文章编号：2095-2163(2021)02-0069-07 中图分类号：TP391 文献标志码：A

基于决策树算法的研究生遴选质量评价

隋雨时[1]，王立明[2]

（1 上海财经大学 信息管理与工程学院，上海 200433；2 哈尔滨工业大学 计算学部，哈尔滨 150001）

摘 要：本文针对如何提高研究生遴选质量、选拔出更多优秀生源的问题，提出一种基于决策树算法的研究生遴选质量评价方法。首先通过分析研究生生源学校以及初试和复试等招生信息，同时结合对研究生的课程学习成绩、参与科研项目情况、硕士毕业论文质量的跟踪，建立了适合于计算机专业研究生质量的评价指标。然后采用经典的 ID3 决策树算法对相关数据进行分析挖掘，以评价现有研究生招生体系中各项指标对研究生培养质量的影响，并通过统计学方法对结论进行逆向分析验证。结果表明在研究生入学考核的各项指标中，面试成绩和上机考试成绩在区分考生能力、优秀研究生遴选中具有关键作用。

关键词：研究生遴选；质量评价；决策树；统计学

Quality evaluation of graduate selection based on decision tree algorithm

SUI Yushi[1], WANG Liming[2]

(1 School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China; 2 Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Aiming at the problem how to improve the quality of graduate student selection in computer science major and select excellent students, a method of graduate student selection quality evaluation based on decision tree algorithm is proposed. First of all, by analyzing the source school of graduate students and the enrollment information, such as the primary and secondary examination, combined with the tracking of the graduate students´ academic performance, participation in scientific research projects, and the quality of the master´s thesis, an evaluation index suitable for the quality of graduate students majoring in computer science is established. Then, the classic ID3 decision tree algorithm is used to analyze and mine the relevant data to evaluate the impact of various indicators in the existing graduate student enrollment system on the quality of graduate student education, and the conclusion is verified by reverse analysis through statistical methods. The results show that the interview scores and computer test scores play a key role in distinguishing the ability of examinees and selecting excellent graduate students.

【Key words】 graduate selection; quality evaluation; decision tree; statistics

0 引 言

研究生教育是高等教育的重要组成部分。研究生的综合素质和培养质量关系到专业型高技术人才的整体水平，是国家基本竞争力的一个重要体现。如何培养出更多高素质、高水平的研究生已成为目前研究生教育的一个普遍关注问题。迄今为止，国内外学者针对研究生的课程设置、培养方式及评价手段等方面已开展了大量的研究及探讨，对提高研究生培养质量起到了促进作用[1-5]。

众所周知，研究生入学的遴选质量也密切影响着高质量研究生的培养。因此，对研究生遴选质量的评价研究具有重要意义。教育发达的西方国家如美国、英国、瑞典等国家都建立了各具特色的质量保障体系[6]。例如，美国实施的是多元主体的研究生教育质量保障体系。其中，部分学校在录取研究生时会更加注重研究生的潜在能力，不完全按照考试分数由高到低录取，而是根据考生各项条件综合选择录取，确保招生质量。除了大学自身提供的质量保障之外，还有一些依赖外部质量保证制度的选拔指标，如学生在大学内的排名、标准化考试成绩等（例如，作为第三方机构主导的研究生入学考试 GRE），来提高研究生选拔的准确性。

目前国内针对研究生招生质量评估的研究，大多集中于生源质量、招生制度建设，或者生源结构的改善等方面[7]。前述研究中，或者强调研究生招生制度建设和研究生生源质量意识对招生质量的作用，或者认为提高招生质量的关键在于生源结构的改善，而没有给出合适的研究生招生质量的评价体系。

文献[8]建立了含有 8 个评价指标的研究生招

作者简介：隋雨时(2000-)，女，本科生，主要研究方向：信息管理及信息系统；王立明(1961-)，男，学士，高级工程师，主要研究方向：计算机应用技术。

通讯作者：王立明 Email：wlm@hit.edu.cn

收稿日期：2020-11-24

生评估指标体系,对目前高校研究生招生质量评价有着十分重要的参考意义。也有学者指出,完整的研究生质量评估制度应包括研究生来源质量和选拔质量,并将研究生招生质量评价指标体系设置为以下三级指标[9]:一级指标由生源质量和选拔质量构成;二级指标为评价的基本要素,包括招录比、上线比、录取率等;三级指标是具体的观测点,即对各种评价要素的进一步细化。文献[10]则提出研究生招生质量的评价指标体系主要包括整体指标和个体指标两个部分。整体指标特指反映高校研究生生源基本情况的自然信息,具体包括招生规模、报考人数、录取情况、录取成绩均值、生源来源、生源学历结构等。个体指标特指具体单个考生的实际情况,由考生的知识结构、思想道德素质、创新素质与能力、心理素质、本科期间学习成绩等指标组成,认为考生的综合素质主要直接地反映在个体指标上,并针对个体指标采用基于灰色聚类评价模型的方法对研究生招生质量进行了研究。

上述研究工作主要针对研究生入学阶段的各项数据提出相应的评价体系或评价指标,没有跟踪研究生在校期间的表现和毕业后的工作情况,事实上从入学前的复试环节,到入学后的学位论文、课业成绩以及研究生发表的高质量期刊文章、专利、获得奖学金情况等都是衡量研究生质量的重要因素[11-13]。为此,本文通过跟踪研究生入学后的培养质量以及毕业工作后的回访调查,提出基于专业基础和综合素质的研究生质量评价指标,并采用机器学习及统计分析方法反向分析各项入学考核指标,得出各项指标对高质量研究生培养的贡献权重,找出具有关键性作用的入学考核指标,为今后选拔优秀生源提供指导性意见和建议。

1 本文评价指标的建立

通过对毕业生就职企业的跟踪回访,用人单位在招聘时重点考虑的要素是应聘者的项目经历、沟通能力、实践经历、学习成绩以及组织协调能力,认为毕业生在研究生期间应当掌握专业基础知识、高级知识、方法论知识、实际知识以及外语计算机等工具类知识,并且应当具备自我学习能力、团队合作能力、实施能力、自我意识能力和沟通能力。毕业生自身普遍认为在校期间掌握的专业基础知识、英语知识、计算机等工具知识、先进知识、方法论知识,以及自我学习能力、自我管理能力、沟通能力、团队合作能力、自我意识能力和执行能力等对个人的未来发展有重要作用。因此,优秀的毕业生不仅要具备扎实的专业基础,还应具备专业理论的实际应用能力、逻辑思维能力以及良好的沟通能力等一些非专业知识和技能[12]。

为此,本文以哈尔滨工业大学计算机学院的硕士研究生为对象,通过跟踪这些学生入学后的培养过程,确定了以下评价指标:

(1)综合素质:主要以毕业论文答辩成绩、发表高质量期刊(会议)论文、发表论文数量为认定。

(2)专业基础:主要以学位课总成绩来认定。其中,毕业论文成绩在 *A* 及以上,或者学位课总成绩在前 20%,或者发表高质量期刊(会议)论文以及发表多篇期刊(会议)论文的学生被评价为优秀。

基于提出的上述评价指标,采用机器学习及统计分析的方法反向分析研究生的各项入学考核指标,得出各项入学考核指标中对最终高质量研究生培养的贡献权重,找出具有关键性作用的入学考核指标,为今后选拔优秀生源提供指导性意见和建议。研究生的入学考核指标包括:

(1)专业基础:包括初试成绩、复试成绩以及总成绩(总成绩 = 初试成绩 + 复试成绩)。

其中,初试成绩包括数学成绩、英语成绩、政治成绩以及专业课成绩;复试成绩包括机试成绩和面试成绩。

(2)综合素质:考生本科院校来源,包括 985 院校、211 院校以及普通院校。

2 基于 ID3 算法的研究生质量评价决策树

2.1 信息增益(ID3)算法

在信息学中,熵用来衡量不确定性。对于分类系统而言,类别 C 是随机变量,具体取值是 $C_1, C_2, \cdots, C_n$,每一个类别出现的概率分别是 $P(C_1), P(C_2), \cdots, P(C_n)$,其中 n 是类别总数,此时分类系统的熵可以表示为:

$$H(C) = -\sum_{i=1}^{n} P(C_i)\log_2 P(C_i), \quad (1)$$

熵越大,样本的不确定性就越大。因此可以使用划分前后集合熵的差值来衡量使用当前特征对于样本集合划分效果的好坏。信息增益(ID3 算法)就是以某个特征划分数据集前后的熵的差值,计算公式为:

$$IG(S \mid T) = HS - \sum_{value(T)} \frac{|S_V|}{S} H(S_V). \quad (2)$$

其中,S 为全部数据样例的集合;$value(T)$ 是对于属性 T 来说所有的取值集合;v 表示 T 的一个属

性值；S_v表示 S 中属性 T 取值为 v 的所有数据样例集合；$|S_v|$表示 S_v中所包含样例的个数。

信息增益（ID3 算法）的特点是使用所有特征划分数据集，得到多个特征划分数据集的信息增益，并根据结果中信息增益最大的特点进行细分，得到数据子集，当数据集中已经找不到新的属性进行节点分割，输出决策树，算法停止。在此过程中，采用自上而下的贪心搜索做出决策。

2.2　样本集合数据特征提取

样本集合包括 2 部分数据。第一部分是参加全国统一入学考试以及复试的研究生入学考核信息，包括考号、姓名、本科院校、政治成绩、英语成绩、数学成绩、专业课成绩、政治英语数学三科成绩、初试成绩、机试成绩、面试成绩、总成绩；另一部分是研究生入学后的在校成绩，包括学位课成绩、毕业论文答辩成绩、发表论文情况。同时考虑到一些硕士研究生出于直博或退学、休学等原因，导致没有毕业论文成绩，因此这部分研究生在校成绩中的相关数据无效。此外，各学校保送的研究生没有入学考试成绩，对这类数据做删除处理。将所有成绩按照分数取值范围规约为“优”、“良”、“差”三档，将毕业院校按照 985、211 及 0（普通院校）分为 3 档。

2.3　基于 ID3 算法生成决策树

首先针对样本集合数据计算特征的信息增益，确定当前分类的最佳特征，计算结果见表 1。显然，面试成绩的信息增益 0.168 9 的值最大，因此确定最优属性为面试成绩，将其选作分裂节点，生成决策树，如图 1 所示。

表 1　样本特征的信息增益计算

Tab. 1　Information gain calculation of sample features

毕业院校	政治分	英语分	数学分	专业课分	三科总分	初试分	机试分	面试分	总成绩
0.023 3	0.077 2	0.027 9	0.020 7	0.017 8	0.131 1	0.076 1	0.107 8	**0.168 9**	0.023 2

基于面试成绩一次划分的决策树可以看出，面试成绩为优秀的学生最终为优秀硕士生的概率最大，而面试成绩为良或差的学生则有不同的可能性，因此需要继续对剩余因素进行挖掘。图 2 是经过 3 次划分后获得的决策树，根据分类规则得出如下结论：在众多影响研究生未来质量的入学考核因素中，面试成绩的影响最大，其次为机试成绩和政治英语数学三科成绩。

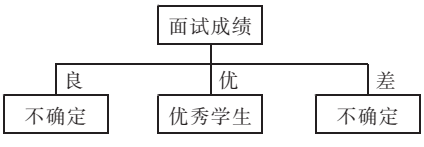


图 1　基于面试成绩一次划分的决策树

Fig. 1　Decision tree based on one-time division of interview results

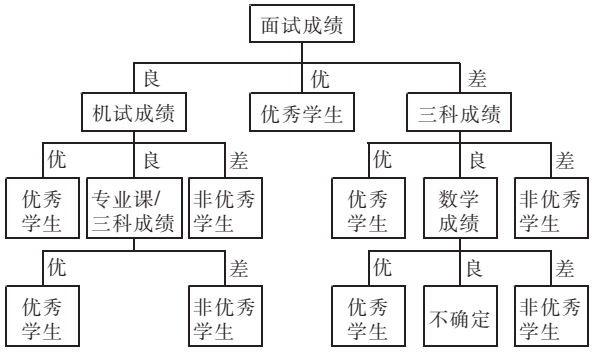


图 2　基于面试成绩三次划分的决策树

Fig. 2　Decision tree based on the three times division of interview results

因此，提高研究生招生质量首先要保证面试环节的专业性、合理性和公平性，制定出一套科学全面的面试考察体系至关重要。而且，在招生环节需要注重考生的政治英语数学三科成绩，这三科成绩突出的学生将来成为高质量研究生的概率较大。对于计算机相关专业的学生而言，上机考试是遴选高质量研究生的一个关键的因素，也是考察学生综合实力的一个重要参考。

3　基于统计学方法的分析验证

3.1　本科毕业院校类型因素分析

将优秀学生与非优秀学生的本科毕业院校类型进行统计对比，见表 2，从统计结果看，本科毕业于 985 院校的学生中，优秀学生与非优秀学生的比例基本相同，分别为 58.82% 和 60.09%，因此，本科毕业院校类型并不是遴选高质量研究生的关键性因素。

表 2　优秀学生与非优秀学生本科毕业院校类型占比

Tab. 2　Proportion of outstanding students and non-outstanding students in the type of undergraduate graduation institutions　%

本科毕业院校	985 院校	211 院校	双非院校
优秀学生	58.82	23.53	17.65
非优秀学生	60.09	10.10	29.81

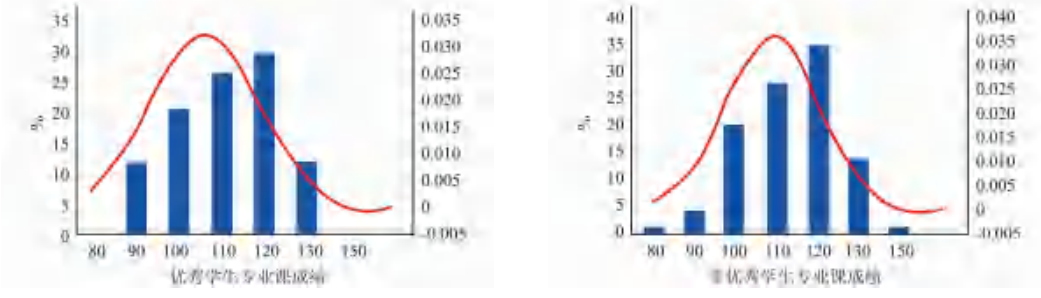
3.2　各项入学考核成绩频率分布直方图和密度曲线分析

采用频率分布直方图和密度曲线对优秀学生和

非优秀学生的各项入学考核指标进行对比分析，如图 3 所示。以初试成绩为例，研究可知，政治成绩达到 75 分及以上的学生人数优秀学生中占比为 65%，非优秀学生中占比为 61%，两者相差不多；英语成绩达到 75 分及以上的学生人数、数学成绩达到 130 分及以上的学生人数两者占据比例基本相同，前者均为 56%，后者均为 27%。由图 3 可以看出，专业课成绩方面，优秀学生中达到 120 分及以上的学生人数占比为 41%，非优秀学生中占比为 49%，高于前者 8%；综合政治英语数学三科成绩来看，达到 280 分及以上的学生人数优秀学生中占比为 21%，非优秀学生中占比为 25%；单看初试总成绩，达到 360 分及以上的学生人数优秀学生中占比为 74%，非优秀学生中占比为 71%，两者相差不大。

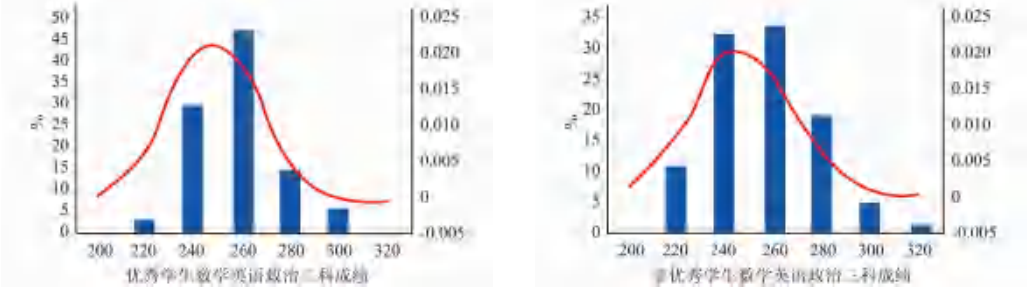
在复试方面，优秀学生中面试成绩达到 80 分及以上的学生人数占比为 85%，非优秀学生中占比为 79%；机试成绩达到 160 分及以上的学生人数优秀学生中占比为 68%，非优秀学生中占比仅为 57%。单独看总成绩，优秀学生中达到 630 分及以上的学生人数占比为 24%，非优秀学生中占比为 20%，两者差距不是很明显。

通过以上分析，可以初步得出结论：优秀研究生与非优秀研究生在初试成绩方面区分不明显，但在复试方面差别较大。如果不区分专业，面试成绩将是一个重要的影响因素。若只针对计算机专业而言，机试成绩将是另一个关键因素。



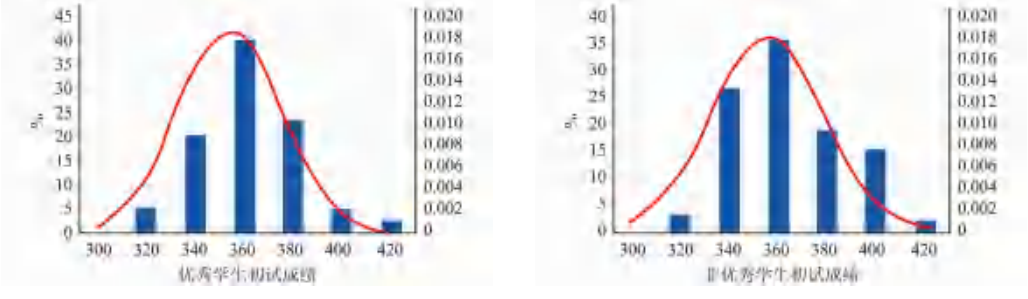
(a)专业课成绩的频率分布直方图和密度曲线

(a) Frequency distribution histogram and density curve of professional course scores



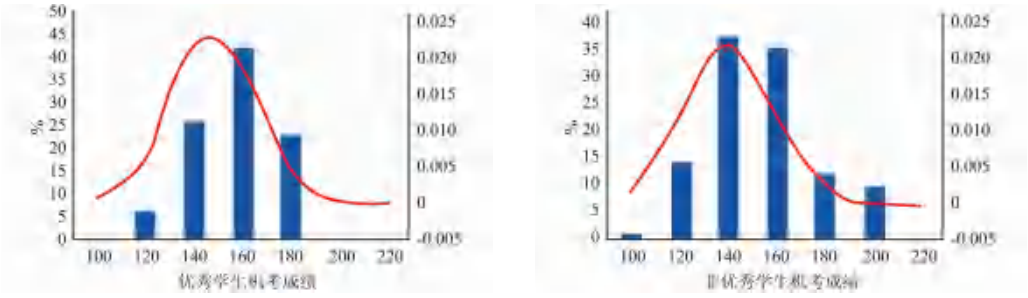
(b)数学英语政治三科成绩的频率分布直方图和密度曲线

(b) Frequency distribution histogram and density curve of the total score of mathematics, English and Politics



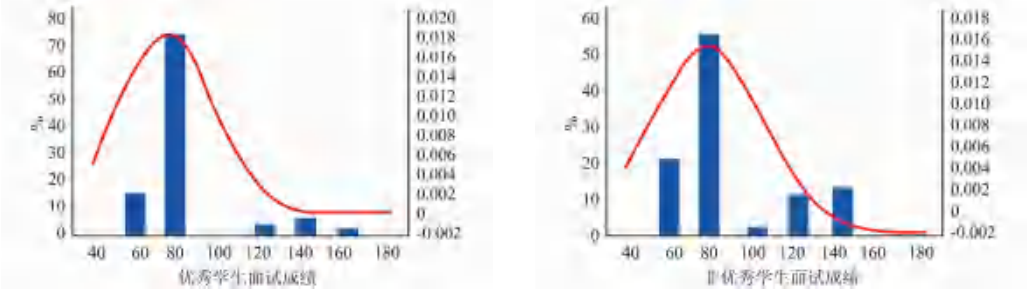
(c)初试成绩的频率分布直方图和密度曲线

(c) Frequency distribution histogram and density curve of the preliminary examination



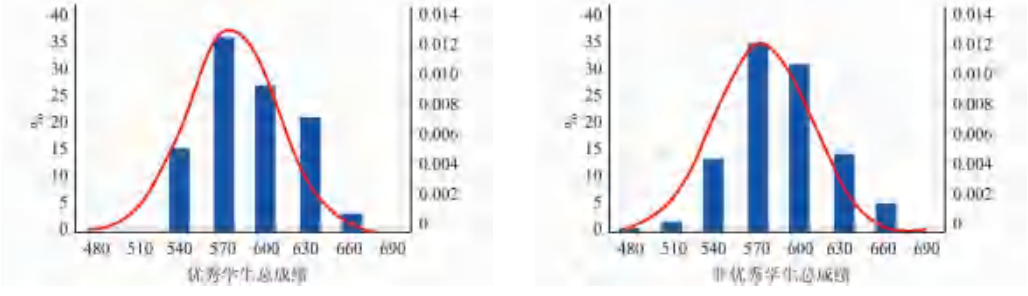
（d）机试成绩的频率分布直方图和密度曲线

（d） Frequency distribution histogram and density curve of computer test results



（e）面试成绩的频率分布直方图和密度曲线

（e） Frequency distribution histogram and density curve of interview results



（f）总成绩的频率分布直方图和密度曲线

（f） Frequency distribution histogram and density curve of total score

图 3　优秀学生与非优秀学生各项入学考核成绩的频率分布直方图和密度曲线

Fig. 3　Frequency distribution histogram and density curve of entrance examination results of excellent students and non excellent students

3.3　各项入学考核成绩的离散度分析

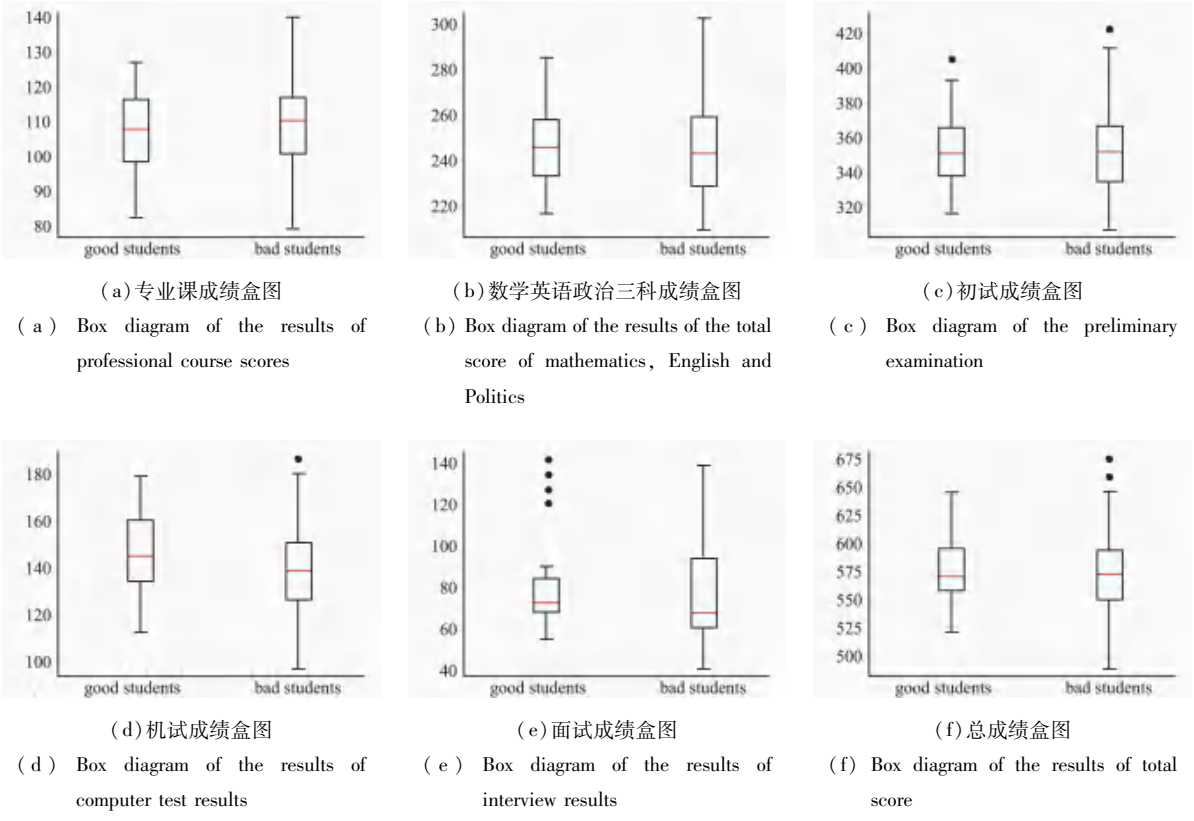
离散程度是衡量差异的一个重要尺度。如果某项指标能使得不同水平的学生在成绩上有着不同的表现，表明该指标更好地区分了候选人。将一种评价指标的优秀学生成绩和非优秀学生成绩绘制在一起可以更方便地看出两者的区别。为此，采用盒图与小提琴图对研究生各项入学考核成绩的离散度作进一步分析。

优秀学生与非优秀学生各项入学考核成绩的盒图如图 4 所示。由图 4 可以直观看出，优秀学生英语成绩、面试成绩的中位数高于非优秀学生。通过盒图对比可以看出优秀学生在机试上也有较好的表现。这说明，英语成绩、面试成绩和机试成绩在各项指标中能够更好的区分考生的优异。

考生各项入学考核成绩的小提琴图如图 5 所示，小提琴图中的垂直线是盒图，边缘是密度曲线。由图 5 可看出，优秀学生的英语成绩大多集中在 75 分段，而非优秀学生大多集中在 65 分段；优秀学生机试成绩的中位数成绩明显高于非优秀学生，而高于 140 分的学生人数占比优秀学生也明显高于非优秀学生；对比面试成绩的中位数，优秀学生的数据同样高于非优秀学生，且大部分优秀学生成绩分布在中位线以上。这表明，面试、机试和英语成绩在区分优秀生源方面发挥了积极作用。

综合研究生各项入学考核成绩的频率分布直方图和密度曲线、盒图以及小提琴图的分析结果，可以得出共性分析结论：复试环节的面试成绩、机试成绩（此项成绩针对计算机相关专业）这两个入学考核指标在最终高质量研究生遴选中具有关键性作用。

原因分析：研究生初试采用全国统一考试的形式，为了取得较高的成绩，有些学生只注重相关课程学习，因而往往可能并不具备全面的知识和能力，发展潜力不足，较难胜任研究生阶段的创新性和挑战性研究工作。因此，初试成绩在优秀研究生的遴选中并没有表现出十分明显的区分度，这与相关调查结论也相吻合，即用人单位以及高校等研究机构更喜欢在知识和能力以及综合素质方面可以充分发展的复合型人才及创新型人才。因此，研究生遴选需要更多地着重考核学生的综合能力和综合素质。



（a）专业课成绩盒图

（a） Box diagram of the results of professional course scores

（b）数学英语政治三科成绩盒图

（b） Box diagram of the results of the total score of mathematics, English and Politics

（c）初试成绩盒图

（c） Box diagram of the preliminary examination

（d）机试成绩盒图

（d） Box diagram of the results of computer test results

（e）面试成绩盒图

（e） Box diagram of the results of interview results

（f）总成绩盒图

（f） Box diagram of the results of total score

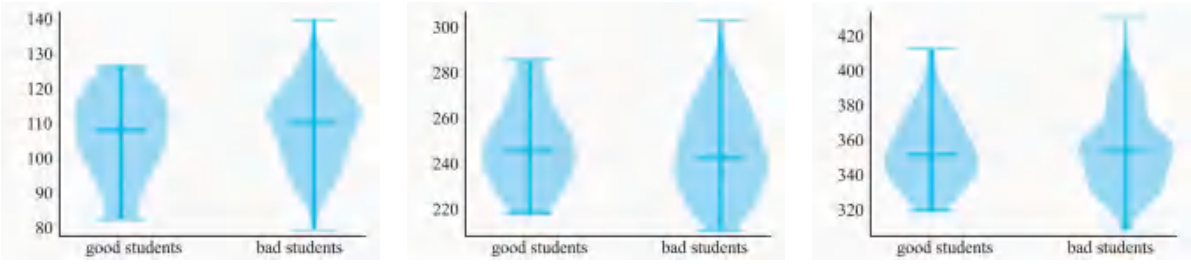
图 4 优秀学生与非优秀学生各项入学考核成绩的盒图

Fig. 4 Box diagram of entrance examination results of excellent students and non excellent students

4 结束语

通过追踪研究生入学后课程学习情况、参与科研项目以及硕士毕业论文完成质量等过程，同时结合对毕业生工作后的回访调查，提出了基于专业基础和综合素质的研究生质量评价指标。基于新的评价指标，首先采用经典的 ID3 算法构建分类决策树，实验结果表明与研究生培养质量相关的众多入学考核因素中，影响最大的是复试环节中的面试考核成绩，其次为复试环节的机试考核成绩和初试环节的政治英语数学三科成绩。再次，采用统计分析方法，利用频率分布直方图和密度曲线、盒图以及小提琴图对学生的各项入学考核指标来进行逆向分析验证，得出复试环节的面试考核成绩和机试考核成绩在高质量研究生培养中具有关键性作用的最终结论。

因此，合理规范研究生考试流程、优化复试方案、科学设置复试比重至关重要，使复试环节在区分考生能力方面的作用更加显著。此外，还要特别注重制定面试过程的标准，在提高面试权重的同时加强面试内容的包容性，使选拔过程更加客观和公正，真正遴选出“综合素质高、专业基础好、实践能力强”的优秀生源。



（a）专业课成绩的小提琴图
（a） Violin chart of professional course results

（b）数学英语政治三科成绩的小提琴图
（b） Violin chart of the total score of mathematics, English and Politics

（c）初试成绩的小提琴图
（c） Violin chart of the results of the preliminary examination

（d）机试成绩的小提琴图
（d） Violin chart of computer test results

（e）面试成绩的小提琴图
（e） Violin chart of interview results

（f）总成绩的小提琴图
（f） Violin chart of total score

图 5　优秀学生与非优秀学生各项入学考核成绩的小提琴图

Fig. 5　Violin chart of all entrance examination results of excellent students and non excellent students

参考文献

[1] HARYANA K, YOGA P N A, SUDIYANTO, et al. Mapping the graduate quality of automotive engineering education (S1) study program FT UNY[C]// Journal of Physics: International Conference on Vocational Education of Mechanical and Automotive Technology. Yogyakarta, Indonesia: IOP Publishing, 2018, 1273: 012034.

[2] TORRES-BARZABAL L M, ORTIZ - CALDERÓN M P, BARCIA-TIRADO D M. Quality indicators for auditing on-line teaching in European universities[J]. TechTrends, 2019, 63 (3): 330-340.

[3] TAI J, AJJAWI R, BOUD D, et al. Developing evaluative judgement: Enabling students to make decisions about the quality of work[J]. Higher Education, 2018, 76(3): 467-481.

[4] PATIL S M, MALIK A K. Correlation based real - time data analysis of graduate students behaviour[M]// SANTOSH K, HEGADI R. Recent trends in image processing and pattern recognition - 2nd International Conference, RTIP2R 2018, Communications in Computer and Information Science. Singapore: Springer, 2018, 1037: 696-706.

[5] DANAHER M, SCHOEPP K, KRANOV A A. Effective evaluation of the non-technical skills in the computing discipline [J]. Journal of Information Technology Education: Research, 2019, 18: 1-18.

[6] MALAGA-TOBOŁA, URSZULA, KWA Ś NIEWSKI, et al. Methods of evaluation of education quality in bologna system [C]// 19th International Multidisciplinary Scientific Geoconference, SGEM 2019. Albena, Bulgaria: EXPO SGEM, 2019 (19): 199-206.

[7] 黄静，屠中华. 高等教育大众化阶段保障研究生招生质量的思考[J]. 学位与研究生教育，2015(11)：51-55.

[8] 侯俊，陈安民. 研究生招生质量评估体系研究[J]. 学位与研究生教育，2007 (7)：22-25.

[9] 王沛. 研究生招生质量评价体系研究[J]. 乐山师范学院学报，2015，30(4)：33-136.

[10] 赵丹，易英欣. 基于灰色聚类评价模型的研究生招生质量研究[J]. 黑龙江高教研究，2014(11)：46-49.

[11] ZAN Peng, XUE Yingjie, CHANG Meihan. Research on the elimination mechanism of postgraduates under the quality assurance system of higher education[C]//Proceedings of the 2019 International Conference on Modern Educational Technology, ICMET 2019. Nanjing, China: Nanjing University of Posts and Telecommunications, 2019: 62-65.

[12] 苏小红，王甜甜，李雪，等. 研究生标准化复试对遴选生源质量的影响分析[J]. 计算机教育，2018(11)：152-159.

[13] SIQUEIRA M B. Sucupira - A platform for the evaluation of graduate education in Brazil[J]. Procedia Computer Science, 2019, 146: 247-255.