

所属类别	2023 年“华数杯”全国大学生数学建模竞赛	参赛编号
本科组		CM2302099

母亲身心健康对婴儿成长影响的分析与探究

摘 要

当前研究表明，母亲的身心健康状况对婴儿的生理和心理会产生一定影响，本文期望在各种客观条件的约束下，研究其相关影响规律，为提高母亲心理健康水平以改善母婴交互质量，提高婴儿睡眠质量等提供有效方案。

本文建立了五个主要模型：模型 I：典型相关分析模型；模型 II：贝叶斯分类器模型；模型 III：非线性整数规划模型；模型 IV：基于熵权 TOPSIS 分析的综合评价模型；模型 V：随机森林回归模型。

针对问题 1，首先对附件中数据进行分类整合与可视化处理，剔除异常数值。接着将母亲身心指标和婴儿行为特征与睡眠质量分成各自变量集合，建立**典型相关分析模型**。最后通过典型相关分析得出母亲身心指标和婴儿睡眠质量相关系数以及标准化相关系数，完整结果见表 3-5，**结果说明婴儿的行为特征与母亲的婚姻状况和分娩方式具有非常大的相关性，且均呈现为正相关型。**

针对问题 2，由于存在不平衡数据，本文基于**贝叶斯分类器模型**构建了婴儿的行为特征与母亲的身体指标与心理指标的关系模型，从而对婴儿的行为特征进行分类，其完整分类结果见表 6，并通过 python 进行求解，**结果准确率为 0.83。**

针对问题 3，由于治疗费用对于患病程度变化率和治疗率均与治疗费用呈正比，因此建立单目标**非线性整数规划模型**，目标函数为最小化治疗费用，分别求解得出婴儿行为特征重矛盾型转化为中等型总花费为 41870 元，矛盾型转化为安静型的总花费为 45943 元，完整结果见表 8 和表 9。

针对问题 4，首先利用**熵权法**求出对应影响指标权重，对婴儿睡眠质量进行打分，并将结果进行**K-means 聚类**。其次将聚类中心按照从小到大依次排序，分别对应到优、良、中、差四个分类类型中，然后建立**随机森林回归模型**，预测得出最后 20 组婴儿综合睡眠质量，得出编号 410、408 等睡眠质量为优，完整结果见表 14。

针对问题 5，通过问题 4 得出的划分结果，将编号 238 的婴儿在随机森林回归模型中得出的结果作为约束条件，置入**单目标非整数线性规划**，引入 RFR 约束条件，将求出结果与问题 3 结果进行对比，结果显示若需要将婴儿的行为特征从矛盾型变成中等型，且睡眠质量变成优，则在原先治疗方案的基础上，在 CBTS 上多降低 1 分的治疗方案。

最后，对模型进行**鲁棒性分析**。本文的模型在具有轻微扰动的原始输入上的预测结果不会显著偏离该原始输入，说明该模型在输入样本存在细微对抗扰动的环境下，模型预测具有不受对抗样本干扰误导的能力。

关键词：典型相关分析；贝叶斯分类器；随机森林；整数规划；熵权 TOPSIS

第 1 章 问题重述

1.1 问题背景

母亲是婴儿成长过程中至关重要的角色之一，她不仅供给了婴儿所需的养分和身体保护，同时也为婴儿提供了情感上的支持和安全感。母亲的心理健康状况，如抑郁、焦虑、压力等问题，可能会对婴儿的认知、情感以及社会行为等方面造成不良影响。当母亲承受过大的压力时，可能会对婴儿的生理和心理发展产生负面影响。研究附带了关于 390 名婴儿及其母亲的相关数据。这些数据包含多个方面，母亲的身体指标包括年龄、婚姻状况、教育程度、怀孕时间、分娩方式，以及产后心理指标，包括 CBTS（分娩相关创伤后应激障碍问卷）、EPDS（爱丁堡产后抑郁量表）、HADS（医院焦虑抑郁量表）。而关于婴儿的睡眠质量则涵盖了整夜睡眠时间、夜间醒来次数以及入睡方式等方面。

1.2 问题描述

问题 1: 建立相关模型，研究母亲的身体指标和心理指标对婴儿的行为特征和睡眠质量是否存在影响和规律。

问题 2: 建立婴儿的行为特征与母亲的身体指标与心理指标的关系模型，并判断数据表中编号 391-410 号婴儿的行为特征信息所属类型。

问题 3: 对母亲焦虑的干预有助于提高母亲的心理健康水平，还可以改善母婴交互质量，促进婴儿的认知、情感和社交发展。CBTS、EPDS、HADS 的治疗费用相对于患病程度的变化率均与治疗费用呈正比，经调研，给出了两个分数对应的治疗费用，详见表 1。现有一个行为特征为矛盾型的婴儿，编号为 238。请你建立模型，分析最少需要花费多少治疗费用，能够使婴儿的行为特征从矛盾型变为中等型？若要使其行为特征变为安静型，治疗方案需要如何调整？

问题 4: 婴儿的睡眠质量指标包含整晚睡眠时间、睡醒次数、入睡方式。对婴儿的睡眠质量进行优、良、中、差四分类综合评判，并建立婴儿综合睡眠质量与母亲的身体指标、心理指标的关联模型，预测编号 391-410 号婴儿的综合睡眠质量。

问题 5: 若要让 238 号婴儿的睡眠质量评级为优，求对问题三的治疗策略最优方法。

第 2 章 思路分析

2.1 问题 1 的思路分析

问题一要求寻找母亲的身体指标和心理指标对婴儿的行为特征和睡眠质量^[2]是否存在影响，因此可以研究这两组变量的整体相关性，采用典型相关分析^[3] (Canonical Correlation Analysis , CCA) 进行相关研究，而母亲的身体指标和心理指标包括 8 个相关变量 (母亲年龄，婚姻状况，教育程度，妊娠时间，分娩方式，CBTS, EPDS, HADS)，婴儿的行为特征和睡眠质量包括 4 个相关变量 (婴儿行为特征，整晚睡眠时间，睡醒次数，入睡方式)，然后通过方差贡献率选取满足条件的典型变量对数。通过分析典型相关系数和典型相关变量对原始指标变量的解释能力，来确定母亲的身体指标和心理指标对婴儿的行为特征和睡眠质量的影响程度。

2.2 问题 2 的思路分析

由于题目要求建立立婴儿的行为特征与母亲的身体指标与心理指标的关系模型，而婴儿的行为特征属于定类变量，因此选取合适的分类模型，可选用贝叶斯分类器^[4]进行分类，以及预测最后 20 组 (编号 391-410 号) 婴儿的行为特征信息。

2.3 问题 3 的思路分析

由于 CBTS、EPDS、HADS^[1]的治疗费用相对于患病程度的变化率均与治疗费用呈正比，因此可以求出每个 CBTS、EPDS、HADS 得分对应的费用，然后建立单目标非线性整数规划模型，求得将编号 238 的婴儿的行为特征从矛盾型转变成中等型。之后再调整部分约束条件，使得编号为 238 的婴儿的行为特征从矛盾性转变成安静型，并找出两种方案的不同点，并对第一种方案进行调整，得出调整结果。

2.4 问题 4 的思路分析

采取熵权 TOPSIS 法对婴儿的睡眠质量进行优、良、中、差四分类综合评判，并得到每个婴儿的综合得分，随后依据综合得分进行 K-Means 聚类，设置 4 个聚类中心，将最终的聚类中心进行从小到大排序，分别对应与优、良、中、差这 4 个类型，之后建立随机森林回归模型，自变量为母亲的身体指标、心理指标，因变量为婴儿的综合睡眠质量得分，并将得到训练模型用于预测最后 20 组婴儿的综合睡眠质量。

2.5 问题 5 的思路分析

在问题 3 的基础上，引入问题 4 已经训练好的随机森林^[5]回归模型，并建立全面的单目标非线性整数回归模型，并进行计算，将计算结果与问题 3 的结果进行对比，得出调整方案。

第 3 章 符号说明与基本假设

3.1 符号说明

表 1: 文章中使用的符号及说明

符号	具体说明
X_j	第 j 个指标的值
w_j	第 j 个指标的权重
C_j	第 j 个指标的信息熵
$f(x_i)$	集合 x_i 对应的得分
$Score_1$	贝叶斯分类器对安静型的分界线
$Score_2$	贝叶斯分类器对中等型的分界线
$RFR(x_i)$	集合 x_i 的随机森林回归得分
\min	综合得分中最小的值
\max	综合得分中最大的值

3.2 基本假设

- 为了简化问题，我们做出以下基本假设，并且每一个假设都是合理的。
- ✓ **假设 1:** 附件数据是随机采样得到的，且每个样本均具有独立性。
理由: 使得可以使用统计方法建模，且数据没有偏或存在系统性误差。
 - ✓ **假设 2:** 母亲的身体指标和心理指标，婴儿的行为特征和睡眠质量均在正常时间段测得。
理由: 避免外界的因素对实验数据进行干扰。
 - ✓ **假设 3:** 治疗的价格不随时间变化而变化，且固定为题目所提供的价格。
理由: 避免随时间变化，治疗价格也出现变化，导致得出的治疗方案过时。

第 4 章 模型的准备

4.1 数据预处理

根据附件提供的数据，对其进行异常值处理和数值化处理。在对整晚睡眠时间指标处理时发现编号为 180 的对象该指标的值为 99:99，明显错误，考虑到样本数量较多，且只有这一个数据异常，为了计算结果的准确性，删除该对象的全部数据。随后对该指标做数值化处理，将时间数据提取小时作为整数部分，将分钟和秒钟转换成对应的小时即（9:00 对应的数值为 9，10:30 对应的数值为 10.5），其次对婴儿行为特征这一指标做独热编码，其中，中等型表示 0，安静型表示 1，矛盾型表示 2。具体见表 2。

表 2 婴儿行为特征的数据分布

指标	类型	例数	百分比
婴儿行为特征	中等型 (0)	225	57.7%
	安静型 (1)	120	30.8%
	矛盾型 (2)	45	11.5%

4.2 我们的工作

为了直观展现本文的建模过程及分析思路，我们制作如下图 1 所示流程图：

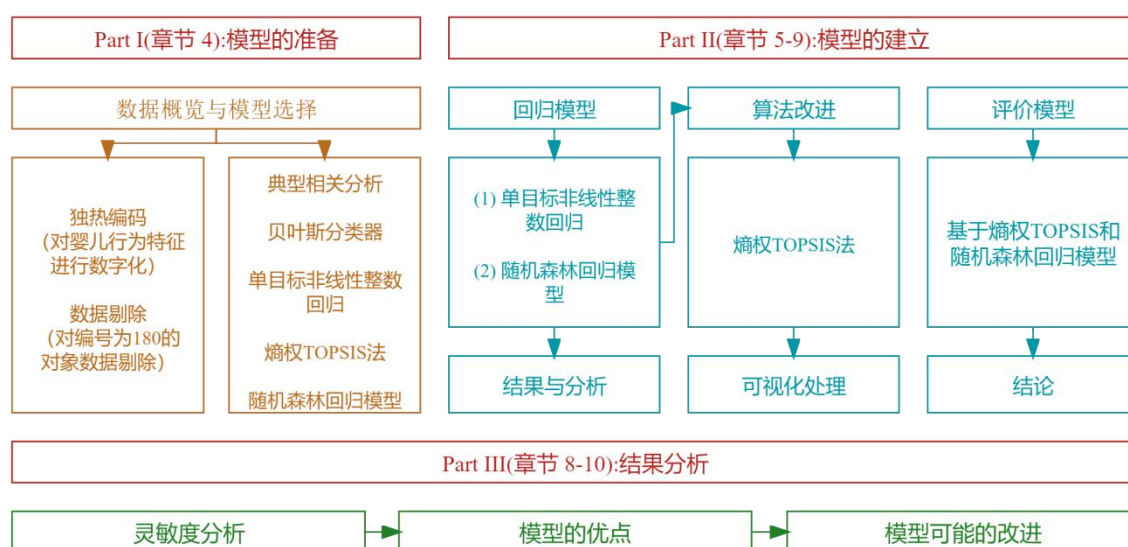


图 1 我们的工作流程图

第 5 章 问题 1 的模型建立与求解

5.1 典型相关分析模型的建立

本文将母亲的身体指标和心理指标和婴儿的行为特征和睡眠质量分成两个变量集合（其中婴儿的行为特征和睡眠质量表示 Y ，母亲的身体指标和心理指标表示 X ）：

$$\begin{aligned} X &= (x_1, x_2, \dots, x_p)^T \\ Y &= (y_1, y_2, \dots, y_q)^T \end{aligned} \quad (1)$$

然后定义两个线性集合 U 和 V ，其中 U 是 X 的线性组合， V 是 Y 的线性组合：

$$\begin{aligned} U_1 &= a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p \\ U_2 &= a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p \\ &\vdots \\ U_p &= a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p \\ V_1 &= b_{11}Y_1 + b_{12}Y_2 + \dots + b_{1q}Y_q \\ V_2 &= b_{21}Y_1 + b_{22}Y_2 + \dots + b_{2q}Y_q \\ &\vdots \\ V_q &= b_{q1}Y_1 + b_{q2}Y_2 + \dots + b_{qq}Y_q \end{aligned} \quad (2)$$

$$\quad (3)$$

这里为了寻求 U_i 和 V_j 中关联最大的线性组合，定义如下 U_i 的方差计算公式和 V_j 的方差计算公式：

$$var(U_i) = \sum_{k=1}^p \sum_{l=1}^p a_{ik}a_{il}cov(X_k, X_l) \quad (4)$$

$$var(V_j) = \sum_{k=1}^q \sum_{l=1}^q b_{jk}b_{jl}cov(Y_k, Y_l) \quad (5)$$

最后得到 U_i 和 V_j 的协方差计算如下：

$$cov(U_i, V_j) = \sum_{k=1}^p \sum_{l=1}^q a_{ik}b_{jl}cov(X_k, Y_l) \quad (6)$$

并采用如下公式来确定 U_i 和 V_j 的相关性 r_{ij} ，得到母亲的身体指标和心理指标和婴儿的行为特征和睡眠质量两组集合的相关系数：

$$r_{ij} = \frac{cov(U_i, V_j)}{\sqrt{var(U_i)var(V_j)}} \quad (7)$$

5.2 问题 1 的结果及分析

通过典型相关分析得到如下数据表 3：

表 3 390 例母亲的身体指标和心理指标与婴儿的行为特征和睡眠质量的相关系数

典型变量	相关系数	特征值	<i>Wilks</i> 值	<i>F</i> 值	<i>P</i> 值
1	0.299	0.089	0.847	1.952	0.001***
2	0.2	0.04	0.93	1.284	0.01**
3	0.165	0.027	0.969	0.971	0.175
4	0.061	0.004	0.996	0.273	0.372

其中，第一对和第二对典型变量的 *p* 值小于 0.01，具有显著的特征，第三对和第四对典型变量虽然 *p* 值并不显著，但从相关系数可以看出，第三对和第四对典型变量对于整体的特征捕捉也是非常明显的，预测选取 4 组典型变量。

母亲的体身体指标和心理指标对婴儿的行为特征和睡眠质量的标准化相关系数表如下表所示：

表 4 390 例婴儿行为特征和睡眠质量的标准化相关系数

变量	U_1	U_2	U_3	U_4
婴儿行为特征 (X_1)	-0.635	0.725	-0.378	-1.309
整晚睡眠时间 (X_2)	0.541	-0.005	0.327	-0.387
睡醒次数 (X_3)	-0.169	-0.345	0.56	-0.054
入睡方式 (X_4)	-0.25	0.47	0.439	0.309

表 5 母亲的体身体指标和心理指标的标准化相关系数

变量	V_1	V_2	V_3	V_4
母亲年龄 (Y_1)	0.073	-0.172	0.016	0.076
婚姻状况 (Y_2)	0.593	-0.061	-1.985	-2.382
教育程度 (Y_3)	-0.308	-0.074	0.211	-0.708
妊娠时间 (Y_4)	0.046	-0.065	0.435	0.008
分娩方式 (Y_5)	0.08	3.423	1.538	3.374

CBTS (Y_6)	0.085	0.139	0.047	-0.021
EPDS (Y_7)	-0.163	-0.156	-0.11	0.062
HADS (Y_8)	-0.027	0.094	0.181	-0.044

由此可写出 4 对标准化典型变量分别为：

$$U_1 = -0.635X_1 + 0.541X_2 - 0.169X_3 - 0.25X_4 \quad (8)$$

$$U_2 = 0.725X_1 - 0.005X_2 - 0.345X_3 + 0.47X_4 \quad (9)$$

$$U_3 = -0.378X_1 + 0.327X_2 + 0.56X_3 + 0.439X_4 \quad (10)$$

$$U_4 = -1.309X_1 - 0.387X_2 - 0.054X_3 + 0.309X_4 \quad (11)$$

$$V_1 = 0.073Y_1 + 0.593Y_2 - 0.308Y_3 + 0.046Y_4 + \quad (12)$$

$$0.08Y_5 + 0.085Y_6 - 0.163Y_7 - 0.027Y_8$$

$$V_2 = -0.172Y_1 - 0.061Y_2 - 0.074Y_3 - 0.065Y_4 + \quad (13)$$

$$3.423Y_5 + 0.139Y_6 - 0.156Y_7 + 0.094Y_8$$

$$V_3 = 0.016Y_1 - 1.985Y_2 + 0.211Y_3 + 0.435Y_4 + \quad (14)$$

$$+ 1.538Y_5 + 0.047Y_6 - 0.11Y_7 + 0.181Y_8$$

$$V_4 = 0.076Y_1 - 2.382Y_2 - 0.708Y_3 + 0.008Y_4 + \quad (15)$$

$$3.374Y_5 - 0.021Y_6 + 0.062Y_7 - 0.044Y_8$$

其中，在第 1 对标准化典型变量中， U_1 （婴儿行为特征和睡眠质量）以 X_1 （婴儿行为特征）的系数绝对值最大，可理解为 X_1 对 U_1 的贡献最大；其中，在第 1 对标准化典型变量中， V_1 （婴儿行为特征）以 Y_2 （婚姻状况）的系数绝对值最大，可理解为 Y_2 对 V_1 的贡献最大；说明在一定范围内，婴儿行为特征与母亲的婚姻状况度有较大相关性。

在第 2 对标准化典型变量中， U_2 （婴儿行为特征和睡眠质量）以 X_1 （婴儿行为特征）的系数绝对值最大，可理解为 X_1 对 U_2 的贡献最大；其中，在第 2 对标准化典型变量中， V_1 （婴儿行为特征）以 Y_5 （分娩方式）的系数绝对值最大，可理解为 Y_5 对 V_2 的贡献最大；说明在一定范围内，婴儿行为特征与母亲的分娩方式具有较大相关性。

在第 3 对标准化典型变量中， U_3 （婴儿行为特征和睡眠质量）以 X_3 （睡醒次数）

的系数绝对值最大，可理解为 X_3 对 U_3 的贡献最大；其中，在第 3 对标准化典型变量中， V_3 （婴儿行为特征）以 Y_2 （婚姻状况）的系数绝对值最大，可理解为 Y_2 对 V_3 的贡献最大；说明在一定范围内，婴儿的睡醒次数与母亲的婚姻状况度具有较大相关性。

在第 4 对标准化典型变量中， U_4 （婴儿行为特征和睡眠质量）以 X_1 （婴儿行为特征）的系数绝对值最大，可理解为 X_1 对 U_4 的贡献最大；其中，在第 4 对标准化典型变量中， V_4 （婴儿行为特征）以 Y_5 （分娩方式）的系数绝对值最大，可理解为 Y_5 对 V_4 的贡献最大；说明在一定范围内，婴儿的行为特征与母亲的分娩方式具有较大相关性。

由上可知：在三对典型变量中，婴儿的行为特征与母亲的婚姻状况和分娩方式具有非常大的相关性，且均为正相关型（对应系数均大于 0），因此母亲的身体指标和心理指标对婴儿的行为特征和睡眠质量有影响。

第 6 章 问题 2 的模型建立与求解

6.1 贝叶斯分类器模型的建立

由于题目要求建立婴儿的行为特征与母亲的身体指标与心理指标的关系模型，而婴儿的行为特征属于定类变量，因此选取合适的分类模型，由于数据存在不平衡数据以及需要后验概率估计对异常检测和风险进行评估，这里选用贝叶斯分类器。

Step1: 数据收集

收集包含特征和对应标签（类别）的训练数据集。

Step2: 特征选择和预处理

对数据进行特征选择、特征提取或特征转换等预处理步骤，以减少噪声和冗余。

Step3: 计算先验概率

对每个类别计算先验概率，即在没有任何数据信息的情况下，某个样本属于某个类别的概率。

Step4: 计算后验概率

使用贝叶斯定理，将先验概率和条件概率结合，计算后验概率，即在给定特征的情况下，属于某个类别的概率。

Step5: 求得贝叶斯分类器模型，并进行预测

对于新的样本，使用计算得到的后验概率，选择具有最高后验概率的类别作为结果。

6.2 贝叶斯分类器模型的求解和检验

通过 python 求解得到贝叶斯分类器的分类准确性如下表 6 所示：

表 6 贝叶斯分类器的分类准确性

准确率	召回率	精确率	F1 值
0.83	0.54	0.62	0.71

对最后有 20 组（编号 391-410 号）婴儿的行为特征分类结果如下表 7 所示：

表 7 婴儿的行为特征分类结果

编号	行为特征	编号	行为特征
391	中等型	401	矛盾型
392	中等型	402	中等型
393	中等型	403	中等型
394	中等型	404	安静型
395	矛盾型	405	中等型
396	矛盾型	406	安静型
397	矛盾型	407	中等型
398	安静型	408	安静型
399	中等型	409	安静型
400	中等型	410	中等型

模型的拟合程度如下图 2 所示

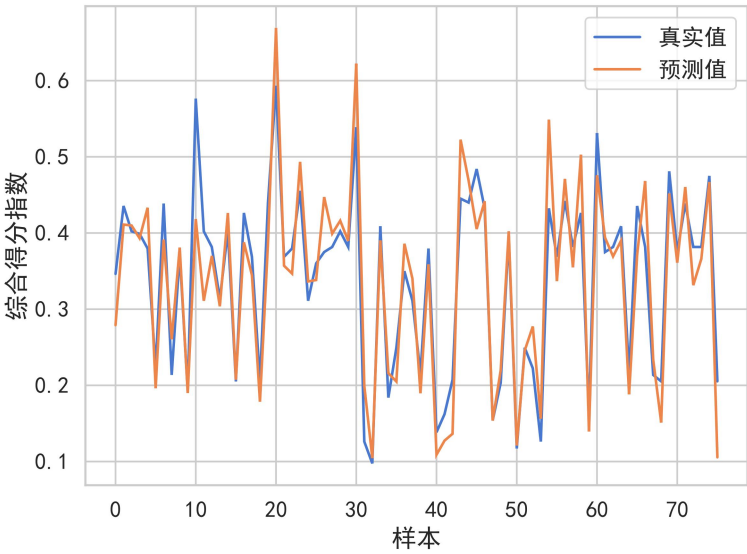


图 2 模型拟合程度对比
第 9 页 共 24 页

第 7 章 问题 3 的模型建立与求解

7.1 治疗费用的单目标非线性整数规划模型建立

Step1: 确定决策变量

首先构建以下决策变量:

- (1) Q 表示治疗所需费用。
- (2) x_i 分别对应于 CBTS, EPDS, HADS 在治疗过程中降低的得分。
- (3) k_i 对应于 x_i , 表示 x_i 前的系数。
- (4) $Score_1$, $Score_2$ 分别表示在问题二中安静型和中等型分界分数, 中等型和矛盾型的分界分数。
- (5) $f(x)$ 表示在问题二中婴儿的综合得分。

Step2: 确定规划的目标

目标函数为最小化治疗所需费用:

$$\min Q = k_1 x_1 + k_2 x_2 + k_3 x_3 \quad (16)$$

其中 $k_1 = (2812 - 200)/3$, $k_2 = (1890 - 500)/2$, $k_3 = (12500 - 300)/5$ 。

Step3: 确定约束条件

- (1) 为确保每项得分在治疗过后不能出现负数以及不能高于对应得分的平均值:

$$\begin{aligned} 0 &\leq 15 - x_1 \leq 6 \\ 0 &\leq 22 - x_2 \leq 9 \\ 0 &\leq 18 - x_3 \leq 8 \end{aligned} \quad (17)$$

其中, 在中等型的婴儿中 CBTS 得分的平均值为 6, 在中等型的婴儿中 EPDS 得分的平均值为 9, 在中等型的婴儿中 HADS 得分的平均值为 8。

- (2) 确保经过治疗后的婴儿的行为特征为中等型:

$$Score_1 \leq f(x_1, x_2, x_3) \leq Score_2 \quad (18)$$

其中, $f(x_1, x_2, x_3)$ 表示经过治疗后的婴儿的综合得分。

(3) 确保 CBTS, EPDS, HADS 数值在治疗过后得分不能提高, 因为分数越高, 症状越严重, 而经过治疗, 理论上数值会降低:

$$x_1, x_2, x_3 \geq 0 \quad (19)$$

(4) 确保 CBTS, EPDS, HADS 数值在治疗过后为整数:

$$x_1, x_2, x_3 \in Z \quad (20)$$

Step4: 修改约束条件的部分参数

为了进而求得婴儿的行为特征从矛盾型转变成安静型, 将公式 17 中的参数进行如下修改:

$$\begin{aligned} 0 &\leq 15 - x_1 \leq 5 \\ 0 &\leq 22 - x_2 \leq 8 \\ 0 &\leq 18 - x_3 \leq 7 \end{aligned} \quad (21)$$

其中, 在安静型的婴儿中 CBTS 得分的平均值为 5, 在安静型的婴儿中 EPDS 得分的平均值为 8, 在安静型的婴儿中 HADS 得分的平均值为 7。

Step5: 对比方案

通过对比两种方案的不同, 找出调整方向。

7.2 问题 3 的求解结果

因此确定单目标非线性整数规划模型为:

$$\begin{aligned} \min \quad & Q = k_1 x_1 + k_2 x_2 + k_3 x_3 \\ \text{s.t.} \quad & \begin{cases} 0 \leq 15 - x_1 \leq 6 \\ 0 \leq 22 - x_2 \leq 9 \\ 0 \leq 18 - x_3 \leq 8 \\ Score_1 \leq f(x_1, x_2, x_3) \leq Score_2 \\ x_1, x_2, x_3 \geq 0 \\ x_1, x_2, x_3 \in Z \end{cases} \end{aligned} \quad (22)$$

通过 python 求解得到结果见表 8 和表 9:

表 8 矛盾型转变成中等型结果表

	1	2	3
系数 (k_i)	937.3	695	2440
自变量 (x_i)	9	13	10

因变量 (Q)	41870
-------------	-------

由表可知, 对与将婴儿的行为特征从矛盾型转变中等型, 需要 CBTS 降低 9 分, EPDS 降低 13 分, HADS 降低 10 分, 总共花费 41870 元。

表 9 矛盾型转变成安静型结果表

	1	2	3
系数 (k_i)	937.3	695	2440
自变量 (x_i)	10	14	11
因变量 (Q)	45943		

由表可知, 对与将婴儿的行为特征从矛盾型转变安静型, 需要 CBTS 降低 10 分, EPDS 降低 14 分, HADS 降低 11 分, 总共花费 45943 元。

由此可得到, 若中途要将婴儿的行为特征从矛盾型转变成安静型, 需要额外在 CBTS, EPDS, HADS 上各降低 1 分, 预计多花费 4030 元 (从 41780 到 45943) 。

第 8 章 问题 4 的模型建立与求解

8.1 基于熵权 TOPSIS 法的综合评价模型建立

Step1: 将原始数据 (X_{i1} = 整晚睡眠时间、 X_{i2} = 睡醒次数、 X_{i3} = 入睡方式) 趋势化并, 根据不同类型的指标按照不同的公式进行正向化处理, 并构建标准化矩阵 Z_{ij} 。

$$Z_{ij} = \frac{X_{ij}}{\sqrt{\sum (X_{ij})}} \quad (23)$$

Step2: 求各指标 (整晚睡眠时间、睡醒次数、入睡方式) 在各方案下的比值 p_{ij} , 其中 $\sum Z_{ij}$ 表示对第 j 列求和。

$$p_{ij} = \frac{Z_{ij}}{\sum Z_{ij}} \quad p_{ij} = \frac{Z_{ij}}{\sum Z_{ij}} \quad (24)$$

Step3: 求各指标 (整晚睡眠时间、睡醒次数、入睡方式) 的信息熵 E_j 。

$$E_j = - \frac{1}{\ln(n)} \sum_{i=1}^n p_{ij} \ln p_{ij} \quad (25)$$

其中，若 $p_{ij} \leq 0$ 则为 $p_{ij} = 0$ 。

Step3: 求各指标（整晚睡眠时间、睡醒次数、入睡方式）的权重 w_j 。

$$w_j = \frac{1 - E_j}{m - \sum E_j} \quad (26)$$

其中 m 表示指标数，这里 $m = 3$ 。

Step4: 计算各评价指标与最优及最劣向量之间的差距 D_i^+, D_i^- 。

$$D_i^+ = \sqrt{\sum_{j=1}^m w_j (Z_j^+ - Z_{ij})^2}, D_i^- = \sqrt{\sum_{j=1}^m w_j (Z_j^- - Z_{ij})^2} \quad (27)$$

Step5: 计算各个对象的最终得分 C_i 。

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \quad (28)$$

其中， C_i 越大表明评价对象越优。

由于要对婴儿的睡眠质量进行优、良、中、差四分类综合评判，这里将采取无监督方法进行聚类，聚类数量为 4，选取的聚类方法为 k-means 聚类，距离计算公式选取绝对值距离法（非欧式距离，由于维度只有 1，计算结果与欧式距离的计算结果完全相同）。

之后，选取四个聚类中心，为了能得到最准确的聚类结果，设 \max 为综合得分的最大值， \min 为综合得分的最小值，因此四个聚类中心如下表所示：

表 10 K-Means 聚类的参数表

聚类算法		聚类数量	距离公式
K-Means 聚类法		4	绝对值距离法
初始 4 个聚类中心			
\min	$\min + (\max - \min)/4$	$\min + (\max - \min)/2$	\max

通过不断更新迭代 4 个聚类中心，由于综合得分越高其对象的入睡质量越差，因此对最终的聚类中心得分进行从小到大排序，其结果分别对应于婴儿的睡眠质量进行优、良、中、差四分类综合评判。

此后，为建立婴儿综合睡眠质量与母亲的身体指标、心理指标的关联模型，由于此前建立的综合睡眠质量模型的最终结果为定量数据，因此该题的解决方案可以转变为定量数据到定量数据的拟合问题，即使用母亲的身体指标、心理指标数据拟合综合睡眠质量得分。

8.2 基于随机森林回归模型建立

由于第一问已经求得母亲的身体指标、心理指标与婴儿的行为特征和睡眠质量有影响，因此这里不再做母亲的身体指标、心理指标和婴儿的综合睡眠质量得分做相关性的论述。

8.3 问题 4 的求解结果

通过熵权法得到权重向量如下表 11 所示：

表 11 熵权法矩阵向量表			
指标对应权重	整晚睡眠时间 (w_1)	睡醒次数 (w_2)	入睡方式 (w_3)
权重值	0.4834	0.1749	0.34163

利用权重，得到各对象的综合得分，部分结果展示如下表 12 所示：

表 12 综合得分部分表					
编号	综合得分	排名	编号	综合得分	排名
1	0.311066201	87	5	0.398961605	54
2	0.375012278	65	6	0.381640682	60
3	0.24943526	104	7	0.395812681	55
4	0.222339049	114	8	0.395812681	55
...

选取 4 各均匀的聚类中心后，最终的聚类中心和聚类结果部分数据如下表 13 所示：

表 13 聚类结果表			
婴儿的睡眠质量优	婴儿的睡眠质量良	婴儿的睡眠质量中	婴儿的睡眠质量差
聚类中心 (0.204)	聚类中心 (0.372)	聚类中心 (0.46)	聚类中心 (0.588)

编号: 216	编号: 286	编号: 39	编号: 163
编号: 129	编号: 296	编号: 143	编号: 184
编号: 276	编号: 342	编号: 36	编号: 259
编号: 255	编号: 229	编号: 108	编号: 238
...

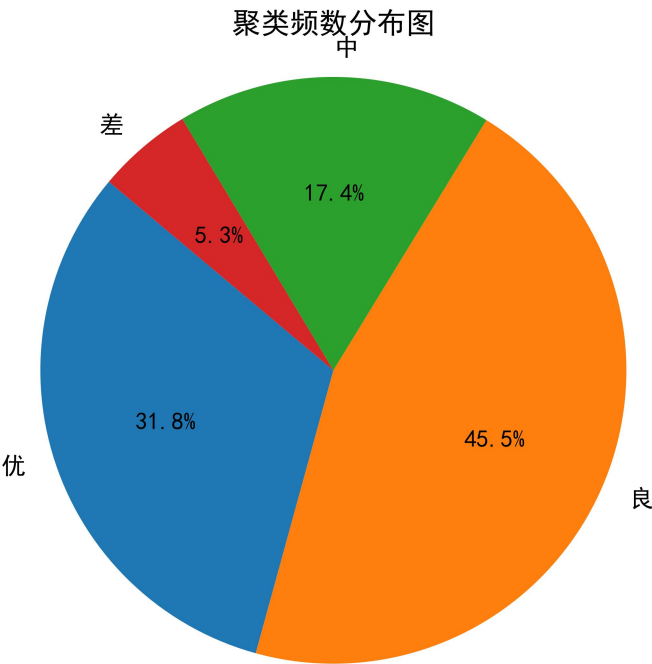


图 3 聚类频数图

从上可知：大部分婴儿的睡眠质量为优（31.8%）和良（45.5%），少部分睡眠质量为差（5.3%）。

随后建立随机森林模型，对母亲的 身体指标、心理指标和婴儿综合睡眠质量进行拟合，此后，对最后 20 组（编号 391-410 号）婴儿的综合睡眠质量进行预测，将其母亲的身体指标、心理指标数据带入随机森林回归模型，得到对应的综合得分，且并入与之距离最近的聚类划分中，从而确定该婴儿的睡眠质量的类型，结果如下表 14 所示：

表 14 睡眠质量预测结果表

编号	睡眠质量分类	睡眠质量所属类型
391	4	差

392	4	差
393	3	中
394	3	中
395	3	在
396	2	良
397	3	中
398	2	良
399	3	中
400	1	优
401	4	差
402	2	良
403	2	良
404	1	优
405	3	中
406	2	良
407	3	中
408	1	优
409	1	优
410	1	优

第 9 章 问题 5 的模型建立与求解

9.1 问题 5 的建立

由问题四的划分结果可知，编号为 238 的婴儿睡眠质量类型为差，而通过治疗手段将婴儿的睡眠质量从差转变成优，其实可以理解为问题五是问题三和问题四的综合，也就是在问题三的基础上引入问题四已经训练好的模型，将编号为 238 的婴儿在随机森林回归模型中的结果作为约束条件一并置入单目标非线性整数规划中，并求得结果。

在问题三的基础上引入一个新的约束条件：

$$RFR_1 \leq RFR(x) \leq RFR_2 \quad (29)$$

其中， RFR 全称为 Random Forest Regression（随机森林回归）其中 RFR_1 表示在评

定睡眠质量类型为良和优的分界线， RFR_2 表示评定在睡眠质量类型为优的上界。

9.2 问题 5 的求解结果

求得问题五的单目标非线性整数规划模型如下：

$$\begin{aligned} \min \quad & Q = k_1 x_1 + k_2 x_2 + k_3 x_3 \\ \text{s.t.} \quad & \begin{cases} 0 \leq 15 - x_1 \leq 6 \\ 0 \leq 22 - x_2 \leq 9 \\ 0 \leq 18 - x_3 \leq 8 \\ Score_1 \leq f(x_1, x_2, x_3) \leq Score_2 \\ x_1, x_2, x_3 \geq 0 \\ x_1, x_2, x_3 \in Z \\ RFR_1 \leq RFR(x_1, x_2, x_3) \leq RFR_2 \end{cases} \end{aligned} \quad (30)$$

从而得到有睡眠质量评定下结果，其中表 15 为无睡眠质量评定结果，表 16 为有睡眠质量评定结果：

表 15 无睡眠质量评定治疗花费

无睡眠质量评定	CBTS 变化量	EPDS 变化量	HADS 变化量	最少花费价格
矛盾型转中等型	9	13	10	41870
矛盾型转安静型	10	14	11	45943

表 16 有睡眠质量评定治疗花费

有睡眠质量评定	CBTS 变化量	EPDS 变化量	HADS 变化量	最少花费价格
矛盾型转中等型	10	13	10	42808
矛盾型转安静型	10	14	12	48383

由表 13 和表 14 对比可知：若需要将婴儿的行为特征从矛盾型变成中等型，且睡眠质量变成优，在原先治疗方案的基础上，在 CBTS 上多降低 1 分即可，也就是对 CBTS 的治疗加强一分，预计将要多花费 958 元；而若需要将婴儿的行为特征从矛盾型变成安静型，且睡眠质量变成优，在原先治疗方案的基础上，在 HADS 上多降低 1 分即可，也就是对 HADS 的治疗加强一分，预计将要多花费 2440 元。

第 10 章 模型的测试

10.1 鲁棒性分析

本文是基于随机森林回归算法建立模型。模型在不同环境扰动影响下得到错误的、甚至特定的结果，危害使用者的安全。正确性衡量模型对未知数据的预测性能，即对于给定未知输入，模型能够正常预测或分类的能力；如图 4 所示，本文的预测模型在具有轻微扰动的原始输入上的预测结果不会显著偏离该原始输入，即鲁棒性较强。说明该模型在输入样本存在细微对抗扰动的情況下，预测具有不受对抗样本干扰或误导的能力。

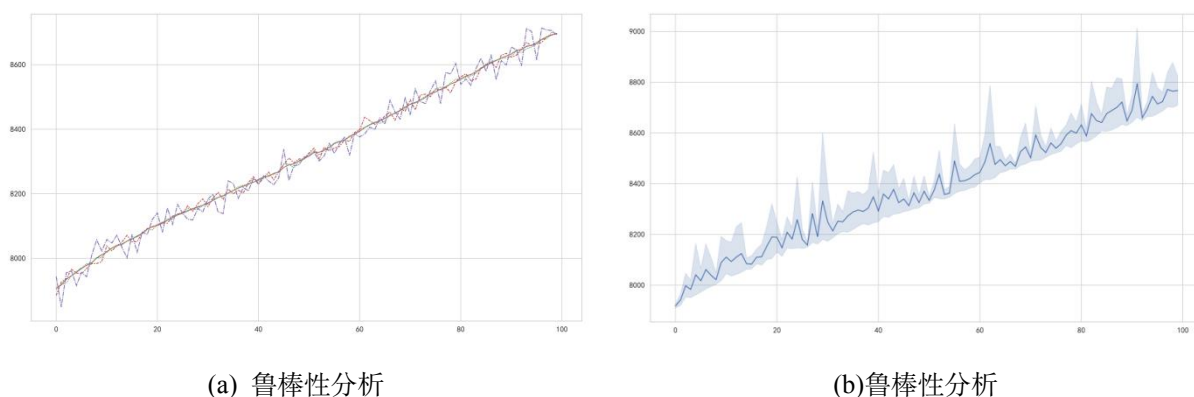


图 4 鲁棒性分析

(a) 中数据的起伏不大，且呈不断上升趋势；(b) 在数据的置信区间变化中向上的置信区间起伏大，而向下的置信区间起伏小，数据整体呈不断上升趋势，由此可以看出该随机森林回归模型具有稳定的鲁棒性。

第 11 章 模型的评价与总结

11.1 模型的优点

- (1) 采用贝叶斯分类器对婴儿的行为特征进行分类和预测，预测效果好，预测准确。
- (2) 对数据进行典型效果分析，避开常规的相关性分析，从研究两组集合之间的相关性。

11.2 可能的改进

- (1) 在采用 K-Means 聚类算法对婴儿的睡眠质量进行划分，聚类中心存在不确定性，会导致后面的划分结果出现误差。
- (2) 没有研究指标之间的非线性相关性，采用的是简单线性相关性。

第 12 章 参考文献

- [1] 孙建芝, 牛晓亚, 韩丽华, 等 . 痰浊证微观辨证指标的实验研究 [J] . 河南中医, 1996, 16 (2) : 21-22.
- [2] 张超, 陈亮, 潘平康, 等 . 缺血性中风中医证型与颈部血管狭窄程度、血脂及 HCY 的相关性 [J] . 陕西中医, 2013, 34 (10) : 1302-1303.
- [3] 李有梅, 梁珣. 典型相关分析综述 [J] . 中国计量大学学报, 2017, 28 (1) : 113-118.
- [4] 李文婷, 妊娠期高血压疾病孕妇胎儿脐动脉血流动力学及新生儿相关指标分析 [J]. 中国妇幼保健, 2020, 35 (4) : 674-677.
- [5] 闫佰忠, 孙剑, 安娜. 基于随机森林模型的地下水水质评价方法 [J] .水电能源科学, 2019, 37 (11) : 66-69.

附录

附录 1

介绍: 工具和软件

通过 Office 2023 编写和生成的论文。

使用 Python 和 Matlab2023a 生成计算图像和数据。

附录 2

介绍: Python 代码的算法

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.cross_decomposition import CCA

# 从 Excel 文件中导入数据

xls = pd.ExcelFile(r'C:\Users\LJY\Desktop\mytest.xlsx')

data = pd.read_excel(xls, xls.sheet_names)

df_X = data['Sheet1'][['母亲年龄', '婚姻状况', '教育程度',

                      '妊娠时间（周数）', '分娩方式', 'CBTS', 'EPDS', 'HADS']]

df_Y = data['Sheet1'][['婴儿行为特征', '整晚睡眠时间（时：分：秒）', '睡醒次数', '入睡方式']]

df_X = df_X.iloc[0:390]

df_Y = df_Y.iloc[0:390]
```

```
# 执行典型相关分析

cca = CCA(n_components=2) # 指定保留的典型变量数量

cca.fit(df_X, df_Y)


# 获取典型变量

X_c, Y_c = cca.transform(df_X, df_Y)


# 计算典型相关系数

cca_corr = np.corrcoef(X_c.T, Y_c.T)[3:, :2]


# 打印典型相关系数

print("Canonical Correlation Coefficients:")

print(cca_corr)


# 绘制典型变量之间的散点图

plt.rcParams['font.sans-serif'] = ['SimHei'] # 设置中文字体为黑体

plt.rcParams['axes.unicode_minus'] = False # 解决负号显示问题

plt.figure(figsize=(6, 4.5))

plt.scatter(X_c[:, 0], Y_c[:, 0], label='典型变量 1', color='blue')

plt.scatter(X_c[:, 1], Y_c[:, 1], label='典型变量 2', color='red')
```

```
plt.xlabel('母亲的身体指标和心理指标')

plt.ylabel('婴儿的行为特征和睡眠质量')

plt.title('典型相关分析')

plt.legend()

plt.grid()

plt.tight_layout()

plt.savefig(r'C:\Users\LJY\Desktop\典型相关分析.svg', format='svg')

plt.show()

import matplotlib.pyplot as plt

import numpy as np

import seaborn as sns


def linear(l, angle):

    a = 30

    angle_radians = np.deg2rad(angle)

    x_start = a*np.cos(angle_radians)

    y_start = a*np.sin(angle_radians)

    x_end = l*np.cos(angle_radians)

    y_end = l*np.sin(angle_radians)
```

```

# 绘制直线

plt.plot([x_start, x_end], [y_start, y_end])


def plot_circle(radius):

    circle = plt.Circle((0, 0), radius, fill=False,

                        label='Circle', color='black')

    plt.gca().add_artist(circle)

n = 5

angle0 = 90-180/n

angle1 = 360/n

for i in range(n):

    linear(85, angle0+i*angle1)

plt.scatter(0, 87, color='b', alpha=0)

plt.xlabel('X-axis')

plt.ylabel('Y-axis')

plt.title('linear')

plt.grid(True)

plt.gca().set_aspect('equal', adjustable='box') # 设置 x 轴和 y 轴的刻度相等，保证图形不会被扭曲

plt.legend(['y1', 'y2', 'y3', 'y4', 'y5'], loc='upper right')

plot_circle(85)

plot_circle(30)

```



```
plt.savefig(r'C:\Users\LJY\Desktop\五射线.svg', format='svg')  
  
plt.show()
```