基于高校招生数据的数据分析及应用研究

摘要

本文主要通过爬虫技术对中国高校网披露的2007年至2017年除国防科技大学外38所985高校在各省文理科专业录取分数情况，并根据实际要求对原数据进行数据清洗、缺失值填补、分组整合等处理，结合特征工程对数据进行探索性分析，统计处各高校近年来开设专业数量的变化趋势、最具竞争力、潜力及最不具竞争力、潜力专业。

在模型方面，本文巧用批次投档分及状元总分为对照目标，利用深度学习算法LSTM（长短记忆神经网络）预测出下年各专业相对门槛，并通过层次分析法模型为考生两两对比所有决策方案，建立一个高考招生推荐系统；为其推荐相应高校和专业。LSTM的预测精度非常高，具有很好的现实意义，且层次分析法中判别矩阵的一致性检验机制能尽量降低因个人主观对比偏差过大而导致模型失去客观性，整体模型实用而科学。

关键词：爬虫；LSTM；层次分析法；高校招生；推荐系统

Abstract

This article mainly uses the crawler technology to disclose the scores of the 985 colleges and universities in the provinces and universities except the National University of Defense Technology in 2007 to 2017. In addition, according to the actual requirements, the original data is cleaned and the missing values are filled. Group integration and other processing, combined with feature engineering to exploratory analysis of data, the statistical offices of various universities in recent years to open a professional number of trends, the most competitive, potential and least competitive, potential professional.

In terms of model, this paper uses the batch voting score and the total score of the champion to be compared with the control target. The deep learning algorithm LSTM (long and short memory neural network) is used to predict the relative threshold of each major in the next year, and the analytic hierarchy process model is used for two pairs of candidates. More than all decision-making programs, establish a college entrance examination recommendation system; recommend corresponding universities and majors. The prediction accuracy of LSTM is very high, and it has good practical significance, and the consistency check mechanism of discriminant matrix in AHP can try to be as far as possible. Reducing the model's loss of objectivity due to excessive subjective contrast bias, the overall model is practical and scientific.

Key Words： reptile; LSTM; analytic hierarchy process; college admissions; recommendation system

目录

1.前言

1.1 研究背景

自 1977 年恢复高考制度以来，各大高校院系不断根据科学体系的发展而调整开设的专业门类，并在至今 40 年间招收了无数来自不同地区、不同类型的学生。这些考生的录取结果共同汇成了一个大型数据库，并每年为下一届的高考考生作为选择大学的参考文库。高校间、各专业间的录取分数对比可以体现出高校和专业间的受欢迎程度，同时也能间接反映出社会与学校间不同专业人才的需求供应关系。

而作为考生，面对历年各大高校数百万条录取情况，不容易发现其中的规律或是专业趋势，导致投档了与自己实力不符合的学校及专业，从而影响了人生格局等。因此为广大考生建立一个高考录取招生分析系统有着现实需求及价值。

1.2 研究意义

招生数据分析也具有一定的理论意义。随着全球信息化的发展，信息技术广泛应用于各个领域。2002 年，我国实现高等学校招生网上录取，为招生工作带来了革命性的变化，同时也使得信息技术为招生工作服务成为可能。十几年来，计算机管理着大量的招生数据。招生管理系统只是对原始的招生数据进行简单的汇总和加工，无法转换成有价值的信息以对下一步的工作进行指导。而数据挖掘技术能够对大型、复杂的数据集进行高效、快速的分析，能够发现隐藏在数据背后的数据之间的关联、趋势及方向。

招生数据分析具有很重大的实际意义，在数百万条考生数据中，蕴藏着丰富的决策信息和知识。如何开发和利用这些宝贵的信息资源，是服务高考录取工作、指导考生科学地填报志愿的一项重要任务，也是目前迫切需要解决的问题。高考志愿分析是高考数据分析的一个重要组成部分，它主要以考生志愿数据为依据，根据志愿相关的各个属性分析高考录取的情况。其中得到的各种知

识，无论是直观的统计信息还是隐含的知识模式，都能够为考生提供丰富的决策支持信息。

1.3 国内外相关研究状况综述

信息技术的飞速发展使得各行各业积累了大量的数据，随着管理信息化的推进和行业业务需求的增大，人们并不满足于对现有数据的简单查询和分析，传统的数据管理方法已不能深入探索数据背后的含义。数据挖掘技术应运而生，该技术就是帮助人们从海量数据中提取有效的、隐含的、潜在有用的知识以优化和促进相应行业的信息化管理和发展。目前，国外数据挖掘发展趋势的研究方面主要有：对知识发现方法的进一步发展，KDD 与数据库的紧密结合等。经过不到二十年的发展，数据挖掘技术已经在诸多领域得到了广泛的应用，也逐步在教育行业中发挥一定作用。目前，高考志愿数据已形成数据仓库，高考相关数据已经便于数据挖掘者使用数据挖掘方法进行多角度的查询与统计、深层次的分析处理以及知识提取。

在当下的中国，将数据挖掘技术应用到高考志愿分析中还在理论研究阶段。每年高考考生由于志愿填报不当而未能就读理想学校、专业甚至落榜的现象十分普遍。目前现有的高考志愿分析主要依赖统计学的方法与技术，而使用数据挖掘技术对高考志愿相关数据进行深入分析，得出相关规则对考生填报志愿进行预测性指导的分析系统在我国尚处于不成熟阶段。

每年高考过后，教育考试部门都会形成大量的高考数据，包括考生信息、考生成绩、报考信息等等。 高考阅卷的信息化管理和实施使得相应的教育考试部门积累了大量的高考数据，这其中包括多年的考生报考数据、考生成绩数据、考生志愿数据等等。目前，教育研究者们对运用计算机及网络技术的高考数据分析的研究在不断进行中。国内对高考数据的分析与挖掘主要集中两大块，一方面是针对高考成绩及考生答题数据的分析与挖掘研究，另一方面研究者们主要从指导高校招生及考生报考的角度来进行研究。而本课题是围绕招生及报考的数据来分析研究，但也会涉及到一些科目的数据。比如一名英语成绩好的学生更有可能报考什么样的大学。

1.4 本文的特色与创新

本文先通过网络爬虫，获取近 38 所 985 高校（剔除国防科技大学）各专业在各个省近 10 年录取情况，字段信息包括有年份、考生地区、高校名称、录取平均分、最低分、最高分等。从数据量来看， 共有近 24 万条样本数据，确保了本次研究结果的科学性、可靠性；从信息维度来看，在正式研究前通过特征工程引入了诸多于专业排名等变量信息，以最大程度程度上解决了不同省份之间高考成绩的差异性。

从研究方法上，本文运用了大量的机器学习方法及统计学上的数据预处理手段，以最大程度上反映出数据内在的逻辑特征，发掘数据间的潜在联系、获取有价值的信息，并对结果进行可视化的分析。

从研究成果上看，本研究通过建立层次分析法结合 LSTM 等数学模型，针对考生自身的高考成绩及对学校和专业的偏好，本着以录取成功率最大为最优目标为该考生进行推荐相关的高校及专业。即本研究最终的目的为构建一个推荐系统，为考生推荐最符合他们的专业及对应高校。这一想法较为新颖，具有一定的创新性。

1.5 本文的主要内容

本文的主要内容分有以下几个部分: 第一章为绪论，主要介绍该课题的背景以及意义，对前人的智慧结晶进行简单的回顾，并述本文的特色和创新。

第二章主要是大致说明了一下数据情况，在进行探索性数据分析的同时进行数据清洗、分组处理，提取中关键的特征，为后面建模做铺垫。并对数据逻辑特征进行可视化显示等。

第三章大致解释 LSTM 算法原理，对各高校专业下年的招生门槛进行预测，并针对现有数据进行建模，构建一个基于层次分析法的推荐系统，针对考生自身的成绩实力及对学校和专业的偏好情况为其推荐相关的专业和学校。

第四章为总结部分，对模型进行优劣评价，并对全文做一个总结，提出出今后研究方向。

2. 高校招生数据分析

2.1 数据说明

本次研究主题数据为 2007 年至 2017 年十一年间 38 所 985 高校（分别为清华大学、北京大学、 厦门大学、中国科学技术大学、南京大学、复旦大学 、天津大学、 哈尔滨工业大学、浙江大学、南开大学、西安交通大学、华中科技大学、 东南大学、武汉大学、上海交通大学 、中国海洋大学、山东大学 、湖南大学、中国人民大学、 北京理工大学、 吉林大学 、中国农业大学、西北农林科技大学 、华东师范大学、中央民族大学 ）在各省（不含港澳台）中对各专业人才的录取情况。数据来源主要来源于网络公开数据网站，如高考网[1]等，利用 python 技术对其进行爬虫，并把爬取得到数据保存到本地进行分析研究。爬取目标 URL 的大致界面如下：

专业名称	高校名称	平均分	最高分	考生地区	科别	年份	批次	专业对比
英语	中国人民大学	--	649	北京	文科	2017	第一批	加入对比
新闻传播学类	中国人民大学	--	650	北京	文科	2017	第一批	加入对比
新闻传播学类	中国人民大学	--	665	北京	理科	2017	第一批	加入对比
国际政治	中国人民大学	--	660	北京	文科	2017	第一批	加入对比
国际政治	中国人民大学	--	662	北京	理科	2017	第一批	加入对比
法学	中国人民大学	--	659	北京	文科	2017	第一批	加入对比
法学	中国人民大学	--	667	北京	理科	2017	第一批	加入对比
社会学类	中国人民大学	--	654	北京	文科	2017	第一批	加入对比
社会学类	中国人民大学	--	662	北京	理科	2017	第一批	加入对比
统计学类	中国人民大学	--	668	北京	理科	2017	第一批	加入对比
农业经济管理类	中国人民大学	--	663	北京	理科	2017	第一批	加入对比
信息资源管理	中国人民大学	--	659	北京	文科	2017	第一批	加入对比
信息资源管理	中国人民大学	--	662	北京	理科	2017	第一批	加入对比
经济学类	中国人民大学	--	657	北京	文科	2017	第一批	加入对比
经济学类	中国人民大学	--	668	北京	理科	2017	第一批	加入对比
政治学、经济学与哲学	中国人民大学	--	655	北京	文科	2017	第一批	加入对比
政治学、经济学与哲学	中国人民大学	--	664	北京	理科	2017	第一批	加入对比
财政学类	中国人民大学	--	653	北京	文科	2017	第一批	加入对比
财政学类	中国人民大学	--	666	北京	理科	2017	第一批	加入对比
金融学类	中国人民大学	--	664	北京	文科	2017	第一批	加入对比
外国语言文学类	北京理工大学	--	625	北京	文科	2017	第一批	加入对比
测控技术与仪器	天津大学	--	639	北京	理科	2017	第一批	加入对比
化学工程与工艺	天津大学	--	667	北京	理科	2017	第一批	加入对比
土木工程	天津大学	--	641	北京	理科	2017	第一批	加入对比
建筑学	天津大学	--	653	北京	理科	2017	第一批	加入对比

图 2.1 数据来源网站界面展示

爬取后的原始数据集共包含 249734 条样本，其中部分样本有重复，且可能存在部分专业数据丢失、无法获取等掉包现象，经初步去重后剩余 238980 个样本，剔除无效样本数占总体样本约 4.3%，认为数据本次研究数据具有的全面性，可以用于科学分析。

在信息维度方面，原始数据共包含 8 个字段，分别为：

- 年份：2007 年～2017 年，共 11 年；
- 学校：除国防科技大学外其余 38 所 985 高校；
- 考生地区：除港澳台外等 31 个省份；
- 文理科分类：文科、理科、综合；
- 专业：高校开设的公开招生专业，共 543 个；
- 批次：专业招生批次；
- 平均分：专业在地区录取的考生高考平均分；
- 最高分：专业在地区录取的考生高考成绩最高分。

2.2 数据清洗及特征工程

2.2.1 数据清洗

对数据初步观察，不难发现部分字段数据出现缺少现象，或个别高校在某年的专业数量突然降低，即某部分高校专业可能在某些年份缺失等。数据质量问题主要有以下几种情况：

1. 部分学校可能只公布了部分专业录取数据，并没有给出所有专业的录取数据，这也可能是爬取时出现掉包现象，或部分专业因为停招了或者在某些省份没招到人，因此也没获取到。
2. 字段数据缺失问题，部分年份的专业录取平均分缺失，或是最高分缺失，甚至最低分和最低分都缺失。如果对于单个成绩分缺失，则可以默认另一个得分为其填充值。但对于两者都缺失的样本则需要借助其他手段填充。
3. 在近年来，高考只分了文科与理科，但在 2007 年出现了综合科，这部分专业对今后的专业学科发展没意义，因此应该剔除。以免造成一定的干扰。

2.2.1 新增内生特征

由于不同年份的试卷难度不一样，不同省份的试题及总分也不尽一致，因此需要建立一个综合评价体系，以对比不同专业在不同学校及不同省份的门槛高度。特别注意的是，专业之间还区分文理科的比较，文科生不能填报理科的专业，因此在如果把文科的专业也与理科的专业放在一起比较，则没有现实意义。因此在数据处理时应该考虑不同专业的可比性。如文理科差异、省份差异、年份差异等。为了消除省份之间的高考成绩之间的差异，先对各个省所有专业录取分归一化转换，转换公司为：

$$x_{ypi} \coloneqq \frac{x_{ypi} - \min\left(x_{yp}\right)}{\max\left(x_{yp}\right) - \min\left(x_{yp}\right)}$$

上式中 y_{pi}为年份y、考生省份p、i专业的某个高校的录取平均分；

$\min\left(x_{yp}\right)$为该高校的所有专业中最低录取平均分；

$\max\left(x_{yp}\right)$为该高校的所有专业中最高录取平均分。

利用上面公式可以得知某学校中各个专业的相对门槛。有了相对指标对比外还应该进行专业排名。对专业录取分进行排名是为了避免专业之间录取分差异较大，特别是对于不同年份的录取分分布变动较大的，此时归一化系数则会带有一定的偏差。举一个极端一点的例子可以说明这种情况，某年报考数学专业的考生特别高分，平均分归一化后为 0.8，但报考物理专业的考生总体平均分相对低，归一化后只有 0.2，如果只从归一化系数去评价，数学专业的排名在所有专业中排名前列 20%，而物理专业为倒数 20%。但实际情况可能为数学专业排名第 2，而物理专业排名第三，其余专业都比物理专业的平均录取分要低。但学校每年开设的专业是大致稳定的，几乎不会因为本年度报考某个专业的学生分数不理想就停止开设该专业。因此如果作为平均录取分进行排序，对比不同专业的排名能更全面反映出专业的真实竞争力水平。

同理，除了对比不同专业在同一高校间的竞争力水平外，还应该对比同一专业在不同高校的竞争力水平。假设有 2 家大学以上在某个省中招取 A 专业，则可以直接对比录取分数线的高低来反映出不同高校对该专业的相对门槛，并排列出竞争力水平，同时为了对比出门槛的相对情况，在此还是对专业平均分进行归一化，计算公式与前面类似，$\min\left(x_{yp}\right)$为该省份中所有开设某专业的最低录取平均分。

同时，为了反映专业在某省份中的竞争力水平，综合各高校在本省中的所有的专业录取平均分进行归一化及增加省份分数排序指标。同一个高校同一专业中录取的考生高考成绩也有差异，一般总体分数分布属于以平均分为均值的正态分布，如果最高录取分与平均分差距太大，则说明该专业录取的生源质量不稳定，往往是由于专业在正常招生中未招满人，无奈把一些高考分数不太理想的考生也招进来，导致拉低平均分。录取平均分与最高分的分差可以在某种程度上反映该高校该专业的冷门火热情况，对未来分数预测具有一定参考价值。录取平均分与最高分的分差属于绝对比较指标，在不同省份、年份之间的对比会出现偏差，因此需要转换成相对指标，转换公式为：

$$r = \frac{x_h - x_a}{x_a}$$

其中x_h为专业录取最高分；x_a为专业录取平均分；r可以理解为增长率。综合上面增加的竞争力评价指标，数据特征维度增加了如下 8 个：

- 本专业在同一大学里对比其他专业竞争力得分；
- 本专业在同一大学里对比其他专业竞争力排名；
- 本高校专业在同一省份里对比其他高校同专业的竞争力得分；
- 本高校专业在同一省份里比其他高校同专业竞争力排名；
- 专业在同一省份里对比其他专业竞争力得分；
- 专业在同一省份里对比其他专业竞争力排名；
- 专业录取最高分与平均分的分差；
- 专业录取最高分比平均分的分差与平均分之比。

其中特别声明的是以上竞争力评价指标都基于录取平均分来评价的。因为评价指标更能代表大众水平。

2.2.3 缺失数据填充及插值

在前面数据清洗时已经过滤了一些肮脏样本、缺失数据，也通过了一些比较直接的方式进行剔除、填充。但对于部分样本录取平均分河最高分都缺失了，一方面，这部分高校专业可能由于在当年没有招到考试，因为缺失了真实参考价值；另一方面，个别年份高校没有对外公布相关专业的录取情况，或暂停对部分专业取消招生，使得这部分数据无法考究。对于类似这种情况，由于

专业录取系统总体上是一个相对稳定、成熟的信息系统，部分缺失的专业录取数据可以通过横向对比同类专业、该校历史的在该专业的录取情况，根据历年专业竞争力变化趋势预计出缺失值。而所需要用到的竞争力等评价指标在上一小节中已经可以构建出来。

正由于不同高校间的专业学科体系大致相近，专业门类在社会中竞争力水平存在一定的关联。因此在本次研究中将基于灰色系统理论应用关联分析对缺失的样本数据进行预测填充。灰色系统理论首先基于对客观系统的新的认识。尽管某些系统的信息不够充分， 但作为系统必然是有特定功能和有序的，只是其内在规律并未充分外露。有些随机量、 无规则的干扰成分以及杂乱无章的数据列，从灰色系统的观点看，并不认为是不可捉摸 的。相反地，灰色系统理论将随机量看作是在一定范围内变化的灰色量，按适当的办法 将原始数据进行处理，将灰色数变换为生成数，从生成数进而得到规律性较强的生成函数[2]。事实上，因素间关联性如何、关联程度如何量化等问题是系统分析的关键和起点。而高校招生系统作为一个年年更新、发展变化的系统，关联分析实际上是动态过程发展态势的量化比较分析。所谓发展态势比较，也就是系统各时期有关统计数据的几何关系的比较。

假设现在高校C在年份为y年中缺失省份为p专业i的录取数据，在存在其他高校在相同年份相同地区有对专业i的录取数据，选取某家这样的高校，参考数列：

$$x_0 = \{x_0(k)|k = 2007,2008,\dots,2017\} = \big(x_0(2007), x_0(2008), \dots, x_0(2017)\big)$$

其中 k 表示年份。假设有 m 个这样的比较数列

$$x_i = \{x_i(k)|k = 2007,2008,\dots,2017\}, i = 1,2,\dots,m$$

记

$$\zeta_i(k) = \frac{\min\limits_s \min\limits_t |x_0(t) - x_s(t)| + \rho \max\limits_s \max\limits_t |x_0(t) - x_s(t)|}{|x_0(t) - x_i(t)| + \rho \max\limits_s \max\limits_t |x_0(t) - x_s(t)|}$$

为比较数列x_i对参考数列x_0在 k 时刻都关联系数，其中$\rho \in [0,1]$为分辨系数。上式中$\min\limits_s \min\limits_t |x_0(t) - x_s(t)|$、$\max\limits_s \max\limits_t |x_0(t) - x_s(t)|$分别为两级最小差及两级最大差。一般来说，分辨率系数ρ越大，分辨率越大；相反亦之。

上面定义的关联系数是描述比较数列与参考数列在某时刻关联程度的一种指 标，由于各个时刻都有一个关联数，因此信息显得过于分散，不便于比较，为此我们给出如下新的定义:

$$r_i = \frac{1}{n}\sum_{k=1}^{n}\zeta_i(k)$$

为数列x_i对参考数列x_0的关联度。只要通过对比缺失的数据列与其他参照列的关联程度，选择最关联度最高的专业在缺失年份的录取分作为填充缺失值的依据。

2.2.4 探索性分析

对比历年不同各个高校开设的专业数量，考虑到前面提及数据质量的问题，即个别年份某些高校没有公布其招生数据，或者在数据爬虫的时候掉包了。如果直接把数据可视化，会发现部分高校在前年开设的专业数量是后年的数倍，造成这种情况大概率是数据缺失，特殊的有剥离学校原二极学院独立剥离，但在本次研究中不考虑高校二级独立学院呢。为了避免这种情况带来统计上的偏差，对各高校专业数量进行移动平均平滑化处理，转换公式如下：

$$x_i(t) := \frac{x_i(t) + x_i(t-1) + x_i(t-2)}{3}, t = 2009, \ldots$$

经过平滑处理后进行可视化，由于数据样本不全，在专业数量上的可能与真实情况有点偏差，但综合整个发展趋势来看，无论是文科或是理科，不同高校间开设的专业数量都有所下降，整个下降趋势对于文科专业来更加明显，这也一定程度上反映出整个社会“重理工，抑文科”的现象，同时也有是学生之间对于文科自然选择，供应与需求互相影响导致这种发展趋势。

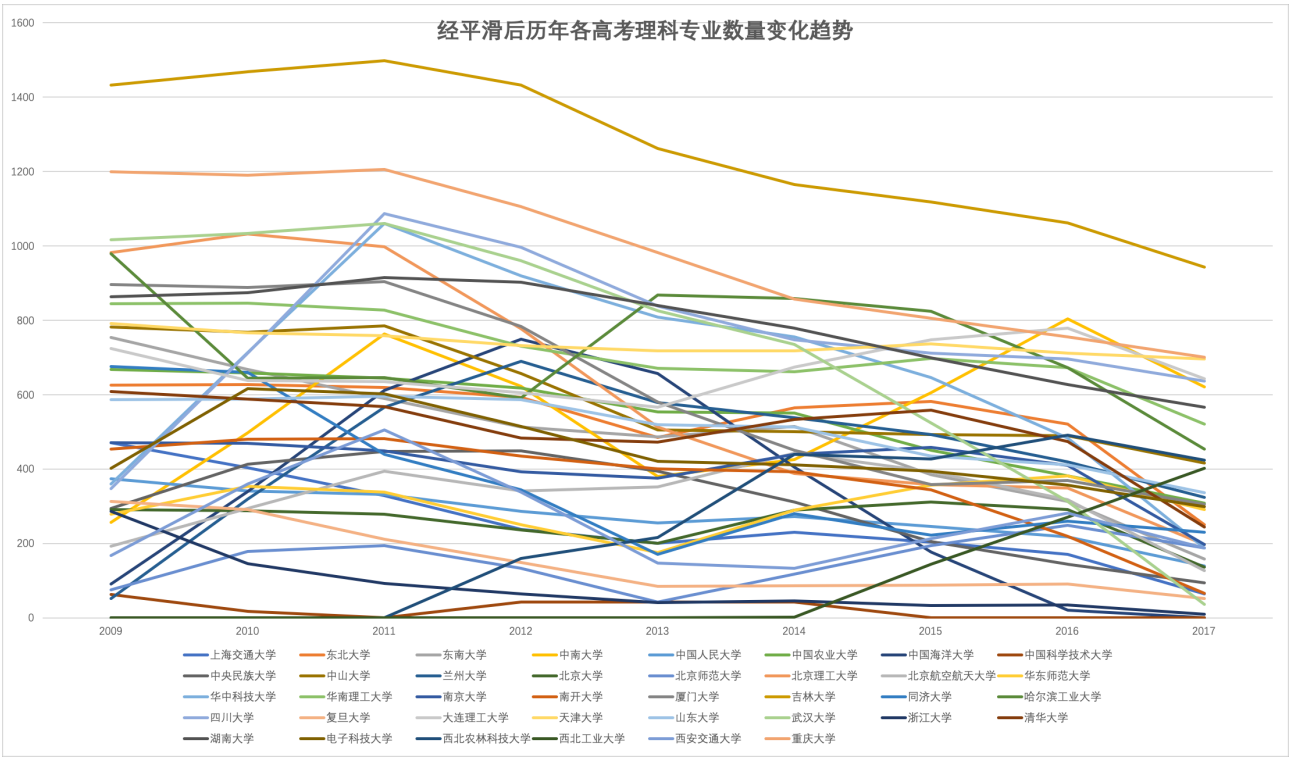

图 2.2 各高校历年理科开设专业数量变化趋势

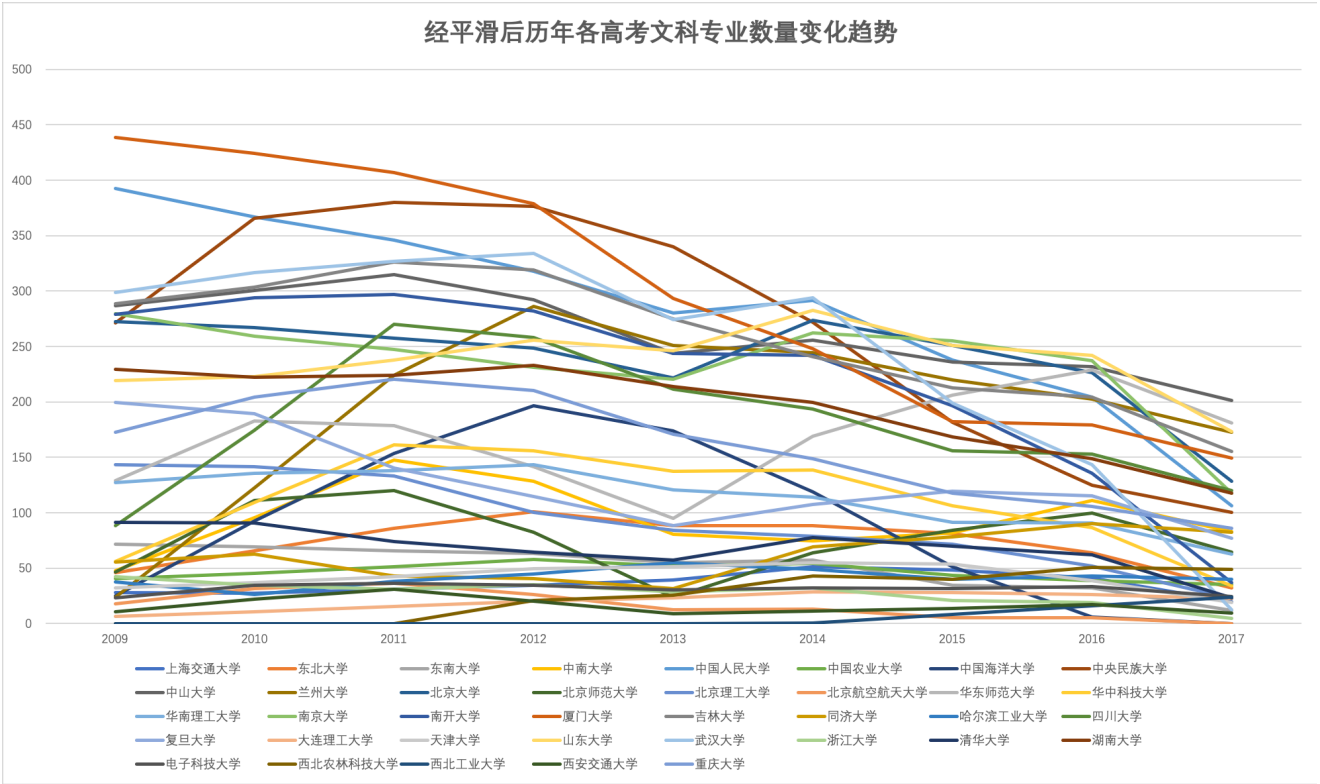

图 2.3 各高校历年文科开设专业数量变化趋势

通过对比各个专业在学校的竞争能力得分，可以得到历年各个专业在整体社会中的综合竞争力水平，把历年综合竞争力得分进行加权回归相加，得到总的评价得分，其中专业竞争力得分越接近 1，表面该专业的门槛越高、越吸引人。考虑专业数量过多，现针对文、理科专业竞争力得分排名前 10 和进行可视化：

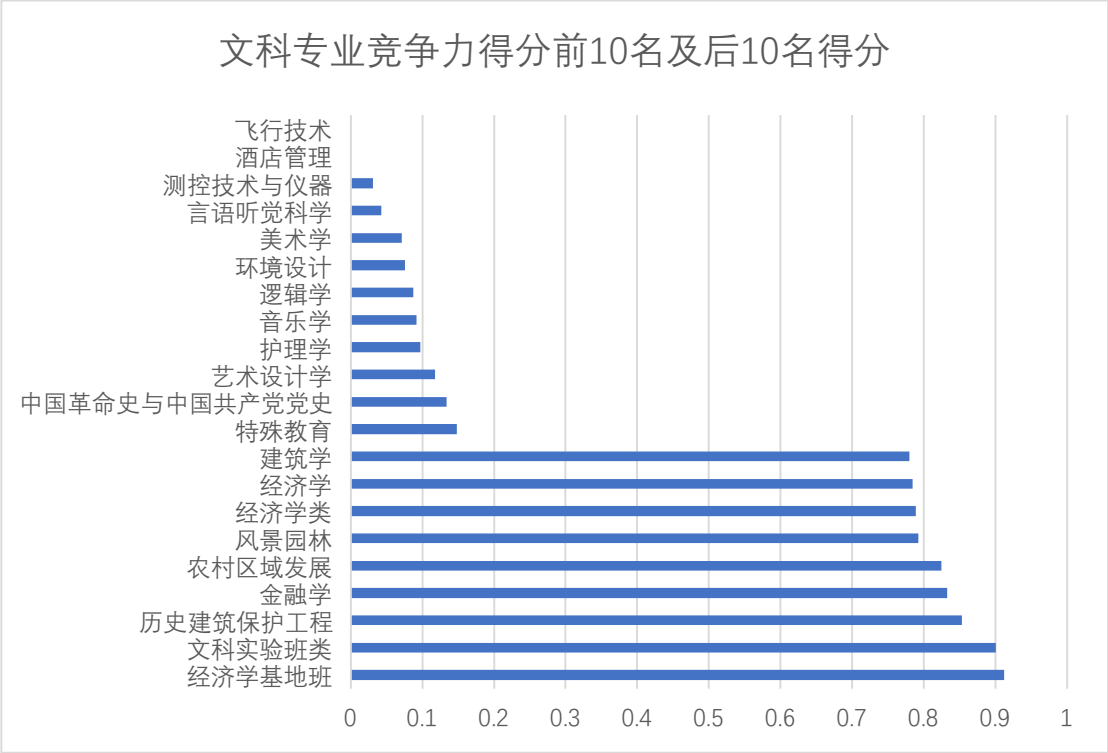

图 2.4 文科专业竞争力得分前 10 前及后 10 名

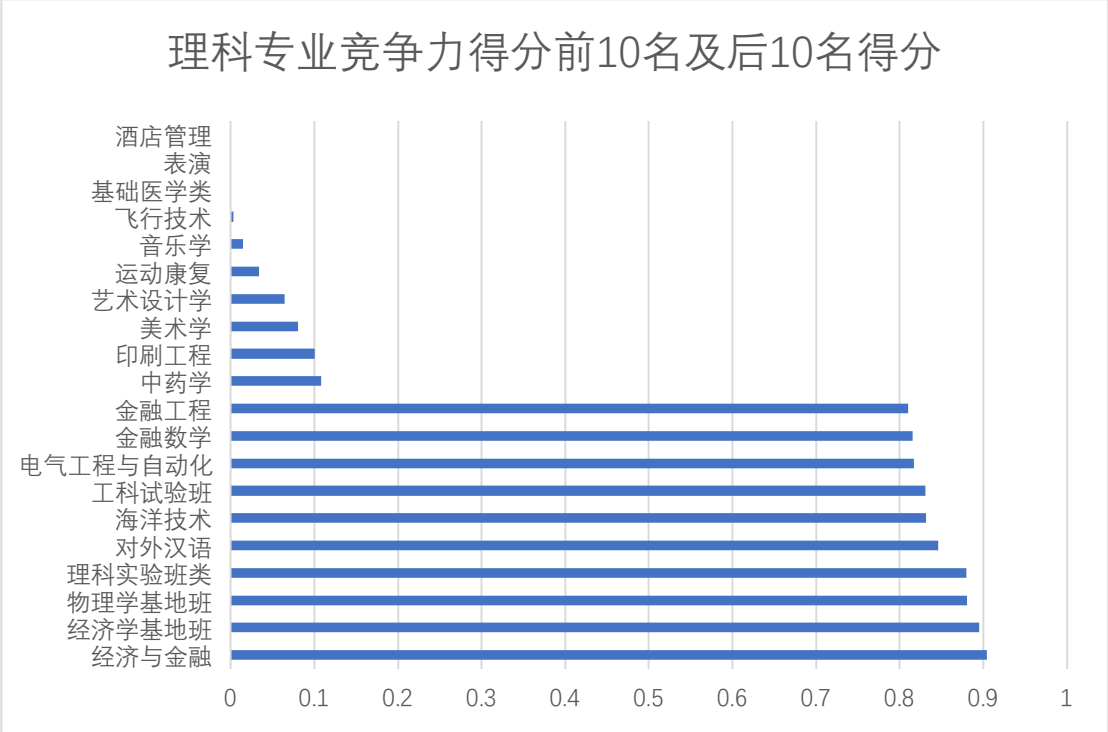

图 2.5 理科专业竞争力得分前 10 前及后 10 名

无论是文科或是理科，排名较前的专业中的关键词为“基地班”、“实验班”、“经济”、“金融”。这个结果属于是情理之内，实验班、基础班等一般以培养科学家为目标，因此对于考生的高考成绩要求会比较高。对于经济、金融相关的专业，受到将来职业规划和薪酬水平等，考生供应过多，竞争力较大。而对于综合竞争力较低的专业，文理科大致一样，属于比较偏门、小众，如飞行技术；又或偏艺术类，比如音乐、艺术设计等。一方面，对于这些小众的学科，我国大多又相关的专业学院，如学音乐的可以去音乐学院，学表演的可以去电影学院，而对于 985 高校而言，基本都为综合性大学，在小众专业上自然比不上一些专业学院。另一方面，对于这些小众的专业，一般不需要太高的知识水平，也即是门槛相对低，因此竞争力水平也自然偏低。

根据专业历年的竞争力得分变化，可以筛选出文理科专业中竞争力得分提升比例最大的和下降比例最大的专业，竞争力得分提高意味着该专业的门槛越来越高，具有较高的发展潜力；反之亦理。造成这种情况一般由供求市场关系影响的。下表列举出文理科最具潜力与最不具潜力排名前 10 的专业名单：

表 2.1 最具潜力与最不具潜力排名前 10 的专业名

排名	文科最具潜力专业	文科最不具潜力专业	理科最具潜力专业	理科最不具潜力专业
1	法学类	农林经济管理	环境科学与工程	广播电视编导
2	金融学类	阿拉伯语	中国语言文学类	新闻学
3	中国语言文学类	乌尔都语	外国语言文学类	历史学
4	历史学类	城乡规划	应用生物科学	口腔医学类
5	传播学	数字出版	自然保护与环境生态类	植物生产类
6	政治学类	财政学	网络工程	社会学
7	法学	人文地理与城乡规划	计算机科学与技术	信息资源管理
8	外国语言文学类	新闻学	知识产权	采矿工程
9	历史建筑保护工程	朝鲜语	软件工程	力学类
10	人类学	文化产业管理	电气类	纺织工程

对于文科考生而言，最具潜力专业主要为实用性专业，入法学、金融学等，而最不具潜力的专业主要为小种语言类或小众专业，造成这种发展趋势主要是受到职业在整体社会的需求变动，大家都追求实用性专业，希望以后在职业工作上能体现自我价值。而对于理科考生而言，主要为工程类计算机相关专业，及个别几个语言文学类专业。语言文学类专业的录取门槛提高可能为部分理科考生在高中是文理科选择错误，希望在大学里得到纠正，而对于这些学生而言，主要受到个人兴趣驱动，有一定的文学功底，以理科考生身份去选择攻读文科专业，属于降维打击现象。而计算机相关的专业具备潜力也很好理解，信息技术一直在改变我们的生活，越来越多的产业都开始数字化，社会就业前景广阔，也使得相关专业具备吸引力。理科最不具潜力专业也有一部分文科系专业，且由于比较小众，只会在社会专业分工精细化中被淘汰掉。

3. 高校及专业推荐系统模型

对于考生而言，志愿填报有两个关键信息：志愿高校及志愿专业。如果考生没有一个目标选择或者对各个高校专业之间没有一个全局的了解，因此在进

行决策的时候会综合各个因素对备选高校和备选专业间两两对比。若备选目标数量较少的情况下，考生还是可以很快通过对比则可作出选择，但当面对上百个目标专业选择的时，则很难进行全局主观对比。因此需要借用一些系统的信息决策理论来进行定量分析。特别是对于高校专业而然，各个学校都有自己的优势和劣势，而且这些优势劣势可以在不同的考生身上有不同的比较效果，如有些考生希望录取高校尽量离家近一点，对于专业选择不太重视，因此在决策过程中先决定高校，再从目标高校中选择该高校开设的专业；而有些考生希望学习自己感兴趣的专业，如果对于一些没有开设该专业的高校则不作考虑，这类考生的逻辑决策则是与前面的颠倒；更有甚者则是综合对比，决策逻辑更加模糊，不确定。而本次研究的目标则是兼顾所有考生，为考生进行个性化推荐。因此必须建立一个可以量化各高校、各专业的数学模型，而层次分析法则在这方面有很好的应用。

3.1 层次分析法量化高校专业匹配程度

层次分析法(Analytic Hierarchy Process，简称 AHP)是对一些较为复杂、较为模 糊的问题作出决策的简易方法，它特别适用于那些难于完全定量分析的问题[2]。

正如前面提到的，在进行决策时需要考虑各个因素，在本次研究中针对目标高校的选择需要考虑到因素分别有：地理位置、学术声誉、毕业生就业率、国际化程度、社会认可程度共 5 个指标，而决策目标分别为 38 所 985 高校，其层次结构模型如下：

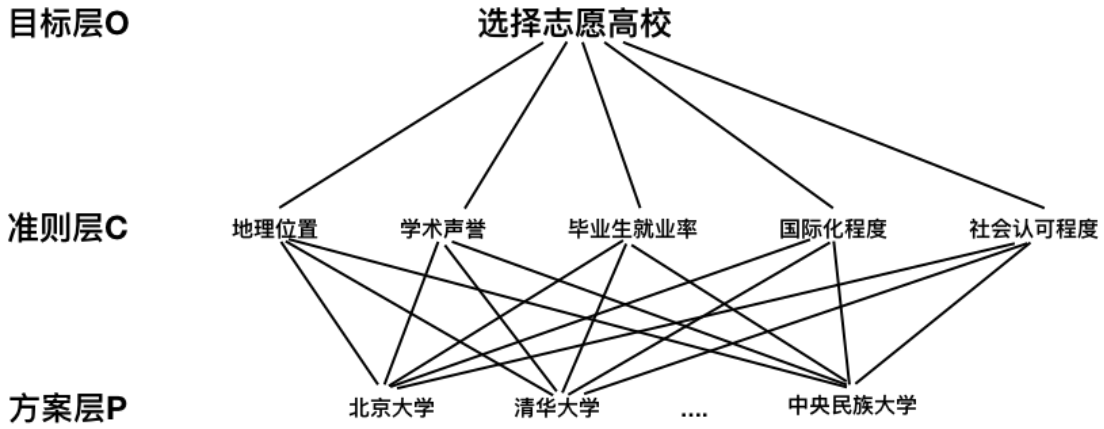

图 3.1 考生选择志愿高校层次决策结构

而对于对专业的层次决策，考虑以下因素：考生兴趣、就业情况、发展潜力、竞争力、学习成本等 5 个因素对比，而决策层则把数百个专业归类为如下门类：哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、军事学、管理学和艺术学 13 大门类。这这样的目标是防止专业之间的对比过于复杂。其层次结构模型如下：

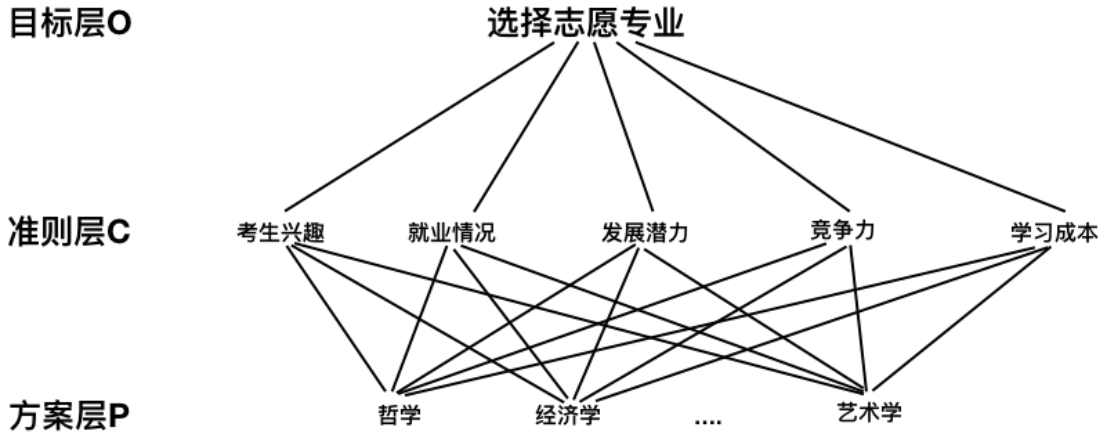

图 3.2 考生选择志愿专业层次决策结构

以上已经完成了层次分析法的第一步：建立递阶层次结构模型；接下来还需要构建各层次中的判别矩阵并进行一致性检验，对于一致性检验不通过的矩阵则让考生重新构建。对于判别矩阵的元素确定值，在本次研究中使用 Saaty 主张引用数字 1~9 及其倒数作为标度，其中 1 表示两个因素对比具有相同重要性，数值越大，说明前者比后者的重要性越强。对于考生而言，判别矩阵的构建主要在目标层对准则层的构建，因为每个考生对于各个因素对关注度不一样，因此对于各个考生而言都能构建出属于自己的独特的判别矩阵。而准则层到方案层到判别矩阵则不完全由考生构建，因为一般而言，考生无法完全掌握各高校或各专业在各个评价因素上的具体情况，因此需要借用一些比较可信的数据使得整个层次分析模型可靠。如各高校的学生声誉在本次研究中引用各高校的论文发表质量指数，高校之间在该因素的对比得分为两校中论文发表质量指数的比值；同理，毕业生就业率则引用中国高等教育学生信息网[3]（学信网）中院校满意度指标；社会认可程度引用高校接受社会捐献资金额度；国际化程度则引用外籍学生比例。而对于专业选择的层次分析模型，发展潜力的对比用用专业的竞争力得分增速；就业情况引用学信网[3]专业满意程度。而对于部分属于比较主观的学习成本则需要结合考试的个人情况进行判别。

决策结构各层次通过一致性检验后并求出各层次间连接的权重值，准则层权值与方案层权值相乘累加后得到总代排序权值，具体的计算示例如下表 3.1 所示。通过对比各备选方案的总排序权值，挑选权值最高的方案作为目标决策。 同时，作为第二志愿和第三专业，则选择权值为次高的方案。

表 3.1 层次总排序

准则		地理位置	学术声誉	毕业生就业率	国际化程度	社会任何度	总排序权值
准则层权值		W1	W2	W3	W4	W5	
方案层排序权值	北京大学	W1,1	W2,1	W3,1	W4,1	W5,1	$\sum W_i \cdot W_{i,1}$
	清华大学	W1,2	W2,2	W3,2	W4,2	W5,2	$\sum W_i \cdot W_{i,2}$
	…	…	…	…	…	…	…
	中央民族大学	W1,38	W2,38	W3,38	W4,38	W5,38	$\sum W_i \cdot W_{i,38}$

同理，对于专业的决策选择也如高校决策一样，先决策出最优科目门类，再针对该门类下各个专业进行再次层析分析，方案层换成该门类下的各专业，即可挑选出最终的决策专业。但目前此方案由一个现实限制则是，考试的高考成绩优秀得足以有资格挑选各个高校和各个专业。一般情况下，因为 38 所高校从批次来看都属于提前批，考试具备投档资格，但由于考生的排名不一致，部分相对落后的考试投档与其实力不符的高校或专业可能会影响到投档成功率，并可能导致考生没被志愿高校录取。因此在进行投档时需要理性确定自己的实力水平及各高校、专业的竞争力、门槛等变动。也即是现在的问题则转化为先精确预测专业及高校的竞争力水平变化及考生实力水平的正确定位。

3.2 基于 LSTM 的高校专业竞争力水平预测

3.2.1 LSTM 原理说明

在介绍 LSTM 网络前先介绍一下 RNN（Recurrent Neural Networks，循环神经网络）。人类并不是每时每刻都从一片空白的大脑开始他们的思考。在你阅读这篇文章时候，你都是基于自己已经拥有的对先前所见词的理解来推断当前词的真实含义。我们不会将所有的东西都全部丢弃，然后用空白的大脑进行思

考。我们的思想拥有持久性。传统的神经网络并不能做到这点，但 RNN 解决了这个问题。RNN 是包含循环的网络，允许信息的持久化。

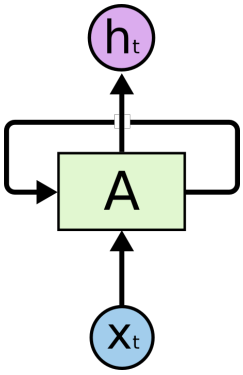

图 3.3 包含循环结构的 RNN 结构图

在上图中，A是神经网络的主体，X_t是输入信息，h_t是输出信息，网络主体的循环结构能使输出信息传递到下一次的网络输入。如果把这单一的网络结构展开，能得到下面更具体的传递机制。这种链式的信息传递机构跟我们思考的方式密切相关，人类每作出一个决策都不是突然发起的，而是根据历史的记忆对周边事物作出判断。举个列子，在电影院看一场电影，你只看到一帧画像，你很难预测出故事将会朝向哪个方向发展，但是如果把前面的动画连贯起来，你拥有了故事前一部分的一段记忆，这时候你预测电影后面的剧情发生的场景准确性就会得到提高。RNN 的基本原理则是在进行下一次计算分析的时候保留前面的“记忆”，学习之前使用到的信息。

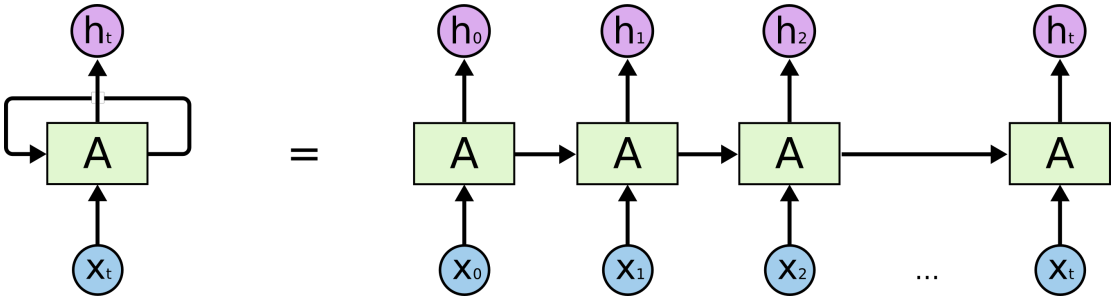

图 3.4 展开的 RNN 结构图

但实际生活中，人们没必要也没可能记住之前的所有记忆，就好比当你思考你明天早餐吃什么时，没必要想起自己昨天没有洗澡一样。因为吃什么早餐和之前有无洗澡的相关性是很弱的，记住太多“无谓的记忆”反而会使生活行动变得好无效率。又或者人们希望每年的同一天吃到的早晨都是一样的，这需要记得去年在这天吃过的早餐，而不必记得昨天或上周吃过的早餐，这样的话相

关信息和需要预测的点的间隔很大，随着这种信息间隔的增大，RNN 链接以往的信息越来越有限。

类似这样的问题称作为长期依赖（Long-term dependencies）问题，理论上的 RNN 是可以有能力处理这种长期依赖问题，人们可以静心选择各种参数或激活函数去解决这类问题，但实际上，Hochreiter(1991)和 Bengio,et al,(1994)发现了一些为什么 RNN 在这些问题上学习相当困难的根本原因[4]。不过 LSMT 并不会出现这类问题。

LSTM 由 Hochreiter 和 Schmidhuber(1997)引入，是一种特殊的 RNN，后来在很多人的努力下变得越来越精炼和流行。它们在大量的问题上有惊人的效果，现在被广泛的使用。LSTM 被明确的设计用来解决长期依赖问题，记住长时间段的信息是他们的必备技能，不像 RNN 那么费力去做还做不好[4]。

所有的递归神经网络都有重复神经网络本身模型的链式形式。在标准的 RNN，这个复制模块只有一个非常简单的结构，例如一个双极性（tanh）层。

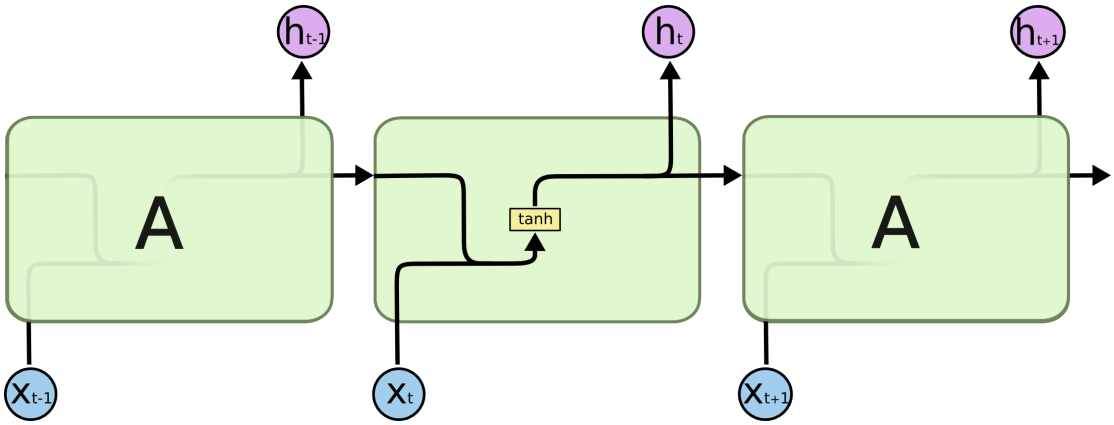

图 3.5 标准 RNN 中的重复模块包含单个模块

LSTM 也有链式结构，但是这个重复模块与上面提到的 RNN 结构不同：LSTM 并不是只增加一个简单的神经网络层，而是四个，它们以一种特殊的形式交互。LSTM 有能力删除或者增加神经元状态中的信息，这一机制是由被称为门限的结构精心管理的。门限是一种让信息选择性通过的方式，它们是由 Sigmoid 神经网络层和逐点相乘器做成的[9]。

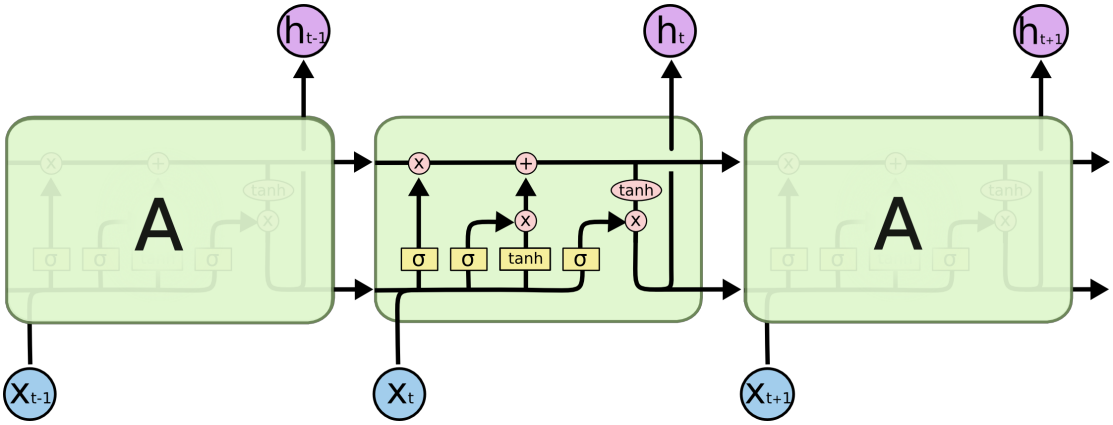

图 3.6 LSTM 中的重复模块包含四个交互层

3.2.2 基于 LSTM 的高校专业竞争力水平预测

高校录取情况为数据序列数列，并且后一年的录取结果时基于前一年的录取情况而作出相应的变化，在时间节点上具有较强联系性，因此 LSTM 对于高校专业招生预测具有很好的应用性。

由于每年高考成绩与各批次的投档线一并给出，且考生知道自己的排名成绩，甚至可以通过媒体得知各省状元的成绩。因此专业录取分数与一本线之比、与高考状元分数只比等可以知道各专业的录取门槛，把这两个指标作为新的特征，结合 2016 年（包括 2016 年）各专业前 5 届招生数据各指标（专业平均分、最高分、排名、竞争力得分等）作为输入特征，以 2017 年各高校专业竞争力得分为输出目标，以 70%的样本作为训练集，剩余 30%作为测试集，来对 LSTM 进行学习。针对学习好的模型，用之预测下年的高校各专业的招生门槛，并对招生门槛分值归一化，对比考生的分值与状元分值对比，考生的成绩排名及总考生数比值，可以得知自己是否达到某高校专业的门槛。对于门槛没有得到的高校和专业在层次分析法中方案层选择中剔除，再对达到门槛的专业进行分析对比选择。根据模型误差结果显示，LSTM 具有较高的拟合效果，训练集竞争力最大误差在 1%内，而测试集的最大误差在 1.5%，均分误差值接近

为 0。

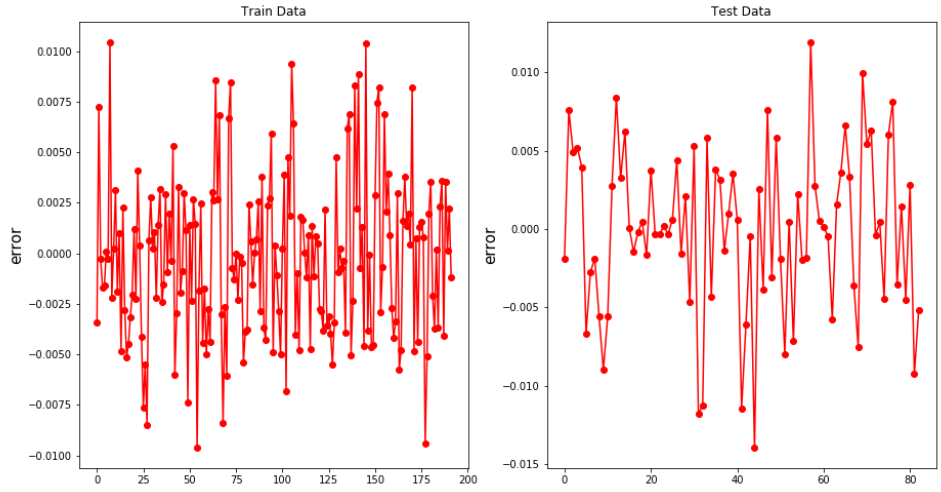

图 3.7 LSTM 预测误差可视化

5.总结与展望

5.1 测试环境

操作系统：mac OS

开发环境：python 3.6、Office Excel

5.2 模型评价

从模型数据角度来看，本次研究收集了约 23 万条专业录取数据，时间跨度 11 年，并通过数据清洗及特征工程使得数据更合理，信息维度更丰富。但在数据获取中，受限于公开信息不全，爬取数据中可能出现掉包现象，因此研究结论与实际上有点偏差，不过在本次研究中使用了大量的统计手段来降低统计偏差。

从数据处理来看，考虑到文理科间不同省份、不同年份的成绩有不同的评价标准，不能直接进行比较，在进行研究分析时，先通过特征工程把绝对值指标换成相对值竞争力指标，更能反映出各专业的真实门槛及竞争力水平。

从模型算法来看，本文从实际出发，巧用批次投档分及状元总分为对照目标，利用深度学习技术预测出下年各专业相对门槛，并通过数学模型为考生推荐相应专业。LSTM 模型的预测精度非常高，具有很好的现实意义，且层次分

析法中判别矩阵的一致性检验能尽量降低因个人主观对比偏差过大而导致模型失去客观性。

5.3 研究展望

数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限的一种手段。在本研究中，只收集到 2017 年前的数据，而对于 2018 年的数据则没有采集到，这是本次研究中的一种遗憾。另外数据没有很完整，若后面在数据能得到所有专业的历年数据，则模型结果及探索性分析更有建设性。

另外对于新开设的专业因为缺少历年数据，未能预测到下年的综合竞争力，因此在进行评比的是没有把新开设专业考虑进去，如果能找到一个模型能预测出新开设专业，能提高模型的全面性。

最后，因为全国高校数量太多，在本次研究中单纯爬取 38 所 985 高校已经有近 24 万条样本数据，考虑到计算资源问题，本次研究模型因为也只局限于 985 高校的专业推荐。而对于非 985 高校则需要获取更多其他高校数据才有意义。

参考文献

[1] http://college.gaokao.com

[2] 司守奎, 孙玺菁. 数学建模算法与应用[M]. 国防工业出版社, 2011.

[3] https://gaokao.chsi.com.cn/zyk/pub/zytj/recommendTop.action

[4] https://blog.csdn.net/yingwei13mei/article/details/53575429