终推荐结果.其中,$Z=<2\cdot R/3>$,($<2\cdot R/3>$为 $2\cdot R/3$ 取整的结果),$X=3\cdot Z$.

2　模型实例分析

以河南省 2020 年高考数据中理科一批的高校数据和考生数据为实验数据,以高考分数为 620 分的某考生 U_1 为对象进行分析,验证模型的可行性.

(1)考生输入注册信息.

(2)系统反馈:考生分差(考生分数和对应批次省控线的差):您超出一本分数线 76 分.考生位次(考生在该省份对应科类的排名):您在本省理科考生中的排名为 27 221.

(3)考生初次填写志愿表单,如表 1 所示.

(4)提取考生偏好. 根据考生志愿表单中填报的高校,挖掘出考生选择志愿的偏好信息.如表 2 所示.根据考生志愿表单中填报的专业,挖掘出其专业特征,如表 3 所示.

计算考生志愿特征对应关键词的词频率.其中,词频率=某个词在词集中出现的次数/词集总数.并针对计算好的志愿特征词频率生成可视化界面,以直观的方式展示给考生.如图 2~图 5 所示.

(5)志愿问题分析.根据志愿填报策略和梯度划分原则对考生初次填写的志愿表单进行深层次分析,列

出考生志愿存在的问题,提醒考生.并给出合理的建议,帮助考生修改和完善志愿表单.如表 4 所示.

表 1 考生 U_1 初次填写的志愿表单

Tab. 1 The first voluntary form filled out by candidate U_1

志愿	高校	专业 1	专业 2	专业 3	专业 4	专业 5
志愿 1	6 000	16	50	18	34	15
志愿 2	6 005	62	63	27	28	67
志愿 3	1 840	51	52	55	12	30
志愿 4	1 240	35	03	18	26	19
志愿 5	1 690	15	17	03	02	16
志愿 6	2 880	32	17	01	26	29

表 2 考生 U_1 的高校关键信息

Tab. 2 Key information of candidate U_1's school

志愿	高校	高校类型	省份	城市	城市级别
志愿 1	6 000	综合类	河南省	郑州市	一线城市
志愿 2	6 005	综合类	河南省	开封市	四线城市
志愿 3	1 840	综合类	江苏省	镇江市	三线城市
志愿 4	1 240	理工类	陕西省	西安市	一线城市
志愿 5	1 690	综合类	辽宁省	沈阳市	一线城市
志愿 6	2 880	理工类	江苏省	常州市	二线城市

表 3 考生 U_1 的专业关键信息

Tab. 3 Key information about candidates U_1's major

志愿	专业 1 类别	专业 2 类别	专业 3 类别	专业 4 类别	专业 5 类别
志愿 1	经济学	管理学	法学	工学	经济学
志愿 2	管理学	管理学	经济学	法学	工学
志愿 3	工学	管理学	工学	理学	工学
志愿 4	管理学	经济学	工学	工学	工学
志愿 5	工学	管理学	经济学	经济学	管理学
志愿 6	工学	工学	经济学	工学	工学

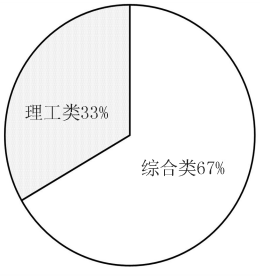

图2 高校类型分布

Fig. 2 Distribution of school types

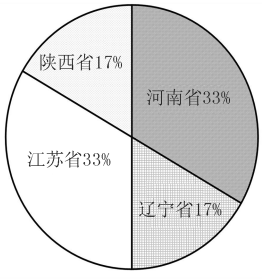

图3 省份分布

Fig. 3 Distribution of provinces

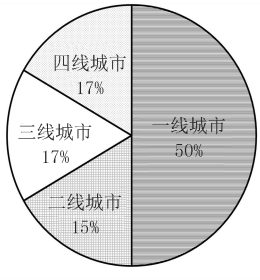

图4 城市级别分布

Fig. 4 Distribution of city levels

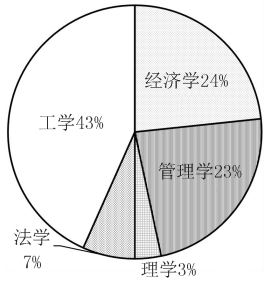

图5 专业分布

Fig. 5 Distribution of major

分析考生志愿表单可知:考生填报的第 1、5 志愿为冲刺类高校,2～4 志愿为稳妥类高校,第 6 志愿为保底类高校.从整体来看,考生填报的志愿表单冲、稳、保分配不太合理,没有按照梯度划分原则从上至下依次概率递增.可见考生志愿部分内容不符合志愿填报准则,建议考生对第 2 志愿和第 5 志愿进行修改.

表 4　考生 U_1 的志愿问题分析

Tab. 4　Analysis of candidate U_1's volunteer problems

志愿	高校	问题分析
志愿 1	6 000	第 1 志愿建议填报冲刺类高校，您填报的高校属于冲刺类，符合志愿填报原则.
志愿 2	6 005	第 2 志愿建议您填报冲刺类高校，您填报的高校属于稳妥类高校，建议您大胆一点，选择录取分数更高的高校，提高分数利用率.
志愿 3	1 840	第 3 志愿建议填报稳妥类高校，您填报的高校属于稳妥类高校，符合志愿填报原则.
志愿 4	1 240	第 4 志愿建议您填报稳妥类高校，您填报的高校属于稳妥类高校，符合志愿填报原则.
志愿 5	1 690	第 5 志愿建议您填报保底类高校，您填报的高校属于冲刺类高校，建议您更稳妥一点，降低志愿填报风险.
志愿 6	2 880	第 6 志愿建议您填报保底类高校，您填报的高校属于保底类高校，符合志愿填报原则.

(6)志愿推荐.计算考生被高校录取的概率，将符合要求的数据加入高校候选集 Q.根据志愿填报规则，考生偏好特征，高校特征结合 Jaccard 系数计算候选集 Q 中所有高校和考生志愿的相似度.部分高校相似度计算如表 5 所示.

表 5　考生志愿与部分高校相似度

Tab. 5　The degree of similarity between candidates' school forms and some colleges

高校	高校类型	省份	城市	城市级别	相似度
2 110	语言类	陕西省	西安市	一线城市	0.15
1 635	理工类	河北省	秦皇岛市	三线城市	0.10
1 550	综合类	北京市	北京市	一线城市	0.27
1 880	综合类	浙江省	宁波市	一线城市	0.27
2 210	综合类	湖北省	武汉市	一线城市	0.27
…	…	…	…	…	…
6 005	综合类	河南省	郑州市	一线城市	0.32

将计算好相似度的高校集按照录取概率划分为 3 个梯度，每个梯度的高校集按照相似度降序排列，过滤掉用户已经选择并且选择合理的高校，在每个梯度中随机选取 I 个高校给目标用户进行推荐，表 6 为当 $I=4$ 时的推荐结果.考生可以根据推荐结果，重新选择高校或者替换掉不合理的高校.

表 6　考生志愿推荐结果

Tab. 6　Results of candidates'college recommendation

第一梯度	第二梯度	第三梯度
1690，1860，2255，2105	1835，6005，1550，3890	2565，5100，6085，6105

3　结　论

为解决高考志愿填报决策困难的问题，提出一个融合智能审核的高考志愿推荐模型.结合志愿填报策略，对考生志愿进行审核，分析考生志愿存在的问题，为考生志愿填报提出建议；并根据考生选择偏好，为考生推荐个性化志愿选择方案，帮助考生修改、完善高考志愿，最后用实例验证模型可行性.由于高考志愿填报受到来自考生、高校、家长等多方面因素的影响，因此下一步将细化高考志愿填报影响因素，从多方面更客观地分析高考志愿填报问题.

参　考　文　献

[1]　李胜.传统高考志愿填报的反思及现实展望[J].教学与管理，2019(10)：30-33.
LI S. Reflection and realistic outlook of traditional college entrance examination voluntary filings[J]. Teaching & Administration，2019(10)：30-33.

[2]　吕开月.高考志愿填报倾向及影响因素研究[D].武汉：华中农业大学，2019.
LYU K Y. College choice tendency and influencing factors[D]. Wuhan：Huazhong Agricultural University，2019.

[3]　杨秀芹，吕开月.社会分层的代际传递：家庭资本对高考志愿填报的影响[J].中国教育学刊，2019(6)：24-29.

YANG X Q,LYU K Y.Intergenerational transmission of social stratification:the impact of family capital on application of college entrance examination[J].Journal of the Chinese Society of Education,2019(6):24-29.

[4] 高昌明.地方高校大学生专业选择的现状分析与对策[J].教育与职业,2015(8):106-108.
GAO C M.Intergenerational transmission of social stratification:the impact of family capital on application of college entrance examination[J].Education and Vocation,2015(8):106-108.

[5] 李敬文,陈志鹏,李宜义,等.组合预测模型在高考数据预测中的应用研究[J].计算机工程与应用,2014,50(7):259-262.
LI J W,CHEN Z P,LI Y Y,et al.Research on combinational model for predicting college entrance examination data[J].Computer Engineering and Applications,2014,50(7):259-262.

[6] 周昱彤,张跃富,刘竞泽,等.Spline 回归在高校录取分数预测及志愿推荐中的应用[J].电子技术与软件工程,2020(10):155-156.
ZHOU Y T,ZHANG Y F,LIU J Z,et al.Application of Spline Regression in College Admission Score Forecast and Volunteer Recommendation[J].Electronic Technology & Software Engineering,2020(10):155-156.

[7] LIU X B,WANG Y Y,ZHANG Z,et al.The application of SVR-based combination algorithm in applying for college in China[C]//2020 IEEE 2nd International Conference on Computer Science and Educational Informatization(CSEI).Xinxiang:[s.n.],2020.

[8] 潘月梅.基于机器学习的智能高考志愿推荐系统[D].南京:南京邮电大学,2019.
PAN Y M.Intelligent college entrance examination volunteer recommendation system based on machine learning[D].Nanjing:Nanjing University of Posts and Telecommunications,2019.

[9] 银虹宇.基于大数据的高考志愿推荐系统的设计与实现[D].成都:电子科技大学,2018.
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.

[10] 沈小娟,孙绍荣.基于统计模型的高考志愿填报决策分析[J].统计与决策,2014(21):57-59.
SHEN X J,SUN S R.Decision Analysis of College Entrance Examination Volunteer Filling Report Based on Statistical Model[J].Statistics and Decision,2014(21):57-59.

[11] 周井芝.基于数据分析的高考志愿决策模型研究[D].济南:山东师范大学,2017.
ZHOU J Z.The study of college choice decision-making model based on data analysis[D].Jinan:Shandong Normal University,2017.

[12] 余奎锋,段桂华,时翔.基于多特征权重模糊聚类的高考志愿推荐算法[J].中南大学学报(自然科学版),2020,51(12):3418-3429.
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

[13] 廖彬,张陶,于炯,等.基于二维划分的杰卡德相似系数批量计算效率优化[J].计算机科学,2017,44(1):219-225.
LIAO B,ZHANG T,YU J,et al.Efficiency optimization of jaccard's similarity coefficient based on two dimensional partition[J].Computer Science,2017,44(1):219-225.

Recommendation model for college application Integrating with intelligent review

Liu Xingbing[a,b], Wang Yingying[a], Sun Qinying[a], Chai Bin[a], Li Ran[a]

(a. College of Computer and Information Engineering; b. Key Laboratory of Henan Province for Educational Artificial Intelligence and Personalized Learning, Henan Normal University, Xinxiang 453007, China)

Abstract: This paper proposes a recommendation model for college application integrating with intelligent review, which solves the problems of the choice of candidates' college and the rationality of the plan for selected college. According to the strategy of college entrance examination and the thought of gradient division, the paper conducts intelligent analysis and reasonable evaluation on the candidates'college entrance examination, screening unreasonable application forms, and points out the problems to warn the candidates. And combined with the candidate's initial application form, the candidate's personalived lablel is extracted, modifing and improving the candidate's application form according to the candidate's choice preference, to realize the intelligent review and recommendation of the college application plan. The example shows that the model can effectively reduce the risk of college application.

Keywords: review; gradient division; Jaccard similarity; college entrance examination; apply to college

[责任编校 陈留院 赵晓华]