

高考志愿录取概率模型研究

陆昌辉¹, 罗永², 黄权^{1,3}, 刘青宝¹, 邓苏¹

LU Chang-hui¹, LUO Yong², HUANG Quan^{1,3}, LIU Qing-bao¹, DENG Su¹

1.国防科技大学 C⁴ISR 技术国防科技重点实验室, 长沙 410073

2.国防科技大学 理学院, 长沙 410073

3.湖南省教育考试院, 长沙 410001

1.Key Lab of Science and Technology for National Defense of C⁴ISR Technology, National University of Defense Technology, Changsha 410073, China

2.College of Science, National University of Defense Technology, Changsha 410073, China

3.Hunan Examination Center of Education, Changsha 410001, China

E-mail: lch829@vip.sina.com

LU Chang-hui, LUO Yong, HUANG Quan, et al. Matriculate probability model of national college entrance examination. *Computer Engineering and Applications*, 2010, 46(21): 14-16.

Abstract: In educational system, the national college entrance examination is one of the best important examinations until today. It affects the benefits of thousands of students. During the entrance examination, choosing fit volent is a very important step. This paper introduces a probability model to describe the situation of a student passing the entrance examination. It analyzes the distribution of the scores of matriculates in every college first, and then discusses the standardization method of the examinees' scores, and deals with the history data through it. Afterwards, the matrix is introduced to store the history data, and how to compute the matriculate probabilities is described. At last, the real matriculate data of Hunan province in 2008 is used to check this model, and the result of this experiment shows the model is rational and feasible.

Key words: national college entrance examination; volent; matriculate probability; model

摘 要: 高考是目前我国教育体制中最重要的考试之一, 它关系到千万考生的切身利益, 而志愿填报是其中的一个重要环节。采用概率模型的形式来描述高考志愿的录取情况, 首先分析了各所高校录取分数的分布情况; 接着探讨了考生成绩的标准化问题, 在此基础上对历史数据进行了处理; 然后提出了用矩阵的形式来存储历史数据, 并给出了录取概率的计算方法; 最后用湖南省 2008 年度的录取真实数据对模型进行了检验, 实验结果表明该模型是合理的和可行的。

关键词: 高考; 志愿; 录取概率; 模型

DOI: 10.3778/j.issn.1002-8331.2010.21.004 文章编号: 1002-8331(2010)21-0014-03 文献标识码: A 中图分类号: TP391

1 引言

高考是关系千万考生利益的一件大事, 志愿填报又是其中很重要的环节。一般而言, 考生在填报志愿时分为志愿明确型和志愿模糊型两类。志愿明确型的考生有自己明确的报考目标, 他们所担心的是自己能否被录取; 志愿模糊型的考生没有明确的报考目标, 只存在一种模糊的倾向, 他们急需的是有人能提供较全面的参考信息, 并推荐几所对自己来说胜算较大的学校或专业, 以便从中筛选出合适的填报方案。不管是志愿明确型还是志愿模糊型的考生, 分析其被某一学校录

取的概率显得非常重要。针对这一热点问题, 借助数理统计的有关知识, 建立了高考志愿录取概率分析模型, 并基于 2008 年湖南省高考录取数据进行了检验, 结果表明理论录取概率与实际的录取情况基本吻合。研究结果有助于提高广大考生填报高考志愿的合理性和科学性。

2 录取概率计算

2.1 录取分数的分布

经研究证明^[1-2], 每年高考考生成绩的分布服从正态分布,

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.70771110); 国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2007AA703109)。

作者简介: 陆昌辉(1976-), 男, 博士, 讲师, 主要研究方向: 数据仓库、数据挖掘、决策支持技术; 罗永(1976-), 男, 博士, 副教授, 主要研究方向: 应用数学、信息安全技术的研究; 黄权(1968-), 博士生, 高级工程师, 主要研究方向: 数据仓库与数据挖掘技术; 刘青宝(1967-), 男, 博士, 研究员, 主要研究方向: 数据仓库与数据挖掘技术; 邓苏(1963-), 男, 博士, 教授, 主要研究方向: 指挥信息系统与决策支持技术。

收稿日期: 2010-03-10 **修回日期:** 2010-05-24

并且这一结论可以通过对历年的高考数据进行分析 and 验证。

定理1 设总体 ξ 服从正态分布 $N(a, \sigma)$, $\xi_1, \xi_2, \dots, \xi_n$ 为 ξ 的子样, 子样的平均值与方差分别记为 $\bar{\xi}$ 与 S^2 , 则 $\bar{\xi}$ 服从正态分布 $N(a, \frac{\sigma}{\sqrt{n}})$, $\frac{nS^2}{\sigma^2}$ 服从自由度为 $n-1$ 的 χ^2 -分布, 简记为 $\frac{nS^2}{\sigma^2} \sim \chi_{(n-1)}^2$, 而且 $\bar{\xi}$ 与 S^2 相互独立。

通过对子样作相应正交线性变换, 运用数理统计的相关知识不难得出上述结论, 这里就不再给出具体证明过程。

根据定理1可知, 各高校各专业录取的考生都可以视为高考成绩总体 ξ 的一个子样 $\xi_1, \xi_2, \dots, \xi_n$, 由于高考成绩总体 ξ 是服从正态的, 从而各高校各专业录取的考生成绩 $\xi_1, \xi_2, \dots, \xi_n$ 也服从正态分布。

2.2 考生成绩标准化

由于每年高考题目的难易程度存在客观的差异, 考生人数和考生的水平也在逐年发生变化, 同时各高校的招生计划

$$I^1 = \begin{bmatrix} \text{清华大学-计算机} & \text{清华大学-生物学} & \dots & \text{清华大学-管理学} \\ \text{北京大学-计算机} & \text{北京大学-生物学} & \dots & \text{北京大学-管理学} \\ \vdots & \vdots & \ddots & \vdots \\ \text{XX大学-计算机} & \text{XX大学-生物学} & \dots & \text{XX大学-管理学} \end{bmatrix}_{p \times q} \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} \begin{array}{l} \text{北京市} \\ \text{广东省} \end{array}$$

每年也都在改变。受这些因素的影响, 同一考分的考生在不同年份的录取结果很可能大相径庭, 就引出同一考分在不同年份的等效问题。每年高考分值的分布都服从正态分布, 可以采用等效处理方法^[3]对其进行处理: 将每年的分值转化为标准正态分布, 然后再用相同方法扩大和平移, 完成历年高考分值的等效。具体描述如下:

以湖南省为例, 设: 某年湖南省有 n 名考生参加高考, 高考分数 X 服从正态分布, 即 $X \sim N(a, \sigma^2)$, 其中 $a = \frac{\sum_{i=1}^n X_i}{n}$ 为所有考生的平均分, $\sigma = \sqrt{\frac{\sum_{i=1}^n (X_i - a)^2}{n}}$ 为标准差。

如果实际考分为 X , 则等效分计算公式为: $T = [100 * \frac{X-a}{\sigma} + 500]$ 。

其中 $y = [x]$ 为取整函数, 即 y 取不超过 x 的最大整数。

若 $T > 900$, 记

$$T = 900 \quad (1)$$

若 $T < 100$, 记

$$T = 100 \quad (2)$$

故 $T \in [100, 900]$ 。

因为 $X \sim N(a, \sigma^2)$, 所以 $\frac{X-a}{\sigma} \sim N(0, 1)$ 。又由 $T \in [100, 900]$ 知 $\frac{X-a}{\sigma} \in [-4, 4]$, 而标准正态分布的密度函数为 $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}$, 于是有 $P = \int_{-4}^4 p(x) dx = 0.9999$ 。这表明, $\frac{X-a}{\sigma}$ 落在区间 $[-4, 4]$ 上的概率为99.99%, 因此式(1)、(2)的近似处理是合理的。

2.3 历史数据处理

取 m 年全省的历史录取数据作为模型计算概率的数据基础。设: 在历史录取数据中, 第1年有 n_1 名考生, 第2年有 n_2 名考生, $\dots$, 第 m 年有 n_m 名考生。将历年高考分数按照上述方法进行标准化, 得出第 j 年第 i 名考生的等效分为 $T_i^{(j)} = [100 * \frac{X_i^{(j)} - a^{(j)}}{\sigma^{(j)}} + 500]$ 。其中, $X_i^{(j)}$ 为第 j 年第 i 名考生的实际高考分数, $a^{(j)} = \frac{\sum_{i=1}^{n_j} X_i^{(j)}}{n_j}$ 为第 j 年全省考生的高考平均分, $\sigma^{(j)} = \sqrt{\frac{\sum_{i=1}^{n_j} (X_i^{(j)} - a^{(j)})^2}{n_j}}$ 为第 j 年全省考生成绩的标准差, $i = 1, 2, \dots, n_j, j = 1, 2, \dots, m$ 。

将高考分值标准化后, 历年各高校各专业录取的等效分就集中到同一个样本空间。

2.4 信息存储矩阵设计

在计算录取概率之前, 需要将处理过的历史录取信息进行特定的数据结构形式进行存储。下面设计了4个矩阵: I^1, I^2, I^3, I^4 , 分别用于存放一批本科, 二批本科, 三批本科以及专科的历史录取信息, 其中矩阵的行指标为学校代码, 列指标为专业代码。以一批本科为例建立矩阵 I^1 如下:

I^1 是一个 $p \times q$ 的矩阵, 由于存放了所有 p 所一批本科学校的所有 q 门专业的录取信息, 其中 I^1 的每个元素均为一个数组单元, 存放被该学校该专业录取的考生的高考成绩等效分。

例如: 元素 $I^1(1, 2)$ 表示清华大学生物学专业的录取信息, 该元素中存放的数组为湖南省 m 年内所有被该学校该专业录取考生的等效高考分值, 记为:

$$I_{i,j}^1 = [T_1^{(1)}, T_2^{(1)}, \dots, T_{n_1}^{(1)}, T_1^{(2)}, T_2^{(2)}, \dots, T_{n_2}^{(2)}, \dots, T_1^{(m)}, T_2^{(m)}, \dots, T_{n_m}^{(m)}]$$

2.5 录取概率计算

下面以一批本科为例详细阐述录取概率的计算, 其他批次的录取概率均与一批本科录取概率的计算方法相同。由于元素 $I_{w,s}^1$ 所存放的数组服从正态分布, 因此, 只需要求出其均值和方差就能完整地描述出该数组的统计规律, 进而计算相应录取概率。

首先根据矩阵 $I_{p \times q}^1$ 求期望矩阵 $E_{p \times q}^1$ 及方差矩阵 $D_{p \times q}^1$ 的估计, 其中元素 $E_{w,s}^1$ 表示第 w 所大学第 s 门专业的录取期望, $D_{w,s}^1$ 表示第 w 所大学第 s 门专业的录取方差。

(1) $E_{w,s}^1$ 的估计

定理2^[4] 总体 ξ 服从正态 $N(a, \sigma^2)$, $\xi_1, \xi_2, \dots, \xi_n$ 为样本, 则统计量: $T = \sqrt{n-1} \frac{\bar{\xi} - a}{S}$, 服从自由度为 $n-1$ 的 t -分布, 记作 $T \sim t_{(n-1)}$ 。

其中 $\bar{\xi} = \frac{\sum_{i=1}^n \xi_i}{n}$ 为样本均值, $S^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$ 为样本方差。

由于一批本科中被第 w 所大学第 s 门专业录取的考生分数 $T_{w,s}^1 \sim N(E_{w,s}^1, D_{w,s}^1)$, 记

$$\bar{T}_{w,s}^1 = \frac{\sum_{l=1}^m \sum_{k=1}^{n_l} T_k^{(l)}}{\sum_{l=1}^m n_l}, S_{w,s}^1 = \sqrt{\frac{\sum_{l=1}^m \sum_{k=1}^{n_l} (T_k^{(l)} - \bar{T}_{w,s}^1)^2}{\sum_{l=1}^m n_l}}$$

分别为 $T_{w,s}^1$ 的样本均值与样本方差。其中, $T_k^{(l)}$ 为第 l 年被第 w 所大学第 s 门专业录取的第 k 名考生的等效分, n_l 为第 l 年第 w 所大

表1 理科610分考生被各学校录取的概率

学校名称	录取概率/(%)	录取平均分	学校名称	录取概率/(%)	录取平均分
中国人民大学	21.54	623.475	天津大学	89.94	595.750
同济大学	40.64	620.262	厦门大学	95.06	597.185
北京航空航天大学	42.51	612.063	哈尔滨工业大学	96.63	596.180
北京邮电大学	67.90	603.443	东南大学	98.58	590.850
华中科技大学	75.28	602.274	中南大学	99.59	580.599
武汉大学	79.97	598.797	湖南大学	99.91	580.618
北京理工大学	82.13	603.040	—	—	—

表2 各学校的热门专业

学校名称	专业名称	最低录取概率/(%)	学校名称	专业名称	最低录取概率/(%)
北京邮电大学	通信工程	73.82	哈尔滨工业大学	电气工程及其自动化	75.99
湖南大学	土木工程	93.86	哈尔滨工业大学	机械设计制造及其自动化	83.82
湖南大学	国际经济与贸易	85.03	华中科技大学	机械设计制造及其自动化	90.62
厦门大学	会计学	95.06	华中科技大学	光电信息工程	79.51
厦门大学	化学类(含化学、化学生物)	87.76	—	—	—

学第s门专业在湖南省的招生数量,m表示共取了m年的历史录取信息。由定理2可知:统计量 $\sqrt{\sum_{l=1}^m n_l - 1} \frac{\bar{T}_{w,s}^1 - E_{w,s}^1}{S_{w,s}^1} \sim t_{(\sum_{l=1}^m n_l - 1)}$,不

难得出 $[\bar{T}_{w,s}^1 - t_{(\sum_{l=1}^m n_l - 1)}(\alpha) \frac{S_{w,s}^1}{\sqrt{\sum_{l=1}^m n_l - 1}}, \bar{T}_{w,s}^1 + t_{(\sum_{l=1}^m n_l - 1)}(\alpha) \frac{S_{w,s}^1}{\sqrt{\sum_{l=1}^m n_l - 1}}]$

是 $E_{w,s}^1$ 的置信水平为 $1 - \alpha$ 的区间估计。取 $E_{w,s}^1$ 的估计为:

$$\hat{E}_{w,s}^1 = \bar{T}_{w,s}^1 \tag{3}$$

(2) $D_{w,s}^1$ 的估计

定理3^[4] 总体 ξ 服从正态 $N(a, \sigma^2)$, $\xi_1, \xi_2, \dots, \xi_n$ 为样本,则统计量: $\frac{nS^2}{\sigma^2} \sim \chi_{(n-1)}^2$ 。其中 $S^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$ 为样本方差。

由定理3可知 $\frac{(\sum_{l=1}^m n_l)S_{w,s}^2}{D_{w,s}^1} \sim \chi_{(\sum_{l=1}^m n_l - 1)}^2$,不难得出 $[\frac{(\sum_{l=1}^m n_l)S_{w,s}^2}{\chi_{(\sum_{l=1}^m n_l - 1)}^2(\frac{\alpha}{2})}, \frac{(\sum_{l=1}^m n_l)S_{w,s}^2}{\chi_{(\sum_{l=1}^m n_l - 1)}^2(1 - \frac{\alpha}{2})}]$ 为方差 $D_{w,s}^1$ 的置信水平为 $1 - \alpha$ 的区间估计。取 $\chi_{(\sum_{l=1}^m n_l - 1)}^2(1 - \frac{\alpha}{2})$

$D_{w,s}^1$ 的估计为:

$$\hat{D}_{w,s}^1 = S_{w,s}^2 = (S_{w,s}^1)^2 \tag{4}$$

由式(3)、(4)的计算结果知:第w所大学第s门专业录取分数的概率密度函数为:

$$p_{w,s}(x) = \frac{1}{D_{w,s}^1 \sqrt{2\pi}} e^{-\frac{(x - E_{w,s}^1)^2}{2D_{w,s}^1}}$$

定义1 当今年考生的高考等效分为T时,他被第w所大学第s门专业的录取概率为 $P_{w,s} = \int_{100}^T p_{w,s}(x) dx$ 。

如果将计算的样本设为各高校录取结果,而没有细化到具体专业,同理可得出考生被相应高校录取的概率。

3 实验验证

下面用湖南省历年的录取数据来进行实验验证:首先根据2004~2007年各所高校在湖南省的录取结果作为样本数据

求得相应的概率密度函数,然后参考定义1计算出各个分数在各所高校的录取概率,最后对照2008年的实际录取结果来进行比较分析。由于篇幅的关系,下面就从实验结果中随机挑选出一部分来进行说明。表1列出610分的理科考生在中国人民大学等12所一批本科院校的录取概率,表2显示的是根据理科考生的最低录取概率计算得出的各所高校的热门专业。

从表1可以看出,2008年610分的理科考生在上述12所学校中被中国人民大学录取的概率最低,被湖南大学录取的概率最高;天津大学录取的平均分虽然较厦门大学和哈尔滨工业大学的低,而610分的考生被天津大学录取的概率相反却比厦门大学和哈尔滨工业大学的都低;中南大学相对湖南大学的情况也是如此。对照2008年的实际录取结果,某一分数考生被各所学校录取的概率基本上与该学校录取的平均分成反比,这与实际情况是吻合的。对于天津大学和中南大学的异常情况,进一步对照具体录取数据,发现在天津大学录取的学生中集中了大量的分数,导致了理科610分这个分数对于天津大学来说其录取概率反而比厦门大学和哈尔滨工业大学的低,中南大学相对湖南大学的情况也如此,这与实际情况也是一致的。

从表2则可以看出,各所学校的一些热门专业相对于该校的录取概率来说均维持在一个较高的水准。结合表1和表2,2008年理科610分的考生在北京邮电大学他上不了通信工程这一热门专业,也不能被华中科技大学的机械设计制造及其自动化、光电信息工程这两个热门专业录取;但对于湖南大学、哈尔滨工业大学来说他可以任意挑选专业;对于厦门大学来说,选择化学类专业基本没问题,但要想选择会计学这一优势专业还存在一定风险。

4 结论

将概率模型应用于高考志愿填报辅助决策中,且在正态分布相关理论的支撑下给出了录取概率的计算方法,理论结果和实际录取情况符合得较好。研究成果对于广大高考考生有效地认识和规避志愿填报的风险,提高志愿填报的科学性和合理性有着重要的辅助作用。但是,在研究中,尚未考虑高考录取中出现的大小年现象^[5]带来的影响,以及招生计划的变化、考生在填报志愿时的博弈心理等因素的影响^[6],这在今后的研究中还有待于进一步深入。

(下转24页)

(5.12, 5.12)、(5.12, -5.12), 函数值为 2 748.78。运行 LINGO 软件和 MATLAB 内部函数, 都会落入上述 4 个局部极值点中。采用有一定技巧的遗传算法在遗传 300 代后才能以 90% 的几率获得最优解^[14]。采用量子竞争决策算法当最大迭代次数超过 80 次时均可获得最优解 3 600, $X=(0, 0)$ 。

为进一步测试算法对高维函数优化的性能, 选取 Rastrigrin 函数进行测试。Rastrigrin 函数是典型的 benchmark 函数, 绝大部分点是连续不可微的, 多峰, 有多个局部极小点。定义如下:

$$f(x) = \sum_{i=1}^n (x_i^2 - 10 \cos(2\pi x_i) + 10)$$

x_i 的取值范围为 $[-5.12, 5.12]$, 全局最优值为 $x_i = 0, f = 0$ 。

表 1 Rastrigrin 函数计算结果表

维数 n	最优值	平均值	标准差
20	0	0	0
30	0	0	0

从对这些典型复杂函数的优化情况来看, 算法能很好地避免陷入局部最优, 达到全局优化和快速收敛的目的。这主要得益于将进化博弈论中学习和调整的动态优化机制引入到算法中, 通过竞争决策及量子门更新, 实现竞争者自学习和自优化的目的, 提高竞争者的优化智能, 增强寻优能力; 同时算法充分利用量子计算中叠加态等理论, 增加竞争群体的多样性, 进一步提高算法的优化能力。竞争决策算法目前仅用于求解离散型组合优化问题, 进化博弈论主要用于经济生物等领域, 而在优化问题中应用很少, 算法扩大了这两者的应用范围, 也为函数优化提供了新的求解方法。

5 结束语

借鉴进化博弈论中的学习和调整的思想, 同时充分利用竞争决策的方法和量子进化计算的特点, 提出了量子竞争决策算法。通过对典型函数的优化实验, 表明算法是一种有效的全局优化算法。将算法应用于多目标优化问题是今后深入研究的课题。

参考文献:

- [1] 宁爱兵, 马良. 竞争决策算法及其在车辆路径问题中的应用[J]. 管理科学学报, 2005, 8(6): 10-18.
- [2] 宁爱兵, 马良. 大规模旅行商问题(TSP)的竞争决策算法[J]. 计算机工程, 2005, 31(9): 23-26.
- [3] 宁爱兵, 马良. 最小比率旅行商(MRTSP)问题竞争决策算法[J]. 计算机工程与应用, 2005, 41(11): 30-32.
- [4] 宁爱兵, 马良. 度约束最小生成树(DCMST)的竞争决策算法[J]. 系统工程学报, 2005, 20(6): 630-634.
- [5] Narayanan A, Moore M. Quantum-inspired genetic algorithm[C]// IEEE Congress on Evolutionary Computation, Nogyaya, 1996: 61-66.
- [6] Han K H, Kim J H. Quantum-inspired evolutionary algorithm for a class of combinatorial optimization[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(6): 580-593.
- [7] 王凌. 量子进化算法研究进展[J]. 控制与决策, 2008, 23(12): 1321-1326.
- [8] 解平, 李斌, 庄镇泉. 一种新的混合量子进化算法[J]. 计算机科学, 2008, 35(2): 166-170.
- [9] Fundenberg D, Levine D K. The theory of learning in games[M]. USA: MIT Press, 1998.
- [10] 付立群. 进化博弈论: 经济学方法论的一次革命[J]. 武警工程学院学报, 2004, 20(4): 17-22.
- [11] 李万, 田盛丰, 黄厚宽. 进化博弈论及 Agent 自组织动力学[J]. 计算机研究与发展, 2006, 43(增刊): 46-50.
- [12] 刘伟兵, 王先甲. 进化博弈中多代理人强化学习模型[J]. 系统理论与实践, 2009, 29(3): 28-33.
- [13] 谢识予. 有限理性条件下的进化博弈论[J]. 上海财经大学学报, 2001, 3(5): 3-9.
- [14] 马良, 朱刚, 宁爱兵. 蚁群优化算法[M]. 北京: 科学出版社, 2008.
- [15] 张葛祥, 李娜, 金炜东, 等. 一种新量子遗传算法及其应用[J]. 电子学报, 2004, 32(3): 476-479.
- [16] 崔玲丽, 肖志权. 实数编码遗传操作机制的研究[J]. 系统仿真学报, 2003, 15(11): 1577-1579.
- [17] 刘勇, 康立山, 陈毓屏. 非数值并行算法(第二册) - 遗传算法[M]. 北京: 科学出版社, 1998.
- [3] 叶宗裕. 高考成绩综合时的标准分研究[J]. 统计与决策, 2005(5): 25-26.
- [4] 菲诗松. 高等数理统计[M]. 北京: 高等教育出版社, 2003.
- [5] 李光辉. 巧用“平均分均值”避免“大年”风险[J]. 高校招生, 2007(6): 41-42.
- [6] 许金龙, 张艺执. 高考志愿的那些事儿: 填报技术手册: 填报志愿四大“疑难”[J]. 招生考试通讯: 高考版, 2009(5): 6-10.

(上接 16 页)

参考文献:

- [1] 舒斌, 谢友芹. 统计分析在学生成绩评估中的应用[J]. 系统工程理论与实践, 2006, 26(7): 134-140.
- [2] 张丹钢. 试卷成绩分布的正态性检验方法[J]. 考试周刊, 2008(48): 5-7.